TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Regressió quantílica i quantílica logística: Una aproximació a l'economia i l'estadística

Autor: Josep Franquet Fàbregas

Director: Montserrat Guillen i Estany

Departament: Departament d'Econometria, Estadística i Economia Aplicada

Convocatòria: Juny 2020

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

RESUM

La part introductòria del treball està formada per la introducció, la revisió de la literatura i la metodologia. En aquesta part introductòria, es justifica el tema (regressió quantílica i quantílica logística), es justifiquen les dues aplicacions amb dades reals, es presenten altres autors que hagin parlat d'aquests temes i s'expliquen els fonaments dels dos tipus de regressions utilitzades.

La primera aplicació utilitza unes dades de l'àmbit de la bioestadística en què la variable resposta és el pes d'un nadó al néixer, mentre que les variables explicatives són característiques de la mare del nadó. Per altra banda, en la segona aplicació, la variable resposta és el percentatge de quilòmetres recorreguts per sobre de la velocitat permesa per un conductor al cap d'un any i les variables explicatives són dades bàsiques de conducció. En la primera aplicació, s'hi utilitza la regressió quantílica i en la segona, la regressió quantílica logística. L'estructura seguida per al desenvolupament de les dues aplicacions és la mateixa: anàlisi descriptiva univariant i multivariant, estimació i anàlisi d'un o diferents models lineals, estimació i anàlisi de models de regressió quantílica o quantílica logística i presentació d'una metodologia basada en els conceptes dels percentils condicionats i els percentils no condicionats. Es finalitza el treball, presentant dos aplicacions *Shiny*, una per a cada aplicació del treball.

PARAULES CLAU

Regressió quantílica, regressió quantílica logística, percentil, quantil, mínims quadrats ordinaris, percentil condicionat, percentil no condicionat, *birthweights* i telemàtics

CLASSIFICACIÓ AMS

62J05 Linear regression

62J12 Generalized linear models

62P05 Applications to actuarial sciences and financial mathematics

62P20 Applications to economics

TITLE

Quantile regression and logistic quantile: An approximation to economics and statistics

SUMMARY

The two main topics of this final degree project are quantile regression and logistic quantile regression. The project starts by justifying the topic and the two real data applications. In the literature review, different research papers (with its authors) are presented. On these papers, its authors have talked about quantile regression, logistic quantile regression or data used in the project. To end this introductory part, the methodology section starts explaining the idea of what a quantile is to end up exploring the characteristics, parameterizations, and estimates of both quantile regression and logistic quantile regression.

The first application uses data from the field of biostatistics in which the response variable is the weight of a baby at birth, while the explanatory variables are characteristics of the baby's mother. On the other hand, in the second application, the response variable is the percentage of kilometres driven above the speed limit by a driver after one year and the explanatory variables are basic driving data. In the first application, quantile regression is used while in the second, logistic quantile regression is applied. The structure followed for the development of the two applications is the same: univariate and multivariate descriptive analysis, estimation and analysis of one or different linear models, estimation and analysis of quantile or logistic quantile regression models and presentation of a methodology based in the concepts of conditional percentiles and unconditional percentiles. The project is completed by presenting two *Shiny* applications, one for each application developed.

KEYWORDS

Quantile regression, logistic quantile regression, percentile, quantile, ordinary least squares, conditional percentile, unconditional percentile, birthweights and telematics

ÍNDEX

1. Introducció

Els mínims quadrats ordinaris resulten ser el mètode més utilitzat a l'hora d'estimar els paràmetres en un model de regressió lineal. Tot i això, aquest mètode necessita tres supòsits que són la base per estimar un model de regressió lineal i garantir unes bones propietats per als estimadors:

1- La mitjana dels errors ha de ser 0 i la seva variància ha de ser constant.
2- Els errors no poden estar correlacionats entre sí.
3- Les variables explicatives han de ser ortogonals als errors, és a dir, no compartir informació.

Moltes vegades, aquests supòsits no es compleixen i cal plantejar altres mètodes d'estimació. D'aquests mètodes alternatius, destaca la regressió quantílica. La diferència principal amb els mínims quadrats ordinaris (MQO a partir d'ara) és que s'està estimant la mediana condicional o altres quantils condicionals de la variable resposta que siguin d'interès, mentre que els mínims quadrats ordinaris estimen la mitjana condicional de la variable resposta donats certs valors de les variables de predicció.

Aquest mètode alternatiu es pot utilitzar quan la variable resposta:

1- És bimodal o multimodal.
2- Presenta asimetries (ja siguin positives o negatives).

Alguns exemples bàsics en els quals es podrien estimar els paràmetres d'un model lineal mitjançant regressió quantílica serien el nivell d'ingrés econòmic en molts països (normalment presenta una asimetria positiva) o els diners gastats pels clients en una botiga (l'interès pot estar en els quantils més alts i no en la mitjana). *(Flom, 2018)*

Aquest treball no es centra només en la regressió quantílica, sinó que va més enllà i estudia, programa i treballa la regressió quantílica logística. La principal diferència és que, en aquest cas, la variable resposta és contínua i presenta un rang de valors entre dos extrems ben definits. Per tant, es pot deduir que, en aquest cas, la variable resposta del model pot ser una probabilitat ($y_{min} = 0, y_{max} = 1$), un tant per cent ($y_{min} = 0, y_{max} = 100$) o una variable que s'estén entre dos valors concrets ($y_{min} = 0, y_{max} = 70$, per exemple).

A causa del fet que aquestes variables estan definides en un interval de valors concret, els mètodes clàssics d'estimació (mínims quadrats ordinaris) són aplicables però, si s'utilitzen, els estimadors no tindran les propietats necessàries per tal que la inferència sigui vàlida. Aquestes propietats són les de tenir estimadors:

1- Consistents
2- Eficients
3- No esbiaixats

4- Asimptòticament normals amb variància constant

Els estimadors compleixen aquestes propietats sempre que es compleixin els supòsits bàsics de l'estimació per mínims quadrats ordinaris anomenats anteriorment. A continuació, es tornen a recordar aquests supòsits i s'expressen de manera algebraica:

1- La mitjana dels errors ha de ser 0, la seva variància ha de ser constant i els errors no poden estar correlacionats entre sí.

$$E[e] = 0, Var\,(e) = E[ee'] = \sigma^2 * I.$$

On $Var\,(e)$ i $E[ee']$ corresponen a la matriu de variàncies i covariàncies de la pertorbació aleatòria: e.

2- Les variables explicatives han de ser ortogonals als errors, és a dir, no compartir informació.

$$E[X' * e] = 0.$$

Aquests supòsits s'han de fer suposant que s'estima un model de regressió lineal del tipus $Y = X\beta + e$. On Y és la matriu amb els valors de la variable resposta a modelitzar; X és la matriu de variables explicatives; β és la matriu amb els estimadors dels paràmetres del model estimats i e és la pertorbació aleatòria o error del model.

En aquest cas, degut a que la variable resposta està limitada en un rang de valors determinat, és molt difícil que es compleixin aquests supòsits especificats.

Durant les darreres dècades, hi ha hagut diferents autors que han intentat proposar altres mètodes per tal d'estimar els paràmetres en regressions en les quals es tingui una variable resposta definida en un interval de valors concret. Per exemple, *Papke and Wooldridge (1996)* van demostrar que la simple estimació per mètodes versemblants pot ser utilitzada per modelitzar l'esperança de variables amb límits definits.

De la mateixa manera, *Lesaffre et al. (2007)* van estudiar l'ús de la transformació logística a una variable normal, la localització i escala de la qual pogués dependre d'un conjunt de covariàncies. Aquesta aproximació pot ser aplicada a variables dependents contínues i discretes definides en un rang de valors determinat. Seguidament, hi ha hagut altres autors[1] que han desenvolupat altres propostes seguint sobretot amb el treball d'aquests autors anomenats (*Papke and Wooldridge i Lesaffre et al.).*

Tant en economia com en estadística, moltes de les variables que són d'anàlisi es mesuren amb tant per cent, i per tant estan definides en un interval de valors concret. D'aquest fet, es pot deduir que la regressió quantílica logística és perfectament vàlida per a aquestes dues ciències i és aplicable a qualsevol cas d'estudi que involucrés

[1] Entre aquests autors, destaquen *Heijtan i Rubin (1991)* i *Kieschnick i McCullough (2003).*

variables dependents mesurades amb tant per cent (o de 0 a 1, fent referència a una probabilitat) o amb un interval de valors concret.

Per tant, l'anàlisi, l'estudi i l'aplicació de la regressió quantílica logística en aquest treball final de grau d'economia i d'estadística permetrà l'exploració d'un concepte modern i molt útil a l'hora d'estimar els paràmetres de regressions que presentin les característiques comentades. Tanmateix, l'exploració, programació i l'anàlisi de la regressió quantílica en exemples aplicats permetrà mostrar el potencial d'aquesta metodologia.

A nivell personal, m'agradaria dir que la modernitat dels conceptes i la poca exploració prèvia que hi ha, són les principals motivacions que m'han portat a decidir-me per a aquest tema. De la mateixa manera, cal destacar que el fet que aquests tipus de regressions siguin aplicables a fets estudiats per les ciències dels dos graus (Economia i Estadística) dels quals realitzo aquest treball final de grau també resulta ser una motivació important.

Una altra motivació que m'agradaria destacar és que l'anàlisi i estudi d'aquests dos conceptes implica anar més enllà dels continguts estudiats en els dos graus que he cursat. És a dir, aquests dos tipus de regressions eren dos conceptes desconeguts per a mi fins que la meva directora d'aquest treball final de grau me'n va parlar i em vaig començar a informar i a llegir sobre ells. Considero que acabar els dos graus anant més enllà dels seus continguts és una de les millors maneres per fer-ho.

Un cop presentat i justificat el tema, es proposa la següent estructura per tal de desenvolupar un bon estudi i desenvolupament. S'inicia el treball amb aquesta introducció en la qual s'explica el que es fa i es justifiquen tant el tema com les dos aplicacions d'estudi. Seguidament, es realitza una revisió de la literatura *(Literature review)* en la que es mostren alguns autors que han utilitzat aquests dos tipus de regressions en els seus estudis. Tanmateix, s'exposen alguns estudis (amb els seus autors) en què s'hagin utilitzat les dos bases de dades que s'utilitzen en aquest treball. Per acabar aquesta part inicial del treball, es fa una explicació del mètode que es segueix, partint del concepte més bàsic de què és un quantil per acabar explorant les característiques tant de la regressió quantílica com de la regressió quantílica logística.

Arribats a aquest punt, ja s'haurà plantejat el tema i es tindrà clar com es treballen, es parametritzen i s'estimen aquestes dues regressions. A continuació, es realitzen dues aplicacions, una per a cada tipus de regressió.

La primera aplicació està relacionada amb la regressió quantílica i la variable dependent en aquest cas és el pes dels nadons al néixer. Es tracta d'un subconjunt de dades detallades de natalitat publicades el juny del 1997 pel Centre Nacional d'Estadístiques

Mèdiques (NIH[2]) dels Estats Units. A partir de variables que fan referència a característiques d'una mare, es podran estimar els diferents quantils de la variable que fan referència al pes del nadó al néixer. En aquest cas, la regressió quantílica és útil per estimar els paràmetres de models de regressió associats a quantils baixos de la variable resposta. En medicina, l'interès està en aquests quantils, ja que un reduït pes del nadó al néixer sol implicar malalties importants durant els primers anys de vida.

Aquesta primera aplicació està molt relacionada amb el camp biològic i mèdic de l'estadística (Bioestadística i Estadística mèdica). En els últims anys, l'estadística ha anat guanyant molt pes i importància en aquests dos camps d'estudi. Gràcies a la seva aplicació, s'han pogut millorar molts dels coneixements que tenien aquestes dos ciències i se li han donat sortides importants a l'estadística. A nivell personal, l'estudi d'aquesta primera aplicació és molt útil per realitzar una bona aproximació a aquesta vessant d'estudi de l'estadística.

La segona aplicació que es realitza es fa utilitzant la regressió quantílica logística a partir de dades proporcionades per una entitat asseguradora. Es tracta de dades recollides durant l'any 2010 i, en aquest cas, la variable dependent és el tant per cent de quilòmetres que al cap d'un any un conductor viatja per sobre de la velocitat permesa de la via en la qual estigui circulant. Les variables independents que s'utilitzen són dades bàsiques de conducció com ara el quilometratge i el tipus de via on es circulava.

Es tracta d'unes dades molt interessants relacionades amb l'àmbit assegurador i, per tant, econòmic. El fet de tractar dades de l'àmbit econòmic, partint d'una base estadística resulta perfecta per tal de combinar els àmbits d'estudi dels dos graus dels quals s'està realitzant aquest treball final de grau.

Cal recordar, tal i com s'ha dit abans, que tant en estadística com amb economia, moltes de les variables estan mesurades amb tant per cent, per tant, l'aplicació de la regressió quantílica logística per a aquest cas seria extrapolable a qualsevol altre cas d'estudi en el qual la variable dependent sigui d'aquest tipus, tenint amb compte sempre les particularitats del cas d'estudi. Aquesta doble aplicació i la particularitat i exclusivitat de les dades han resultat les dues principals motivacions a l'hora d'incloure aquesta aplicació en el meu treball final de grau.

Un cop finalitzades aquestes dues exploracions, es va més enllà de la seva simple parametrització i estimació i es realitza una aplicació "*Shiny*" utilitzant el seu paquet d'R. Gràcies a aquesta aplicació interactiva, l'usuari serà capaç de realitzar estimacions relacionades amb els dos casos d'estudi a partir d'escenaris en els que caldrà donar valors a les variables predictores de manera fàcil, simple i interactiva.

[2] National Institute of Health

En definitiva, l'estudi, l'anàlisi i l'aplicació d'aquests dos conceptes: la regressió quantílica i la regressió quantílica logística, em serà útil per anar més enllà dels continguts dels graus dels quals estic realitzant aquest treball final de grau. Per tant, m'ajudarà a ampliar els coneixements d'aquestes dues ciències de les quals en seré futurament graduat. Per últim, m'agradaria aclarir i remarcar que un altre objectiu que busco és que la realització d'aquest treball em serveixi per tal de decidir-me a l'hora de triar el màster universitari que m'agradaria cursar el proper curs.

2. Revisió de la literatura

La regressió quantílica resulta ser un mètode emergent, l'ús de la qual ha crescut molt en els últims anys. Tanmateix, cal remarcar que l'àmbit econòmic és un dels camps en què s'hi poden trobar més articles científics en els quals s'hagi utilitzat aquest tipus d'estimació (veure *Pitarque, 2019*). Per exemple, *Baker et al. (2020)* utilitzen la regressió quantílica per tractar un tema molt important en els nostres dies: els elevats preus dels habitatges a les ciutats. Mitjançant la regressió quantílica, demostren que aquests preus elevats tenen un impacte negatiu en la salut mental de les persones. Tanmateix, demostren que a persones amb una salut mental inicial inferior els afecta més negativament que d'altres amb una salut mental inicial superior. Aquest fet destaca perquè l'observen gràcies a la regressió quantílica, mentre que amb altres anàlisis longitudinals clàssics no és observable.

Per altra banda, *Atsalakis et al. (2020)*, mitjançant la regressió quantílica, analitzen l'impacte dels desastres naturals sobre el creixement econòmic. D'aquesta manera, descobreixen com diferents quantils de desastres naturals (en funció del nivell catastròfic o intensitat del desastre natural) afecten a diferents quantils del creixement del Producte Interior Brut (PIB a partir d'ara) d'un país. Arriben a la conclusió que la relació entre la intensitat dels desastres naturals i el creixement econòmic és principalment negativa. Tot i així, admeten que, depenent del quantil que s'estigui examinant, aquesta relació pot arribar a ser positiva. Finalment, obtenen resultats diferents quan s'estimen regressions quantíliques per grups de països que difereixen amb: clima i context econòmic o democràtic.

Per altra banda, la presència d'articles científics de caràcter econòmic en què s'hagi utilitzat la regressió quantílica logística resulta ser escassa o inexistent. Aquest fet es deu principalment a l'elevat grau de novetat que presenta aquest concepte. En el camp de la bioestadística, *Bottai et al. (2010)* utilitzen la regressió quantílica logística per analitzar el nivell de depressió (mesurat amb un índex que va de 0 a 60) en adolescents a partir de característiques de la persona com ara: sexe, convivència amb els pares o no, si li havien succeït esdeveniments inesperats propers, entre d'altres. Mitjançant aquesta anàlisi, demostren que la regressió quantílica logística és molt útil per estimar els paràmetres de models en què la variable resposta es trobi limitada amb un interval de valors concret i altres mètodes d'estimació no siguin vàlids (MQO, per exemple). Una de les conclusions a les que arriben és la presència d'una relació inversa entre un índex de cohesió familiar i l'índex de depressió (variable resposta del model).

Per a la primera aplicació d'aquest treball, també s'han utilitzat unes dades de l'àmbit de la bioestadística. En aquest cas, es tracta de dades detallades de natalitat. Aquestes mateixes dades han sigut utilitzades per *Koenker et al. (2001)* per explicar els fonaments de la regressió quantílica. Aquest llibre es pot considerar com l'origen i la primera

exposició clara i detallada del concepte de regressió quantílica. La principal conclusió a la que arriben aquests autors és la potencialitat que té la regressió quantílica enfront altres mètodes clàssics d'estimació.

Tal i com s'ha dit anteriorment, es recorda que les dades que s'utilitzen en la segona aplicació d'aquest treball són de l'àmbit assegurador. *Pitarque (2019)* utilitza aquestes dades per ajustar un model inspirat amb una mesura del risc coneguda com a *TVaR (Tail Value at Risk).* Precisament, utilitza la regressió quantílica per estimar aquests models. Per a l'estimació d'aquest concepte del món assegurador, obté que les variables Edat i percentatge de quilòmetres amb horari nocturn han estat les més rellevants a l'hora d'estimar els paràmetres dels diferents models.

Per últim, *Pérez et al. (2019)* utilitzen aquestes dades per modelitzar el nombre de quilòmetres recorreguts per sobre de la velocitat permesa mitjançant models de regressió quantílica. Descobreixen que el risc de conduir per sobre de la velocitat permesa és heterogeni i depèn del percentil al que s'estigui fent referència. Tanmateix, arriben a la conclusió que les campanyes per reduir la velocitat dels conductors haurien d'anar dirigides als homes que condueixen per carreteres no urbanes i, especialment a aquells que condueixen un elevat percentatge de quilòmetres amb horari nocturn.

3. Metodologia

En aquesta secció del treball, s'expliquen clarament els conceptes que s'analitzen, s'estudien i es programen al llarg del treball per tal que es tingui ple coneixement sobre les eines utilitzades. Per tant, s'inicia aquest apartat partint i explicant la idea de què és un quantil per acabar explorant les característiques, parametritzacions i estimacions tant de la regressió quantílica com de la regressió quantílica logística.

3.1. Què és un quantil?

El concepte de quantil és el concepte bàsic que s'ha d'entendre per tal de comprendre els dos tipus de regressions que s'utilitzen en aquest treball final de grau.

Es defineix el quantil d'ordre p suposant que es disposa d'una mostra d'observacions d'una variable Y amb N observacions definides amb el subíndex t i amb funció de distribució $F\ (Y)$:

$$Y_t: t = 1,2, \dots, N\ amb\ F(Y).$$

Es defineix el quantil d'ordre p de la mostra, on $0 < p < 1$, com el valor que deixa una proporció p d'observacions per sota i una proporció $1 - p$ per sobre. Alguns exemples concrets són els següents:

- La mediana (quantil d'ordre $0{,}50$ o percentil 50) deixa una proporció del 50% de les observacions per sota i una proporció del 50% de les observacions per sobre. Es pot deduir que $F(Mediana) = 0{,}50$ i es defineix:

$$Mediana = F^{-1}(0{,}50).$$

- El quantil d'ordre $0'15$ o percentil 15 deixa una proporció del 15% de les observacions per sota i una proporció del 85% de les observacions per sobre. En aquest cas, la funció de distribució pren el valor de $F(Y) = 0{,}15$ i es defineix:

$$Quantil\ 0{,}15 = F^{-1}(0{,}15).$$

Una forma alternativa de definir un quantil és mitjançant la següent expressió:

$$\min_{b \epsilon \mathbb{R}} \left[\sum_{y_i \geq b} p|y_i - b| + \sum_{y_i \leq b} (1-p)|y_i - b| \right].$$

On p és el quantil, y els diferents valors de les observacions de la mostra per a la variable Y i b el valor que minimitza la expressió. És fàcilment demostrable que el valor b que minimitza l'expressió és el que deixa una proporció p de la mostra per sota i una proporció $(1 - p)$ per sobre, on $p \epsilon [0,1]$.

(Vicéns i Sánchez, 2012)

3.2. La regressió quantílica

El primer aspecte que cal tenir clar és que els orígens de la regressió quantílica (*Koenker i Basset, 1978)* estan molt relacionats amb la regressió lineal. Tot i així, el model de regressió quantílica s'estima mitjançant la minimització asimètrica ponderada dels errors absoluts, fet que representa una alternativa als mínims quadrats ordinaris en què, per estimar els paràmetres, es minimitza la suma dels errors al quadrat. Tal i com s'ha remarcat anteriorment, l'estimació de la regressió quantílica està relacionada amb la presència entre les dades a modelitzar de valors atípics, heteroscedasticitat o canvi estructural. En aquests casos, la mitjana condicional de la variable resposta (estimada per MQO) donada una sèrie de valors de les variables exògenes no és la millor estimació i és molt útil l'estimació d'una regressió quantílica per millorar la qualitat de les estimacions.

En aquestes situacions, la regressió quantílica permet estimar els paràmetres de diferents models lineals associats a diferents quantils de la variable resposta del model. D'aquesta manera, l'estimació està menys perjudicada pels inconvenients citats: valors atípics, canvi estructural... L'especificació del model de regressió quantílica associat al quantil p de la variable Y presenta la següent forma:

$$y(p) = X * \beta_p + u_p.$$

On $y(p)$ és el vector amb els valors del quantil p de la variable dependent o endògena del model, que també denotem per $Quant_p(y|X) = X * \beta_p$, X és la matriu amb els valors de les s variables independents o exògenes del model, β_p és el vector dels paràmetres del model associats al quantil p de la variable Y i u_p és la pertorbació aleatòria corresponent a aquest model de regressió. En aquest cas, es té que $Quant_p(y|X) = X * \beta_p$, que implica:

$$Quant_p(u_p|X) = 0.$$

Aquesta és l'única suposició que es fa sobre la pertorbació aleatòria del model de regressió quantílica associat al quantil p de la variable Y. Per tant, es poden tenir tants models de regressió com quantils de la variable Y s'estiguin considerant.

Per tal d'explicar com s'estimen els coeficients dels diferents models de regressió, es parteix de la definició alternativa de quantil plantejada en l'apartat anterior. En aquesta, el valor b correspon al quantil p d'Y que minimitza la funció. Si es considera que el valor b de l'expressió anterior és una simplificació del producte $X * \beta_p$ quan $X = 1$ on 1 és el vector unitari, llavors es té que el problema de l'estimació de paràmetres en regressió quantílica és:

$$\min_{\beta_p \epsilon \mathbb{R}^s} \left[\sum_{y \geq X*\beta_p} p|y - X*\beta_p| + \sum_{y < X*\beta_p} (1-p)|y - X*\beta_P| \right].$$

Tal i com s'ha dit abans, es pot veure que en aquesta expressió, es porta a terme una minimització de les desviacions absolutes ponderades amb pesos asimètrics. És a dir, a cada observació i, se li dona més o menys pes en funció del quantil d'Y (p) del qual s'estiguin estimant els paràmetres del model de regressió quantílica.

(Vicéns i Sánchez, 2012)

Per últim, és important remarcar que la idea que es pot estimar una regressió quantílica mitjançant la segmentació de Y en subconjunts d'acord amb la seva distribució no condicional d'y i després estimar les diferents regressions per MQO és errònia i no dona lloc als mateixos resultats que en el cas d'estimacions de models de regressió quantílica. *(Cortés, 2019)*

3.3. La regressió quantílica logística

La regressió quantílica logística és un cas especial i concret de regressió quantílica. Tal i com s'ha dit abans, la particularitat que té aquesta regressió és que la variable dependent està definida amb un interval de valors concret, és a dir:

$$y \epsilon [y_{min}, y_{max}].$$

A continuació, s'explica la terminologia d'aquest tipus de regressió, com es parametritza i com es procedeix, un cop estimats els paràmetres, a la inferència.

En primer lloc, cal tenir clares les característiques especials de la variable dependent y que s'acaben de comentar. Seguidament, es defineix la matriu X, com la matriu que conté els valors de les s variables independents del model per a les n observacions del model:

$$\mathrm{X} = (\mathrm{x_{i1}}, \dots, \mathrm{x_{is}}).$$

On $i = 1, \dots, n$ i fa referència a cadascuna de les observacions utilitzades per estimar el model. Es defineix el quantil p d'Y donada la matriu de covariables X com $\mathrm{Q_y(p)}$, on es recorda que p és una proporció entre 0 i 1. Un exemple concret és la mediana condicional ($\mathrm{Q_y}(0'5)$), la qual fa referència al valor de y que divideix la distribució condicional de la variable dependent del model amb dues parts amb la mateixa probabilitat.

Tanmateix, s'assumeix que per a cada quantil p d'Y que s'estigui considerant, existeix un vector amb les estimacions dels paràmetres del model ($\beta_p = \{\beta_{\mathrm{p},0}, \beta_{\mathrm{p},1}, \dots, \beta_{\mathrm{p,s}}\}$). Les estimacions d'aquests paràmetres es realitzen d'una manera semblant al cas de la regressió quantílica amb algunes diferències. Per a mostrar aquestes diferències, es

parteix de l'expressió que ha servit per explicar l'estimació dels paràmetres en el cas de la regressió quantílica i s'aplica al cas de la regressió quantílica logística:

$$\min_{\beta_p \epsilon \mathbb{R}^s} \left[\sum_{y \geq h^{-1}(X*\beta_p)} p\left|y - h^{-1}(X*\beta_p)\right| + \sum_{y < h^{-1}(X*\beta_p)} (1-p)\left|y - h^{-1}(X*\beta_P)\right| \right].$$

On h és una funció no decreixent coneguda (h) que va de l'interval d'y $(\mathrm{y_{min}}, \mathrm{y_{max}})$ a la recta real. Respecte aquesta funció (h), hi ha diferents opcions per les quals es pot optar. Entre aquestes opcions, destaquen la funció pròbit, la funció $loglog$ o la transformació logística. La transformació logística és la que s'utilitza en aquest treball i es defineix de la següent manera:

$$\mathrm{h\,(y) = logit\,(y) = log\,(\frac{y - y_{min}}{y_{max} - y})}.$$

Es recorda que per tal de poder calcular sempre aquest quocient, s'ha de garantir que ni el numerador ni el denominador prenguin el valor de 0. Això es pot aconseguir realitzant els següents canvis en el quocient:

$$y_{min} = \min(y) - \varepsilon.$$

$$y_{max} = \max(y) + \varepsilon.$$

On ε és un valor suficientment petit que faci que $\min(y) \neq y_{min}$ i $\max(y) \neq y_{max}$. D'aquesta manera, es garanteix que ni el numerador ni el denominador prenguin el valor 0 i no es tinguin problemes a l'hora de calcular el log.

Un cop estimats els paràmetres β_p, s'estima el quantil condicional p de la variable Y segons la següent expressió:

$$\mathrm{h\{Q_y(p)\} = \beta_{p,0} + \beta_{p,1}x_{i1} + \cdots + \beta_{p,s}x_{is}}.$$

Si es realitza la transformació inversa d'aquesta expressió per obtenir el valor de $\mathrm{Q_y}(p)$, es té que:

$$\mathrm{Q_y(p) = h^{-1}(\beta_{p,0} + \beta_{p,1}x_{i1} + \cdots + \beta_{p,s}x_{is})}.$$

La qual implica,

$$\mathrm{Q_y}(p) = \frac{\exp\left(\beta_{p,0} + \beta_{p,1}x_{i1} + \cdots + \beta_{p,s}x_{is}\right)y_{max} + y_{min}}{\exp\left(\beta_{p,0} + \beta_{p,1}x_{i1} + \cdots + \beta_{p,s}x_{is}\right) + 1}.$$

(Bottai et al., 2010)

4. Regressió quantílica: Modelització del pes dels nadons al néixer

4.1. Descripció de les dades i descriptiva inicial

Tal i com s'ha dit anteriorment, en aquest primera aplicació, s'estudia, s'analitza, s'estima i s'interpreta la regressió quantílica. Per fer-ho, s'utilitza una base de dades publicada pel centre nacional d'estadístiques mèdiques dels Estats Units l'any 1997, la qual fa referència a unes dades detallades de natalitat. Per conveniència, no s'utilitzen totes les dades, sinó que es procedeix a partir d'una selecció de 50.000 mares de la mateixa. Aquesta selecció s'ha adoptat de la referència de la qual s'han obtingut les dades (*Koenker et al., 2001*). Les observacions que contenien dades mancants havien estat prèviament esborrades.

Per tant, la mostra amb la que es treballa en aquesta primera aplicació està formada per 50.000 dones que han donat a llum als Estats Units, l'edat de les quals està compresa entre els 18 i els 45 anys. Les variables d'aquesta base de dades es presenten, s'expliquen i es codifiquen en la següent taula 4.1.1:

Taula 4.1.1: Variables de la base de dades (nom, descripció i codificació)

Nom	**Descripció**
Pes	Variable quantitativa que mesura amb $grams$ el pes del nadó al néixer.
$Etnicitat$	Variable categòrica que determina l'etnicitat de la mare: - Dona de raça negra: 1 - Dona de raça blanca: 0
$Casada$	Variable categòrica que determina si la dona estava casada en el moment del naixement: - Dona casada: 1 - Dona no casada: 0
Nen	Variable categòrica que determina si el nounat és nen o nena: - Nen: 1 - Nena: 0
$Edat$	Variable quantitativa que indica l'edat de la dona en $anys$ en el moment en què es va produir el naixement.
$Fumadora$	Variable categòrica que indica si la mare era fumadora durant l'embaràs: - Mare fumadora: 1 - Mare no fumadora: 0
$Cigarrets/dia$	Variable quantitativa que indica el nombre de cigarrets al dia que fumava la mare durant l'embaràs.
$Pes\ guanyat$	Variable quantitativa que indica el pes guanyat amb $quilograms$ per la mare durant l'embaràs.
$Visita\ prenatal$	Variable categòrica que indica si la mare va realitzar visites mèdiques prenatals (En cas que n'hagi realitzat vàries, s'indica la última): - Sense visites prenatals: 0 - Visita prenatal 1r. trimestre embaràs: 1 (Categoria de referència) - Visita prenatal 2n. trimestre embaràs: 2

	- Visita prenatal 3r. trimestre embaràs: 3
$Educació$	Variable categòrica que indica el nivell educatiu de la mare: - Menys del graduat escolar: 0 - Graduat escolar: 1 (Categoria de referència) - Estudis universitaris (sense graduat universitari): 2 - Graduat universitari: 3

Un cop presentades i explicades les variables que s'utilitzen per al desenvolupament d'aquesta primera aplicació, cal tenir amb compte tres consideracions:

- Respecte la variable $Etnicitat$ no queda clar si hi havia dones d'altres races o simplement van ser excloses de l'estudi.
- En la base de dades original, la variable $Edat$ estava centrada en la seva mediana ($27\ anys$). Per tal de facilitar la interpretació dels resultats, s'ha decidit descentrar aquesta variable segons la següent expressió:

$$Edat\ (anys) = Edat(BB.DD.Original) + 27.$$

- Tanmateix, en la base de dades original, la variable $Pes\ Guanyat$ estava expressada amb lliures i centrada en la seva mediana ($30\ lliures$). Seguint amb el procediment emprat per a la variable $Edat$, s'ha decidit descentrar aquesta variable i passar-la a $quilograms\ (kg)$ segons la següent expressió:

$$Pes\ guanyat\ (kg) = (Pes\ guanyat\ (BB.DD.Original) + 30) * 0{,}45.$$ [3]

- En cas que la variable $Fumadora$ sigui igual a '0', indicant que la mare no ha sigut fumadora durant l'embaràs, la variable $Cigarrets/dia$ també pren el valor de '0'.

A continuació, es procedeix a fer una anàlisi descriptiva individual per a cadascuna de les variables. En primer lloc, es comencen mostrant en la següent taula 4.1.2 els estadístics bàsics per a cadascuna d'elles obtinguts utilitzant la funció *summary ()* d'R:

Taula 4.1.2: Estadístics descriptius univariants bàsics de les diferents variables

Variable	Min.	1er. Quartil	Mediana	Mitjana	3er. Quartil	Max.
Pes	240,00	3062,00	3402,00	3371,00	3720′00	6350,00
$Etnicitat$	0,00	0,00	0,00	0,16	0,00	1,00
$Casada$	0,00	0,00	1,00	0,71	1,00	1,00
Nen	0,00	0,00	1,00	0,51	1,00	1,00
$Edat$	18,00	23,00	27,00	27,42	32,00	45,00
$Fumadora$	0,00	0,00	0,00	0,13	0,00	1,00
$Cigarrets/dia$	0,00	0,00	0,00	1,48	0,00	60,00
$Pes\ guanyat$	0,00	9,90	13,50	13,82	17,55	44,10
$Visita\ prenatal$	0,00	3,00	3,00	2,70	3,00	3,00
$Educació$	0,00	0,00	1,00	1,22	2,00	3,00

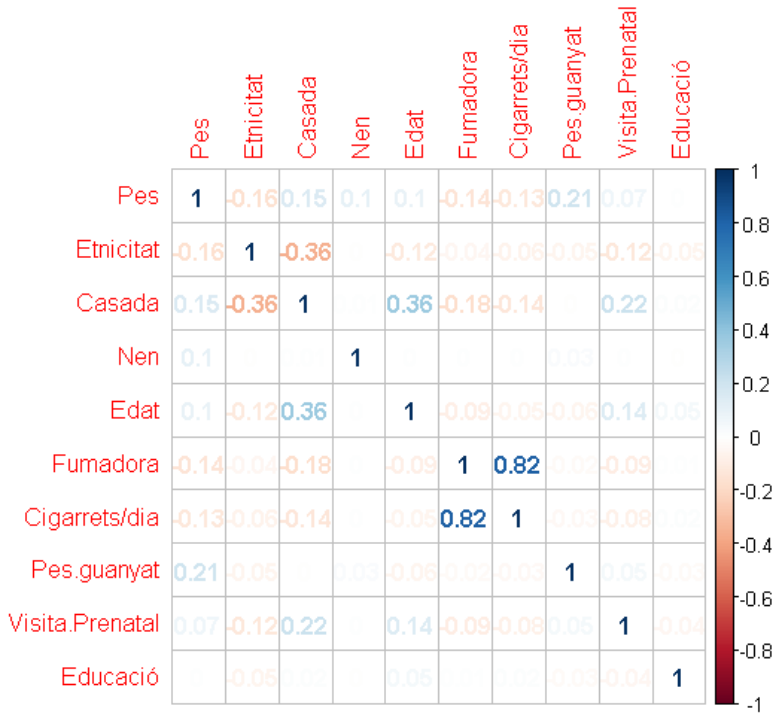
[3] Es recorda que una lliure (*lbs* en anglès) equival a 0,45 quilograms.

Aquests són els estadístics bàsics que formarien l'anàlisi descriptiva univariant inicial. Algunes consideracions a tenir amb compte respecte la taula són:

- Tots els nombres es mostren amb dos xifres decimals per tal de mantenir un criteri d'uniformitat.
- Per a les variables categòriques binàries[4], la mitjana que es mostra a la taules representa la proporció de valors '1' que té aquella variable.
- Respecte els estadístics descriptius mostrats de la variable $Pes\ Guanyat$, cal destacar el valor màxim ($44{,}10\ kg$) i el valor mínim ($0{,}00\ kg$) mostrats. Es tracta de valors molts extrems (tant per excés com per defecte).
- Tal i com s'ha pogut deduir, la variable Pes és la variable dependent o endògena dels diferents models estimats. Un apunt que cal fer al respecte és el reduït valor que presenta per a l'estadístic que especifica el valor mínim de la variable (Min.) i l'elevat valor que presenta per al valor màxim de la variable (Max.). Tal i com es pot veure, aquests valors resulten ser de 240,00 i 6350,00. Tal i com succeeix amb la variable $Pes\ Guanyat$, es tracta de valors molts extrems (tant per excés com per defecte).

Seguint amb l'anàlisi descriptiva iniciada, es calculen les correlacions lineals entre les diferents variables presentades i es mostren en la següent taula 4.1.3. S'utilitzen les funcions *cor ()* i *corrplot ()* d'R:

Taula 4.1.3: Taula de correlacions lineals entre les diferents variables



[4] Les variables categòriques binàries poden prendre només dos valors: ′0′ i ′1′. En el nostre cas, aquestes variables són *Etnicitat, Casada, Nen i Fumadora.*

De la taula anterior, la primera fila o la primera columna són importants ja que mostren les correlacions lineals entre la variable dependent i les variables exògenes dels models que s'estimen. Per tal de tenir clars aquestes valors, es mostren en la següent taula 4.1.4:

Taula 4.1.4: Correlacions lineals de les variables exògenes amb la variable Pes

Variable	Correlació lineal amb *Pes*
Pes	1,00
$Etnicitat$	$-0{,}16$
$Casada$	0,15
Nen	0,10
$Edat$	0,10
$Fumadora$	$-0{,}14$
$Cigarrets/dia$	$-0{,}13$
$Pes\ guanyat$	0,21
$Visita\ Prenatal$	0,07
$Educació$	$< 0{,}01$

Arribats a aquest punt, caldria recordar que correlació no implica causalitat. Es pot afirmar que correlació simplement significa associació de fet. Degut a que en les correlacions mostrades en la taula anterior, la variable dependent del model (Pes) hi està involucrada, es considera que una correlació superior a $0{,}20$ en valor absolut s'ha de tenir amb compte abans de procedir a l'estimació dels diferents models. Des d'aquest punt de vista, cal destacar la correlació lineal entre les variables Pes i $Pes\ Guanyat$, la qual pren un valor de $0{,}21$.

Tanmateix, hi ha altres coeficients de correlació lineals de la taula 4.1.3 que cal tenir amb compte. Aquests fan referència als valors de les correlacions lineals entre variables explicatives del model. En aquest cas, es considera que un coeficient de correlació lineal superior a $0{,}30$ amb valor absolut, cal tenir-lo amb compte. Es mostren en la següent taula 4.1.5 les correlacions lineals a tenir amb compte amb les variables que hi estan implicades:

Taula 4.1.5: Correlacions a tenir amb compte entre variables explicatives del model

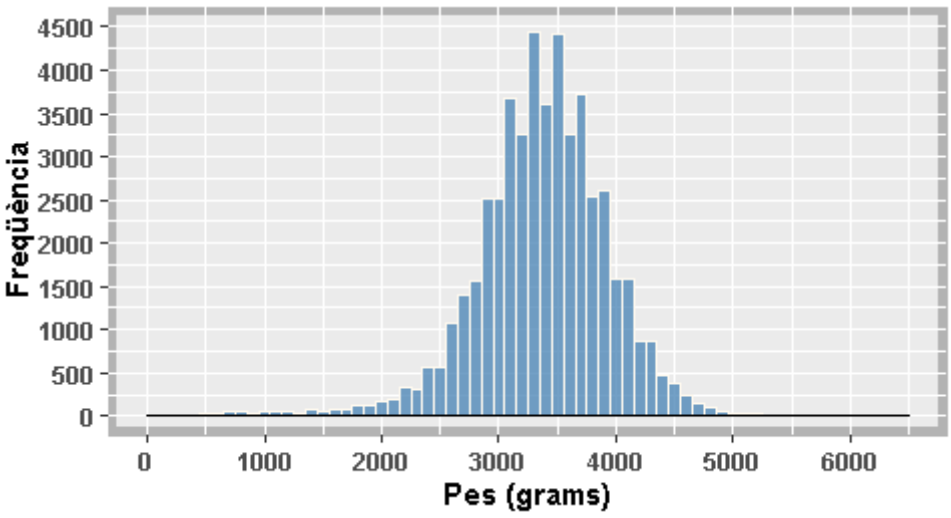
Variables	Correlació lineal
$Etnicitat - Casada$	$-0{,}35$
$Edat - Casada$	0,36
$Cigarrets/dia - Pes$	0,82

Un cop realitzada l'anàlisi descriptiva inicial de les dades que formen part de l'estudi, a continuació, es procedeix a realitzar una anàlisi gràfica, tot mostrant els gràfics més

adequats tenint amb compte el tipus de variable i quina és la característica concreta que s'està mesurant en cada cas.

En primer lloc, es comença per la variable dependent dels models estimats: Pes. Degut a que es tracta d'una variable quantitativa contínua, la millor manera de visualitzar-ne la seva distribució és mitjançant un histograma. Aquest es pot veure en el següent gràfic 4.1.6:

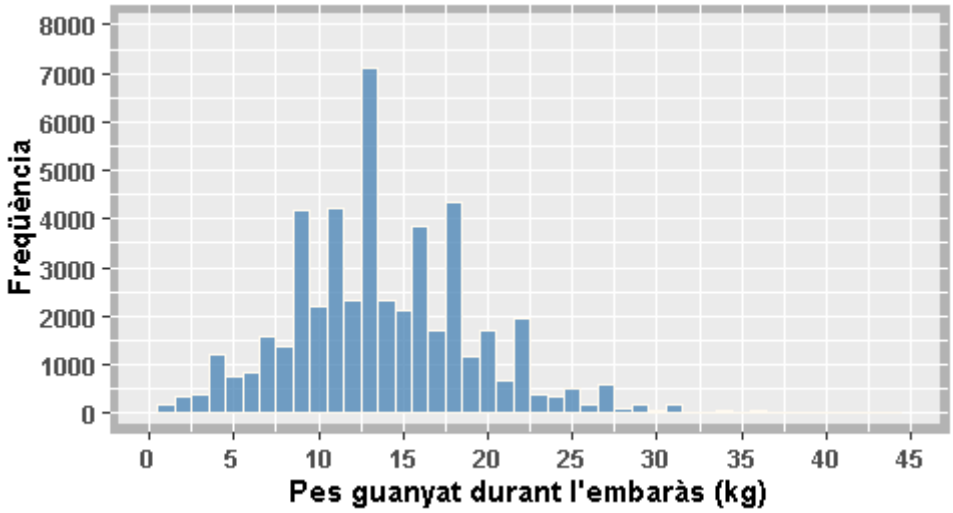
Gràfic 4.1.6: Histograma de la variable Pes



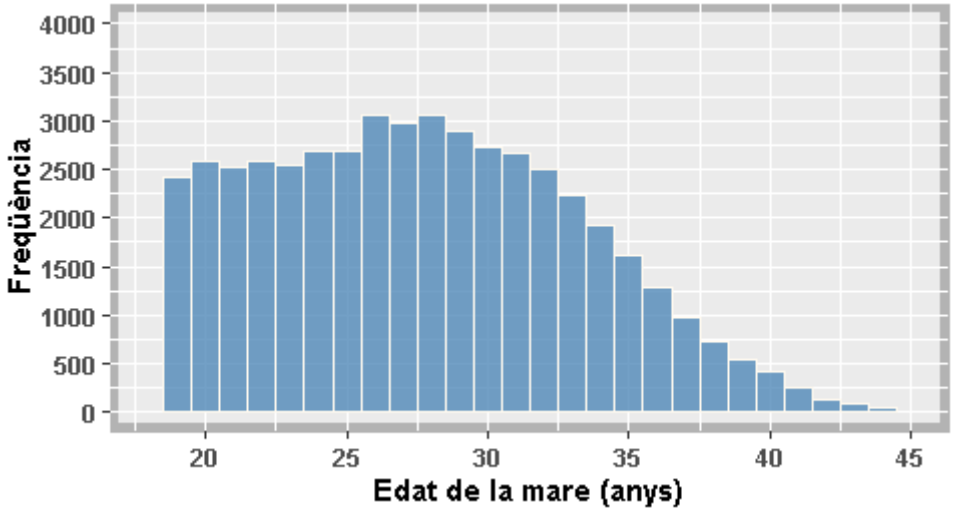
Mitjançant l'observació d'aquest histograma, es pot concloure que la variable dependent del model segueix aproximadament una distribució normal en què una gran proporció de les dades es troben entre els valors de 3000 i 4000 $grams$. En relació amb aquesta afirmació, si es visualitza la taula 4.1.2, es pot observar que els estadístics primer quartil, mitjana, mediana i tercer quartil d'aquesta variable prenen valors que es troben en aquest interval de valors. En l'histograma, s'hi poden observar alguns valors *outliers*, els quals estan relacionats amb el mínim i el màxim d'aquesta variable.

Seguint amb aquesta anàlisi gràfica, es procedeix realitzant gràfics per tal de visualitzar la distribució d'algunes de les variables explicatives. En primer lloc, es fa referència a les variables quantitatives. Per tant, es procedeixen a realitzar histogrames per a les variables $Pes\ Guanyat\ i\ Edat$ i un diagrama de barres per a la variable $Cigarrets/dia$. Amb aquests gràfics, es visualitza la distribució dels valors que presenten aquestes tres variables. Cal remarcar que en el cas de la variable $Cigarrets/dia$, es realitza un diagrama de barres ja que es tracta d'una variable quantitativa discreta, la qual només pot prendre valors enters no negatius. Aquests tres gràfics es mostren en les següents regions gràfiques 4.1.7, 4.1.8 i 4.1.9:

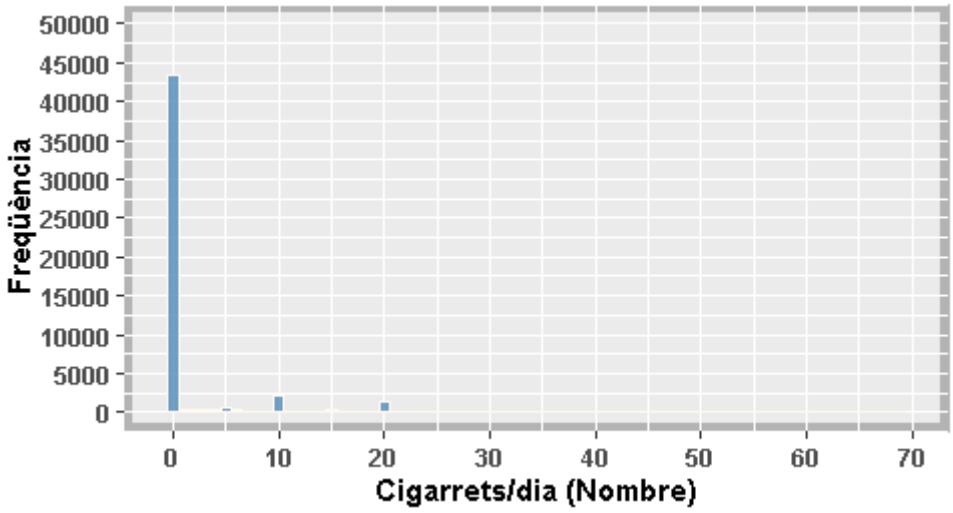
Gràfic 4.1.7: Histograma de la variable $Pes\ Guanyat$



Gràfic 4.1.8: Histograma de la variable $Edat$



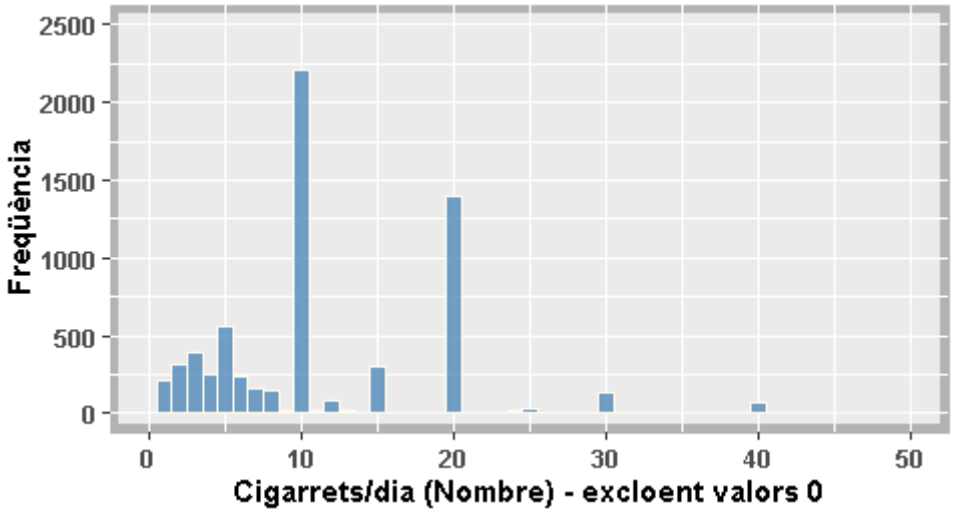
Gràfic 4.1.9: Diagrama de barres per a la variable $Cigarrets/dia$



Cal recordar que la mida de la mostra que s'ha utilitzat per a realitzar aquests tres gràfics és de $n = 50.000\, mares$, la qual és la mida de la mostra amb la que s'està treballant. Observant els dos histogrames, les hipòtesis de distribució normal haurien de ser rebutjades per als valors de les variables $Pes\ Guanyat$ i $Edat$. En ells, s'observa asimetria i una variabilitat associada que no és la característica d'una distribució normal.

Fent referència al diagrama de barres associat a la variable $Cigarrets/dia$, aquest no és gaire informatiu ja que aquesta variable presenta molts valors 0, els quals fan referència a les mares que no fumen. Per tal de tenir una millor visualització de les dades d'aquesta variable, es presenta el següent diagrama de barres com el gràfic 4.1.10, en el que s'han exclòs les observacions en què $Cigarrets/dia$ pren el valor de 0:

Gràfic 4.1.10: Diagrama de barres per a la variable $Cigarrets/dia$ (excloent valors 0)



Si s'exclouen els valors en què la variable $Cigarrets/dia$ és 0, es pot observar que la majoria de dades es troben entre 0 i 20: al voltant del 75% de les dades. Per altra banda, es poden observar alguns valors anòmals (*outliers*). Per a aquest segon diagrama de barres, s'ha utilitzat una mida de mostra de $n = 6533\, mares$, el qual representa un $13'07\%$ del total de la mostra. Per últim, s'ha comprovat si per a totes les mares no fumadores ($Fumadora$ pren el valor de 0), la variable $Cigarrets/\, dia$ també pren el valor 0. El resultat ha estat positiu i des d'aquest punt de vista, es pot dir que la base de dades no és incongruent en relació a aquestes dues variables.

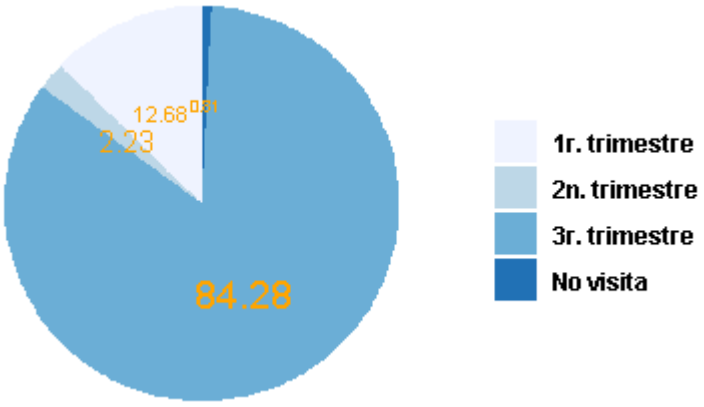
Havent analitzat gràficament les variables quantitatives, a continuació es procedeix a realitzar el mateix tipus d'anàlisi per a les variables categòriques. En primer lloc, es mostra en la següent taula 4.1.11, la taula de freqüències associada a la variable $Visita\ Prenatal$:

Taula 4.1.11: Taula de freqüències per a la variable *Visita Prenatal*

	Visita Prenatal			
	No visita	1r. Trimestre	2n. Trimestre	3r. Trimestre
Freqüència	403	6.339	1.114	42.144
Percentatge (%)	0,81%	12,68%	2,23%	84,28%

Per tal de veure aquestes dades d'una forma gràfica, es mostra el següent diagrama de pastís. Degut al tipus de variable que es tracta (variable categòrica amb poques categories), és el gràfic més correcte per a la seva representació gràfica. Aquest es mostra en el següent gràfic 4.1.12:

Gràfic 4.1.12: Diagrama de pastís per a la variable *Visita Prenatal*



Es pot observar que la majoria de mares han realitzat la última visita prenatal a l'últim trimestre de l'embaràs (fins a un 84'28%). De la mateixa manera, resulta estrany el fet que hi hagi mares que hagin realitzat la última visita prenatal durant el primer o segon trimestre d'embaràs. S'haurien d'estudiar les circumstàncies per les quals aquestes mares no han realitzat aquesta última visita en el tercer trimestre d'embaràs. Respecte les mares que no han assistit a cap visita prenatal, es creu que aquesta ha estat la seva preferència i així ho han volgut o bé no han pogut assistir-hi per algun tipus d'impediment de caràcter econòmic, social o assistencial.

Respecte les mares que no han realitzat la última visita prenatal en l'últim trimestre d'embaràs (sobretot les que ho han fet en el 2n. trimestre), es té la presumpció que el seu nadó va néixer abans de temps i no van ser a temps de realitzar una visita prenatal en l'últim trimestre d'embaràs. Si fos així, es considerarà que, el determinat pes del nounat es deu al moment en el que s'ha produït el naixement i no a característiques de la mare. Si s'arriba a aquesta conclusió, caldrà tenir-ho amb consideració a l'hora de procedir amb les diferents estimacions.

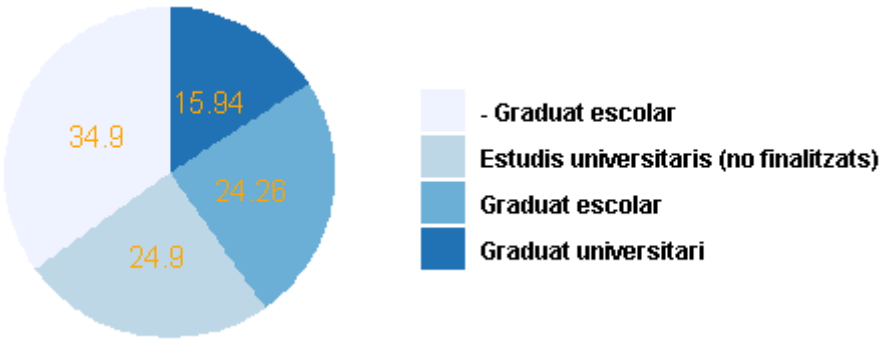
L'altre variable categòrica amb més d'una categoria de la qual es disposa és la variable *Educació*, la qual reflecteix l'educació de la mare. Es mostra en la següent taula 4.1.13, la taula de freqüències per als valors d'aquest variable:

Taula 4.1.13: Taula de freqüències per a la variable *Educació*

	Educació			
	− Graduat	*Graduat*	*Universitat (NG)*	*Universitat*
Freqüència	17.449	12.129	12.449	7.973
Percentatge (%)	34′90%	24′26%	24′90%	15′95%

De la mateixa manera que en el cas anterior, es mostra el següent diagrama de pastís en la següent regió gràfica 4.1.14 en el que es pot veure la distribució de valors que presenta aquesta variable:

Gràfic 4.1.14: Diagrama de pastís per a la variable *Educació*



Es pot observar que la mostra està força equilibrada pel que fa el nivell educatiu de la mare que donarà a llum entre les quatre categories possibles. De tota manera, cal ressaltar el nombre més alt de mares sense el graduat escolar (34,90% de la mostra).

Per tal de finalitzar l'anàlisi descriptiva univariant, es tenen amb compte les quatre variables categòriques binàries de les que es disposa: *Etnicitat*, *Nen*, *Fumadora* i *Casada*. Es presenten en la següent taula 4.1.15 les proporcions dins la mostra de cadascuna de les categories d'aquestes variables:

Taula 4.1.15: Proporcions de les categories de les variables categòriques binàries

	Valor	
Variable	0	1
Etnicitat	83′72%	16′28%
Nen	48′42%	51′58%
Fumadora	86′93%	13′07%
Casada	28′74%	71′26%

En aquesta taula 4.1.15, es pot observar que a la mostra amb la que s'està treballant, hi abunden les mares d'etnicitat de raça blanca, no fumadores i casades. Pel que fa al sexe dels nounats, la mostra està força equilibrada pel que fa al nombre de nens i nenes.

4.2. Descripció multivariant: *Profiling* de la variable resposta

Abans de procedir a l'estimació dels diferents models de regressió que s'estimen en aquesta primera aplicació, s'estudia, en aquest apartat, si hi ha diferències en els valors que pren la variable Pes en funció dels valors que prenen les diferents variables explicatives. En estadística, aquest procés rep el nom de realitzar un *profiling* de la variable resposta.

Per tal de procedir, cal tenir amb compte que l'anàlisi és diferent en funció de si la variable explicativa a la qual s'està condicionant la distribució de la variable resposta és quantitativa o categòrica. Es comença, en primer lloc, amb les variables categòriques. En la següent taula 4.2.1, es mostren diferents estadístics bàsics per als valors de la variable resposta (Pes) en funció de les diferents categories que presenten les variables categòriques de la mostra:

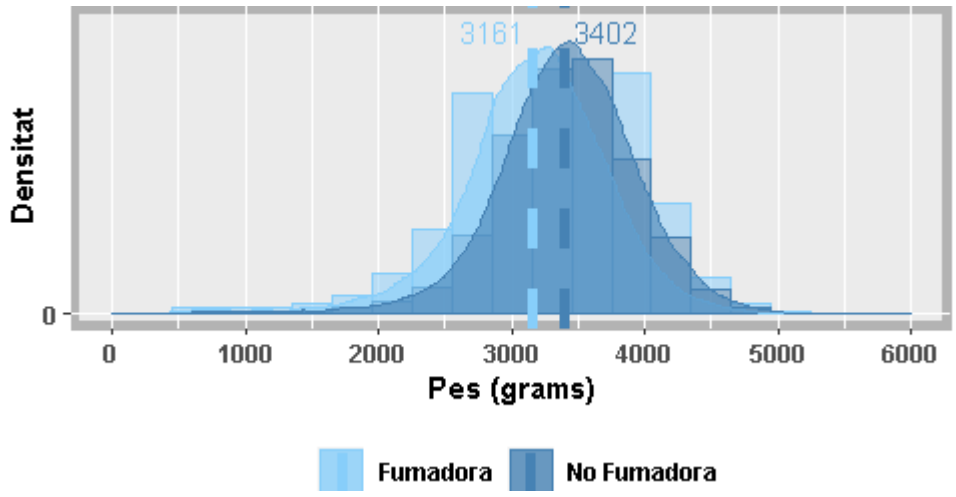
Taula 4.2.1: Estadístics bàsics dels valors de la variable Pes en funció de les categories de les variables categòriques

Variable/Estadístic	Mitjana	Desv. Est.	Mín.	Màx.
Sexe del nadó				
Nen	3.427,00	577,68	284,00	5.970,00
Nena	3.311,00	547,74	240,00	6.350,00
Etnicitat				
Negra	3.163,00	613,68	240,00	6.350,00
Blanca	3.411,00	547,62	284,00	5.970,00
Estat civil				
Casada	3.426,00	551,78	240,00	5.970,00
No Casada	3.234,00	579,00	284,00	6.350,00
Hàbit tabàquic: Fumadora				
Sí	3.161,00	576,77	312,00	5.245,00
No	3.402,00	558,03	240,00	6.350,00
Última visita prenatal				
No visita	3.055,00	748,79	284,00	5.330,00
1r. trimestre	3.302,00	562,35	369,00	5.220,00
2n. trimestre	3.276,00	508,53	454,00	5.415,00
3r. trimestre	3.387,00	564,53	240,00	6.350,00
Educació				
-Graduat	3.337,00	579,66	240,00	5.415,00
Graduat	3.394,00	564,16	330,00	6.350,00
Universitat (NG)	3.467,00	531,05	322,00	5.642,00
Universitat	3.259,00	567,35	284,00	5.330,00

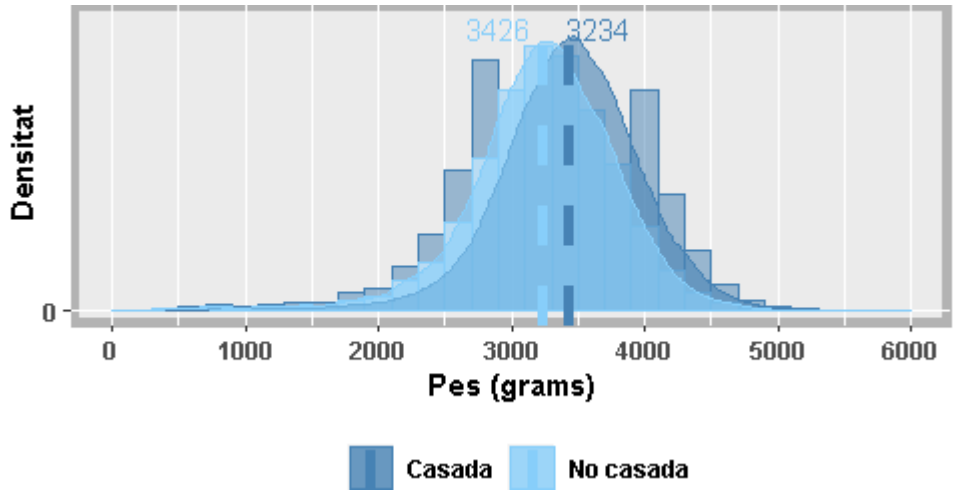
En els següents gràfics 4.2.2, 4.2.3 i 4.2.4, es mostren les distribucions condicionals dels valors de la variable Pes en funció de les categories de les variables Nen, $Etnicitat$ i

Casada. Es fa per a aquestes tres variables ja que són les que presenten més variabilitat en els valors de la variable *Pes* en funció de les seves categories.

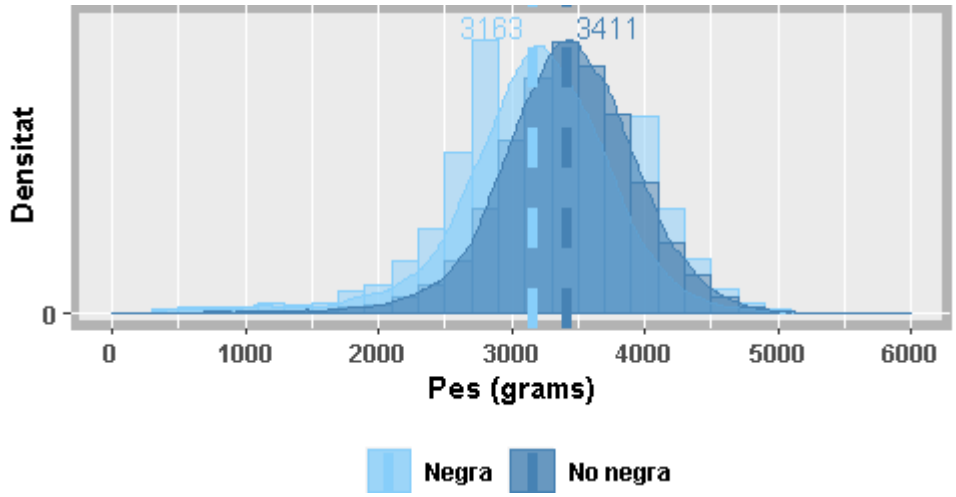
Gràfic 4.2.2: Histogrames i corbes de densitat dels valors de *Pes* en funció de *Fumadora*



Gràfic 4.2.3: Histogrames i corbes de densitat dels valors de *Pes* en funció de *Casada*



Gràfic 4.2.4: Histogrames i corbes de densitat dels valors de *Pes* en funció d'*Etnicitat*



Es pot veure que s'han utilitzat histogrames i corbes de densitat per tal de veure d'una forma més clara les diferències en la distribució dels valors de Pes en funció de les diferents categories. S'observa que les mares casades, de raça blanca i no fumadores solen presentar valors de la variable Pes més alts. Per altra banda, les variabilitats que presenten aquestes distribucions condicionades solen ser molt semblants independentment de la categoria que s'estigui considerant.

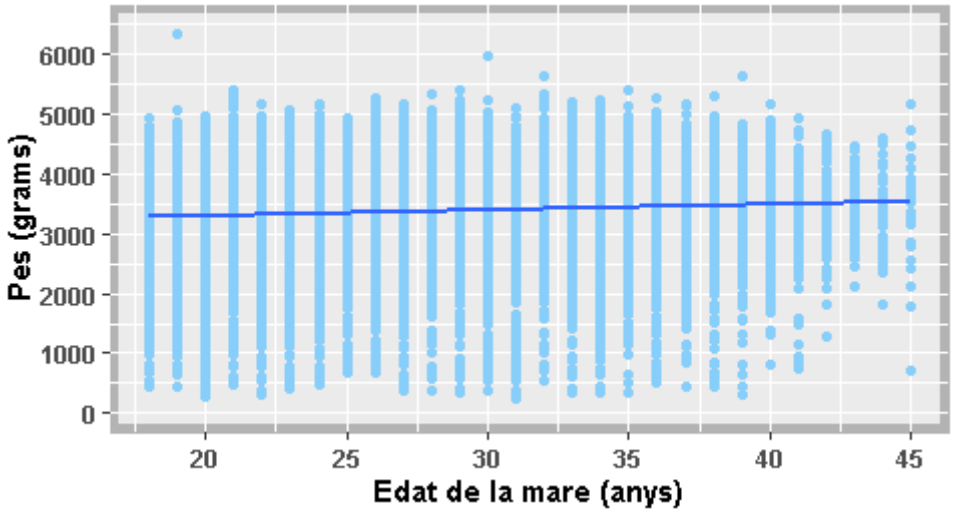
Seguidament, es procedeix amb el *profiling* de la variable Pes en funció de les variables quantitatives de la mostra. En primer lloc, es recorden en la següent taula 4.2.5 les correlacions lineals d'aquestes variables amb la variable resposta:

Taula 4.2.5: Correlacions de les variables quantitatives amb la variable resposta (Pes)

Variable	Correlació lineal amb Pes
$Edat$	$0'10$
$Pes\ Guanyat$	$0'21$
$Cigarrets/dia$	$-0{,}13$

Seguidament, es mostra en la següent regió gràfica 4.2.6 un diagrama de dispersió i la corresponent recta de regressió amb els valors de les variables Pes i $Edat$.

Gràfic 4.2.6: Diagrama de dispersió i recta de regressió per a $Pes\ i\ Edat$



La recta de regressió mostrada prové del model lineal estimat $Pes = \beta_0 + \beta_1 * Edat$. Es mostren els detalls d'aquesta estimació en la taula 4.2.7:

Taula 4.2.7: Model de regressió lineal $Pes = \beta_0 + \beta_1 * Edat$

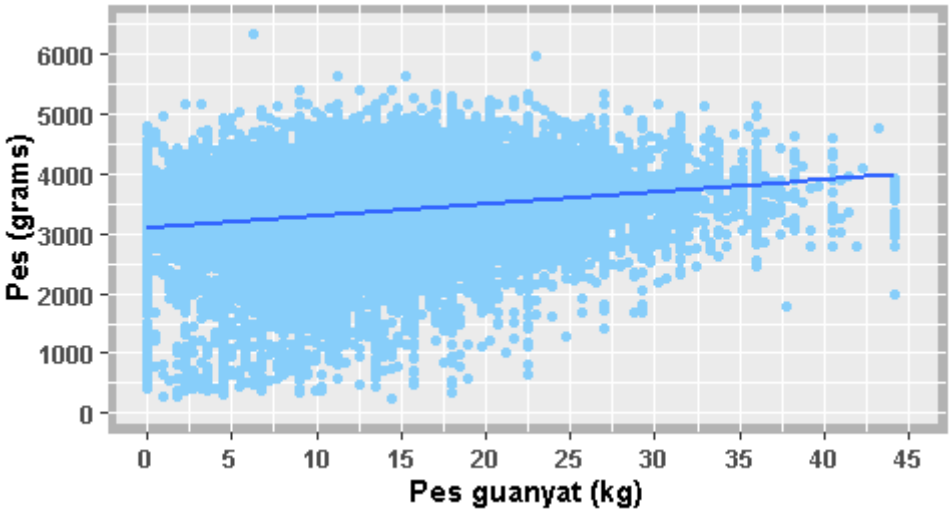
Variable	Coeficient	Error est.	Valor t	P-valor ($< \lvert t \rvert$)
$Constant$	3106,46	12,33	252,03	$< 0{,}01$
$Edat$	9,64	0,44	21,91	$< 0{,}01$
$Error\ estàndard\ residual = 563{,}7\ (49998\ g.ll.)$				

$R^2 = 0{,}01$	$Estadístic\ F = 479{,}9\ (1\ i\ 49998\ g.ll.)$
$R^2\ ajustat = 0{,}009$	$P - valor(< F) =< 0{,}01$

Es pot observar que el coeficient de bondat de l'ajust de la regressió resulta ser molt petit ($R^2 = 0{,}01$). Per tant, es pot dir que mitjançant una recta no es pot capturar la relació (o se'n pot capturar molt poca) entre les variables Pes i $Edat$ de la mostra.

Es realitza el mateix procediment utilitzant la variable quantitativa: $Pes\ Guanyat$. Es mostra el diagrama de dispersió i la recta de regressió en la següent regió gràfica 4.2.8:

Gràfic 4.2.8: Diagrama de dispersió i recta de regressió per a $Pes\ i\ Pes\ Guanyat$



La recta de regressió mostrada prové del model lineal estimat $Pes = \beta_0 + \beta_1 * Pes\ Guanyat$. Es mostren els detalls d'aquesta estimació en la taula 4.2.9:

Taula 4.2.9: Model de regressió lineal $Pes = \beta_0 + \beta_1 * Pes\ Guanyat$

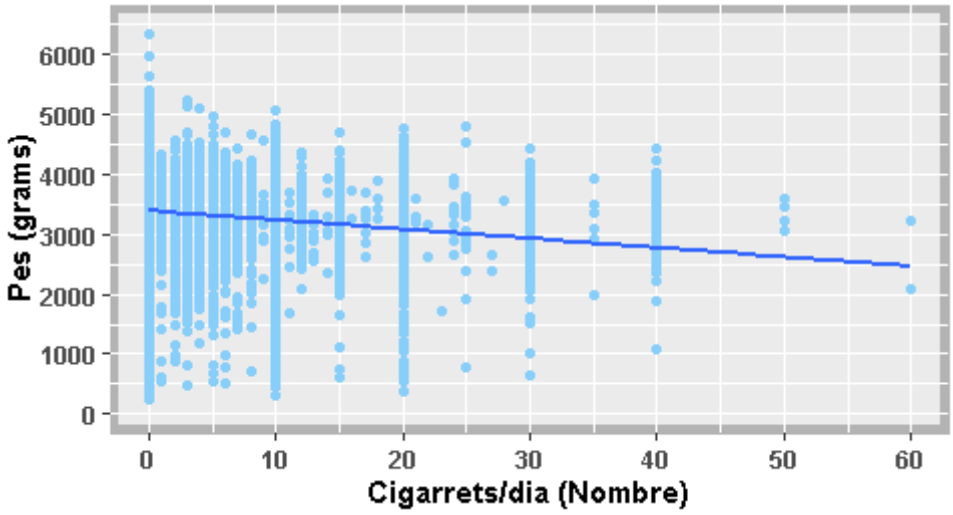
Variable	Coeficient	Error est.	Valor t	P-valor ($< \|t\|$)
$Constant$	3089,97	6,41	482,24	$< 0{,}01$
$Pes\ Guanyat$	20,32	0,43	47,52	$< 0{,}01$
$Error\ estàndard\ residual = 554\ (49998\ g.ll.)$				
$R^2 = 0{,}04$		$Estadístic\ F = 2258\ (1\ i\ 49998\ g.ll.)$		
$R^2\ ajustat = 0{,}04$		$P - valor(< F) =< 0{,}01$		

En aquest cas, sí que s'observa que es té un coeficient de bondat de l'ajust una mica més elevat que en el cas anterior ($R^2 = 0'04$), d'aquesta manera es pot dir que mitjançant una recta es pot capturar un cert grau de relació entre aquestes variables. Tanmateix, es pot veure que la recta de regressió estimada indica que existeix una relació positiva entre les variables $Pes\ Guanyat\ i\ Pes$. Respecte aquesta variable, es creu que pot ser poc informativa ja que un guany de pes de la mare segurament vingui explicat per un pes més elevat del nadó. Si és així, aquesta variable no estaria indicant una característica

de la mare sinó que, en aquest cas, es faria referència a una conseqüència del valor de la variable Pes. Més tard, es decidirà què es fa amb aquesta variable a l'hora d'estimar els paràmetres dels diferents models.

Per últim, es realitza el mateix procediment per a la variable $Cigarrets/dia$. Es mostra el corresponent diagrama de dispersió i la recta de regressió en la següent regió gràfica 4.2.10:

Gràfic 4.2.10: Diagrama de dispersió i recta de regressió per a $Cigarrets/dia\ i\ Pes$



La recta de regressió mostrada prové del model lineal estimat $Pes = \beta_0 + \beta_1 * Cigarrets/dia$. Es mostren els detalls d'aquesta estimació en la taula 4.2.11:

Taula 4.2.11: Model de regressió lineal $Pes = \beta_0 + \beta_1 * Cigarrets/dia$

Variable	Coeficient	Error est.	Valor t	P-valor ($< \|t\|$)
$Constant$	3393,59	2,64	1287,46	$< 0{,}01$
$Cigarrets/dia$	$-15{,}46$	0,54	$-28{,}64$	$< 0{,}01$
$Error\ estàndard\ residual = 561{,}8\ (49998\ g.ll.)$				
$R^2 = 0{,}02$		$Estadístic\ F = 820{,}4\ (1\ i\ 49998\ g.ll.)$		
$R^2\ ajustat = 0{,}02$		$P - valor(< F) =< 0{,}01$		

En aquest cas, segons la recta de regressió estimada, s'indica que hi ha una relació negativa entre les variables $Pes\ i\ Cigarrets/dia$. Si s'observa el valor del coeficient de bondat de l'ajust d'aquesta recta de regressió, es pot dir que mitjançant una recta només es pot capturar un cert grau de la relació existent entre els valors d'aquestes dos variables.

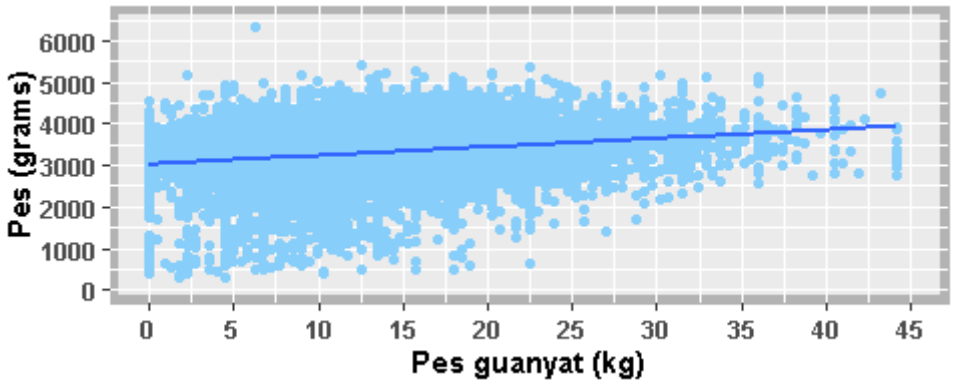
Per últim, es recorda la hipòtesi feta anteriorment per la qual s'ha dit que la variable $Pes\ Guanyat$ pot ser poc informativa ja que el $Pes\ Guanyat$ per una mare és una conseqüència del Pes del nounat i no n'és una causa. Relacionat amb aquesta hipòtesi, es creu que la relació entre les variables Pes i $Pes\ Guanyat$ pot variar en funció de

l'$Edat$ de la mare. A continuació, s'analitza si això succeeix en la mostra amb la que s'està treballant. Per fer-ho, es comença dividint la mostra en tres grups segons els valors de la variable $Edat$:

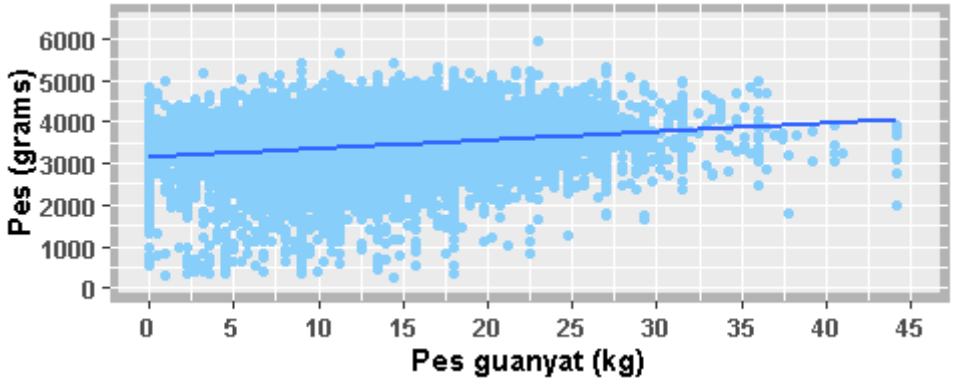
- Grup 1: Aquesta submostra inclou les mares amb una $Edat$ d'entre 18 i 27 anys. Es tenen 22929 observacions en aquest grup.
- Grup 2: Aquesta submostra inclou les mares amb una $Edat$ d'entre 27 i 36 anys. Es tenen 22568 mares en aquest grup.
- Grup 3: Aquesta submostra inclou les mares amb una $Edat$ d'entre 36 i 45 anys. Es tenen 4503 observacions en aquest grup.

Es realitza aquesta divisió ja que és la que es considera més oportuna tenint amb compte el rang de valors que pren la variable $Edat$ en la mostra amb la que s'està treballant. A continuació, es realitzen tres diagrames de dispersió amb els valors de les variables $Pes\ i\ Pes\ Guanyat$ per a cadascuna de les tres submostres establertes. Es mostren aquests diagrames de dispersió en les següents regions gràfiques 4.2.12, 4.2.13 i 4.2.14.

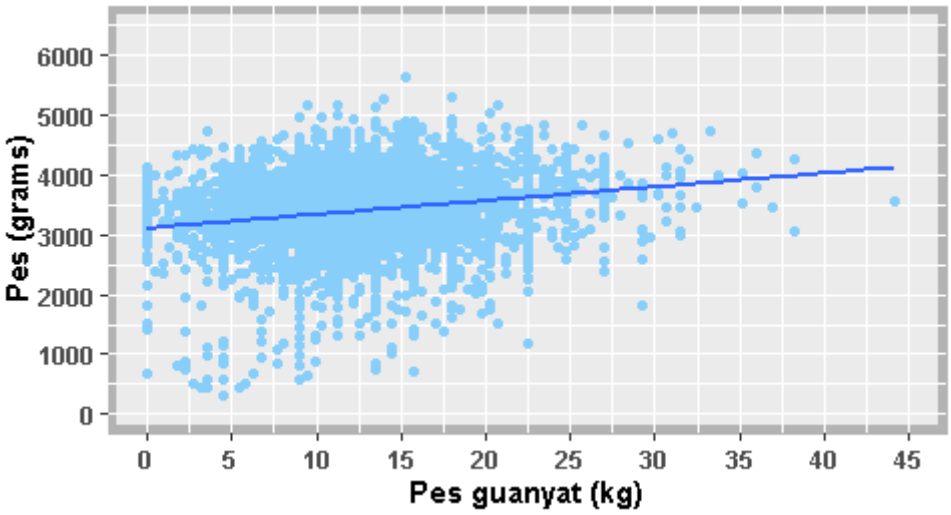
Gràfic 4.2.12: Diagrama de dispersió i recta de regressió per als valors de Pes i $Pes\ Guanyat$ de les observacions en què l'$Edat$ es troba entre 18 i 27 anys



Gràfic 4.2.13: Diagrama de dispersió i recta de regressió per als valors de Pes i $Pes\ Guanyat$ de les observacions en què l'$Edat$ es troba entre 27 i 36 anys



Gràfic 4.2.14: Diagrama de dispersió i recta de regressió per als valors de Pes i $Pes\ Guanyat$ de les observacions en què l'$Edat$ es troba entre 36 i 45 anys



Observant aquests tres diagrames de dispersió, es pot veure a simple vista que no hi ha diferència entre els tres grups establerts pel que fa a la relació entre els valors de les variables $Pes\ i\ Pes\ Guanyat$. L'única diferència observable és la diferent mida mostral de cadascun dels grups (S'observa aquest fet segons el número de punts que es veuen en cada diagrama de dispersió).

Es recorda que si s'hi haguessin vist diferències, caldria incloure la interacció $Edat * Pes\ Guanyat$ als models que s'estimen properament. Per acabar, cal tenir amb compte que es podrien realitzar contrastos d'hipòtesi en els que es determinaria si amb un cert grau de confiança, el paràmetre associat a la interacció entre les variables Pes i $Pes\ Guanyat$ de cadascun dels models lineals és estadísticament diferent de 0 o no.

Per últim, com a conclusió d'aquest apartat i utilitzant-se d'enllaç amb el següent apartat, es defineix el primer model lineal que s'estimarà:

- De les variables categòriques binàries, s'estima aquest primer model amb les variables de la mostra $Etnicitat, Casada\ i\ Nen$.
- De les variables quantitatives, s'estima el model amb $Edat\ i\ Cigarrets/dia$.
- $Pes\ guanyat$ no s'inclou en aquest primer model ja que es considera que és poc informativa. D'aquesta manera, es confirma la hipòtesi plantejada anteriorment mitjançant la qual s'ha dit que la variable $Pes\ Guanyat$ no indica una característica de la mare sinó que fa referència a una conseqüència del Pes del nounat.
- La variable $Fumadora$ no s'inclou en el model ja que la variable $Cigarrets/dia$ és molt més informativa. Per altra banda, la variable $Fumadora$ no indica cap altre tipus d'informació que $Cigarrets/dia$ no contingui.

- Per a les variables $Educació$ i $Visita\ Prenatal$, s'estableixen variables dicotòmiques per a cadascuna de les seves categories.

S'utilitza la següent taula 4.2.15 per a mostrar com es defineixen aquestes variables dicotòmiques. En aquest cas, es defineix la variable dicotòmica per a la categoria $3r.trimestre$ de la variable $Visita\ Prenatal$ de la següent manera:

Taula 4.2.15: Definició de la variable dicotòmica $Visita\ Prenatal$: $3r.trimestre$

Categories de la variable dicotòmica	Definició de la categoria de la variable dicotòmica
1	La mare va realitzar la última $Visita\ Prenatal$ en el 3r. trimestre d'embaràs.
0	La mare va realitzar la última $Visita\ Prenatal$ en qualsevol de les altres categories d'aquesta variable.

Mitjançant la definició d'aquestes variables dicotòmiques, es podrà veure com afecta cadascuna de les categories al valor de la variable Pes. Es recorda que es realitza l'estimació d'aquest primer model lineal amb totes les variables dicotòmiques possibles (8, una per cada categoria i s'aniran descartant a mesura que es vagin estimant les diferents regressions lineals i quantíliques).

4.3. Primera estimació: Model lineal (MQO)

Un cop s'ha realitzat l'anàlisi descriptiva inicial univariant i multivariant per al conjunt de variables de les que es disposa, ja s'està en una bona posició per tal de començar l'estimació dels diferents paràmetres associats als models de regressió que s'estimaran. En primer lloc, es procedeix amb un model lineal estimat per mínims quadrats ordinaris amb les variables explicatives especificades en l'apartat anterior. Per a l'estimació, s'utilitza la funció *lm()* d'R. En la següent taula 4.3.1, s'hi poden veure les estimacions dels diferents paràmetres i d'altres característiques associades a aquest model de regressió lineal:

Taula 4.3.1: Model lineal de la primera estimació proposada

Variable	Coeficient	Error est.	Valor t	P-valor ($< \lvert t \rvert$)
$Constant$	3176,10	13,69	232,04	$< 0{,}01$
$Etnicitat$	$-211{,}47$	7,17	$-29{,}49$	$< 0{,}01$
$Casada$	69,56	6,43	10,81	$< 0{,}01$
Nen	116,15	4,90	23,72	$< 0{,}01$
$Edat$	3,51	0,49	7,19	$< 0{,}01$
$Cigarrets/dia$	$-14{,}05$	0,54	$-25{,}83$	$< 0{,}01$
VP_0	$-194{,}49$	27,54	$-7{,}06$	$< 0{,}01$
VP_1	$-5{,}76$	7,62	$-0{,}76$	0,45
VP_2	$-16{,}32$	16,81	$-0{,}97$	0,33

VP_3	–	–	–	–
Ed_MG	37,24	7,58	4,91	$< 0{,}01$
Ed_G	59,99	8,37	7,17	$< 0{,}01$
Ed_EU	77,63	8,98	8,64	$< 0{,}01$
Ed_GU	–	–	–	–
$Error\ estàndard\ residual = 547{,}2\ (49988\ g.ll.)$				
$R^2 = 0{,}067$		$Estadístic\ F = 325{,}3\ (11\ i\ 49988\ g.ll.)$		
$R^2\ ajustat = 0{,}067$		$P - valor(< F) =< 0{,}01$		

Per a aquesta primera estimació que s'ha realitzat, s'han utilitzat les variables explicatives més oportunes segons les conclusions a les que s'ha arribat en l'apartat anterior. Tal i com es pot veure, no s'han estimat els coeficients associats a les variables dicotòmiques relacionades amb $Visita\ Prenatal = 3r.trimestre$ i $Educació = Graduat\ Universitari$, ja que R ha pres aquestes categories com a categories de referència o categories base de les variables dicotòmiques creades. Les conclusions que es poden extreure d'aquesta estimació són:

- Si es realitza el test de significació individual per als diferents coeficients estimats, el qual presenta les següents hipòtesis:

  $$H_0: \beta_i = 0\ vs. H_1: \beta_i = 0\ on\ i = 0, \dots, 13.$$

  I l'estadístic de contrast es calcula de la següent manera:

  $$t = \frac{\widehat{\beta_i}}{S.d.(\widehat{\beta_i})} \underset{sota\ H_0}{\sim} t_{n-k}\ on\ n = 50000\ i\ k = 14.$$

  Realitzant aquest contrast d'hipòtesi per als diferents coeficients estimats, es pot veure que per als coeficients de les variables dicotòmiques relacionades amb $Visita\ prenatal = 1r.trimestre$ i $Visita\ prenatal = 2n.trimestre$ no es pot rebutjar la hipòtesi nul·la del contrast amb un nivell de confiança del 95%. Amb els p-valors obtinguts de $0'45$ i $0'33$, no es pot rebutjar amb aquest nivell de confiança que aquests paràmetres siguin estadísticament iguals a 0.

- Si es realitza el test de significació conjunta del model, el qual presenta les següents hipòtesis:

  $$H_0: \beta_1 = \dots = \beta_{13} = 0\ vs. H_1: Algun\ \beta_i \neq 0\ on\ i = 1, \dots, 13.$$

  I l'estadístic de contrast es calcula de la següent manera:

  $$F = \frac{\frac{(SQR - SQL)}{m}}{\frac{SQL}{n-k}} \underset{sota\ H_0}{\sim} F_{m,n-k}\ on\ m = 13, n = 50000\ i\ k = 14.$$

  On SQL és la suma dels quadrats dels errors de la regressió lineal ampliada o estimada; SQR és la suma dels quadrats dels errors del model restringit ($y_t =$

$\beta_0 + e_{2i}$); m és el nombre de restriccions lineals o el nombre de coeficients del model (excloent la constant: β_0), n és el nombre d'observacions que s'han utilitzat per estimar el model de regressió lineal i k és el nombre de variables predictores utilitzades per estimar el model de regressió lineal.

Si es realitza aquest contrast, s'accepta la hipòtesi alternativa del contrast i per tant, es conclou que amb un nivell de confiança del 99%, existeix algun paràmetre β_i $(i = 1, \dots, 13)$ estadísticament diferent de 0.

- Si s'observa el coeficient de bondat de l'ajust (R^2) del model estimat, aquest resulta ser de $0{,}067$, el qual està lluny del valor de 1, el qual indicaria un ajust del model òptim. D'aquesta manera, es pot afirmar que mitjançant una recta només es pot capturar un cert grau de relació entre les variables exògenes i la variable dependent del model.

Davant d'aquestes conclusions, una possibilitat que es pot plantejar és la d'estimar un nou model lineal sense incloure les variables dicotòmiques relacionades amb les variables $Visita\ Prenatal = 1r.trimestre\ i\ 2n.trimestre$ ja que, a part que per als coeficients associats a les mateixes no es pot rebutjar la hipòtesi nul·la del contrast de significació individual, es considera que no aporten informació de característiques de la mare ni de quant pesarà el nounat. En el cas de $Visita\ Prenatal = 2n.trimestre$, es contempla que és més aviat una informació de quan s'ha produït el part. Per altra banda, no s'inclouen les variables dicotòmiques relacionades amb $Visita\ Prenatal = 3r.trimestre$ i $Educació = Graduat\ Universitari$ ja que, com s'ha dit anteriorment, s'han pres com a categories base de les variables dicotòmiques creades. Seguidament, s'estima aquest nou model i en la següent taula 4.3.2, es mostren les estimacions dels diferents paràmetres del model:

Taula 4.3.2: Model lineal de la segona estimació proposada

Variable	Coeficient	Error est.	Valor t	P-valor ($< \lvert t \rvert$)
$Constant$	3173,06	13,39	237,00	$< 0{,}01$
$Negra$	$-211{,}86$	7,16	$-29{,}58$	$< 0{,}01$
$Casada$	70,29	6,40	10,99	$< 0{,}01$
Nen	116,13	4,90	23,72	$< 0{,}01$
$Edat$	3,52	0,49	7,23	$< 0{,}01$
$Cigarrets/dia$	$-14{,}06$	0,54	$-25{,}85$	$< 0{,}01$
VP_0	$-192{,}71$	27,49	$-7{,}01$	$< 0{,}01$
Ed_MG	38,24	7,53	5,08	$< 0{,}01$
Ed_G	61,30	8,29	7,39	$< 0{,}01$
Ed_EU	79,17	8,88	8,92	$< 0{,}01$
$Error\ estàndard\ residual = 547{,}2\ (49990\ g.ll.)$				
$R^2 = 0{,}067$		$Estadístic\ F = 397{,}4\ (9\ i\ 49990\ g.ll.)$		
$R^2\ ajustat = 0{,}067$		$P - valor(< F) =< 0{,}01$		

En aquest nou model estimat, s'hi pot observar una millora respecte el model anterior:

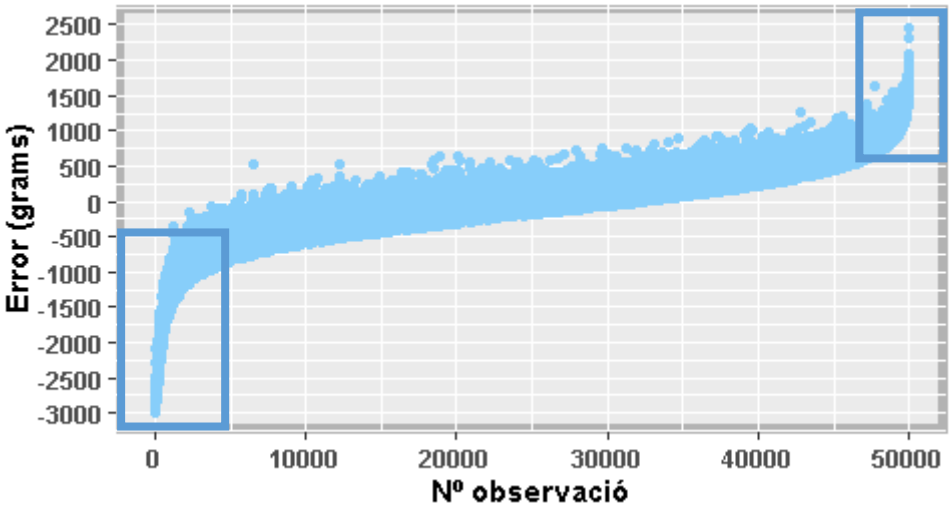
- Si es realitzen els diferents tests de significació individual per a tots els coeficients obtinguts, amb una confiança del 95%, en tots els casos, s'accepta la hipòtesi alternativa del contrast. D'aquesta manera, es pot dir que amb un nivell de confiança del 95%, es pot acceptar que tots els paràmetres són estadísticament diferents de 0.

Per altra banda, es té que:

- El test de significació conjunta segueix sent significatiu a l'1%.
- L'R^2 ajustat del model ha augmentat (De 0,06659 a 0,0666). Degut al diferent nombre de variables explicatives que conté cadascun dels dos models, cal utilitzar aquest coeficient de bondat de l'ajust per tal de comparar els coeficients de bondat de l'ajust dels dos models estimats.

Per tant, es considera que el segon model és més adequat per complir l'objectiu que s'està buscant en aquest estudi. Si es té amb compte aquest últim model estimat i es procedeix amb l'anàlisi dels errors del model, es mostra el següent gràfic 4.3.3 en el qual es relaciona cada observació amb el seu error en $grams$:

Gràfic 4.3.3: Errors del segon model estimat segons l'observació

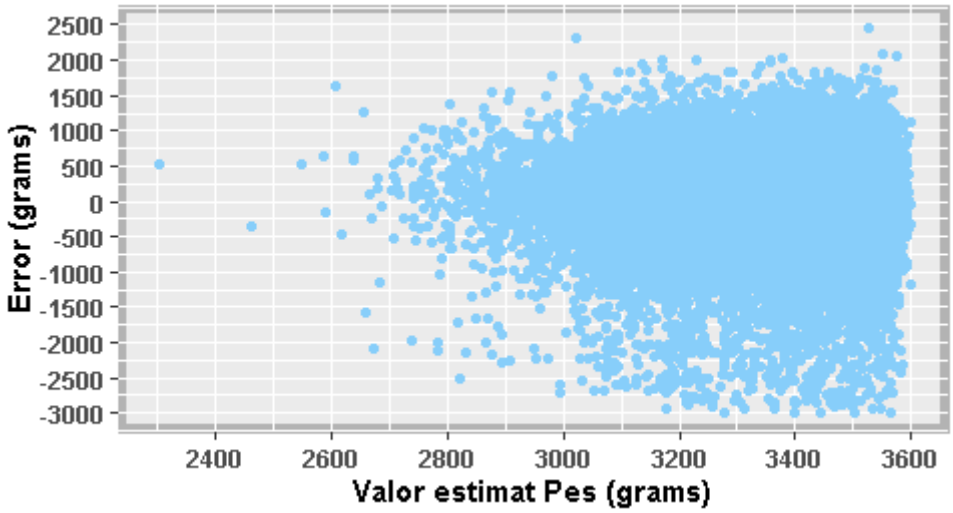


En aquest punt, cal recordar que les dades estan ordenades de menor a major segons el valor real de la variable resposta que s'està utilitzant: Pes. Tanmateix, cal recordar que es defineix l'error d'una observació i ($Error_i$) com la diferència entre el valor estimat de la variable Pes ($\widehat{Pes_i}$) i el valor real de la variable Pes (Pes_i) d'aquesta observació:

$$Error_i = \widehat{Pes_i} - Pes_i$$

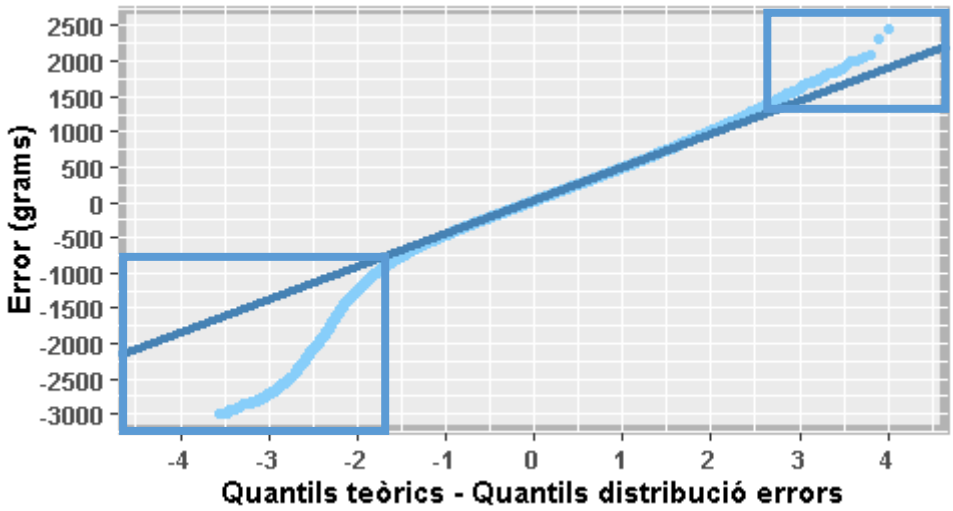
De la mateixa manera, per tal de seguir amb l'anàlisi dels errors del model, es mostra el següent gràfic 4.3.4 en què es mostren els valors estimats pel model de la variable *Pes* en funció dels errors comesos en cada cas:

Gràfic 4.3.4: Errors del segon model estimat segons valor estimat de *Pes*



A continuació, s'analitza la normalitat dels errors del model utilitzant el següent *plot probabilístic normal* (*PPN*). Aquest es mostra en la següent regió gràfica 4.3.5 i es comparen els quantils d'una distribució normal, els paràmetres de la qual han estat estimats a partir dels errors obtinguts (Anomenats quantils teòrics mostrats amb un color blau fort) i els quantils de la distribució dels errors obtinguts (Anomenats quantils distribució errors mostrats amb un color blau clar):

Gràfic 4.3.5: *PPN* per als errors del model



En aquests gràfics, s'observa que mitjançant aquest model es té un problema molt gran a l'hora d'estimar els valors més petits i més grans de la variable *Pes* (Es remarquen amb requadres blaus aquests valors en els gràfics). Aquests valors tenen uns errors molt

elevats en valor absolut i no segueixen una distribució normal. Es recorda que la no normalitat d'aquests errors causarà problemes a l'hora de procedir amb la inferència per part del model. A continuació, s'utilitza el test de *Jarque-Bera* per tal de contrastar la normalitat dels errors. Es recorda que les hipòtesis del test de *Jarque-Bera* per tal de contrastar la normalitat d'una variable aleatòria X són:

$$H_0: X \sim Normal \; vs. \; H_1: X \; no \; Normal.$$

Com és de suposar, en el nostre cas, X és la variable aleatòria relacionada amb els errors del model. L'estadístic de contrast, per contrastar la normalitat d'unes dades (associades a una variable aleatòria X), es calcula de la següent manera:

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right) \sim \chi_2^2.$$

On S es la desviació típica de les dades de les quals es vol comprovar la normalitat i K és el coeficient de curtosi associat a aquestes dades. Havent definit el contrast de *Jarque-Bera*, es realitza el mateix utilitzant la funció *jarque.bera.test()* d'R obtenint el corresponent resultat que es mostra en la següent taula 4.3.6:

Taula 4.3.6: Test de *Jarque-Bera* per als errors del model

$Estadístic\ JB = 22421$	$Graus\ de\ llibertat = 2$	$P - valor\ (> \chi_2^2) =< 0{,}01$

Degut al p-valor tan reduït que s'obté, amb una confiança del 99%, es pot acceptar la hipòtesi alternativa del contrast i es conclou que els errors del model no segueixen una distribució normal.

Per tant, observant els gràfics i procediments anteriors, es pot concloure que:

- Amb una confiança del 99%, es pot acceptar que els errors del model no segueixen una distribució normal.
- Els errors del model no resulten ser homoscedàstics, aspecte que queda confirmat observant el gràfic errors vs. valors estimats.
- Es pot observar que es tenen problemes per als errors (valors elevats i no normals) de les observacions en què els valors reals de la variable Pes pertanyen a valors extrems (tant valors grans com petits). Aquest fet es pot observar en el gràfic en què es mostren els errors en funció de l'observació[5] i en el *PPN* per als errors. Pel que fa a aquests errors, es pot dir que resulten ser molt elevats i no segueixen una distribució Normal[6].

[5] Es recorda que les observacions estan ordenades de manera ascendent segons els valors de la variable *Pes*, la qual fa referencia al valor del pes del nounat per a una observació concreta.
[6] Aquests errors han estat els que han provocat que el test de *Jarque-Bera* resultés ser no significatiu.

Aquestes conclusions porten a determinar que l'estimació per mínims quadrats ordinaris no resulta ser efectiva, ja que s'estan incomplint moltes de les hipòtesis bàsiques per utilitzar-la per poder realitzar inferència. Es recorda que si no es compleixen les hipòtesis bàsiques del model, no es garanteixen unes propietats bàsiques per als estimadors del model, i això donarà lloc a una inferència incorrecta. S'hauria d'optar per un altre tipus d'estimació si es volen estimar d'una millor manera els paràmetres d'aquests models, i en conseqüència els pesos dels nadons al néixer a partir de característiques de la mare.

4.4. Segona estimació: Regressió quantílica

Havent estimat un model lineal, s'ha vist que es tenen problemes a l'hora de procedir amb la inferència i estimació de valors extrems de la variable resposta (Pes), tant per excés com per defecte. Fent referència a la introducció del treball, s'hi ha remarcat que la regressió quantílica és útil quan es té l'interès en determinats quantils de la variable resposta. Aquesta sentència hauria de fer pensar que la regressió quantílica serà útil per a l'estimació dels diferents paràmetres per a aquest cas d'estudi.

En aquest cas, l'interès està posat en els quantils extrems (tant alts com baixos) de la variable Pes. Relacionat amb aquest interès, es recorda que per a aquest cas d'estudi, es tindran problemes mèdics per als nadons que presentin valors extrems de la variable resposta molt alts (obesitat infantil, trastorns infantils...) o molt baixos (malalties cròniques, anèmies...), fet que confirma l'interès en aquests quantils de la variable Pes.

Per tant, es comença aquesta segona estimació, mitjançant l'estimació d'una regressió quantílica associada al desè percentil de la variable Pes. S'utilitzen les variables que s'han considerat més oportunes per estimar la variable resposta mitjançant un model lineal. És a dir, s'utilitzen les mateixes variables que en el segon model lineal estimat. Per fer-ho, s'utilitza la funció *qr()* d'R del paquet *quantreg*. Es mostren els detalls d'aquesta estimació en la següent taula 4.4.1:

Taula 4.4.1: Model de regressió quantílica per al quantil 0,1 de la variable Pes

Variable	Coeficient	Error est.	Valor t	P-valor ($< \lvert t \rvert$)
$Constant$	2686,41	27,17	98,85	$< 0{,}01$
$Etnicitat$	$-261{,}59$	16,77	$-15{,}60$	$< 0{,}01$
$Casada$	82,12	13,27	6,19	$< 0{,}01$
Nen	73,06	9,65	7,57	$< 0{,}01$
$Edat$	$-0{,}82$	1,04	$-0{,}79$	0,43
$Cigarrets/dia$	$-17{,}63$	1,16	$-15{,}24$	$< 0{,}01$
VP_0	$-330{,}94$	52,36	$-6{,}32$	$< 0{,}01$
Ed_MG	41,59	15,61	2,66	$< 0{,}01$
Ed_G	71,00	16,25	4,37	$< 0{,}01$
Ed_EU	114,24	18,01	6,34	$< 0{,}01$

Respecte el model de regressió quantílica estimat, es pot dir que:

- Degut a que s'han utilitzat les variables explicatives que s'han considerat més adequades després d'estimar els diferents models lineals, s'observa que només hi ha un coeficient pel qual si es realitza el test de significació individual, amb una confiança del 95%, no s'accepta la hipòtesi alternativa del contrast. Per tant, es pot dir que amb aquest nivell de confiança, el paràmetre associat a la variable $Edat$ no és estadísticament diferent de 0 ($P - valor = 0{,}42 > 0{,}05$).
- La interpretació dels coeficients obtinguts per a les variables quantitatives es realitzaria tal que per cada augment d'una unitat del predictor numèric, s'espera que el quantil $0'1$ de la variable resposta augmenti (o disminueixi) el valor del coeficient.
- Per a les variables categòriques, el fet de passar del valor 0 al valor 1 de la mateixa, s'espera que el quantil 0,1 de la variable resposta es modifiqui (tant per excés com per defecte) el valor del coeficient.

Un cop estimada la regressió quantílica relacionada amb el quantil 0,1 de la variable Pes, a continuació, es mostra la següent taula 4.4.2 en què s'hi poden veure diferents estimacions dels paràmetres d'aquests models en funció del quantil de la variable Pes que s'estigui considerant (Tanmateix, es mostra l'estimació per MQO). Cal recordar que totes aquestes regressions quantíliques s'han estimat utilitzant la funció *qr()* d'R.

Taula 4.4.2: Estimacions dels paràmetres per a diferents models de regressió quantílica per als quantils: $0{,}05 - 0{,}25 - 0{,}50 - 0{,}75 - 0{,}95$ $(+MQO)$ de la variable Pes

	$Quantil$					MQO
	0,05	0,25	0,50	0,75	0,95	
$Constant$	2559,14 (< 0,01)	2921'29 (< 0,01)	3169,47 (< 0,01)	3431,00 (< 0,01)	3807,00 (< 0,01)	3173,07 (< 0,01)
$Etnicitat$	−364,50 (< 0,01)	−205,00 (< 0,01)	−184,21 (< 0,01)	−179,00 (< 0,01)	−152,00 (< 0,01)	−211,86 (< 0,01)
$Casada$	104,57 (< 0,01)	62,71 (< 0,01)	67'47 (< 0,01)	61,00 (< 0,01)	85,00 (< 0,01)	70,29 (< 0,01)
Nen	36,57 (0,02)	101,59 (< 0,01)	125,32 (< 0,01)	141,00 (< 0,01)	143,00 (< 0,01)	116,13 (< 0,01)
$Edat$	−4,79 (< 0,01)	2,12 (< 0,01)	4,26 (< 0,01)	6,00 (< 0,01)	9,50 (< 0,01)	3,52 (< 0,01)
$Cigarrets/dia$	−19,64 (< 0,01)	−14,94 (< 0,01)	−12,98 (< 0,01)	−12,83 (< 0,01)	−11,83 (< 0,01)	−14,06 (< 0,01)
$Visita\ Prenatal\_0$	−520,00 (< 0,01)	−230,88 (< 0,01)	−135,74 (< 0,01)	−152,00 (< 0,01)	−37,00* (0,50)	192,71 (< 0,01)
$Educació\_MG$	31,57 (0,21)	39,88 (< 0,01)	40,89 (< 0,01)	46,00 (< 0,01)	46,00 (< 0,01)	38,24 (< 0,01)
$Educació\_G$	65,93 (0,01)	73,59 (< 0,01)	58,47 (< 0,01)	60,00 (< 0,01)	37,50 (0,02)	61,30 (< 0,01)
$Educació\_EU$	141,79 (< 0,01)	97,35 (< 0,01)	72,84 (< 0,01)	67,00 (< 0,01)	19,00* (0,25)	79,17 (< 0,01)

Cal remarcar que els valors entre parèntesi que es troben a sota dels valors dels coeficients fan referència al p-valor del contrast d'hipòtesi individual del corresponent coeficient.

Per tant, en aquesta taula, s'hi pot veure com es modifiquen les estimacions dels paràmetres associats a una mateixa variable en funció del quantil de la variable resposta per al qual s'estigui estimant la regressió quantílica. Algunes observacions que es poden realitzar d'aquesta taula són les següents:

- Per als coeficients marcats amb el superíndex $*$, no es pot rebutjar amb una confiança del 95% que els paràmetres siguin estadísticament iguals a 0.
- Si una variable té un coeficient negatiu, és a dir influeix negativament a l'hora d'estimar un quantil determinat de la variable Pes, aquesta tendència negativa es manté al llarg dels diferents quantils d'aquesta variable.
- Per a les variables dicotòmiques creades a partir de les variables originals $Visita\ Prenatal$ i $Educació$, s'observa que la tendència (positiva o negativa) es manté relativament constant (en magnitud) al llarg dels diferents quantils.
- És important observar la diferència entre les estimacions dels paràmetres mitjançant MQO i mitjançant els models de regressió quantílica associats als diferents quantils de la variable Pes considerats.

Un cop s'han estimat els diferents models de regressió quantílica i es tenen els valors dels diferents coeficients en funció del quantil de la variable Pes, es poden realitzar contrastos d'hipòtesi per tal de contrastar si:

- Hi ha diferències estadístiques en les estimacions dels paràmetres associats a una mateixa variable en funció del quantil de la variable Pes que s'estigui considerant.
- Dins d'un mateix model de regressió quantílica (associat a un cert quantil de la variable Pes), hi ha diferències estadístiques en les estimacions obtingudes dels paràmetres associats a les diferents variables.
- Hi ha diferències estadístiques entre els coeficients d'un model associat a un quantil de la variable Pes amb els coeficients d'un altre model de regressió quantílica associat a un altre quantil de la variable Pes.

Aquests resultarien ser els contrastos d'hipòtesi principals, tot i que se'n podrien realitzar d'altres. A continuació, en la següent taula 4.4.3, es mostra el resultat del contrast d'hipòtesi en el qual es contrasta la igualtat estadística de tots els coeficients obtinguts en els models de regressió quantílica associats als quantils $0{,}25$ i $0{,}75$ de la variable Pes:

Taula 4.4.3: Contrast d'hipòtesi per contrastar si hi ha diferències estadístiques entre els coeficients dels models de regressió quantílica associats als quantils 0,25 i 0,75 de la variable Pes

Graus de llibertat	G. ll. Residuals	Estadístic F	P-valor ($> F$)
9	99991	10,32	$< 0{,}01$

Les hipòtesis d'aquest contrast d'hipòtesi són les següents:

$$H_0: \beta_{0,25} = \beta_{0,75} \dots \beta_{10,25} = \beta_{10,75} vs. H_1: Algun\ \beta_{i,25} \neq \beta_{i,75}\ per\ a\ i = 1, \dots, 10.$$

Cal remarcar que la nomenclatura mostrada per a definir aquestes hipòtesis és la definida en la secció *metodologia*. Tenint amb compte el reduït p-valor mostrat a la taula associat a l'estadístic calculat ($< 0{,}01$), es pot dir que amb una confiança del 99%, es pot acceptar la hipòtesi alternativa del contrast. D'aquesta manera, es pot dir que amb una confiança del 99%, existeix algun $\beta_{i,25} \neq \beta_{i,75}$ per $i = 1, \dots, 10$.

A continuació, es tenen amb compte les dos variables quantitatives utilitzades per estimar els diferents models de regressió quantílica: $Cigarrets/dia$ i $Edat$. Per a aquestes dues variables, s'estimen diferents models de regressió quantílica associats a diferents quantils p ($p = 0{,}1 - 0{,}2 - 0{,}3, \dots, 0{,}9$) de la variable Pes. Aquests models de regressió quantílica associats al quantil p de la variable Pes poden ser especificats segons:

$$Pes(p) = \beta_{0,p} + \beta_{1,p} * Cigarrets/dia.$$

$$Pes(p) = \beta_0 + \beta_{1,p} * Edat.$$

Es recorda que aquests models de regressió quantílica s'han estimat mitjançant la funció *qr()* d'R. Un cop estimats aquests models, l'objectiu és contrastar si hi ha diferències estadístiques en el valor de l'estimació del paràmetre $\beta_{1,p}$ per a un quantil p de Pes amb l'estimació del paràmetre β_1 si s'estimessin els mateixos models de regressió mitjançant MQO:

$$Pes = \beta_0 + \beta_1 * Cigarrets/dia.$$

$$Pes = \beta_0 + \beta_1 * Edat.$$

S'estimen aquests models de regressió lineal mitjançant la funció *lm()* d'R. Per a procedir amb l'objectiu, un cop estimats aquests models lineals, s'estableixen intervals de confiança del 95% sobre els coeficients β_1 segons:

$$IC(\beta_1)_{95\%} = \left(\beta_1 - z_{\frac{0.05}{2}} * \sqrt{se_{\beta_1}}, \beta_1 - z_{\frac{0.05}{2}} * \sqrt{se_{\beta_1}}\right).$$

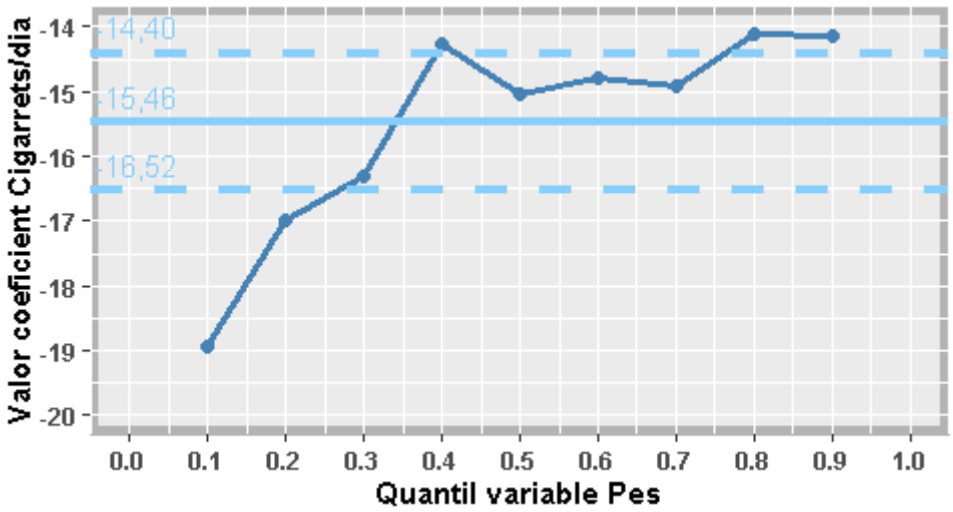
On β_1 s'estima mitjançant l'estimació d'aquest paràmetre ($\widehat{\beta_1}$) un cop estimats aquests models de regressió lineal, se_{β_1} és la desviació estàndard del coeficient β_1 i s'estima mitjançant $\widehat{se_{\beta_1}}$. Per últim, $z_{\frac{0.05}{2}}$ fa referència al valor z d'una distribució normal estàndard que deixa una cua a la dreta d'una probabilitat de $0{,}025$ ($z_{0.025} = 1{,}96$). A continuació, es mostra la següent taula 4.4.4 en la que es resumeixen els càlculs dels intervals de confiança dels coeficients β_1 per als dos models de regressió lineal especificats:

Taula 4.4.4: Càlcul dels intervals de confiança del coeficient β_1 dels models lineals $Pes = \beta_0 + \beta_1 * Cigarrets/dia$ i $Pes = \beta_0 + \beta_1 * Edat$

	$\widehat{\beta_1}$	$\widehat{se_{\beta_1}}$	IC (inferior)	IC (superior)
$Cigarrets/dia$	$-15{,}46$	0,54	$-16{,}52$	$-14{,}40$
$Edat$	9,64	0,44	8,78	10,50

Un cop calculats aquest intervals de confiança i estimats els diferents models de regressió quantílica especificats anteriorment, es mostren dos gràfics en els que es determina per a quins models de regressió quantílica (associats als quantils p de la variable Pes), el coeficient $\beta_{1,p}$ es troba dins d'aquests intervals de confiança especificats. En primer lloc, es mostra en la següent regió gràfica 4.4.5, la contrastació estadística entre els coeficients $\beta_{1,p}$ i β_1 associats amb els models on apareix la variable $Cigarrets/dia$:

Gràfic 4.4.5: Contrastació estadística entre els coeficients $\beta_{1,p}$ i β_1 associats amb els models on apareix la variable $Cigarrets/dia$ en funció del quantil (p) de la variable Pes

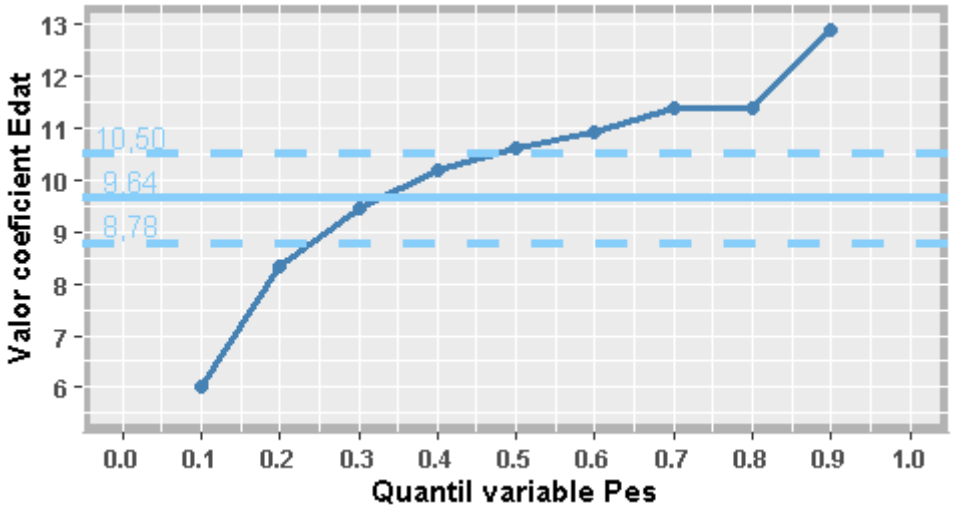


Observant aquest gràfic 4.4.5, s'arriba a la conclusió que amb una confiança del 95%, les estimacions dels paràmetres $\beta_{1,p}$ dels models de regressió quantílica de la forma

$Pes(p) = \beta_{0,p} + \beta_{1,p} * Cigarrets/dia$ associats als quantils $0{,}1 - 0{,}2 - 0{,}4 - 0{,}8\ i\ 0{,}9$ de la variable Pes són estadísticament diferents de l'estimació del paràmetre β_1 obtinguda en un model de regressió lineal de la forma $Pes = \beta_0 + \beta_1 * Cigarrets/dia$.

A continuació, es realitza el mateix procediment i es mostra el corresponent gràfic 4.4.6 en el que es contrasten estadísticament els valors dels coeficients $\beta_{1,p}$ i β_1 associats amb els models on apareix la variable $Edat$:

Gràfic 4.4.6: Contrastació estadística entre els coeficients $\beta_{1,p}$ i β_1 associats amb els models on apareix la variable $Edat$ en funció del quantil (p) de la variable Pes



Observant aquest gràfic 4.4.6, s'arriba a la conclusió que amb una confiança del 95%, les estimacions dels paràmetres $\beta_{1,p}$ dels models de regressió quantílica de la forma $Pes(p) = \beta_{0,p} + \beta_{1,p} * Edat$ associats als quantils $0{,}1 - 0{,}2 - 0{,}5 - 0{,}6 - 0{,}7 - 0{,}8\ i\ 0{,}9$ de la variable Pes són estadísticament diferents de l'estimació del paràmetre β_1 obtinguda en un model de regressió lineal de la forma $Pes = \beta_0 + \beta_1 * Edat$.

Aquesta última anàlisi s'ha fet per demostrar les diferències en les estimacions dels mateixos paràmetres que es donen en funció de si s'estima un model de regressió lineal o diferents models de regressió quantílica per a certs quantils de la variable resposta. Relacionat amb el que s'ha comentat anteriorment, és important destacar les diferències que es donen en les estimacions d'aquests paràmetres quan s'estan considerants quantils baixos i alts de la variable Pes ja que resulten ser d'interès per a aquest cas d'estudi.

4.5. Percentils condicionats vs. Percentils no condicionats

S'acaba aquesta primera aplicació en la que s'ha utilitzat, analitzat i parametritzat la regressió quantílica fent una petita reflexió i proposant una metodologia a la qual s'hi tornarà més tard i s'adaptarà a la regressió quantílica logística en la segona aplicació del treball. En primer lloc, es defineixen dos conceptes:

- Percentil no condicionat: El percentil no condicionat d'una observació correspon al percentil al qual pertany la variable resposta real d'aquella observació respecte la distribució de probabilitats de la variable resposta de la mostra amb la que s'estigui treballant.
- Percentil condicionat: El percentil condicionat d'un cas concret correspon al percentil al qual correspon en funció dels valors que prenguin les variables explicatives.

Per tal de deixar clars aquests dos conceptes, es mostra un exemple de càlcul, en el que es determinen aquests dos conceptes, per a un cas concret i d'aquesta manera es determina com es calculen els mateixos per a la resta d'observacions. En aquest cas, s'utilitza l'observació 30412 per a realitzar els corresponents càlculs. En primer lloc, es mostren en la següent taula 4.5.1 els valors de les variables explicatives per a aquest cas concret:

Taula 4.5.1: Valors de les variables explicatives de l'observació 30412

	$Etnicitat$	$Casada$	Nen	$Edat$	$Cigarrets/dia$	$Visita\ Prenatal\_0$	$Educació\_G$	$Educació\_MG$	$Educació\_EU$
Valor	0	0	0	24	10	0	0	0	0

1- Percentil no condicionat: Per a aquesta observació, la variable resposta pren el valor de $3525\ g$. Per altra banda, de la funció de distribució empírica de la variable resposta ($\widehat{F(y)}$) de la mostra amb la que s'està treballant, es mostra en la següent taula 4.5.2 la següent informació:

Taula 4.5.2: Valors de la funció de distribució empírica de la variable Pes ($\widehat{F(y)} = 0{,}60$ i $\widehat{F(y)} = 0{,}61$)

	$Probabilitat\ acumulada\ o\ \widehat{F(y)}$	
	$0'60$	$0'61$
$Valor$	$3515\ g$	$3534\ g$

Es recorda que $\widehat{F(y)}$ és la funció de distribució empírica dels valors reals de la variable resposta de la mostra. Per tant, observant aquestes dades, es pot concloure que el percentil 61 és el percentil no condicionat de la observació 30412, ja que la variable resposta es troba entre els dos valors que acumulen una probabilitat de la distribució de la variable resposta de $0'60$ i $0'61$.

2- El percentil condicionat per a aquesta observació es calcula de la següent manera:

- En primer lloc, s'estimen els diferents models de regressió quantílica associats als quantils p de la variable Pes (Es considera $p = 0{,}01 - 0{,}02, \dots, 0{,}99$). Per a fer-ho, s'utilitza la funció *rq()* d'R.
- Seguidament, s'estima el valor de la variable resposta d'aquesta observació concreta utilitzant les estimacions dels paràmetres proposades pels models estimats i els valors de les variables explicatives de l'observació d'estudi.
- Per últim, per a tots els valors estimats (un per a cada model de regressió quantílica considerat), es troba el valor que s'apropa més a la resposta real de l'observació (ja sigui per excés o per defecte). El percentil de la variable resposta associat al model utilitzat per trobar aquest valor més pròxim constituirà el percentil condicionat de l'observació.

Per exemple, per a l'observació 30412, si es regressen les variables explicatives en funció del model de regressió quantílica associat al quantil $0{,}80$ de la variable Pes, es troba que el valor estimat per a aquesta observació seria de $3528{,}95\ g$. Es resumeix aquest càlcul en la següent taula 4.5.3.

Taula 4.5.3: Càlcul del percentil condicionat per a l'observació 30412 (variables explicatives i coeficients model de regressió quantílica quantil $0{,}80$ de la variable Pes)

	$Etnicitat$	$Casada$	Nen	$Edat$	$Cigarrets/dia$	$Visita\ Prenatal\_0$	$Educació\_MG$	$Educació\_G$	$Educació\_EU$
Valor	0	0	0	24	10	0	0	0	0
Coef.	$-187{,}75$	58,00	143,25	7,00	$-12{,}58$	$-132{,}75$	49,00	64,00	64,00

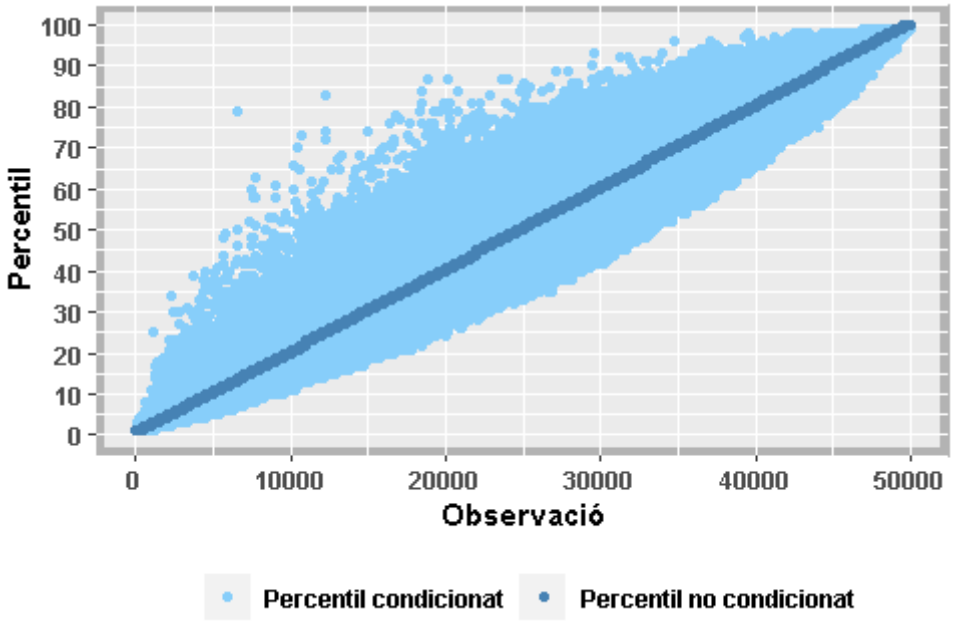
$$\hat{y} = 3486{,}75 + 7{,}00 * (24) + 10 * (-12{,}58) = 3528{,}95\ g.$$

Aquest valor estimat es correspon perfectament amb el pes real del nounat per a aquesta mare concreta. En conclusió, per a l'observació 30412, s'ha trobat que el percentil no condicionat és el 61, mentre que el percentil condicionat correspon al percentil 80.

Segons aquestes conclusions, aquesta metodologia proposada permet extrapolar més enllà de la mostra amb la que s'està treballant i afirmar que si es tingués una altra mostra composada per n mares amb les mateixes característiques (mateixos valors per a les variables explicatives) que la observació considerada (observació 30412 de la mostra amb la que s'està treballant), aquesta mare concreta es trobaria en el percentil 80 (percentil condicionat) de la distribució de valors de la variable Pes. Per tant, si es calcula el percentil condicionat de cada observació, permet situar a una mare concreta dins d'una població de referència (mares amb les mateixes característiques).

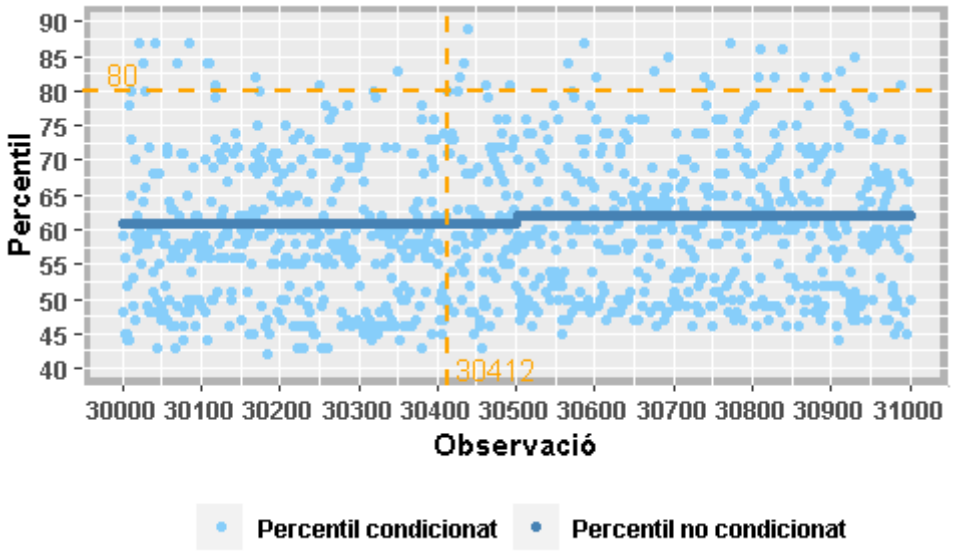
A continuació, es realitzen aquests dos càlculs per a totes les observacions de la base de dades amb la qual s'està treballant i es mostren aquests valors en la següent regió gràfica 4.5.4:

Gràfic 4.5.4: Percentils condicionats vs. percentils no condicionats per a totes les observacions



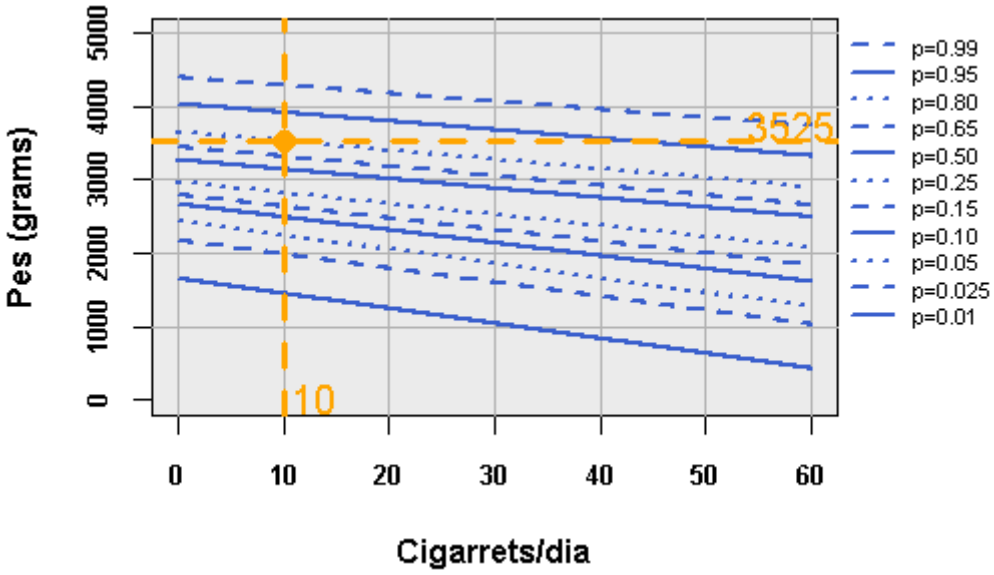
Es pot observar que hi ha una gran dispersió relacionada amb els percentils condicionats i no hi ha correspondència entre els dos tipus de percentils. Seguidament, en el següent gràfic 4.5.5, es pot veure quina és la poca correspondència entre els dos tipus de percentils en dos percentils no condicionats concrets $(61 - 62)$:

Gràfic 4.5.5: Percentils condicionats vs. percentils no condicionats per als percentils no condicionats $61 - 62$ de la variable Pes



La intersecció de les dos línies taronges discontínues que es poden veure en el gràfic marquen els valors de l'observació 30412, la qual s'ha utilitzat d'exemple per a mostrar el càlcul dels percentils condicionats i no condicionats. Retornant a la idea que si es tingués una població de referència, aquesta mare concreta es situaria en el percentil 80 dels valors de la variable Pes d'aquesta població, es mostra el següent gràfic 4.5.6 per mostrar gràficament aquesta idea:

Gràfic 4.5.6: La mare 30412 en la seva població de referència



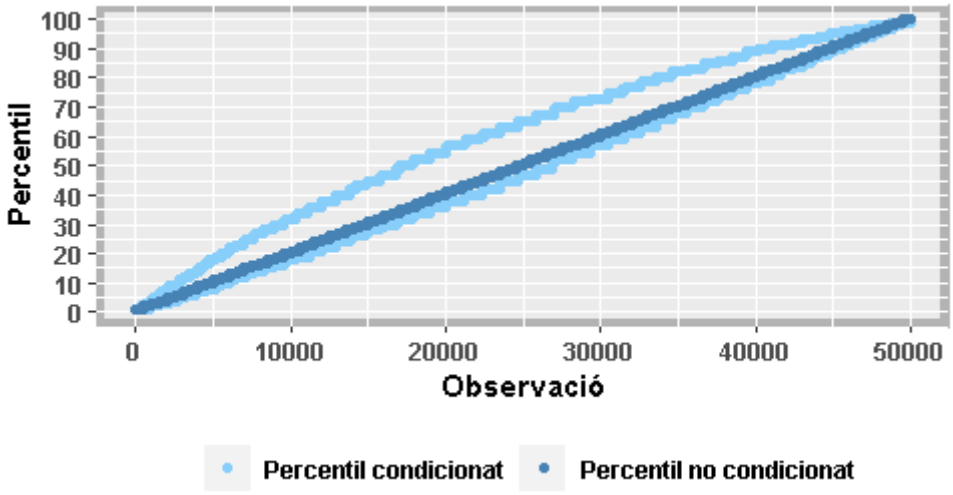
Respecte aquest gràfic, cal mencionar que els diferents models de regressió quantílica associats als diferents quantils p de la variable Pes estimats per a realitzar-lo es poden especificar de la següent forma:

$$Pes(p) = \beta_{0,p} + \beta_{1,p} * Etnicitat + \beta_{2,p} * Casada + \beta_{3,p} * Nen + \beta_{4,p} * Edat + \beta_{5,p} * Cigarrets/dia + \beta_{6,p} * VP\_0 + \beta_{7,p} * Ed\_MG + \beta_{8,p} * Ed\_G + \beta_{9,p} * Ed\_EU.$$

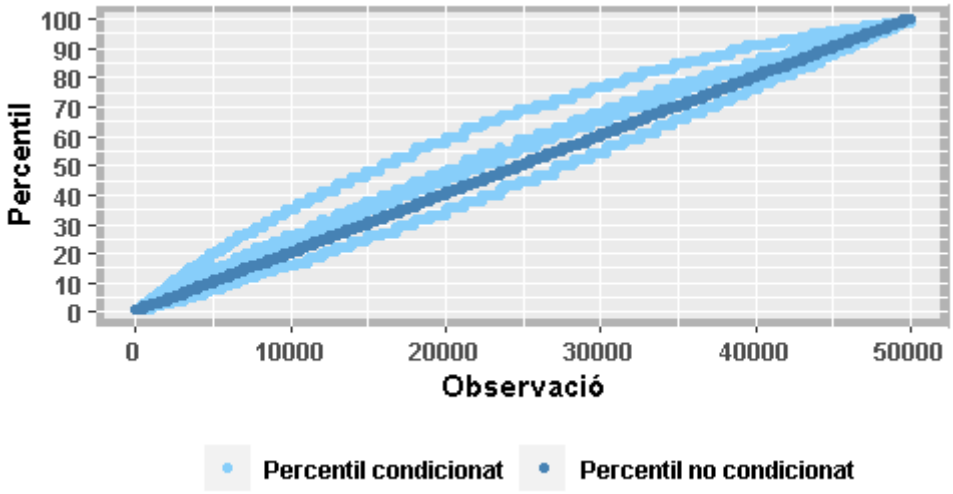
Per tant, es pot veure que s'han estimat aquests models de regressió quantílica amb les mateixes variables explicatives que s'han estimat els anteriors models. S'ha estimat aquest model de regressió quantílica per als quantils p de la variable Pes que s'hi poden veure en la llegenda del gràfic. Un cop estimats els diferents paràmetres $\beta_{i,p}$ associats als diferents quantils considerats, s'han substituït les variables explicatives pels valors de les variables explicatives de la mare de referència (observació 30412), deixant lliure la variable $Cigarrets/dia$. D'aquesta manera, s'estimen els diferents quantils de la variable Pes per a la població de referència (amb les característiques de la mare considerada). Tal i com es pot veure en el gràfic, s'estimaran aquests quantils de Pes en funció del valor que prengui la variable $Cigarrets/dia$. Un cop finalitzats aquests procediments, es situa la mare considerada en el gràfic ($Cigarrets/dia = 10$ i $Pes = 3525\ g$) i s'observa clarament que aquesta es troba en el quantil 0,8 de la variable Pes en aquesta població de referència.

Un cop mostrada, explicada i comentada aquesta metodologia proposada, ens plantegem com evoluciona la dispersió entre els percentils condicionats i els no condicionats de les 50000 observacions de la mostra a mesura que s'afegeixen variables explicatives als diferents models de regressió quantílica estimats. Per tant, a continuació, s'especificaran i s'estimaran un seguit de models de regressió quantílica (cada cop s'afegirà una variable explicativa més) i es mostraran un seguit de gràfics (4.5.7,...,4.5.14) realitzats a partir del càlcul dels percentils condicionats i no condicionats per a cada model de regressió quantílica estimat:

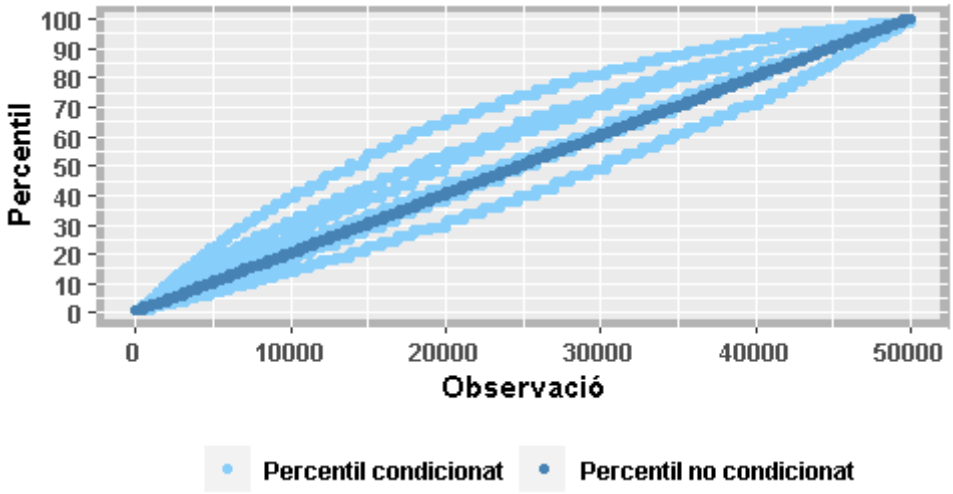
Gràfic 4.5.7: Percentils condicionats vs. percentils no condicionats segons models de regressió quantílica $Pes(p) = \beta_{0,p} + \beta_{1,p} * Etnicitat$



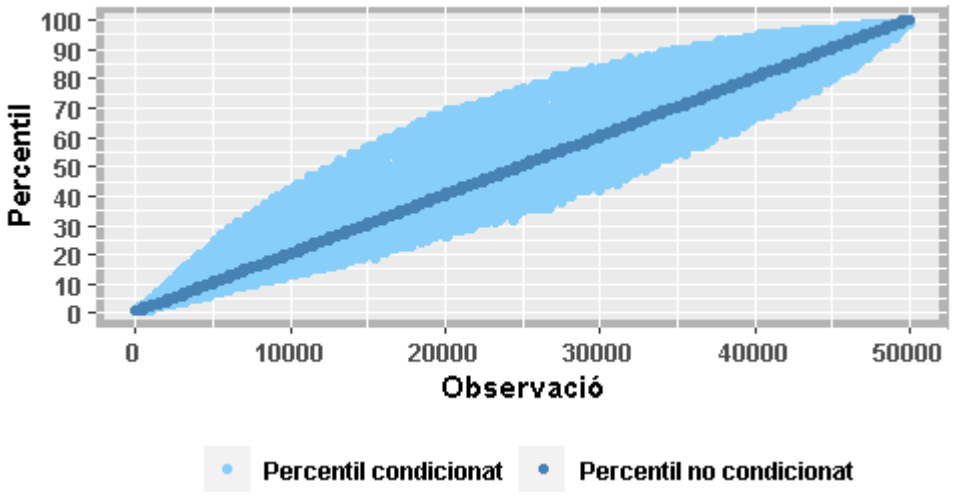
Gràfic 4.5.8: Percentils condicionats vs. percentils no condicionats segons models de regressió quantílica $Pes(p) = \beta_{0,p} + \beta_{1,p} * Etnicitat + \beta_{2,p} * Casada$



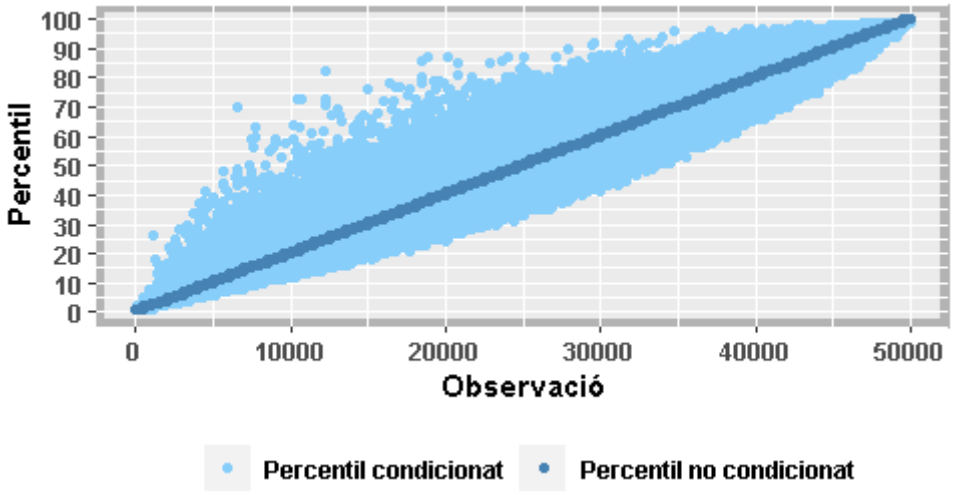
Gràfic 4.5.9: Models: $Pes(p) = \beta_{0,p} + \beta_{1,p} * Etnicitat + \beta_{2,p} * Casada + \beta_{3,p} * Nen$



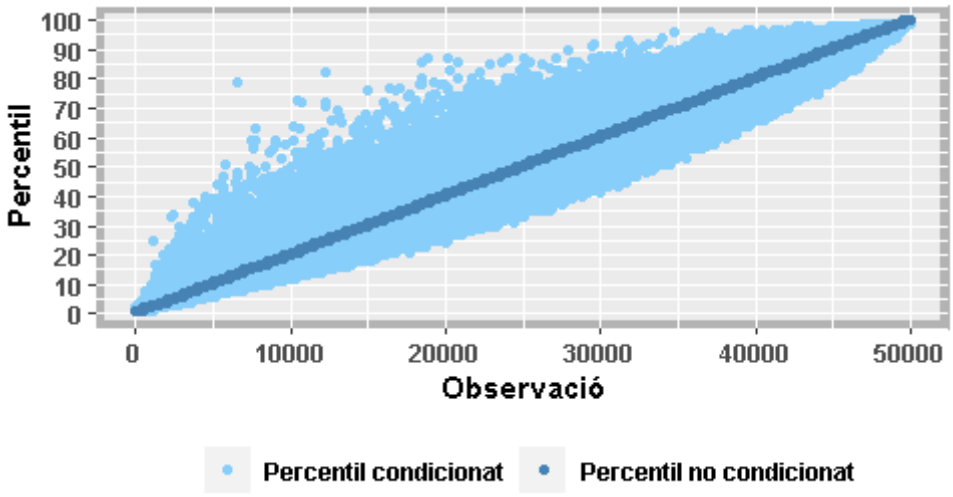
Gràfic 4.5.10: Models: $Pes(p) = \beta_{0,p} + \beta_{1,p} * Etnicitat + \beta_{2,p} * Casada + \beta_{3,p} * Nen + \beta_{4,p} * Edat$



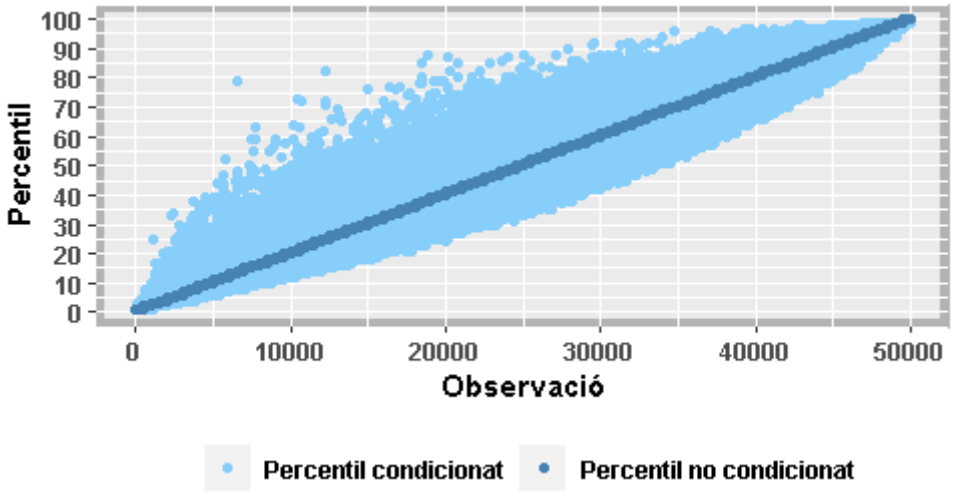
Gràfic 4.5.11: Models: $Pes(p) = \beta_{0,p} + \beta_{1,p} * Etnicitat + \beta_{2,p} * Casada + \beta_{3,p} * Nen + \beta_{4,p} * Edat + \beta_{5,p} * Cigarrets/dia$



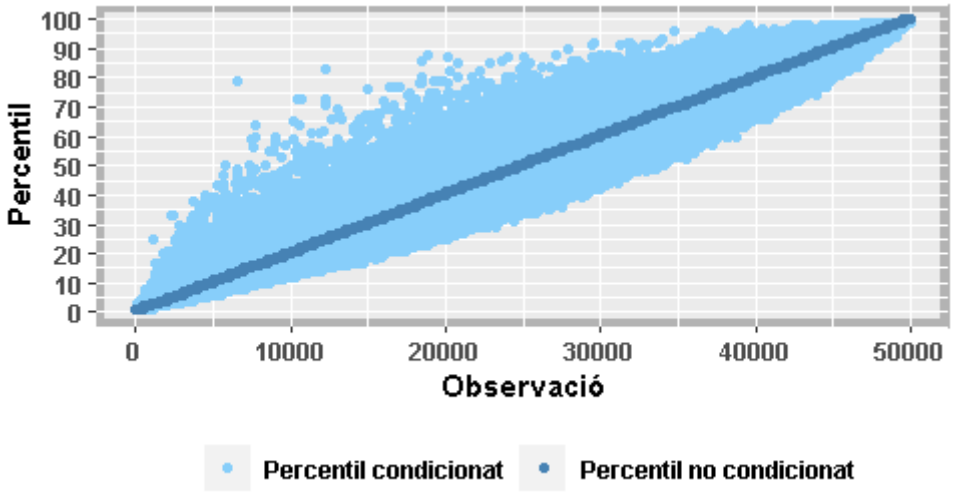
Gràfic 4.5.12: Models: $Pes(p) = \beta_{0,p} + \beta_{1,p} * Etnicitat + \cdots + \beta_{6,p} * VP\_0$



Gràfic 4.5.13: Models: $Pes(p) = \beta_{0,p} + \beta_{1,p} * Etnicitat + \cdots + \beta_{7,p} * Ed\_MG$



Gràfic 4.5.14: Models: $Pes(p) = \beta_{0,p} + \beta_{1,p} * Etnicitat + \cdots + \beta_{8,p} * Ed\_G$



Finalitzada aquesta primera aplicació del treball, cal realitzar una sèrie de reflexions i de consideracions finals:

- S'han pogut veure les limitacions que tenen els mínims quadrats ordinaris a l'hora d'estimar els paràmetres de models de regressió lineal quan no es compleixen els supòsits bàsics d'estimació: normalitat dels errors, homoscedasticitat dels errors, etc. En aquestes situacions, no es garanteixen les propietats bàsiques dels estimadors, i aquest fet causarà problemes quan es procedeixi amb la inferència.
- S'ha demostrat la potencialitat de la regressió quantílica per a estimar aquests paràmetres quan els mínims quadrats ordinaris no poden garantir les propietats bàsiques dels estimadors. En aquesta aplicació, la regressió quantílica ha estat molt interessant ja que es tenia un interès en els quantils extrems (tant elevats com baixos) de la variable resposta. Tanmateix, es recorda que la regressió quantílica també és molt útil en altres situacions: asimetries o canvi estructural, per exemple.
- S'han pogut observar les diferències que hi ha en les estimacions dels paràmetres relacionats amb les diferents variables explicatives en funció del quantil de la variable resposta que s'estigui considerant. Tanmateix, s'han pogut veure les diferències en les estimacions d'aquests paràmetres en comparació amb els de la regressió lineal estimada per mínims quadrats ordinaris.
- En la part final de l'aplicació, s'ha proposat una metodologia basada en els conceptes de percentil condicionat i percentil no condicionat. Aquesta permet anar més enllà de la mostra amb la que s'està treballant i situar una observació concreta en un percentil determinat dintre de la seva població de referència (observacions amb els mateixos valors de les variables explicatives).

  El valor d'aquest percentil en el qual quedi situada una mare dintre de la seva població de referència, pot ser una mesura de si realment el nounat té un reduït pes o no, ja que s'estarà comparant a aquesta mare amb altres mares amb les mateixes característiques. En aquest sentit, es pot afirmar que el percentil no condicionat d'una observació respecte una mostra formada per observacions amb característiques diferents no serà cap indicador de si realment el pes del nounat és petit o gran.

- Per últim, s'ha pogut veure com augmenta la dispersió entre els percentils condicionats i els percentils no condicionats a mesura que s'afegeixen variables explicatives als diferents models de regressió quantílica estimats.

5. Regressió quantílica logística: Modelització del percentatge de quilòmetres recorreguts al cap d'un any per sobre de la velocitat permesa

5.1. Descripció de les dades i descriptiva inicial

En aquesta segona aplicació, s'estudia, s'analitza i es parametritza l'estimació de la regressió quantílica logística. Es recorda que la regressió quantílica logística és una variant de la regressió quantílica en què per a la seva estimació, es precisa que la variable dependent o resposta del model estigui limitada amb un interval de valors concret. Concretament, es considera que:

$$y \in [y_{min}, y_{max}].$$

Aquest fet provoca canvis metodològics respecte la regressió quantílica, els quals es podran veure al llarg del desenvolupament d'aquesta aplicació i que ja han estat explicats i resumits de manera teòrica en l'apartat de la *metodologia*. Per a aquesta aplicació, s'utilitzen unes dades provinents d'una entitat asseguradora que, per motius de privacitat, omet la seva identitat. Aquestes són dades de l'any 2010, any en el qual aquesta entitat asseguradora va demanar als seus clients que instal·lessin un petit aparell als seus cotxes mitjançant el qual es recollien dades bàsiques de conducció com ara la velocitat, el quilometratge i el tipus de via on circulaven. La base de dades original disposava de 9614 observacions i 19 variables. Per motius de privacitat, per a la realització d'aquest treball, s'ha pres una submostra de la mateixa, la qual està composada per 7691 conductors i 6 variables. Es tracta d'una selecció aleatòria d'individus presa de la referència de la qual s'han obtingut les dades (*Pitarque, 2019*). El procediment que es seguirà per al desenvolupament d'aquesta aplicació és exactament el mateix que s'ha seguit per a la primera aplicació. D'aquesta manera, es poden veure de forma clara les diferències en els resultats que s'obtenen i s'estarà en una bona posició per tal de realitzar una comparació dels dos tipus de regressions en l'apartat de *conclusions*.

Per tant, es comença mostrant la següent taula 5.1.1 en què es presenten les variables que s'utilitzen en aquesta aplicació amb una breu descripció per a cadascuna d'elles:

Taula 5.1.1: Variables de la base de dades (nom, descripció i codificació)

Nom	Descripció
$Perc\_km$	Variable quantitativa que mesura el percentatge (%, en tant per cent) de quilòmetres recorreguts per part d'un conductor per sobre de la velocitat permesa durant l'any 2010.
Ln_km	Variable quantitativa que mesura amb logaritme el nombre de quilòmetres recorreguts per part d'un conductor durant l'any 2010.
$Perc\_urb$	Variable quantitativa que mesura el percentatge (%, en tant per cent) de quilòmetres recorreguts per part d'un conductor en zones urbanes respecte

	el total de quilòmetres conduïts.
$Perc\_noc$	Variable quantitativa que mesura el percentatge (%, en tant per cent) de quilòmetres recorreguts per part d'un conductor en horari nocturn respecte el total de quilòmetres conduïts.
$Edat$	Variable quantitativa que indica l'edat del conductor en $anys$ a l'inici de l'any 2010.
$Sexe$	Variable categòrica que indica el sexe del conductor: - Home: 1 - Dona: 0

Respecte aquestes variables presentades, cal fer les següents consideracions:

- La variable dependent del model ($Perc\_km$) s'ha obtingut a partir de dues variables de la base de dades original segons la següent transformació:

$$Perc\_km = \frac{Nº\ de\ km\ recorreguts\ per\ sobre\ de\ la\ velocitat\ permesa\ (any\ 2010)}{Nº\ de\ km\ recorreguts\ (any\ 2010)} * 100.$$

- La variable independent Ln_km s'ha obtingut a partir d'una variable de la base de dades original que indica el nombre de quilòmetres recorreguts per un conductor durant l'any 2010 segons la següent transformació:

$$Ln\_km = \ln(Nº\ de\ km\ recorreguts\ (any\ 2010)).$$

- La variable $Perc\_noc$ s'ha obtingut a partir de dues variables de la base original segons la següent transformació:

$$Perc\_noc = \frac{Nº\ de\ km\ recorreguts\ en\ horari\ nocturn\ (any\ 2010)}{Nº\ de\ km\ recorreguts\ (any\ 2010)} * 100.$$

- La variable $Perc\_urb$ s'ha obtingut a partir de dues variables de la base original segons la següent transformació:

$$Perc\_urb = \frac{Nº\ de\ km\ recorreguts\ en\ zones\ urbanes\ (any\ 2010)}{Nº\ de\ km\ recorreguts\ (any\ 2010)} * 100.$$

Es realitza la transformació logarítmica en la variable Ln_km prenent com a referència els treballs anteriors que existeixen amb dades d'aquest tipus. Per altra banda, tenint amb compte la primera transformació presentada (variable dependent) i recordant les característiques de la regressió quantílica logística, es té que la variable resposta del model està limitada en l'interval de valors que pot prendre un percentatge mesurat en tant per cent:

$$Perc\_km\ \epsilon\ [0, 100].$$

Procedint de la mateixa manera que en la primera aplicació, a continuació es mostra la següent taula 5.1.2 amb els estadístics bàsics univariants de les variables que s'utilitzen en aquesta segona aplicació:

Taula 5.1.2: Estadístics descriptius univariants bàsics de les diferents variables

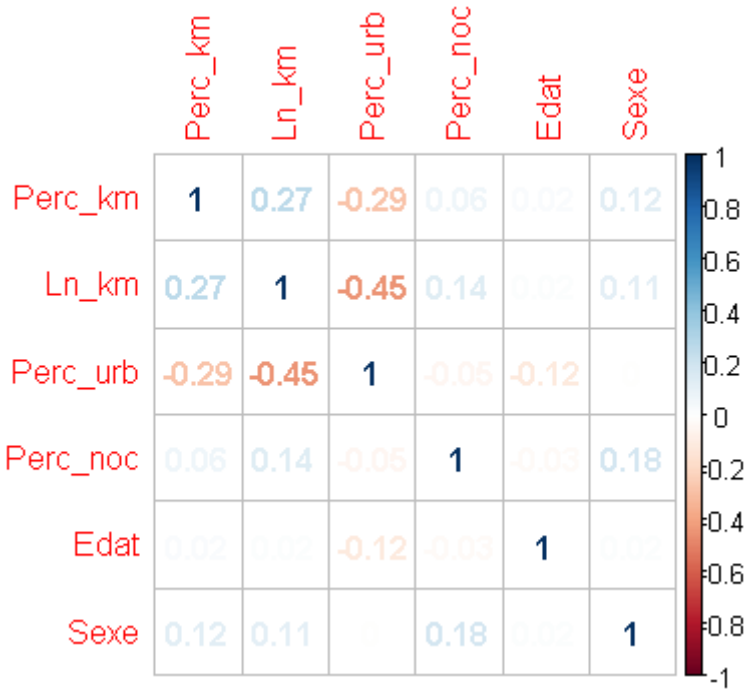
Variable	Min.	1er. Quartil	Mediana	Mitjana	3er. Quartil	Max.
$Perc\_km$	0,00	3,08	6,14	9,11	12,22	64,12
Ln_km	−0,37	8,92	9,37	9,26	9,76	10,96
$Perc\_urb$	0,00	15,60	23,47	26,36	34,46	100,00
$Perc\_noc$	0,00	2,46	5,30	7,01	9,89	46,34
$Edat$	18,11	22,65	24,61	24,77	26,87	31,56
$Sexe$	0,00	0,00	1,00	0,51	1,00	1,00

Un cop presentats els estadístics bàsics univariants que formen l'anàlisi descriptiva inicial, es realitzen alguns comentaris respecte aquesta taula:

- De la variable dependent dels models que s'estimaran ($Perc\_km$), cal destacar que hi ha 25 conductors del total de la mostra que no han realitzat cap quilòmetre per sobre de la velocitat permesa. Per a aquests conductors, aquesta variable pren el valor mínim (Mín.) mostrat a la taula: $0{,}00$.
- Respecte $Perc\_km$, també és important destacar la diferència numèrica que hi ha entre el valor del tercer quartil ($12{,}22$) i el valor màxim de la variable ($64{,}12$); es recorda que entre aquests dos valors, s'hi troben 1922 conductors. Per contra, entre el valor mínim de la variable ($0{,}00$) i el tercer quartil ($12{,}22$), s'hi troben 5769 conductors. Es veurà més endavant que els conductors amb valors de la variable resposta alts seran els que prendran importància en la nostra anàlisi.
- Per a la única variable categòrica de la que es disposa: $Sexe$, el valor de la mitjana de $0{,}51$ indica que la mostra està balancejada si es tenen amb compte el nombre d'homes i de dones.
- Per a la variable $Perc\_urb$, destaca el seu valor màxim (Màx.): $100{,}00$; aquest indica que 5 persones han realitzat la totalitat dels quilòmetres en carreteres localitzades en zones urbanes.
- Les edats dels conductors indiquen que tots són joves. No hi ha conductors per sobre de 32 anys. Aquest fet ve provocat per com es va comercialitzar aquest producte. No es va oferir a conductors més experimentats. Aquest fet tindrà implicacions en les conclusions ja que no es podrà extrapolar el què es trobi a tota una població més gran de conductors.

A continuació, es calcularan les correlacions lineals presents entre les variables que s'utilitzen en l'anàlisi. Es mostren aquestes correlacions lineals en la següent taula 5.1.3:

Taula 5.1.3: Correlacions lineals entre les diferents variables



A continuació, degut a la importància que pren la variable resposta en els models que s'estimen en aquesta segona aplicació, es mostra la següent taula 5.1.4 en la qual s'hi poden veure les correlacions lineals expressades en la primera fila o columna de la taula anterior:

Taula 5.1.4: Correlacions lineals entre la variable $Perc\_km$ i les diferents variables exògenes dels models

Variable	Correlació lineal amb $Perc\_km$
Ln_km	0,27
$Perc\_urb$	$-0,29$
$Perc\_noc$	0,06
$Edat$	0,02
$Sexe$	0,12

D'aquesta taula, caldria destacar els elevats valors relacionats amb les correlacions lineals de les variables Ln_km i $Perc\_urb$ amb la variable resposta. Tenint en compte aquest fet i com a hipòtesi inicial, es pot dir que es creu que aquestes variables independents seran les que prendran més importància a l'hora d'estimar els paràmetres associats als diferents models que s'estimen en aquesta aplicació. Tot i així, aquesta hipòtesi haurà de ser confirmada més tard quan s'estimin els diferents models. L'explicació que es podria donar per a aquestes correlacions lineals elevades és que aquells conductors que fan més quilòmetres són els que recorren més distàncies i normalment ho fan per vies interurbanes. Certament, podria haver-hi excepcions.

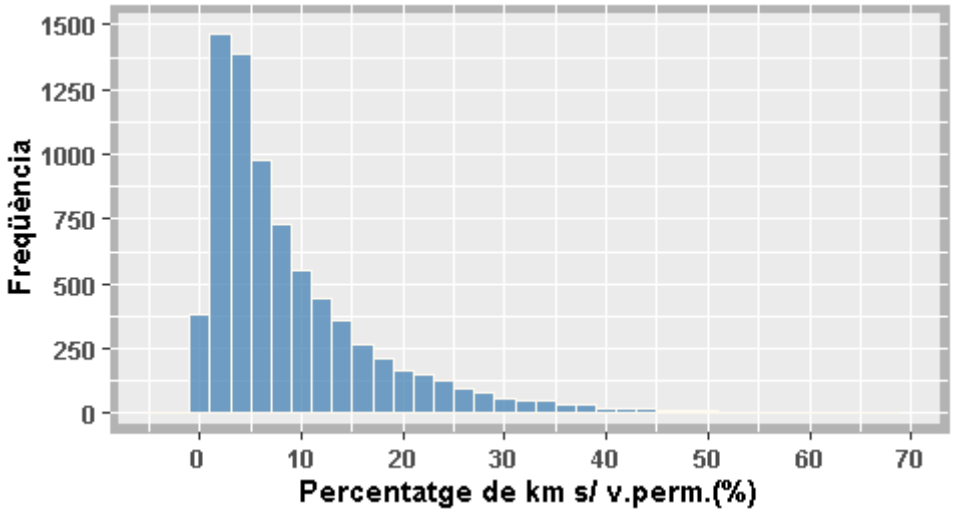
De la mateixa manera, cal tenir amb compte altres correlacions entre variables explicatives del model. En aquest cas, degut al menor nombre de dades de les que es disposa (en comparació amb la primera aplicació), es mostren en la següent taula 5.1.5 aquells coeficients de correlació lineal, amb les variables implicades en aquestes correlacions, superiors a 0,10 amb valor absolut:

Taula 5.1.5: Correlacions a tenir amb compte entre les variables explicatives del model

Variables	Correlació lineal
$Ln\_km - Perc\_urb$	$-0{,}45$
$Sexe - Perc\_noc$	0,18
$Ln\_km - Perc\_noc$	0,14
$Edat - Perc\_urb$	$-0{,}12$
$Ln\_km - Sexe$	0,11

Havent realitzat l'anàlisi descriptiva inicial, a continuació, es procedeix amb l'anàlisi gràfica mostrant els gràfics adients per a cadascuna de les variables proposades. S'inicia aquesta anàlisi amb la variable dependent del model. Es recorda que es tracta d'una variable quantitativa limitada en l'interval de valors: $[0, 100]$. Per tant, es realitza un histograma per veure'n la seva distribució. Aquest es mostra en la següent regió gràfica 5.1.6:

Gràfic 5.1.6: Histograma de la variable resposta ($Perc\_km$)

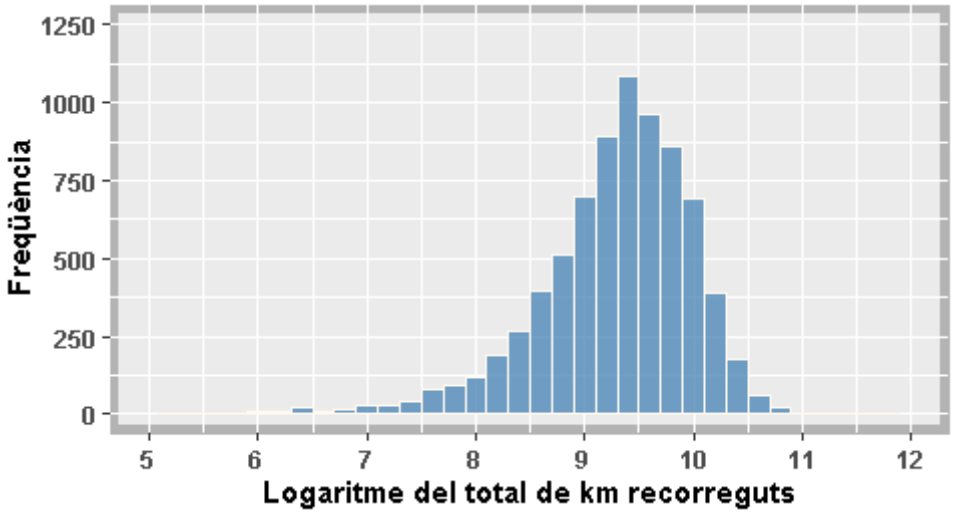


Mitjançant l'observació d'aquest histograma, es confirma un fet comentat anteriorment: es tracta d'una variable resposta formada principalment per valors baixos. Es pot veure que la majoria de valors es troben concentrats entre els valors de 0 i 20. Es recorda que en contraposició amb la primera aplicació, on prenien importància les observacions amb nadons amb un pes reduït, en aquest cas, les observacions amb valors alts de la variable resposta seran les que prendran més importància per a la nostra anàlisi. Des d'aquest

punt de vista, es pot dir que valors alts de la variable resposta (més quilòmetres recorreguts per sobre de la velocitat permesa) poden ser senyal d'una major propensió a la sinistralitat i, en definitiva, una major probabilitat de patir un accident.

Per a la variable *Ln_km*, es recorda que els valors de la mateixa, s'han calculat a partir de la transformació logarítmica del nombre total de quilòmetres recorreguts. Aquesta transformació permet reduir la dispersió de les dades[7] comprimint els valors molt alts i ajustant els valors baixos. Després d'aquesta transformació, la variable transformada es pot visualitzar en el següent histograma, el qual es mostra en la següent regió gràfica 5.1.7:

Gràfic 5.1.7: Histograma de la variable *Ln_km*

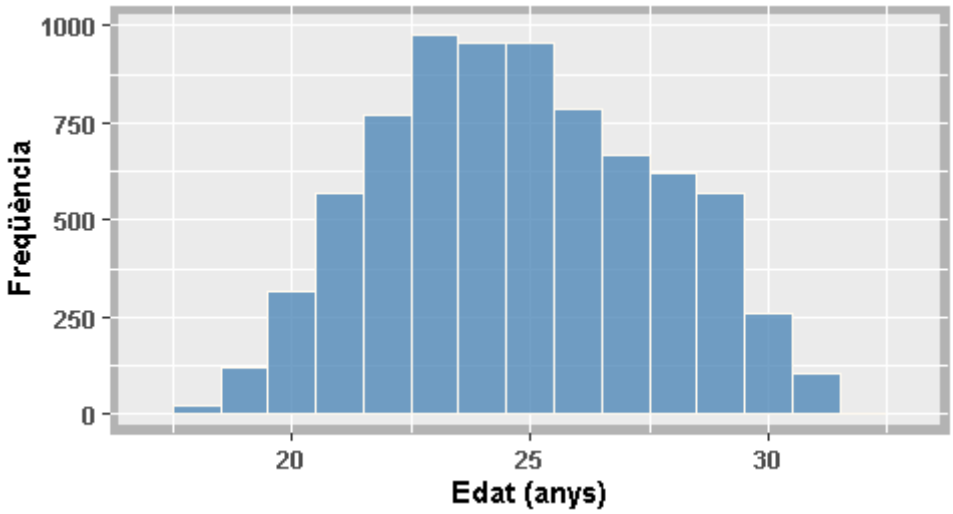


En aquest histograma, s'observa com s'ha reduït la variabilitat inicial de les dades originals, quedant la majoria concentrades entre els valors de 8 i 11. Un aspecte a destacar d'aquest histograma és que, degut a la seva dimensió, no es mostra el valor mínim de la variable presentat en la taula descriptiva inicial univariant (-0,37). Aquest valor està relacionat amb una persona que només ha recorregut 0,69 quilòmetres durant l'any 2010.

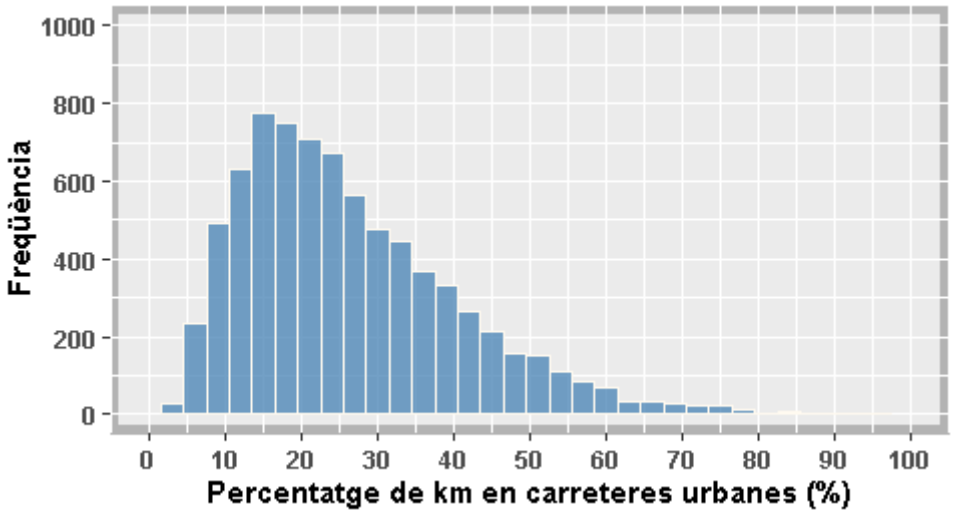
Seguidament, es realitzen histogrames per a les tres variables quantitatives restants de la base de dades: $Edat, Perc\_urb$ i $Perc\_noc$. Degut a que es tracta de variables quantitatives, l'histograma és el gràfic més adient per visualitzar-ne la seva distribució de valors. Es mostren aquests histogrames en les següents regions gràfiques 5.1.8, 5.1.9 i 5.1.10:

[7] Les variable que expressa el nombre total de quilòmetres recorreguts de la base de dades original té una $s^2 = 59701962$.

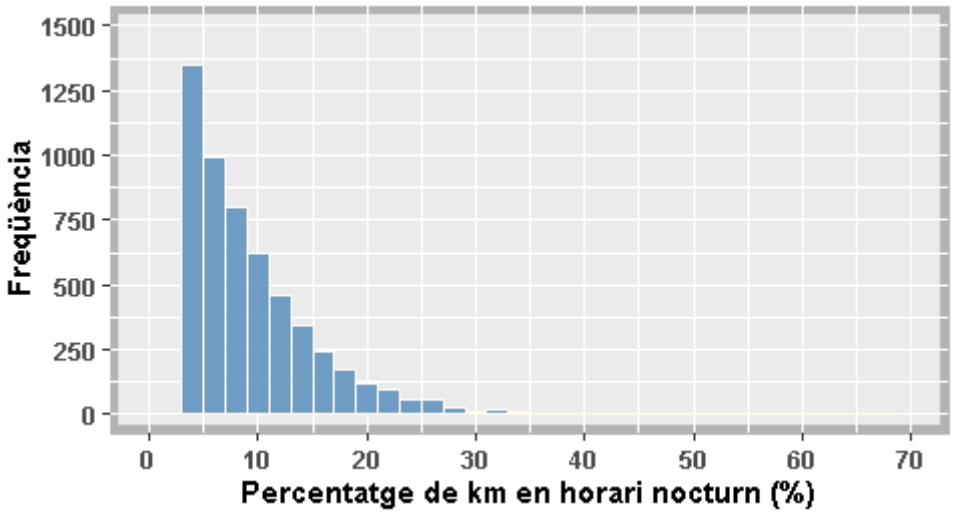
Gràfic 5.1.8: Histograma de la variable $Edat$



Gràfic 5.1.9: Histograma de la variable $Perc\_urb$



Gràfic 5.1.10: Histograma de la variable $Perc\_noc$



Mitjançant l'observació de l'histograma en què es mostra la distribució de valors de la mostra de la variable $Edat$, cal remarcar que es tracta d'una mostra jove: tots els conductors es troben concentrats entre les edats de 18 i 32 anys. A aquest fet, se li pot associar una avantatge i una desavantatge:

- Desavantatge: La consideració de si els resultats que s'obtindran serien els mateixos per una altra mostra en què el rang d'una variable que definís l'edat dels conductors fos molt més ampli.
- Avantatge: Diversos estudis associen una sinistralitat en la carretera major en aquesta franja d'edat "juvenil". Per tant, l'estudi d'aquest grup d'edat pot ser considerat com l'estudi d'una població objectiu i d'un interès alt.

Fent referència a la variable que mesura el percentatge de quilòmetres conduïts en carreteres urbanes ($Perc\_urb$), es pot observar que hi ha una tendència a no conduir gaires quilòmetres en aquest tipus de carreteres. Concretament, si es miren amb més detall les dades, es pot veure que dues terceres parts de la mostra ($66{,}27\%$ dels conductors), han conduït en carreteres urbanes un percentatge de quilòmetres situat en l'interval $[0\%, 30\%]$.

Per altra banda, si es mira el tercer histograma mostrat, encara es pot veure que hi ha una tendència negativa més alta (en comparació amb $Perc\_urb$) a conduir pocs quilòmetres en horari nocturn. En aquest cas, es pot veure que el $89{,}52\%$ dels conductors han conduït un percentatge de quilòmetres en hores nocturnes comprès en l'interval $[0\%, 15\%]$. Per tant, es pot dir que aquestes són dues variables amb una tendència a tenir més valors petits que alts, presentant d'aquesta manera una asimetria positiva o cap a la dreta.

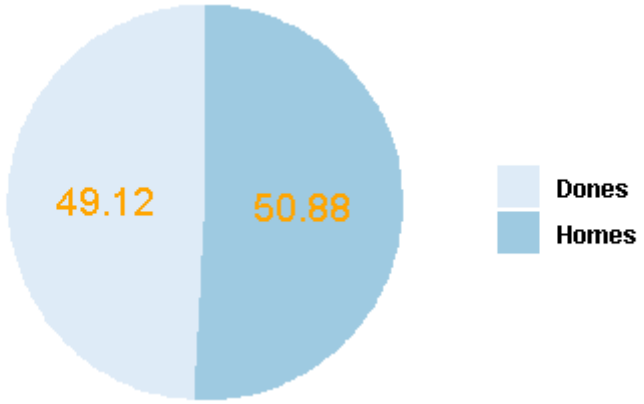
Per últim, es fa referència a la variable categòrica de la base de dades ($Sexe$) i es mostrarà en la següent taula 5.1.11, una taula de freqüències per veure'n la seva distribució:

Taula 5.1.11: Taula de freqüències per als valors de la variable categòrica $Sexe$

	$Sexe$	
	Homes	Dones
Freqüència	3913	3778
Percentatge (%)	50,88%	49,12%

Seguidament, es mostra el següent diagrama de pastís per visualitzar-ne la seva distribució gràficament. Aquest es mostra en la següent regió gràfica 5.1.12:

Gràfic 5.1.12: Diagrama de pastís per als valors de la variable categòrica $Sexe$



Per tant, amb aquest diagrama de pastís, es confirma un fet comentat anteriorment en què s'ha dit que la mostra amb la que es treballa es troba equilibrada pel que fa al nombre d'homes i de dones.

5.2. Descripció multivariant: *Profiling* de la variable resposta

Seguidament, abans de procedir a l'estimació dels diferents models que s'estimen en aquest apartat, es fa una descripció multivariant de la variable resposta mitjançant la qual es condiciona la distribució de valors d'aquesta variable en funció dels diferents valors que puguin prendre les diferents variables explicatives. Tal i com correspon, es fa una anàlisi diferenciada en funció de si es tracta de variables explicatives categòriques o quantitatives. S'inicia aquesta anàlisi amb la variable explicativa categòrica del model: $Sexe$ del conductor. Es mostra en la següent taula 5.2.1, la distribució de valors de la variable $Perc\_km$ en funció de les categories d'aquesta variable:

Taula 5.2.1: *Profiling* de la variable resposta en funció de la variable categòrica $Sexe$

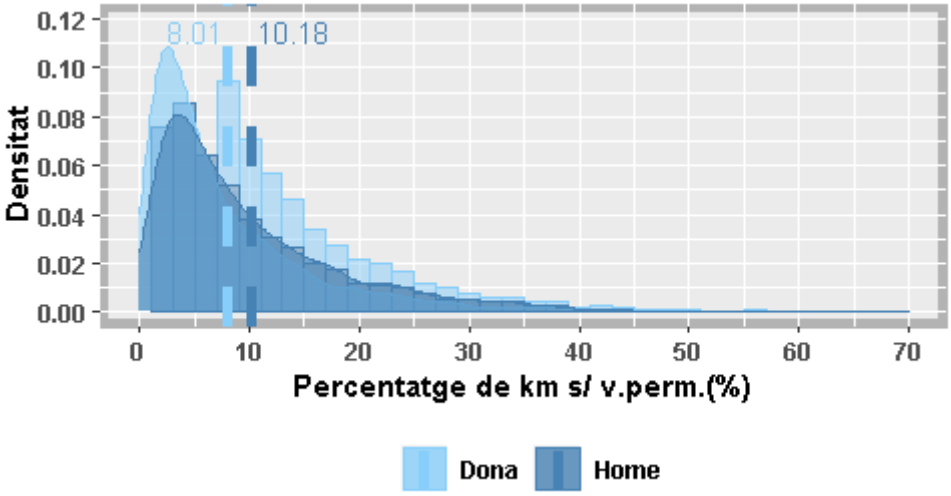
Variable/Estadístic	Mitjana	Desv. Est.	Min.	Max.
Sexe				
Home	10,18	9,11	0,00	64,12
Dona	8,01	8,23	0,00	57,32

Mitjançant l'observació d'aquesta taula, es pot dir que a simple vista no es veuen diferències pel que fa als valors de la variable resposta en funció de si el conductor és home o dona. Si es tenen amb compte els diferents estadístics mostrats, es pot deduir que aquests prenen valors semblants independentment del $Sexe$ del conductor. Es recorda que per tal de decidir si hi haguessin diferències estadístiques entre aquestes dues submostres, es podria realitzar un contrast d'hipòtesi mitjançant el qual es realitzaria una comparació de mitjanes per a dues mostres independents[8]. Seguidament,

[8] Es tracta d'un exemple de contrast d'hipòtesi que es podria realitzar. En aquest cas, es faria un contrast de diferència de mitjanes.

tal i com s'ha realitzat en la primera aplicació del treball, es mostra el següent gràfic 5.2.2 en què es presenta la distribució de valors condicionada (mitjançant histogrames i corbes de densitat) de la variable $Perc\_km$ en funció de les categories de la variable $Sexe$:

Gràfic 5.2.2: Histogrames i corbes de densitat dels valors de $Perc\_km$ en funció de $Sexe$



En aquest cas, es pot observar que tant els histogrames com les corbes de densitat obtinguts relacionats amb els valors de $Perc\_km$ són força semblants tenint amb compte el $Sexe$ del conductor. Tot i així, de manera generalitzada, caldria destacar uns valors més elevats de la variable $Perc\_km$ si el conductor és un home.

A continuació, es procedeix amb la descripció multivariant de la variable resposta en funció de les variables quantitatives de la base de dades. Tal i com s'ha fet en la primera aplicació, en primer lloc, es recorden en la següent taula 5.2.3 les correlacions lineals de les mateixes amb la variable resposta:

Taula 5.2.3: Correlacions lineals de les variables quantitatives amb la variable resposta ($Perc\_km$)

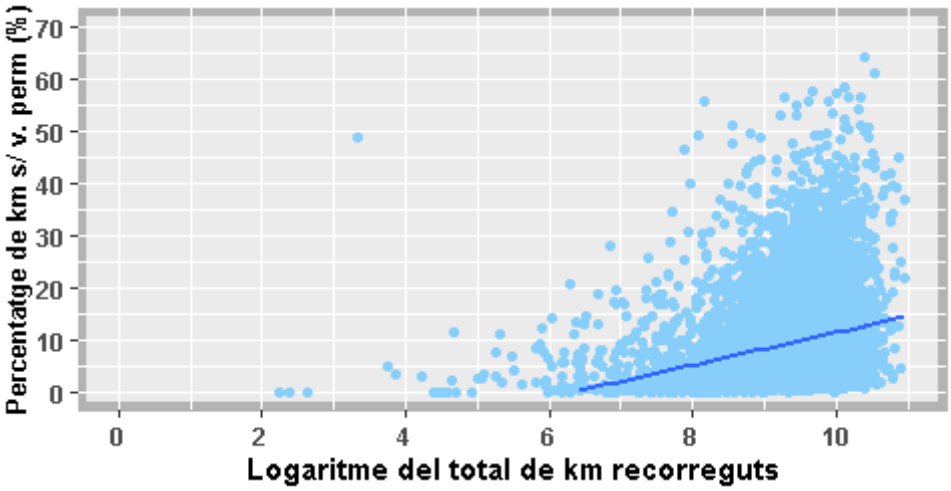
	Correlació lineal amb $Perc\_km$
Ln_km	0,27
$Perc\_urb$	−0,29
$Perc\_noc$	0,06
$Edat$	0,02

L'anàlisi d'aquestes variables quantitatives es fa procedint de la mateixa manera en què s'ha realitzat per a la primera aplicació. En primer lloc, es mostren diferents diagrames de dispersió en què s'hi poden veure els punts que relacionen els valors de la variable resposta en funció dels valors de la variable explicativa concreta. Tanmateix, s'estima

una recta de regressió que relaciona aquestes dues variables. Per últim, es mostra el model lineal que ha donat lloc a la recta de regressió mostrada i es realitza la corresponent anàlisi.

Per tant, s'inicia aquesta anàlisi mostrant el següent diagrama de dispersió i la corresponent recta de regressió en què es relacionen els valors de les variables $Perc\_km$ i Ln_km. Aquest es mostra en la següent regió gràfica 5.2.4:

Gràfic 5.2.4: Diagrama de dispersió i recta de regressió per a $Perc\_km$ i Ln_km



La recta de regressió mostrada prové del model lineal estimat $Perc\_km = \beta_0 + \beta_1 * Ln\_km$. Es mostren els detalls d'aquesta estimació en la taula 5.2.5:

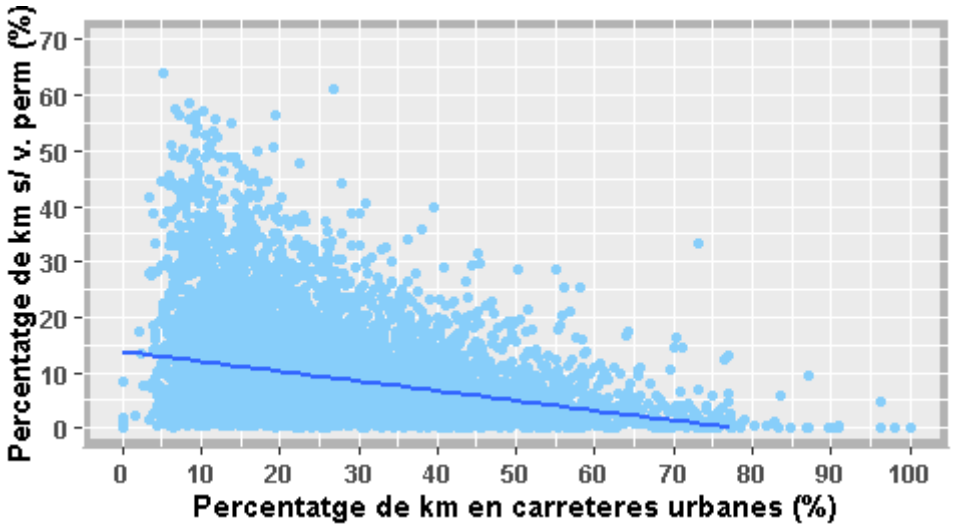
Taula 5.2.5: Model de regressió lineal $Perc\_km = \beta_0 + \beta_1 * Ln\_km$

Variable	Coeficient	Error est.	Valor t	P-valor ($< \|t\|$)
$Constant$	$-19{,}39$	1,18	$-16{,}48$	$< 0{,}01$
Ln_km	3,08	0,13	24,31	$< 0{,}01$
$Error\ estàndard\ residual = 8{,}44\ (7689\ g.ll.)$				
$R^2 = 0{,}07$		$Estadístic\ F = 590{,}9\ (1\ i\ 7689\ g.ll.)$		
$R^2\ ajustat = 0{,}07$		$P - valor(< F) =< 0{,}01$		

En aquest cas, es pot veure que el coeficient de la variable Ln_km resulta ser significativament diferent de 0 a un nivell de significació de l'1% i que s'ha obtingut un coeficient de bondat de l'ajust de 0,07. Aquest R^2 indica que mitjançant una recta no es pot capturar (o es pot capturar molt poca) relació entre aquestes variables: Ln_km i la variable dependent del model ($Perc\_km$). De la mateixa manera, es pot observar que la recta de regressió mostrada, té un pendent positiu confirmant la correlació lineal positiva abans calculada entre aquestes dues variables.

A continuació, es realitza el mateix procediment per a la variable numèrica $Perc\_urb$. Es mostra en la següent regió gràfica 5.2.6 el diagrama de dispersió i la corresponent recta de regressió que relaciona els valors d'aquestes dos variables:

Gràfic 5.2.6: Diagrama de dispersió i recta de regressió per a $Perc\_km$ i $Perc\_urb$



La recta de regressió mostrada prové del model lineal estimat $Perc\_km = \beta_0 + \beta_1 * Perc\_urb$. Es mostren els detalls d'aquesta estimació en la taula 5.2.7:

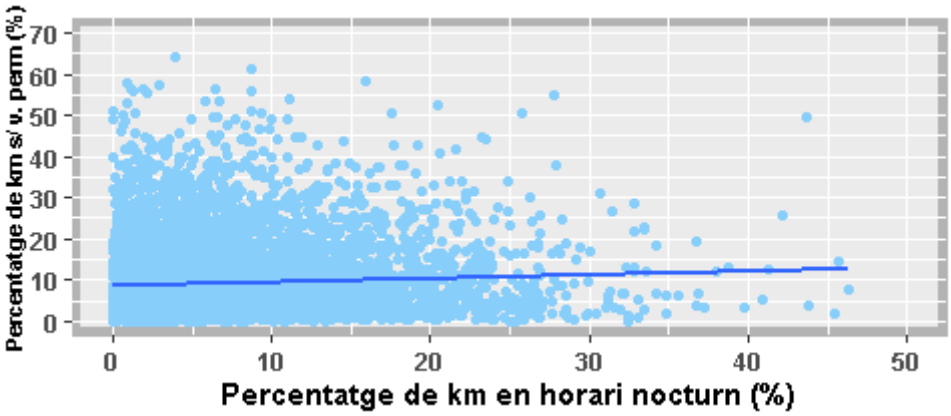
Taula 5.2.7: Model de regressió lineal $Perc\_km = \beta_0 + \beta_1 * Perc\_urb$

Variable	Coeficient	Error est.	Valor t	P-valor ($< \|t\|$)
$Constant$	13,77	0,20	68,64	$< 0{,}01$
$Perc\_urb$	$-0{,}18$	0,01	$-26{,}42$	$< 0{,}01$
$Error\ estàndard\ residual = 8{,}38\ (7689\ g.ll.)$				
$R^2 = 0{,}08$		$Estadístic\ F = 697{,}9\ (1\ i\ 7689\ g.ll.)$		
$R^2\ ajustat = 0{,}08$		$P - valor(< F) =< 0{,}01$		

El coeficient de la variable explicativa ($Perc\_urb$) segueix sent significativament diferent de 0 a un nivell de significació de l'1%. En aquest cas, es pot veure que s'ha obtingut un R^2 que pren un valor de 0,08. De la mateixa manera que per al cas anterior, es pot dir que aquest indica que la recta de regressió estimada pot capturar molt poca relació entre aquestes dues variables. Per a aquesta variable explicativa, la relació es de caràcter negatiu. Es pot mencionar aquest fet fent referència al pendent de la recta de regressió o recordant el coeficient de correlació lineal negatiu anteriorment calculat entre aquestes dues variables.

Seguidament, es té amb compte la variable numèrica $Perc\_urb$ i es porta a terme el mateix procediment seguit per a les dues variables quantitatives anteriors. Es mostra el corresponent diagrama de dispersió i la corresponent recta de regressió en la següent regió gràfica 5.2.8:

Gràfic 5.2.8: Diagrama de dispersió i recta de regressió per a $Perc\_km$ i $Perc\_noc$



La recta de regressió mostrada prové del model lineal estimat $Perc\_km = \beta_0 + \beta_1 * Perc\_noc$. Es mostren els detalls d'aquesta estimació en la taula 5.2.9:

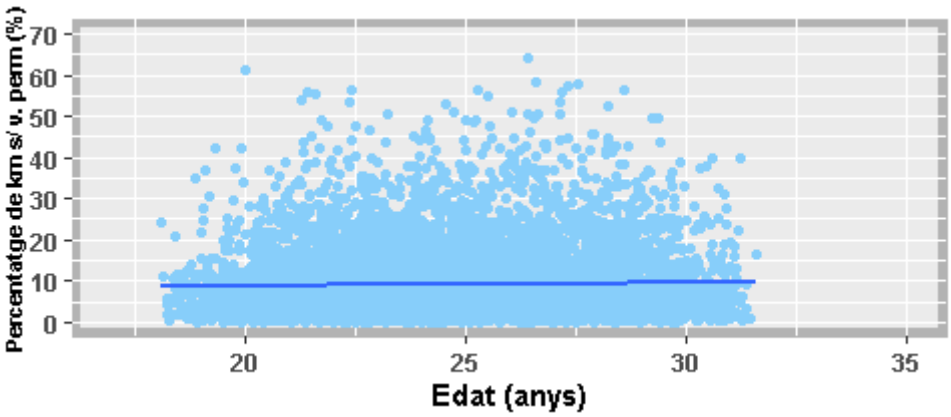
Taula 5.2.9: Model de regressió lineal $Perc\_km = \beta_0 + \beta_1 * Perc\_noc$

Variable	Coeficient	Error est.	Valor t	P-valor ($< \|t\|$)
$Constant$	8,49	0,15	55,93	$< 0{,}01$
$Perc\_noc$	0,09	0,02	5,41	$< 0{,}01$
$Error\ estàndard\ residual = 8{,}74\ (7689\ g.ll.)$				
$R^2 =< 0{,}01$		$Estadístic\ F = 29{,}23\ (1\ i\ 7689\ g.ll.)$		
$R^2\ ajustat =< 0{,}01$		$P - valor(< F) =< 0{,}01$		

Per a aquest model de regressió, s'ha obtingut un coeficient de bondat de l'ajust molt petit: $R^2 < 0{,}01$. Aquest valor es correspon perfectament amb la correlació lineal petita calculada anteriorment entre aquestes dues variables. Tanmateix, es podria categoritzar aquesta petita relació lineal de caràcter positiu.

Per últim, es té amb compte la variable numèrica que falta per analitzar ($Edat$) i es mostra el següent diagrama de dispersió amb la corresponent recta de regressió en la següent regió gràfica 5.2.10:

Gràfic 5.2.10: Diagrama de dispersió i recta de regressió per a $Perc\_km$ i $Edat$



La recta de regressió mostrada prové del model lineal estimat $Perc\_km = \beta_0 + \beta_1 * Edat$. Es mostren els detalls d'aquesta estimació en la taula 5.2.11:

Taula 5.2.11: Model de regressió lineal $Perc\_km = \beta_0 + \beta_1 * Edat$

Variable	Coeficient	Error est.	Valor t	P-valor ($< \|t\|$)
$Constant$	7,37	0,88	8,36	$< 0{,}01$
$Edat$	0,07	0,03	1,99	0,05
$Error\ estàndard\ residual = 8{,}75\ (7689\ g.ll.)$				
$R^2 =< 0{,}01$		$Estadístic\ F = 3{,}96\ (1\ i\ 7689\ g.ll.)$		
$R^2\ ajustat =< 0{,}01$		$P - valor(< F) = 0{,}05$		

Es recorda que aquesta variable és la que presenta una correlació lineal més petita amb la variable resposta. Aquest fet pot ser confirmat si s'observa el coeficient de bondat de l'ajust obtingut per a aquesta recta de regressió ($R^2 < 0{,}01$). Amb aquest R^2 tan petit, es pot afirmar que mitjançant una recta es pot capturar molt poca relació entre aquestes dues variables. De la mateixa manera, es pot confirmar que la petita relació que pugui existir entre aquestes dues variables és de caràcter positiu.

Per últim, havent finalitzat l'anàlisi descriptiva multivariant, s'arriba a les següents conclusions:

- En un principi, la variable $Edat$ serà exclosa dels models que s'estimaran degut a la petita correlació lineal que manté amb la variable resposta del model. Tanmateix, si es recorda que tots els conductors de la base de dades es troben amb una franja d'edat concreta, s'arriba a la conclusió que la seva supressió no portarà a errors en l'estimació dels diferents paràmetres degut, principalment, a que l'efecte de la variable $Edat$ resulta ser mínim trobant-se els seus valors tan propers. De tota manera, aquesta idea es pot reconsiderar si es busqués un efecte quadràtic de l'edat, ja que la gràfica podria indicar que per a quantils elevats de la variable $Perc\_km$, es podria tenir un efecte de paràbola.
- Es considera que totes les altres variables seran importants a l'hora d'estimar els diferents paràmetres dels models. Per tant, s'estimaran aquests models amb totes les variables explicatives que formen part de la base de dades amb la que s'està treballant.
- S'ha observat que els coeficients de bondat de l'ajust resulten ser molt petits en els diferents models de regressió estimats. Es pot dir que una raó darrere d'aquest fet és la presència d'altres tipus de relacions[9] entre les variables explicatives i la variable resposta del model.

[9] Entre aquests altres tipus de relacions, s'inclouen, per exemple, relacions quadràtiques.

5.3. Primera estimació: Model lineal (MQO)

Arribats a aquest punt, ja es té una visió detallada de les dades amb les quals s'està treballant. En primer lloc, s'ha realitzat una anàlisi descriptiva univariant de cadascuna de les variables de la base de dades mitjançant la qual es pot tenir una visió estadística individualitzada de cadascuna d'elles. Seguidament, s'ha realitzat una anàlisi descriptiva multivariant de la variable resposta mitjançant la qual s'ha caracteritzat aquesta variable en funció de les variables explicatives del model. Realitzats aquests dos procediments, s'està amb una bona posició per iniciar-se en l'estimació de models, els quals estimaran el percentatge de quilòmetres que recorrerà un conductor per sobre de la velocitat permesa al cap d'un any. Es comença aquesta anàlisi plantejant un model lineal estimat per mínims quadrats ordinaris utilitzant la funció *lm()* d'R. Es mostren els detalls d'aquesta estimació en la següent taula 5.3.1:

Taula 5.3.1: Model lineal de la primera estimació proposada

Variable	Coeficient	Error est.	Valor t	P-valor ($< \|t\|$)
$Constant$	$-5{,}13$	1,38	$-3{,}72$	$< 0{,}01$
$Sexe$	1,82	0,19	9,49	$< 0{,}01$
Ln_km	1,81	0,14	12,92	$< 0{,}01$
$Perc\_urb$	$-0{,}13$	0,01	$-18{,}21$	$< 0{,}01$
$Perc\_noc$	0,02	0,02	0,99	0,32
$Error\ estàndard\ residual = 8{,}22\ (7686\ g.ll.)$				
$R^2 = 0{,}12$		$Estadístic\ F = 257{,}9\ (4\ i\ 7686\ g.ll.)$		
$R^2\ ajustat = 0{,}12$		$P - valor(< F) =< 0{,}01$		

En aquest cas, s'han inclòs les variables explicatives segons les conclusions proposades en l'anterior apartat del treball. Segons aquesta estimació, s'arriba a les següents conclusions:

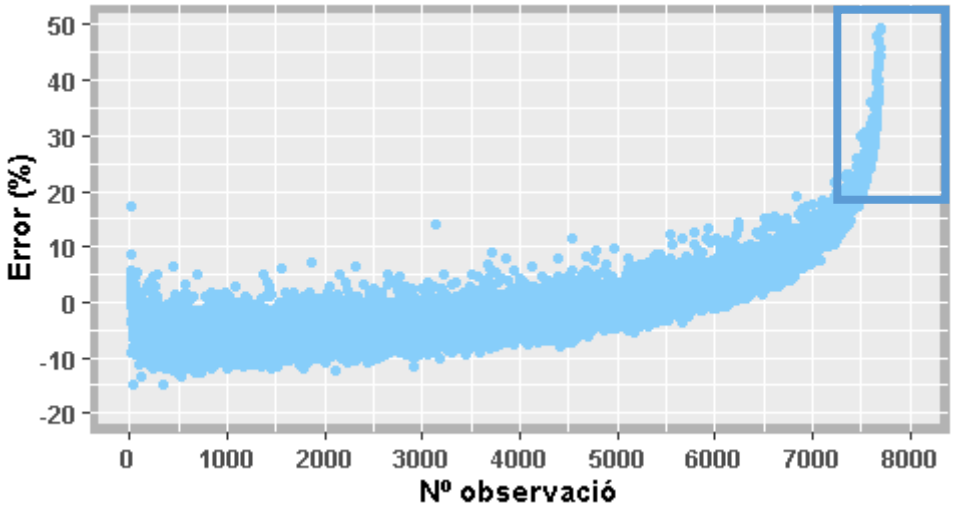
- Si es realitza el contrast de significació individual de cadascun dels coeficients del model (veure secció 4.3.), només per al coeficient relacionat amb la variable $Perc\_noc$ no es rebutja la hipòtesi nul·la del contrast. Es confirma aquest fet si es té amb compte el p-valor obtingut en aquest contrast $(0{,}32)$, mitjançant el qual no es pot rebutjar amb una confiança del 95% que el paràmetre sigui estadísticament igual a 0. Es recorda que $Perc\_noc$ és la variable categòrica que presenta una menor correlació lineal amb la variable resposta de les que s'han inclòs en el model.
- Per altra banda, si es realitza el contrast de significació conjunta de tots els coeficients relacionats amb les variables explicatives del model (veure secció 4.3.), el p-valor obtingut $(< 0{,}01)$ indica que acceptem que existeix almenys algun coeficient $\beta_i\ on\ i = 1, 2, 3, 4$ diferent de 0 a un nivell de significació de l'1%.

- S'observa que l'R^2 pren un valor de $0{,}118$, tal i com succeïa en la primera aplicació, aquest valor està lluny del valor que indica un model d'ajust òptim (1). D'aquesta manera, es pot afirmar que mitjançant una recta, es pot capturar molt poca relació entre les variables explicatives i la variable endògena del model.

Havent finalitzat aquesta primera estimació, una millora que es podria proposar consisteix en eliminar la variable $Perc\_noc$ de l'anàlisi (es recorda que el coeficient relacionat amb la mateixa no és estadísticament diferent de 0) i estimar un nou model. Tot i així, es rebutja aquesta opció degut al reduït nombre de variables que conté el model.

Per tant, es considera que el model estimat és el model lineal adequat per tal de complir amb l'objectiu que s'està buscant. Es tenen amb compte els errors obtinguts del model i es realitza la corresponent anàlisi. En primer lloc, es mostra el següent gràfic 5.3.2 en què s'hi poden veure els errors del model (en percentatge) segons la observació:

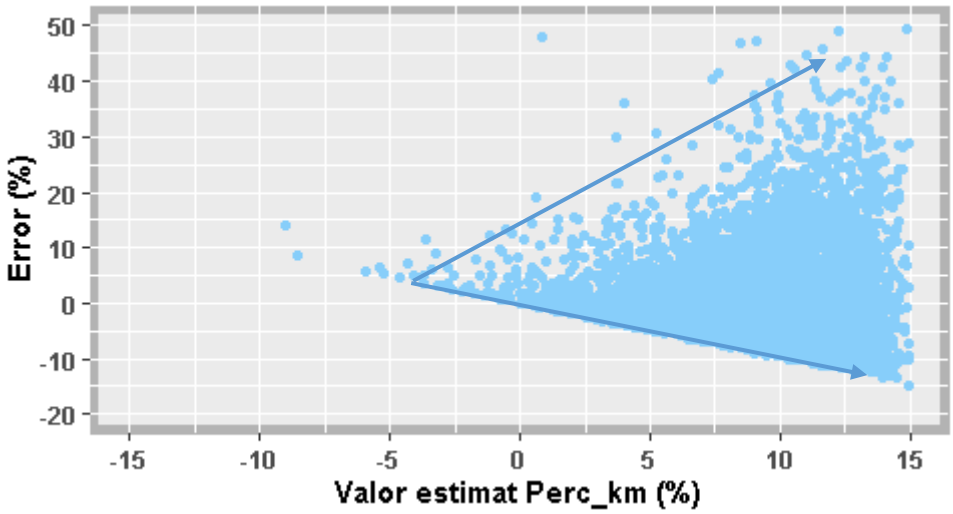
Gràfic 5.3.2: Errors del model segons la observació



Respecte aquest gràfic, cal recordar que les dades s'han ordenat de menor a major segons el valor real de la variable resposta en cada observació ($Perc\_km$). Gràcies a aquest procediment, es pot veure que els errors del model creixen amb majors valors reals de la variable resposta. D'aquesta manera, es pot afirmar que el model estimat tindrà problemes i, per tant, realitzarà una inferència incorrecta per a les observacions en què la variable resposta prengui valors alts. Es marquen aquestes observacions amb un requadre blau.

Per seguir amb l'anàlisi dels errors del model, a continuació, es pot veure el següent gràfic 5.3.3 en què es relacionen els valors estimats pel model per a cada observació de la variable $Perc\_km$ en funció dels seus corresponents errors (també en percentatge):

Gràfic 5.3.3: Errors del model estimat segons el valor estimat de $Perc\_km$



Respecte aquest gràfic, cal destacar una sèrie d'aspectes:

- La distribució dels errors del model no resulta ser homoscedàstica: la variància va augmentant a mesura que va augmentant el valor estimat pel model de $Perc\_km$ per a cada observació. S'observa una forma de piràmide marcada amb fletxes blaves que indica aquest augment de la variabilitat.
- Es pot veure que alguns dels valors estimats no es troben en l'interval de la variable resposta[10]. Concretament, si s'analitza la distribució dels valors estimats de la variable $Perc\_km$ pel model, es mostra la següent taula 5.3.4 amb els corresponents estadístics:

Taula 5.3.4: Estadístics univariants dels valors estimats pel model de $Perc\_km$

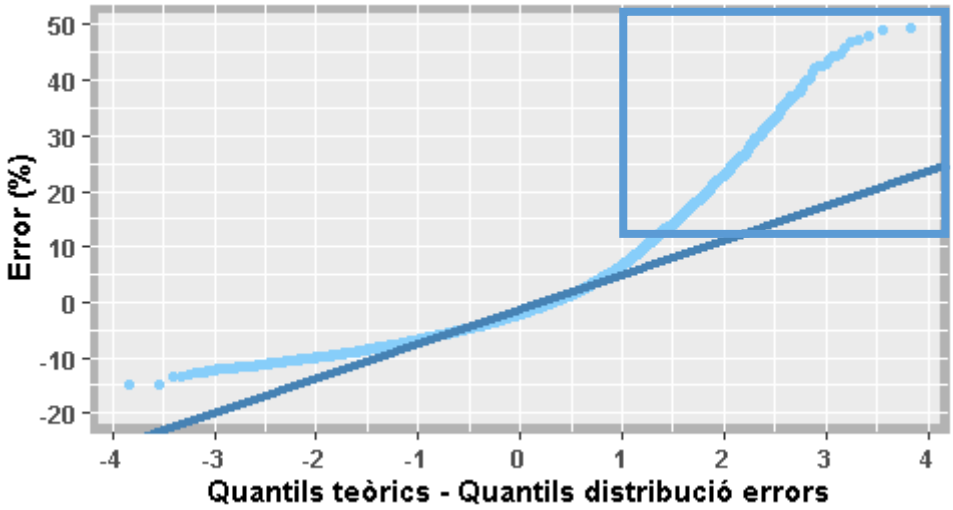
Estadístic	Mitjana	Desv. Est.	Min.	Max.
	9,11	3,02	−17,33	15,86

D'aquesta taula, cal destacar el valor màxim d'aquests valors estimats, el qual és de 15,86. Aquest valor resulta estar lluny dels valors alts que prenia la variable resposta entre les dades amb les que s'està treballant (al voltant de $50 - 60$). D'aquest fet, es pot afirmar que el model lineal no és un bon model per a realitzar aquest estudi. Tanmateix, és necessari comentar que una possible millora seria la d'estimar un model de regressió logístic, el qual forçaria que la variable resposta estigués limitada amb un interval de valors concret: $[0, 1]$.

Per últim, es mostra el següent *plot probabilístic normal* mitjançant el qual es pot tenir una idea gràfica de si els errors del model seguiran o no una distribució normal. Es mostra el mateix en la següent regió gràfica 5.3.5:

[10] Es recorda que aquest interval és el següent: $[0, 100]$.

Gràfic 5.3.5: PPN per als errors del model



Si s'observa aquest gràfic, es pot tenir una primera idea que els errors del model no segueixen una distribució normal. A més, si es realitza el test de *Jarque-Bera* per contrastar la normalitat dels errors (veure secció 4.3.), s'arriba al següent resultat que es mostra en la taula 5.3.6:

Taula 5.3.6: Test de *Jarque Bera* per als errors del model

$Estadístic\ JB = 9711{,}5$	$Graus\ de\ llibertat = 2$	$P - valor\ (> \chi^2_2) =< 0{,}01$

Després de la realització d'aquest contrast, el reduït p-valor obtingut fa acceptar la hipòtesi alternativa del contrast, aspecte que implica que amb una confiança del 99%, es pot acceptar que els errors del model no segueixen una distribució normal. De la mateixa manera, es pot confirmar que els errors que contribueixen amb una importància més gran a la no normalitat dels errors, són els que prenen valors més elevats. Tal i com s'ha realitzat anteriorment, es marquen aquests errors en els diferents gràfics 5.3.3 i 5.3.5 amb un requadre blau.

Després de la finalització d'aquesta primera anàlisi en la que s'ha estimat un model lineal, s'arriba a les següents conclusions:

- Els errors del model no segueixen una distribució normal ni resulten ser homoscedàstics.
- Per a aquesta aplicació, es pot dir que es tenen problemes a l'hora d'estimar valors alts de la variable resposta. Amb aquest model, aquests casos quedaran sub-estimats.
- Els valors estimats de la variable $Perc\_km$ per aquest model no es corresponen amb els valors que hauria de tenir la variable resposta (en la taula 5.3.4, se'n pot

veure la seva distribució). Es recorda que una forma de millorar aquest aspecte seria mitjançant l'estimació d'un model de regressió logística.

Aquest seguit de conclusions portarien a determinar que, tampoc per a aquesta aplicació, l'estimació per mínims quadrats ordinaris seria adequada. Degut a l'incompliment de les hipòtesis bàsiques d'aquest model, no es garanteixen les propietats bàsiques per als estimadors. Tanmateix, aquest aspecte donarà lloc a una inferència incorrecta. En aquest cas, el model lineal resulta ser més inadequat que en el cas anterior degut a que la variable resposta es troba limitada en un interval de valors concret. Tal i com s'ha realitzat en el cas anterior, s'hauria d'optar per altres tipus d'estimacions si es vol estimar el percentatge de quilòmetres que recorre un conductor per sobre de la velocitat permesa al cap d'un any.

5.4. Segona estimació: Regressió quantílica logística

Tal i com ha passat en la primera aplicació, amb l'estimació del model lineal que s'acaba de fer, s'ha vist que es tenen molts problemes per a analitzar observacions amb valors reals alts de la variable resposta. En aquest punt, es recorda que la variable resposta fa referència al percentatge de quilòmetres conduïts per sobre de la velocitat permesa per una persona al cap d'un any. A causa d'això, valors alts implicaran una major propensió a la sinistralitat i, en definitiva, una major probabilitat de patir un accident. Per tant, degut a que l'interès està en aquests conductors, seguint una metodologia semblant que en la primera aplicació, es proposa un altre tipus d'estimació: la regressió quantílica logística. Es recorda que aquesta regressió utilitza una metodologia semblant a la regressió quantílica i és estimada per quantils de la variable resposta. Per altra banda, la regressió quantílica logística implica que la variable resposta es situï amb un interval de valors concret, tal i com succeeix en aquest cas. Es recorda que $y \in [0, 100]$.

Per a desenvolupar aquest tipus d'estimació, s'utilitzarà la funció *Log.lqr()* del paquet *lqr* d'R. S'inicia aquesta anàlisi estimant la regressió quantílica logística que fa referència al percentil 50 de la variable resposta. En aquest cas, s'afegeixen les variables $Edat$ i $Edat$ al quadrat ($Edat^2$) com a variables explicatives al model[11]. Es realitza d'aquesta manera ja que es recorda que quan s'ha fet la descriptiva multivariant de la variable resposta, s'ha observat un efecte quadràtic d'aquesta variable per als quantils més elevats de la variable resposta. En primer lloc, es mostra l'equació del model especificat:

$$Q_{logit(Perc\_km)}(p) = F\left(\beta_{p,0} + \beta_{p,1} * Ln\_km + \beta_{p,2} * Perc\_urb + \beta_{p,3} * Perc\_noc + \beta_{p,4} * Sexe + \beta_{p,5} * Edat + \beta_{p,6} * Edat^2\right).$$

[11] Cal recordar que ja que s'ha inclòs la variable $Edat$ al quadrat ($Edat^2$), també s'inclou la variable $Edat$ en el model.

On Q representa el p quantil condicional de la variable resposta donat un vector de predictors i F és la transformació lineal del predictor lineal que garanteix que la predicció es troba entre 0 i 100. Per a aquest cas concret , $p = 0{,}50$, que implica analitzar la mediana de la variable $Perc\_km$. El subíndex $logit$ implica que s'ha realitzat aquesta transformació definida en la metodologia sobre el predictor lineal per a obtenir el quantil estimat de la variable resposta.

Es mostra aquesta estimació en la següent taula 5.4.1:

Taula 5.4.1: Model de regressió quantílica logística per al quantil 0,5 de $Perc\_km$ (mediana)

Variable	Coeficient	Error est.	Valor z	P-valor ($< \|t\|$)
$Constant$	$-8{,}22$	1,81	$-4{,}54$	$< 0{,}01$
Ln_km	0,46	0,02	22,85	$< 0{,}01$
$Perc\_urb$	$-0{,}02$	0,001	$-12{,}04$	$< 0{,}01$
$Perc\_noc$	0,001	0,004	0,39	0,70
$Sexe$	0,02	0,05	5,19	$< 0{,}01$
$Edat$	0,13	0,14	0,88	0,38
$Edat^2$	$-0{,}003$	0,003	$-0{,}94$	0,35

Respecte aquest model estimat, es pot dir que:

- Si es realitza el contrast de significació individual per al seguit de coeficients obtinguts, per als coeficients relacionats amb les variables explicatives $Perc\_noc$, $Edat$ i $Edat^2$, la conclusió d'aquest test és que no es pot rebutjar la hipòtesi nul·la del contrast. Mitjançant l'observació dels p-valor d'aquests contrastos ($0{,}70$, $0{,}38$ i $0{,}35$), no es pot rebutjar amb una confiança del 95% que cadascun d'aquests paràmetres siguin estadísticament iguals a 0.
- Tal i com s'acaba de dir, respecte el coeficient de la variable $Edat^2$, no es pot rebutjar amb una confiança del 95% que el paràmetre sigui estadísticament igual a 0. Tot i així, es deixa en el model ja que, tal i com s'ha comentat, aquest efecte quadràtic de l'$Edat$ es podria donar per a quantils alts de la variable $Perc\_km$.
- La interpretació dels diferents coeficients estimats és equivalent al cas de la regressió logística. Per a realitzar aquesta interpretació, cal utilitzar el concepte d'*Odds ratio* (a partir del risc relatiu), el qual pot ser definit de la següent manera per a un cert percentatge p:

$$Odd = \frac{p}{1-p} \; on \; p = \frac{e^{\beta_0+\beta_1*x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1*x_1+\cdots+\beta_k x_k}}.$$

Si es substitueix p en la primera expressió, s'arriba a la conclusió que el risc relatiu és:

$$Odd = e^{\beta_0+\beta_1*x_1+\cdots+\beta_k x_k}.$$

Si es tenen amb compte dos individus amb les mateixes variables explicatives, excepte una (x_i), la qual pren els següents valors:

- La variable x_i per al primer individu pren el valor de $x_{i1} = x_{i1}$.
- La variable x_i per al segon individu pren el valor de $x_{i2} = x_{i1} + 1$.

I si es realitza el quocient dels seus *Odds*, s'obté l'*Odds ratio*:

$$\frac{Odd_2}{Odd_1} = \frac{e^{\beta_0+\beta_1*x_1+\cdots+\beta_i(x_{i1}+1)+\cdots+\beta_k x_k}}{e^{\beta_0+\beta_1*x_1+\cdots+\beta_i(x_{i1})+\cdots+\beta_k x_k}} = e^{\beta_i}.$$

Per tant, un augment d'una unitat en el predictor x_i, s'espera que modifiqui el quantil de la variable resposta amb un $\left(e^{\beta_i} - 1\right)\%$. Per tant, cal calcular els valors e^{β_i}, els quals es mostren en la següent taula 5.4.2:

Taula 5.4.2: Valors e^{β_i} per als diferents coeficients del model

Coeficient relacionat amb la **variable**	Valor e^{β_i}
Ln_km	1,59
$Perc\_urb$	0,98
$Perc\_noc$	1,00
$Sexe$	1,30

Per a iniciar aquestes interpretacions, cal tenir clar que:

- Un coeficient $\beta_i > 0$ fa augmentar el quantil de la variable resposta.

- Un coeficient $\beta_i < 0$ fa reduir el quantil de la variable resposta.

Seguidament, es posaran dos exemples (un per a una variable categòrica i un per a una variable quantitativa) de quins són els efectes sobre el quantil de la variable resposta que té un augment o decrement d'una unitat del predictor. Per al quantil 0,5 de la variable resposta:

- Un augment d'una unitat del predictor $Perc\_urb$ fa reduir un 2% ($0{,}98 - 1 = -0{,}02$) el quantil 0,50 de la variable resposta.
- El fet de passar de ser dona a ser un home ($Sexe = 1$) fa augmentar un 30% ($1{,}30 - 1 = 0{,}30$) el quantil 0,50 de la variable resposta.

Per al predictor $Edat$, l'augment d'una unitat en el mateix, és més difícil de ser interpretat i s'ha de calcular la següent expressió:

$$e^{\beta_6+(((Edat_1)^2-(Edat_2)^2)*\beta_7)}.$$

On $Edat_1$ i $Edat_2$ són els valors de l'edat per als quals es vol calcular l'efecte sobre el quantil 0,50 de la variable resposta. Sempre succeirà que $Edat_1 > Edat_2$. Es posa l'exemple concret de quin seria l'efecte sobre el quantil 0,50 de la variable resposta el

fet d'augmentar el predictor $Edat$ amb una unitat (de 21 a 22, per exemple). Per a aquest cas, es calcula:

$$e^{\beta_5+((22^2-21^2)*\beta_6)} = 1{,}00.$$

On $\beta_5 = 0{,}127$ i $\beta_6 = -0{,}003$. En aquest cas, el fet d'augmentar el predictor $Edat$ de 21 a 22 no modifica el quantil 0,50 de la variable resposta. Tot i així, es pot deduir que a majors valors d'$Edat$, el terme $(Edat_1)^2 - (Edat_2)^2$ serà més gran i això farà incrementar (cap a positiu) l'efecte d'aquesta variable sobre el quantil 0,50 de la variable resposta.

Per tant, un cop s'ha estimat la regressió quantílica logística per al quantil 0,50, seguidament se n'estimaran d'altres relacionades amb altres quantils de la variable resposta. A continuació, es mostra la següent taula 5.4.3 en què s'hi poden veure els diferents valors dels coeficients per a un seguit de quantils de la variable $Perc\_km$. Tanmateix, s'afegeixen els valors dels coeficients d'un model lineal amb aquestes variables per tal de poder comparar-ne els valors:

Taula 5.4.3: Coeficients segons variables per a models de regressió quantílica logística per als quantils de la variable $Perc\_km$: $0{,}05 - 0{,}25 - 0{,}50 - 0{,}75 - 0{,}95$ $(+MQO)$

	Quantil de Perc_km					*MQO*
	0,05	0,25	0,50	0,75	0,95	
Constant	−13,67 (< 0,01)	−11,65 (< 0,01)	−8,22 (< 0,01)	−6,35 (< 0,01)	−4,99 (< 0,01)	−18,74 (< 0,01)
Ln_km	1,43 (< 0,01)	0,79 (< 0,01)	0,46 (< 0,01)	0,27 (< 0,01)	0,11 (< 0,01)	1,79 (< 0,01)
Perc_urb	−0,03 (< 0,01)	−0,02 (< 0,01)	−0,02 (< 0,01)	−0,02 (< 0,01)	−0,03 (< 0,01)	−0,13 (< 0,01)
Perc_noc	−0,005 (< 0,01)	−0,002* (0,457)	0,001* (0,698)	0,001* (0,758)	0,001* (0,381)	0,01* (0,36)
Sexe	0,20 (< 0,01)	0,26 (< 0,01)	0,26 (< 0,01)	0,22 (< 0,01)	0,11 (< 0,01)	1,83 (< 0,01)
Edat	−0,33 (< 0,01)	0,09* (0,462)	0,13* (0,380)	0,20* (0,101)	0,30 (< 0,01)	1,14 (0,032)
$Edat^2$	0,006 (< 0,01)	−0,002* (0,393)	−0,003* (0,349)	−0,004* (0,094)	−0,006 (< 0,01)	−0,02 (0,028)

Cal remarcar que els valors entre parèntesi que es troben a sota dels valors dels coeficients fan referència al p-valor del contrast d'hipòtesi individual del corresponent coeficient.

Respecte aquesta taula, es poden realitzar els següents comentaris:

- Per als coeficients marcats amb el superíndex $*$, no es pot rebutjar amb una confiança del 95% que els paràmetres siguin estadísticament iguals a 0.

- Respecte els coeficients relacionats amb la variable $Edat$, s'ha confirmat la hipòtesi plantejada anteriorment: per a quantils alts de la variable resposta, amb una confiança del 95%, es pot rebutjar que el paràmetre sigui estadísticament igual a 0. Tanmateix, es confirma aquesta hipòtesi per a quantils reduïts de la variable resposta. Per tant, aquesta variable és un clar exemple de variable que canvia el seu efecte segons el quantil de la variable resposta al qual s'estigui fent referència.
- Tal i com succeïa en la primera aplicació, quan s'ha estimat la regressió quantílica, els efectes dels diferents predictors solen tenir el mateix signe (positiu o negatiu) independentment del quantil de la variable resposta al qual es faci referència. Tot i així, les magnituds dels coeficients sí que canvien.
- Les interpretacions dels diferents efectes que tenen aquests predictors sobre el corresponent quantil de la variable resposta s'han de fer de la mateixa manera que s'han realitzat per al cas anterior: la regressió quantílica logística per al quantil $0{,}50$.

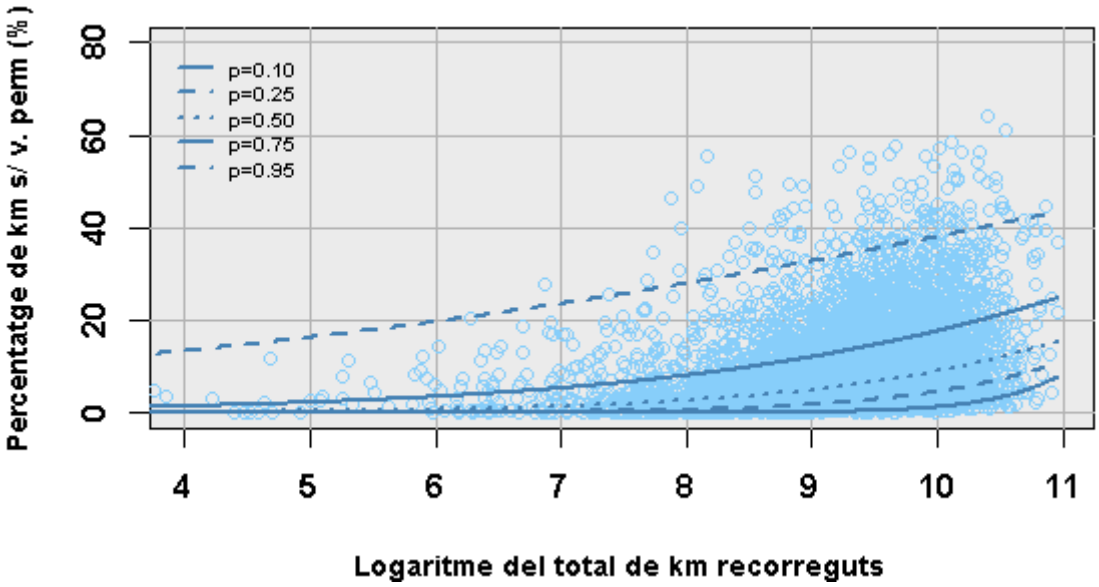
Si s'observa la taula anterior, es pot veure que els coeficients relacionats amb la variable Ln_km són els que varien més amb els quantils de la variable resposta. Per tant, a continuació, s'analitza la relació entre aquestes dues variables.

Per a iniciar aquesta anàlisi, s'estimen un seguit de regressions quantíliques logístiques de la forma:

$$Q_{logit(Perc\_km)}(p) = F\left(\beta_{p,0} + \beta_{p,1} * Ln\_km\right).$$

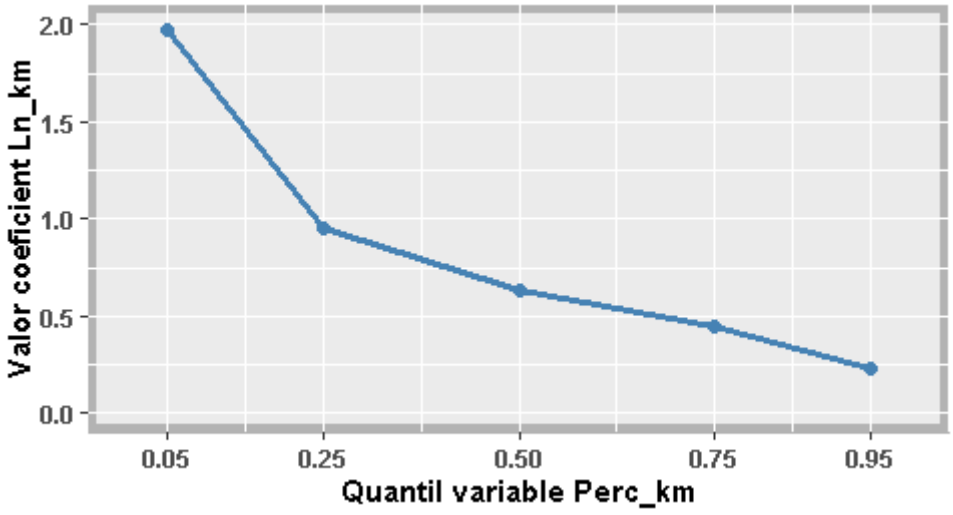
Aquest procediment es realitza per a un seguit de quantils de la variable resposta: $0{,}05 - 0{,}25 - 0{,}50 - 0{,}75 - 0{,}95$. D'aquesta manera, es podrà veure l'efecte diferenciat que té aquesta variable en funció del quantil de la variable resposta que s'estigui considerant. Un cop estimades aquesta sèrie de regressions, es mostra la següent regió gràfica 5.4.4 en què es mostren gràficament els diferents quantils de $Perc\_km$ estimats a partir de diferents valors que pugui prendre la variable Ln_km:

Gràfic 5.4.4: Regressions quantíliques logístiques de la forma $Q_{logit(Perc\_km)}(p) = F(\beta_{p,0} + \beta_{p,1} * Ln\_km)$ per als quantils $p = 0{,}05 - 0{,}25 - 0{,}50 - 0{,}75 - 0{,}95$ de la variable $Perc\_km$



Amb aquest gràfic, es pot veure com varia l'efecte del predictor Ln_km en funció del quantil de la variable resposta al qual s'estigui fent referència. Tanmateix, es mostra el següent gràfic 5.4.5 en què s'hi poden veure els diferents valors que pren el coeficient $\beta_{p,1}$ en funció del quantil de la variable $Perc\_km$ que s'estigui considerant:

Gràfic 5.4.5: Valors dels coeficients $\beta_{p,1}$ (Eix d'ordenades) en funció dels quantils de la variable resposta (eix d'abscisses)



Amb aquest gràfic 5.4.5, es pot veure el diferent valor que pren el coeficient $\beta_{p,1}$ dels anteriors models de regressió quantílica logística estimats en funció del quantil de la variable $Perc\_km$ que s'estigui considerant.

A continuació, es té amb compte la importància que té la variable $Sexe$ a l'hora d'estimar els paràmetres dels diferents models estimats. La importància que pren aquesta variable es pot veure en:

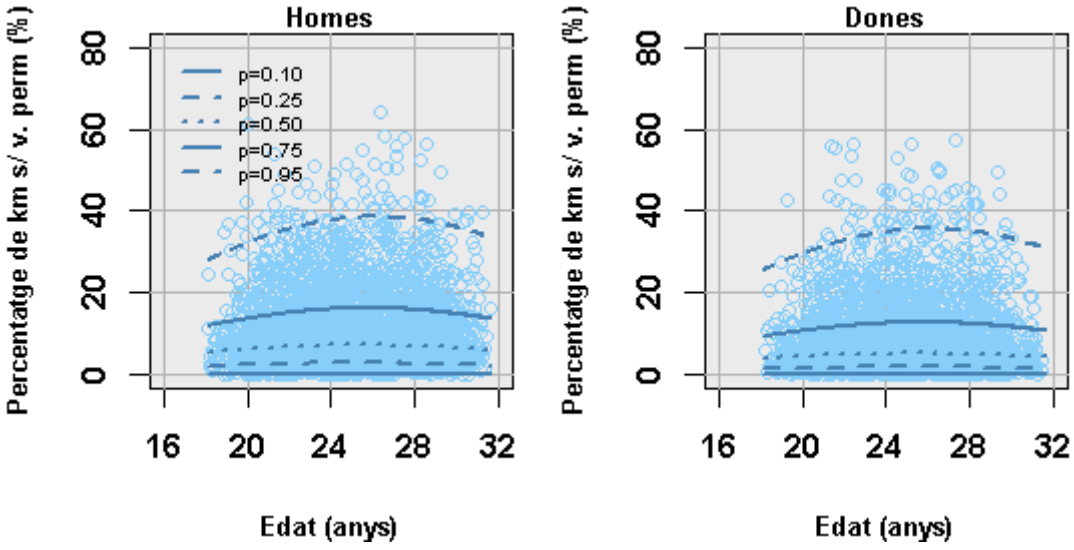
- La regressió quantílica logística estimada associada al quantil 0,50 de la variable $Perc\_km$, el fet de ser home o dona feia variar força la variable resposta.
- Tanmateix, si s'observa la taula anterior on es presenten els diferents coeficients relacionats amb diferents estimacions, es pot veure que els coeficients relacionats amb la variable $Sexe$ solen prendre valors alts i positius, indicant d'aquesta manera que els homes prendran valor més elevats de la variable resposta.

Per altra banda, quan s'ha explicat com s'ha de fer la interpretació dels diferents coeficients estimats, s'ha vist que l'efecte sobre la variable resposta del model d'un augment d'una unitat de la variable $Edat$, resultava força difícil de ser interpretat degut a la inclusió de la variable $Edat$ al quadrat en el model. Per a aclarir, aquestes consideracions, s'estimen un seguit de models de la forma:

$$Q_{logit(Perc\_km)}(p) = F\left(\beta_{p,0} + \beta_{p,1} * Sexe + \beta_{p,2} * Edat + \beta_{p,3} * Edat^2\right).$$

Per iniciar aquesta anàlisi, s'estimen un seguit de models de regressió quantílica logística, relacionats amb diferents quantils de la variable $Perc\_km$, de la forma definida anteriorment. A continuació, es representen en la següent regió gràfica 5.4.6 els valors estimats de $Perc\_km$ en funció del valor de la variable $Edat$ i del $Sexe$ del conductor:

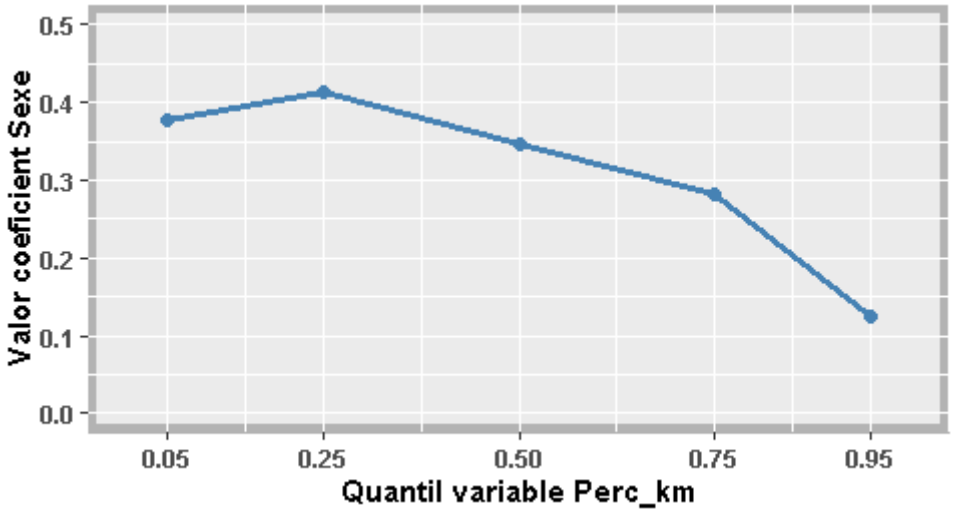
Gràfic 5.4.6: Models de regressió quantílica logística de la forma $Q_{logit(Perc\_km)}(p) = F(\beta_{p,0} + \beta_{p,1} * Sexe + \beta_{p,2} * Edat + \beta_{p,3} * Edat^2)$ per als quantils $p = 0{,}05 - 0{,}25 - 0{,}50 - 0{,}75 - 0{,}95$ de la variable $Perc\_km$



Amb aquests gràfics, es confirma l'efecte quadràtic de l'edat a quantils elevats de la variable resposta. Tanmateix, es pot dir que no es visualitza aquest efecte en quantils baixos de la variable resposta tal i com s'havia mencionat anteriorment. De la mateixa manera, un altre aspecte a remarcar és la tendència que tenen els homes a tenir valors més alts de la variable resposta amb la que s'està treballant.

Per últim, degut a que s'està considerant la importància de la variable $Sexe$ a l'hora d'estimar els diferents quantils de la variable $Perc\_km$, es mostren en el següent gràfic 5.4.7, els valors del coeficient $\beta_{p,1}$ del model anterior en funció del quantil de la variable resposta que s'estigui considerant.

Gràfic 5.4.7: Valors dels coeficients $\beta_{p,1}$ (Eix d'ordenades) en funció dels quantils de la variable resposta (eix d'abscisses)



5.5. Percentils condicionats vs. Percentils no condicionats

Tal i com s'ha anunciat anteriorment, es finalitza aquesta segona aplicació de la mateixa manera que s'ha fet per a la primera aplicació. Per tant, a continuació s'adapta la metodologia proposada anteriorment per al cas de la regressió quantílica logística. Per tant, en primer lloc, es recorden les definicions dels conceptes de percentil condicionat i percentil no condicionat:

- Percentil no condicionat: El percentil no condicionat d'una observació correspon al percentil al qual pertany la variable resposta real d'aquella observació respecte la distribució de probabilitats de la variable resposta de la mostra amb la que s'estigui treballant.
- Percentil condicionat: El percentil condicionat d'un cas concret correspon al percentil al qual correspon en funció dels valors que prenguin les seves variables explicatives.

De la mateixa manera que s'ha fet anteriorment, es mostra el càlcul d'aquests dos conceptes per a una observació concreta. En aquest cas, es tria l'observació 3897 de la mostra amb la que s'està treballant per a exemplificar aquests dos càlculs. S'inicia aquesta anàlisi presentant els valors de les variables explicatives d'aquesta observació en la següent taula 5.5.1:

Taula 5.5.1: Valors de les variables explicatives de l'observació 3897

Variable	Ln_km	$Perc\_urb$	$Perc\_noc$	$Sexe$	$Edat$	$Edat^2$
Valor	8,95	44,98	8,68	0,00	23,38	546,43

1- Percentil no condicionat: La variable resposta real d'aquesta observació pren el valor de 6,25%. Es recorda que aquest valor indica que aquesta dona ($Sexe = 0$) ha recorregut, durant el 2010, un 6,25% dels quilòmetres totals recorreguts per sobre de la velocitat permesa. Seguidament, es remet a la funció de distribució empírica ($\widehat{F(y)}$) dels valors observats de la variable resposta i es mostra la següent taula 5.5.2 amb la següent informació:

Taula 5.5.2: Valors de la funció de distribució empírica dels valors observats de la variable resposta ($\widehat{F(y)} = 0{,}50\ i\ \widehat{F(y)} = 0{,}51$)

	$Probabilitat\ acumulada\ o\ \widehat{F(y)}$	
	0′50	0′51
$Valor$	6,15 %	6,30 %

On $\widehat{F(y)}$és la funció de distribució empírica dels valors reals o observats de la variable resposta ($Perc\_km$). Tenint amb compte els valors mostrats en aquesta taula, s'arriba a la conclusió que el percentil no condicionat de la observació 3897 és el percentil 51.

2- Per a realitzar el càlcul del percentil condicionat d'aquesta observació, es segueix un procediment semblant al que s'ha seguit en la primera aplicació, el qual es resumeix a continuació:

- Primer, s'estimen els diferents models de regressió quantílica logística relacionats amb cadascun dels percentils de la variable resposta considerats ($p = 0{,}01 - \cdots - 0{,}99$).
- En segon lloc, s'estima el valor del predictor lineal de la observació en funció dels diferents models estimats i els seus valors de les variables explicatives.
- Seguidament, s'aplica la transformació lineal *logit* al predictor lineal per tal de garantir que el quantil de la variable resposta estimat es trobi entre 0 i 100. Aquesta transformació lineal es defineix com:

$$Q_y(p) = \frac{e^{\beta_{p,0}+\beta_{p,1}*x_1+\beta_{p,2}*x_2+\cdots+\beta_{p,k}*x_k} * y_{max} - y_{min}}{1 + e^{\beta_{p,0}+\beta_{p,1}*x_1+\beta_{p,2}*x_2+\cdots+\beta_{p,k}*x_k}}.$$

On $\beta_{p,0} + \beta_{p,1} * x_1 + \beta_{p,2} * x_2 + \cdots + \beta_{p,k} * x_k$ és el predictor lineal, y_{min} és 0, y_{max} és 100 i $Q_y(p)$ és el quantil de la variable $Perc\_km$ estimat.

- Per últim, després d'aplicar aquesta transformació lineal als diferents predictors lineals calculats, de tots els valors $Q_y(p)$ estimats, es troba el més proper (per excés o per defecte) a la resposta real de l'observació. El percentil de la variable resposta associat al model que ha estimat el $Q_y(p)$ més proper serà el percentil no condicionat de l'observació.

Per a l'observació que s'està analitzant, es calcula el valor del predictor lineal utilitzant els coeficients estimats del model de regressió quantílica logística associats al percentil 69 de la variable resposta. Aquests valors es defineixen en la següent taula 5.5.3:

Taula 5.5.3: Valors de les variables explicatives de l'observació 3897 i valors dels coeficients estimats del model de regressió quantílica associat al percentil 69 de la variable $Perc\_km$

	β_0	Ln_km	$Perc\_urb$	$Perc\_noc$	$Sexe$	$Edat$	$Edat^2$
Valor		8,95	44,98	8,68	0,00	23,38	546,43
Coef.	−6,73	0,31	−0,02	0,001	0,24	0,18	−0,004

El predictor lineal és pot calcular de la següent manera:

$$-6{,}73 + 0{,}31 * Ln_{km} - 0{,}02 * Perc\_urb + \cdots - 0{,}004 * Edat^2 = -2{,}70$$

I si se li aplica la corresponent transformació lineal:

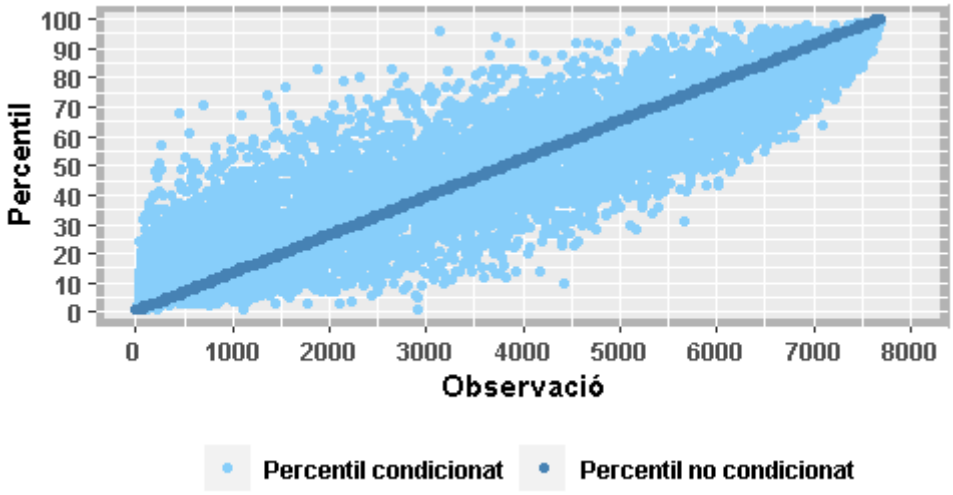
$$Q_y(0{,}69) = \frac{e^{-2{,}70} * 100 - 0}{1 + e^{-2{,}70}} = 6{,}30\%$$

Es pot apreciar que aquest valor és molt proper al valor observat de la variable resposta per a aquesta observació, el qual es recorda que és de 6,25%. Per tant, després de la realització d'aquests dos càlculs, s'arriba a la conclusió que per a l'observació 3897, el percentil no condicionat de la variable resposta és el 51 mentre que el percentil condicionat és el 69. Es recorda que aquest percentil condicionat està calculat en funció dels valors de les variables explicatives d'aquesta conductora.

Tal i com succeïa en la primera aplicació d'aquest treball, aquesta metodologia permetria extrapolar més enllà de la mostra amb la que s'està treballant. D'aquesta manera, es pot afirmar que si es tingués una població de referència (amb els mateixos valors de les variables explicatives que l'observació 3897), aquesta conductora es situaria en el percentil 69 de la variable $Perc\_km$.

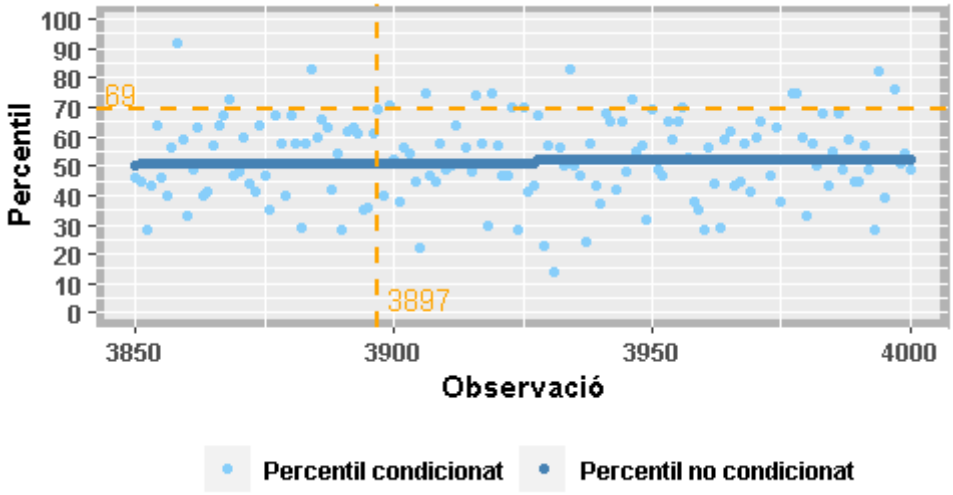
Seguidament, es realitzen aquests dos càlculs per a totes les observacions de la base de dades i es mostren els valors dels percentils condicionats i els percentils no condicionats en el següent gràfic 5.5.4:

Gràfic 5.5.4: Percentils condicionats vs. percentils no condicionats per a totes les observacions de la mostra



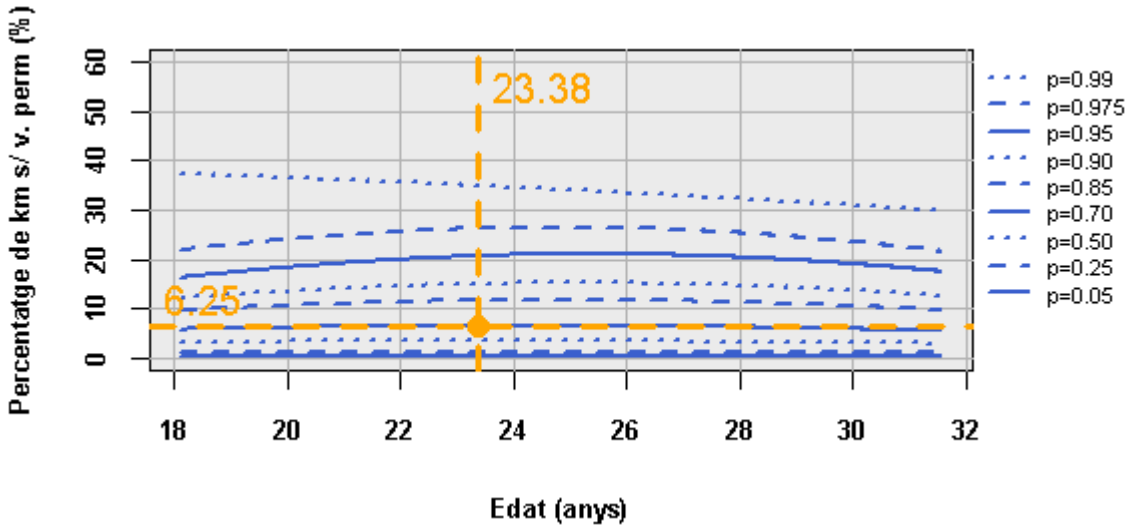
En aquest cas, s'observa encara més dispersió que per al cas de la primera aplicació. Es pot veure que hi ha molt poca correspondència entre els percentils condicionats i els no condicionats de la variable resposta per a cada observació. Aquesta dispersió es pot veure de forma més clara si es miren dos percentils concrets de la variable resposta. Per tant, en el següent gràfic 5.5.4 es mostren els valors d'aquests dos tipus de percentils per als percentils no condicionats 50 i 51 de la variable resposta:

Gràfic 5.5.4: Percentils condicionats vs. percentils no condicionats per als percentils no condicionats $50 - 51$ de la variable resposta



On la intersecció de les dos línies taronges discontínues marquen els valors dels dos percentils per a l'observació 3897, utilitzada anteriorment per a exemplificar el càlcul d'aquests dos tipus de percentils. Seguidament, si es té amb compte la població de referència d'aquesta conductora (formada per conductores amb els mateixos valors de les variables explicatives que l'observació 3897), es mostra el següent gràfic 5.5.5 amb els valors estimats de $Perc\_km$ en funció de valors d'$Edat$ per a diferents quantils de la variable $Perc\_km$ d'aquesta població de referència:

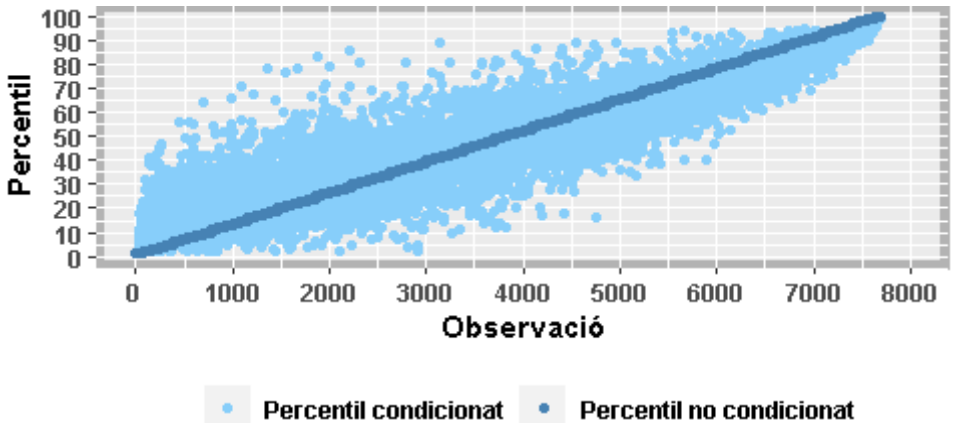
Gràfic 5.5.5: La conductora 3897 en la seva població de referència



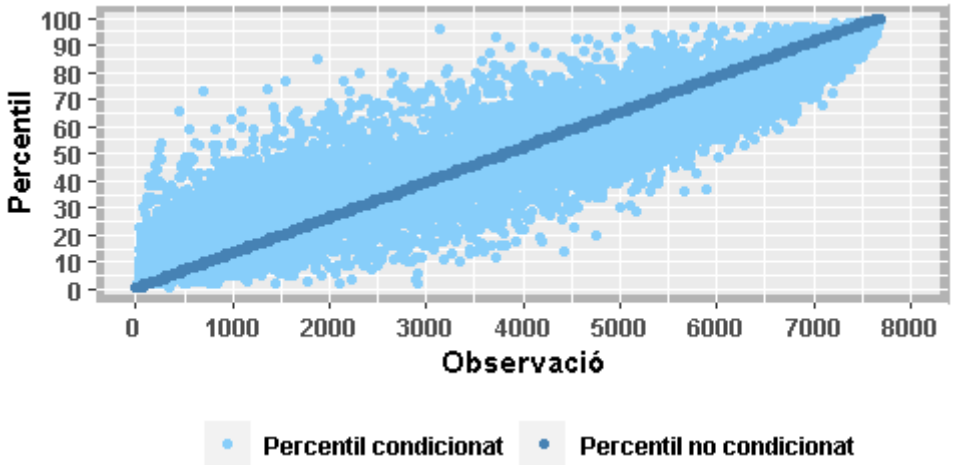
Mitjançant l'observació d'aquest gràfic 5.5.5, es pot veure que si es situa la conductora 3897 en la seva població de referència, aquesta es situaria en el percentil condicionat(69) de la distribució de valors de la variable $Perc\_km$ d'aquesta població.

Presentada aquesta metodologia i adaptada al cas de la regressió quantílica logística, es finalitza aquesta aplicació de la mateixa forma que s'ha fet en la primera. A continuació, s'estimen un seguit de models de regressió quantílica logística (cada cop s'afegeix una variable explicativa més) i es mostren una sèrie de gràfics (5.5.6,...,5.5.10) en què es pot veure com evoluciona la dispersió entre els dos tipus de percentils a mesura que s'afegeixen variables explicatives als models de regressió quantílica logística estimats:

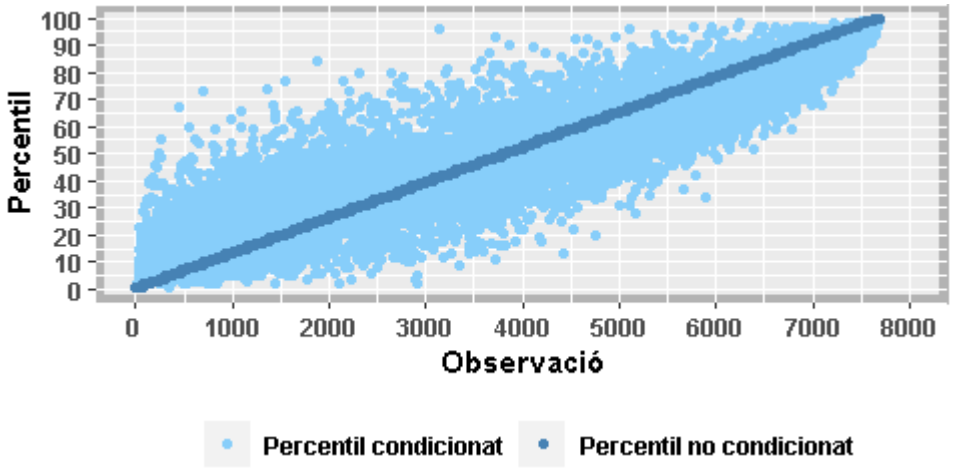
Gràfic 5.5.6: Models: $Q_{logit(Perc\_km)}(p) = F(\beta_{p,0} + \beta_{p,1} * Ln\_km)$



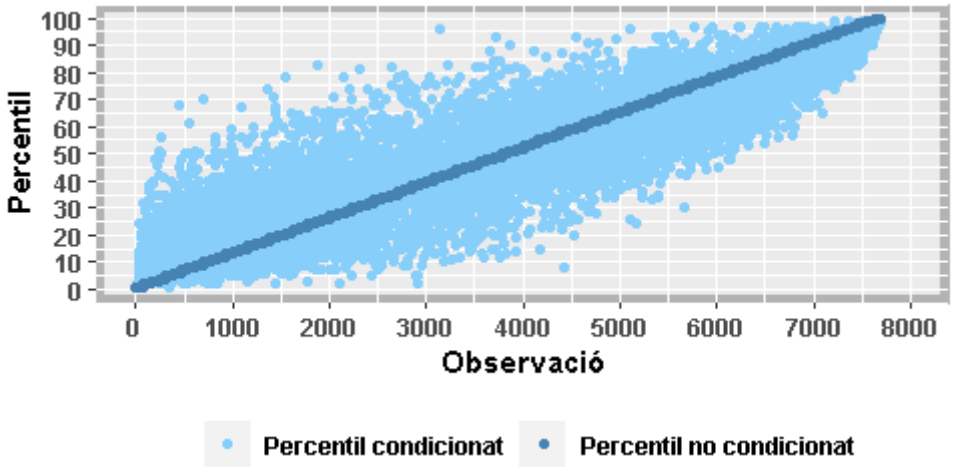
Gràfic 5.5.7: Models: $Q_{logit(Perc\_km)}(p) = F(\beta_{p,0} + \beta_{p,1} * Ln\_km + +\beta_{p,2} * Perc\_urb)$



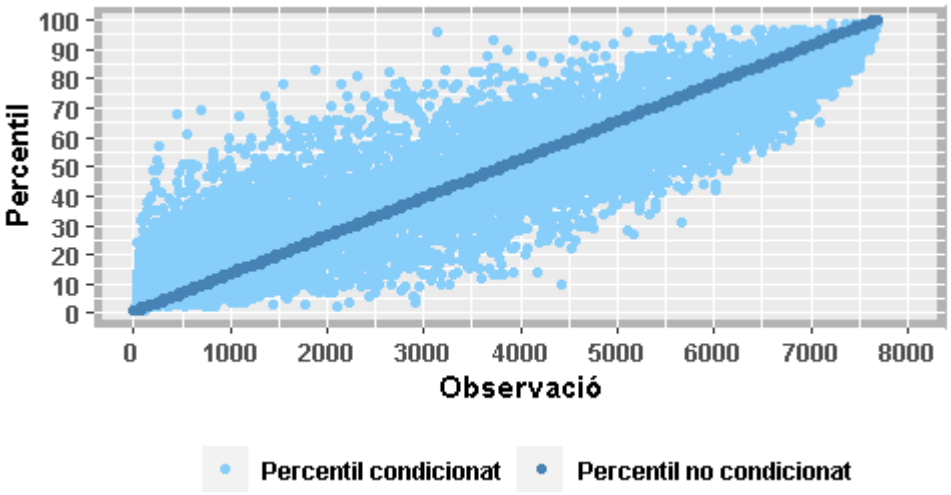
Gràfic 5.5.8: Models: $Q_{logit(Perc\_km)}(p) = F(\beta_{p,0} + \beta_{p,1} * Ln\_km + +\beta_{p,2} * Perc\_urb + \beta_{p,3} * Perc\_noc)$



Gràfic 5.5.9: Models: $Q_{logit(Perc\_km)}(p) = F(\beta_{p,0} + \beta_{p,1} * Ln\_km + +\beta_{p,2} * Perc\_urb + \beta_{p,3} * Perc\_noc + +\beta_{p,4} * Sexe)$



Gràfic 5.5.10: Models: $Q_{logit(Perc\_km)}(p) = F(\beta_{p,0} + \beta_{p,1} * Ln\_km + +\beta_{p,2} * Perc\_urb + \beta_{p,3} * Perc\_noc + +\beta_{p,4} * Sexe + \beta_{p,5} * Edat)$



Finalitzada aquesta segona aplicació del treball, a continuació es realitzaran una sèrie de consideracions:

- Tal i com ha succeït en la primera aplicació del treball, s'ha demostrat la limitació dels mínims quadrats ordinaris a l'hora d'estimar els paràmetres associats a diferents models de regressió lineals quan no es compleixen els supòsits bàsics per a la seva estimació. Aquest fet donarà lloc a forces problemes per portar a terme la inferència basada en aquests estimadors, els quals no tenen unes bones propietats.
- S'ha pogut veure la potencialitat de la regressió quantílica logística a l'hora d'estimar els paràmetres associats als diferents models de regressió segons els quantils de la variable resposta del model quan aquesta està limitada en un interval de valors concret. Aquest mètode resulta ser molt efectiu sempre que es tingui un interès en certs quantils de la variable resposta.
- Després de l'adaptació de la metodologia proposada en la primera aplicació per al cas de la regressió quantílica logística, s'ha pogut demostrar la utilitat que té la mateixa per a aquest cas d'estudi o altres de semblants. Es recorda que aquesta permet anar més enllà de la mostra amb la què s'està treballant i extrapolar cap a una hipotètica població de referència formada per n conductors amb les mateixes característiques que un conductor considerat.

  El percentil condicionat al qual quedaria situat un determinat conductor en la seva població de referència podria ser perfectament una mesura del risc utilitzada per entitats asseguradores per a classificar els conductors. En aquest cas, s'està comparant a un conductor considerat amb altres conductors amb les mateixes característiques. Des d'aquest punt de vista, resulta ser molt més correcta aquesta comparació que si se'l comparés amb altres conductors amb unes característiques diferents.

- Per últim, després de veure com evoluciona la dispersió entre els dos tipus de percentils considerats a mesura que s'afegeixen variables explicatives als diferents models de regressió quantílica logística estimats, s'ha pogut observar que, en aquest cas, s'assoleix la dispersió final entre els dos tipus de percentils d'una forma molt més ràpida. S'atribueix aquest fet a que, des d'un principi, s'està treballant amb variables explicatives quantitatives, fet que provoca que es situï més ràpidament a cada conductor en el seu percentil condicionat que li correspongui. Es recorda que en la primera aplicació, s'iniciava aquesta anàlisi afegint variables categòriques als diferents models de regressió quantílica estimats.

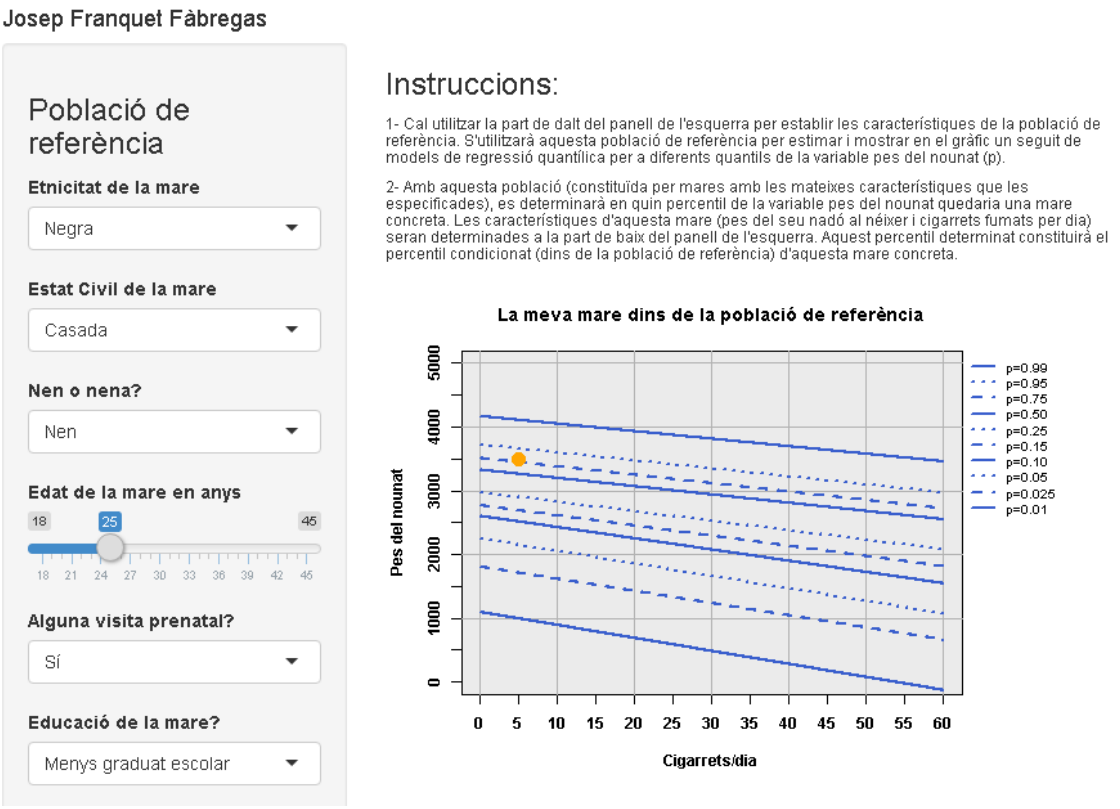
6. Aplicacions *Shiny*

Tal i com s'ha remarcat en la secció *introducció*, finalitzat el treball, s'ha decidit anar més enllà del seu simple desenvolupament. S'han programat, desenvolupat i presentat dues aplicacions interactives desenvolupades amb el paquet *Shiny* d'R, una per a cada aplicació del treball. Amb aquestes aplicacions interactives, l'usuari pot determinar una població de referència en la qual es situarà una mare (en la primera aplicació) o un conductor (en la segona aplicació) concrets. S'utilitza aquest apartat del treball per tal de realitzar una explicació i presentar aquestes dues aplicacions.

6.1. Aplicació *Shiny*: Pes dels nadons al néixer

En aquesta primera aplicació interactiva, l'usuari pot determinar una població de referència formada per n mares amb una sèrie de característiques. Aquestes característiques vindran determinades pels valors que prenguin les variables explicatives d'aquestes mares. Un cop especificada aquesta població de referència, es situarà en la mateixa a una mare en un percentil concret de la variable Pes. Es mostra una captura de la part superior de l'aplicació creada en la següent figura 6.1.1:

Figura 6.1.1: Part superior de l'aplicació *Shiny* per a la primera aplicació

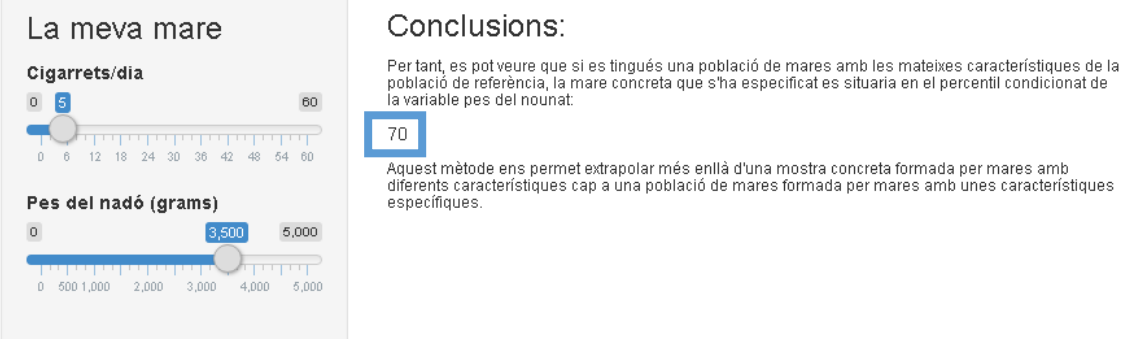


El panell de l'esquerra de la pantalla serveix per a determinar la població de referència. Aquesta es determina a partir d'especificar valors de les diferents variables explicatives dels models. Un cop especificada aquesta població de referència, s'han d'haver estimat un seguit de models de regressió quantílica per a certs quantils (es poden veure en la

llegenda) de la variable $Pes\ del\ nounat$. Cal remarcar que les variables explicatives que s'han utilitzat són les més adequades finalitzada la primera aplicació del treball. Estimats aquests models, es mostren en el gràfic, els valors estimats de la variable $Pes\ del\ nounat$ (eix d'ordenades) per a cadascun dels quantils considerats d'aquesta variable en funció de diferents valors que pugui prendre la variable $Cigarrets/dia$ (eix d'abscisses).

Seguidament, es mostra una captura de pantalla de la part inferior de l'aplicació en la següent figura 6.1.2:

Figura 6.1.2: Part inferior de l'aplicació *Shiny* per a la primera aplicació



Amb el panell de l'esquerra de la part inferior de l'aplicació, l'usuari pot determinar les característiques d'una mare concreta que es vulgui situar en la població de referència. Concretament, l'usuari pot determinar el nombre de $Cigarrets/dia$ fumats per la mare durant l'embaràs i el Pes del nadó al néixer amb $grams$. Un cop especificades aquestes característiques, el punt taronja que es pot veure en el gràfic de la part superior situarà a aquesta mare en aquesta població de referència. De la mateixa manera, el nombre que es troba dins del requadre blau indicarà el percentil condicionat d'aquesta mare en la seva població de referència. Es recorda que aquest percentil condicionat indica el percentil de la variable $Pes\ del\ nounat$, en el qual estaria aquesta mare concreta (amb aquestes característiques) dintre d'una hipotètica població formada per n mares d'idèntiques característiques. Aquesta aplicació interactiva està publicada en línia i es pot visitar utilitzant el següent enllaç:

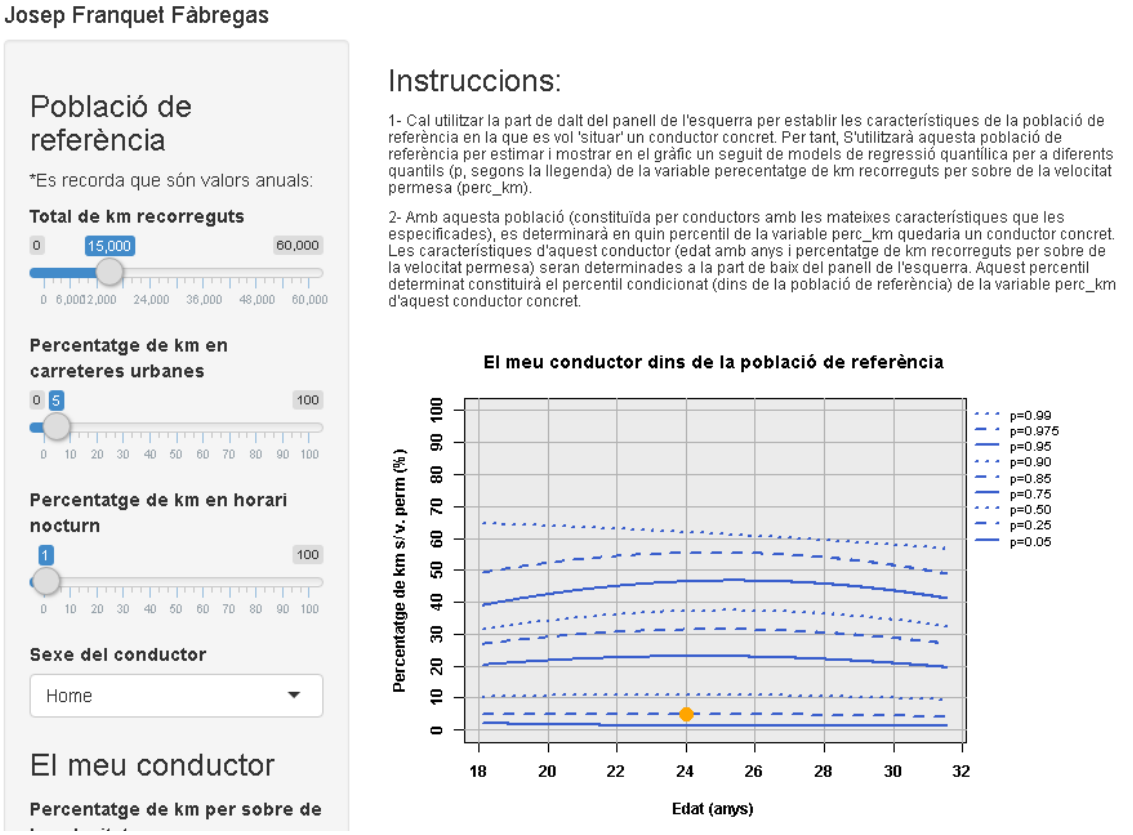
https://aplicacionstfgjosepfranquet.shinyapps.io/Shiny_publicat/

6.2. Aplicació *Shiny*: Percentatge de quilòmetres recorreguts per sobre de la velocitat permesa al cap d'un any

A continuació, s'utilitza aquest apartat del treball per explicar i presentar la segona aplicació *Shiny* relacionada amb el percentatge de quilòmetres recorreguts per un conductor per sobre de la velocitat permesa al cap d'un any. Tal i com es veurà, l'estructura és semblant que en el cas de l'anterior aplicació. En la part superior de l'aplicació, l'usuari disposa d'un panell per a determinar les característiques de la

població de referència. Tal i com s'ha fet en l'anterior aplicació, aquesta es determinarà a partir de donar valors a les variables explicatives dels models de regressió quantílica logística que s'estimaran. En aquest cas, les variables explicatives són les que s'han utilitzat quan s'han estimat els diferents models de regressió quantílica logística en la segona aplicació del treball. Seguidament, es mostra en la següent figura 6.2.1 una captura de pantalla de la part superior d'aquesta aplicació:

Figura 6.2.1: Part superior de l'aplicació *Shiny* per a la segona aplicació

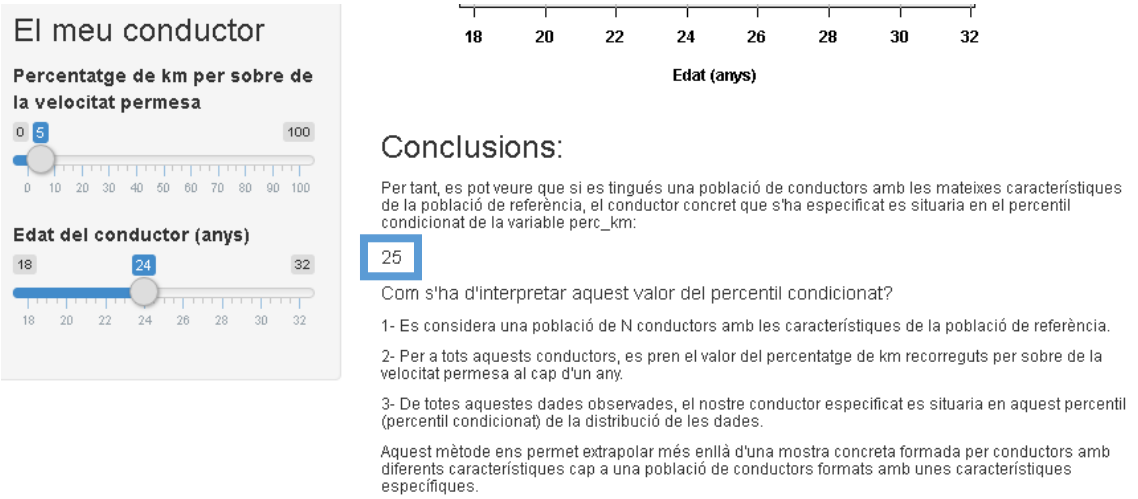


Un cop especificada aquesta població de referència, un seguit de models de regressió quantílica logística seran estimats per a certs quantils de la variables resposta dels models (Es poden veure aquests quantils en la llegenda del gràfic). Estimats aquests models, es mostren en el gràfic una sèrie de valors de la variable $Perc\_km$ estimats a partir d'aquests models per a cadascun dels quantils de la variable considerats en funció de valors que pugui prendre la variable $Edat$.

En la part inferior de l'aplicació, l'usuari disposa d'un panell per a determinar les característiques d'un conductor concret que l'usuari vulgui comparar amb la seva població de referència. Aquest panell es situarà en la part inferior esquerra de l'aplicació i permet a l'usuari determinar el percentatge de quilòmetres recorreguts per sobre de la velocitat permesa i l'edat d'aquest conductor. Igual que en l'anterior aplicació, un cop determinats aquests valors, el punt taronja indicarà en el gràfic on quedarà aquest

conductor en la seva població de referència. A continuació, es mostra en la següent figura 6.2.2, una captura de pantalla de la part inferior de l'aplicació:

6.2.2: Part inferior de l'aplicació *Shiny* de la segona aplicació



Tal i com es pot veure, a la part inferior dreta de l'aplicació, l'usuari podrà visualitzar quin és el percentil condicionat d'aquest conductor en la seva població de referència (Es marca aquest valor amb un requadre blau). Per tant, si es tingués una població de referència com l'especificada i un conductor amb les característiques especificades, aquest es situaria en el percentil 25 de la variable $Perc\_km$ en una hipotètica població formada per n conductors d'idèntiques característiques. Aquesta segona aplicació interactiva no s'ha pogut publicar en línia degut a la privacitat que s'ha de mantenir amb les dades que s'està treballant.

7. Conclusions

En aquest treball final de grau, s'han explorat, analitzat i aplicat dos mètodes de regressió nous, moderns i, tal i com s'ha vist, amb un gran potencial: la regressió quantílica i la regressió quantílica logística. Es tracta de dos mètodes de regressió que no solen ser explicats en l'àmbit acadèmic del grau en estadística i del grau en economia, els quals prefereixen centrar-se en altres mètodes d'estimació clàssics, com ara els mínims quadrats ordinaris. Després de la finalització d'aquest treball, es pot afirmar que aquests dos mètodes de regressió són una molt bona alternativa enfront aquests mètodes clàssics d'estimació quan l'interès no es troba en la predicció de la mitjana de la variable resposta en funció dels valors de les variables explicatives, sinó en els seus quantils.

La revisió de la literatura ha permès veure la modernitat d'aquests dos conceptes. En ella, es destaca la poca presència d'articles científics en què s'hagin utilitzat, sobretot, models de regressió quantílica logística. Tanmateix, fent referència a la regressió quantílica, s'hi destaca l'augment, en els últims anys, d'articles científics basats en aquesta metodologia. Per últim, s'hi remarca que l'àmbit econòmic és un dels camps d'estudi on s'ha utilitzat més aquest tipus de regressió, presentant i resumint dos articles recents de caràcter econòmic en què s'ha utilitzat la regressió quantílica.

Abans de procedir a presentar les conclusions més importants de les dues aplicacions reals desenvolupades, és necessari remarcar que la integritat del treball té un fort component economètric. Al llarg del mateix, s'estimen diferents models lineals (estimats per MQO) i, estimats aquests models, es procedeix amb una anàlisi dels mateixos: contrastos d'hipòtesi sobre els coeficients obtinguts, interpretació d'aquests coeficients, interpretació del coeficient de bondat de l'ajust del model, entre d'altres. Aquestes estimacions i les seves corresponents anàlisis, juntament amb la tipologia de les dades de la segona aplicació del treball, són les que formen la part econòmica del treball.

En la primera aplicació del treball, s'han estimat diferents models de regressió quantílica per tal d'estimar els paràmetres de diferents models lineals associats a certs quantils de la variable resposta, la qual és el pes d'un nadó acabat de néixer. Es tracta, per tant, de dades del camp de la bioestadística o de l'estadística mèdica. Concretament, es tracta de dades detallades de natalitat de l'any 1997 obtingudes del *National Institute of Health* dels Estats Units. La mostra amb la que es treballa està formada per 50.000 mares d'entre 18 i 45 anys que han donat a llum als Estats Units durant aquell mateix any.

Amb les dades disponibles, en primer lloc, s'ha realitzat una anàlisi descriptiva univariant de cadascuna de les variables de la base de dades. Respecte aquesta primera anàlisi, es pot afirmar que la variable resposta del model segueix aproximadament una

distribució normal. Tanmateix, cal destacar que la mostra està formada principalment per mares d'ètnia no negra, no fumadores durant l'embaràs i casades. La majoria d'aquestes mares havien realitzat la última visita prenatal durant l'últim trimestre d'embaràs i respecte el nivell d'estudis de les mares, la mostra estava força equilibrada pel que fa als diferents nivells d'estudis que podia prendre la variable categòrica que fa referència al nivell educatiu de la mare. Per últim, respecte els nounats, la mostra estava força equilibrada pel que fa al nombre de nens i al nombre de nenes.

Seguidament, s'ha realitzat una anàlisi descriptiva multivariant de la variable resposta. S'ha pogut observar que les mares d'ètnia no negra, casades i no fumadores solien presentar valors més alts de la variable resposta (Pes). De la mateixa manera, s'ha pogut observar la poca influència que té l'edat de la mare sobre el pes real del nounat. Després de prendre un seguit de decisions: eliminació de variables i creació d'una sèrie de variables dicotòmiques, s'ha procedit amb la primera estimació: un model lineal estimat per mínims quadrats ordinaris. Estimat aquest primer model lineal, s'ha comprovat que no s'han complert els supòsits bàsics d'estimació per tal de garantir unes bones propietats per als estimadors i, per tant, una inferència correcta.

Les observacions que han presentat més problemes són les que tenen valors reals de la variable Pes extrems (tant alts com baixos). Aquestes són les que han presentat errors (diferència entre valor real i valor estimat pel model) més alts en valor absolut i han fet rebutjar la hipòtesi de normalitat dels errors del model. En l'àmbit mèdic, són precisament aquests casos els que resulten tenir un major interès; nadons amb un reduït pes són més propensos a desenvolupar anèmies i malalties cròniques mentre que nadons amb un elevat pes al néixer són més propensos a presentar obesitat infantil amb totes les seves implicacions.

Degut a què l'interès està en certs quantils de la variable resposta, aquest ha resultat ser un bon cas d'estudi per a aplicar-hi la regressió quantílica. Per tant, s'ha procedit estimant certs models de regressió quantílica associats a diferents quantils de la variable pes del nounat al néixer. Amb aquests models estimats, s'han pogut observar les diferències en les estimacions dels paràmetres en funció del quantil de la variable resposta que s'estigui estimant i amb els que s'estimarien per mínims quadrats ordinaris.

D'aquesta manera, s'ha pogut veure la influència que tenen les diferents variables explicatives en funció del quantil de la variable resposta que s'estigui considerant. Per exemple, si es fa referència a quantils baixos de la variable resposta, es pot observar que les variables més importants per a estimar aquests quantils són l'etnicitat de la mare i si la mare ha realitzat alguna visita prenatal o no. Mentre que si es volen estimar quantils alts de la variable resposta, l'etnicitat de la mare seguirà sent important i guanyarà importància el sexe del nadó.

Per últim, s'ha finalitzat aquesta primera aplicació presentant una metodologia basada en els conceptes de percentil condicionat i percentil no condicionat d'una observació. Aquesta permet anar més enllà de la mostra amb la que s'està treballant i extrapolar cap a una hipotètica població formada per n mares d'idèntiques característiques. Mitjançant el concepte de percentil condicionat, aquesta metodologia permet situar a una determinada mare en un percentil concret de la variable resposta del model dins de la seva població de referència. Aquest fet resulta ser molt útil ja que, amb aquesta metodologia, s'està comparant el pes del nadó al néixer d'una mare concreta amb els pesos dels nadons d'altres dones d'idèntiques característiques.

Per altra banda, en la segona aplicació del treball, s'han utilitzat unes dades de l'àmbit de les assegurances, i per tant, econòmic. Aquestes són dades bàsiques de conducció: velocitat, quilometratge i tipus de via on es circulava de 7691 conductors durant l'any 2010. En aquest cas, la variable resposta dels diferents models ha estat el percentatge de quilòmetres recorreguts per sobre de la velocitat permesa al cap d'un any. Degut a que la variable resposta està limitada amb un interval de valors concret, s'han estimat certs models de regressió quantílica logística associats a diferents quantils de la variable resposta.

Seguint amb la metodologia i procediments emprats en la primera aplicació del treball, en primer lloc, s'ha realitzat una anàlisi descriptiva univariant de cadascuna de les variables de la base de dades. Amb aquesta primera anàlisi, s'ha pogut observar que la variable resposta del model prenia sobretot valors baixos, estant la majoria de valors entre 0% i 20%. Tanmateix, s'ha trobat que la mostra estava equilibrada pel que fa al nombre d'homes i de dones i que les edats dels mateixos es troben entre els 18 i els 32 anys. Per últim, s'ha observat una baixa tendència per part dels conductors a conduir tant amb horari nocturn com en zones urbanes.

Seguidament, s'ha realitzat un *profiling* de la variable resposta en funció dels diferents valors que puguin prendre les variables explicatives. Amb aquesta anàlisi descriptiva multivariant, s'han pogut observar valors més alts de la variable resposta a major nombre de quilòmetres recorreguts. De la mateixa manera, s'ha pogut veure que els valors reals de la variable resposta no es veuen gaire modificats ni per l'edat del conductor ni per el percentatge de quilòmetres recorreguts amb horari nocturn. Per últim, és necessari destacar que existeix una relació inversa o negativa entre els valors reals de la variable resposta i el percentatge de quilòmetres recorreguts en zones urbanes.

Arribats a aquest punt, s'ha estimat el primer model lineal. Estimat aquest model, s'ha vist que, tal i com succeïa en l'anterior aplicació, no s'han complert els supòsits bàsics necessaris per a la correcta estimació dels paràmetres per mínims quadrats ordinaris. D'aquesta manera, no es poden garantir unes bones propietats per als estimadors i això

porta problemes a la inferència a partir d'aquest model. En aquest cas, les observacions que han provocat la no normalitat dels errors i han tingut uns errors més alts en valor absolut han estat les que tenien valors reals més elevats de la variable resposta.

En aquest cas d'estudi, són precisament aquests conductors els més interessants en ser estudiats i analitzats ja que valors alts de la variable resposta (més quilòmetres recorreguts per sobre de la velocitat permesa) poden ser senyal d'una major propensió a la sinistralitat i, en definitiva, una major probabilitat de patir un accident. Per tant, degut a que l'interès està en aquests quantils de la variable resposta i aquesta està limitada en un interval de valors concret, aquest és un bon cas d'estudi per a aplicar-hi la regressió quantílica logística.

A continuació, s'han estimat diferents models de regressió quantílica logística per a certs quantils de la variable resposta. En aquest cas, s'ha inclòs la variable $Edat$ del conductor al quadrat ja que en l'anàlisi descriptiva multivariant de la variable resposta, s'ha observat un efecte quadràtic de l'$Edat$ per a quantils elevats de la variable resposta. Amb aquests models estimats, s'han pogut observar les diferències en les estimacions dels paràmetres en funció del quantil de la variable resposta que s'estigui considerant i els estimats per mínims quadrats ordinaris. Fent referència a l'estimació dels quantils elevats de la variable resposta, s'ha pogut veure que les variables més importants per a la seva estimació han estat l'edat del conductor i el nombre total de quilòmetres recorreguts al cap d'un any.

Per últim, s'ha adaptat la metodologia proposada en la part final de la primera aplicació per al cas de la regressió quantílica logística. Aquesta permet situar a un conductor concret en un percentil de la variable resposta en la seva població de referència. Igual que abans, aquest fet resulta ser molt útil ja que se l'està comparant amb altres conductors d'idèntiques característiques. Cal remarcar que el percentil condicionat d'un conductor en la seva població de referència podria ser perfectament una mesura del risc utilitzada per les assegurances a l'hora d'avaluar la propensió a la sinistralitat d'aquest conductor.

Finalitzades les dues aplicacions, s'han realitzat dues aplicacions interactives utilitzant el paquet *Shiny* d'R, una per a cada aplicació del treball. Amb aquestes aplicacions interactives l'usuari és capaç de determinar una població de referència, a partir de donar valors a les variables predictores de manera fàcil, simple i interactiva. Determinada aquesta població de referència, l'usuari pot especificar les característiques d'una mare o conductor concrets per tal de determinar-ne el percentil condicionat i "situar-lo" en aquest percentil de la corresponent variable resposta de la seva població de referència.

8. Bibliografia

En aquest últim apartat del treball final de grau, s'hi mostren les referències que s'han utilitzat per al seu desenvolupament:

- Atsalakis, G.; Bouri, E. i Pasiouras, F. (2020). *Natural disasters on economic growth: a quantile on quantile approach.* Annals of Operations Research, 1-27.
- Baker, E.; Ngoc, P.; Daniel, L. i Bentley, R. (2020). *New evidence on mental health and housing affordability in cities: A quantile regression approach.* Cities, 96.
- Bottai M.; Cai, B. i McKeown, RE. (2010). *Logistic quantile regression for bounded outcomes.* Statistics in Medicine, 29.2, 309-317.
- Cortés, A. (2019). *Regresión cuantílica para la cuantificación del riesgo.* Barcelona: Màster Interuniversitari en Estadística i Investigació Operativa UB-UPC.
- Flom, P. (2018). *An introduction to quantile regression.* Recuperat de: https://towardsdatascience.com/an-introduction-to-quantile-regression-eca5e3e2036a.
- Heitjan D. i Rubin D. (1991). *Ignorability and coarse data.* Annals of Statistics, 19, 2244-2253.
- Kieschnick R. i McCullough BD. (2003). *Regression analysis of variates observed on (0, 1): percentages, proportions and fractions.* Statistical Modelling, 3, 193-213.
- Koenker, R. i Bassett, G. (1978). *Regression quantiles.* Econometrica: journal of Econometric Society, 33-50.
- Koenker, R. i Hallock, K. (2001). *Quantile regression.* Journal of economic perspectives, 15.4, 143-156.
- Lesaffre, E.; Rizopoulos, D. i Tsonaka R. (2007). *The logistic transform for bounded outcome scores.* Biostatistics, 8.1, 72-85.
- Papke, L. i Wooldridge, J. (1996). *Econometric methods for fractional response variables with an application to 401 (k) plan participation rates.* Journal of applied Econometrics, 11.6, 619-632.
- Pérez, A.; Guillen, M.; Alcañiz, M. i Bérmudez, L. (2019). *Quantile regression with telematics information to assess the risk of driving above the posted speed limit.* Risks, 7.3, 80.
- Pitarque, A. (2019). *La regressió quantílica per a les mesures de risc.* Barcelona: Màster Interuniversitari en Estadística i Investigació Operativa UB-UPC.
- Vicéns, J. i Sánchez, B. (2012). *Regresión cuantílica: estimación y contrastes.* Madrid: Facultat de CC.EE. i EE., Universitat Autònoma de Madrid.