MAESTRÍA EN DIRECCIÓN DE NEGOCIOS
UNIVERSIDAD NACIONAL DE CÓRDOBA
FACULTAD DE CIENCIAS ECONÓMICAS
ESCUELA DE GRADUADOS

Estimación de tarifas de fletes a través del uso del método de la Regresión Cuantílica

Javier Martin

2015

Tutor: MBA Germán Tisera

Agradecimientos

A la Escuela de Graduados de la Facultad de Ciencias Económicas de la Universidad Nacional de Córdoba, institución en la que no sólo cursé la Maestría en Dirección de Negocios (MBA) que dio origen a este trabajo, sino que también me abrió las puertas del programa GCLOG (Graduate Certificate in Logistics and Supply Chain Management) del Massachusetts Institute of Technology (MIT).

A mi tutor de tesis, el contador Germán Tisera, que, con su calidez, experiencia y sapiencia, me guio en cada una de las etapas por las que transitó el presente trabajo.

Al Massachusetts Institute of Technology, que me abrió sus puertas a través del programa GCLOG, y en donde por primera vez entre en contacto con la Regresión Cuantílica, luego de realizar investigaciones lideradas por el Dr. Chris Caplice y orientadas por el Dr. Roberto Pérez Franco. Estas investigaciones constituyen la base del presente trabajo.

A la Dra. Isabel Martínez Silva, que con sus amplios conocimientos y enorme predisposición me ayudó a comprender y profundizar ciertos aspectos claves de la Regresión Cuantílica. De igual manera a la Dra. Jessica Logan, el Dr. Roger Koenker y el Dr. Fang Chen.

A mis padres, tíos y abuelos, que me apoyan incondicionalmente desde diferentes lugares y desde que tengo memoria, en todos los proyectos que deseo emprender.

A mi novia, quien me acompañó durante todo el cursado de la Maestría en Dirección de Negocios, y me apoyó durante la realización de este trabajo.

A mis amigos de la vida, compañeros de facultad, y todos aquellos que en algún momento confiaron en mí.

Simplemente, gracias.

Contenido

Índice de Ilustraciones

Índice de Tablas

A. Presentación del Proyecto

1. Resumen

El presente trabajo se articula como Trabajo Final de Aplicación correspondiente a la Maestría en Dirección de Negocios de la Escuela de Graduados de la Facultad de Ciencias Económicas de la Universidad Nacional de Córdoba. El mismo, tiene por objeto estimar tarifas de fletes a través del uso del método de la Regresión Cuantílica, utilizando información correspondiente a Estados Unidos de América. Es importante resaltar que durante la realización del trabajo no fue posible acceder a fuentes confiables de la República Argentina en materia de transporte, que fueran capaces de proveer una base de datos con el nivel de detalle requerido. Es por este motivo que se utilizó como ambiente de prueba la base de datos correspondiente a los Estados Unidos de América.

El método de regresión de los Mínimos Cuadrados Ordinarios, MCO[1] de ahora en adelante, es uno de los métodos más utilizados debido a la sencillez de las hipótesis que lo sustentan y su facilidad de cálculo. A pesar de ello, las hipótesis de partida necesarias para su aplicación suelen incumplirse, especialmente cuando se trabajan con grandes bases de datos procedentes de encuestas. La presencia de heterocedasticidad[2], cambio estructural o datos atípicos son algunas de las circunstancias que dan a lugar tales incumplimientos. La Regresión Cuantílica representa una solución a estos problemas a través de un método de estimación basado en la minimización de desviaciones absolutas ponderadas con pesos asimétricos, las cuales no se ven afectadas por datos extremos.

En el presente trabajo se desarrollarán modelos de estimación de tarifas del Costo por Carga para movimientos de camiones, donde el Costo por Carga es una función de la ciudad de origen, la ciudad de destino, la distancia a recorrer, el volumen a transportar, el tipo de carga y la geografía. Una vez obtenidas las estimaciones a través de los dos métodos de regresión antes mencionados,

[1] En inglés, Ordinary Least Squares (OLS).
[2] En estadística se dice que un modelo de regresión lineal presenta heterocedasticidad cuando la varianza de las perturbaciones no es constante a lo largo de las observaciones.

se comparará y contrastará la calidad de los diferentes acercamientos, en pos de poder ofrecer una metodología de trabajo para futuros estudios.

2. Marco teórico

Este Trabajo Final de Aplicación se enmarca dentro del área de conocimiento del Supply Chain Management[3] y tiene por objeto plantear y desarrollar un problema de transporte haciendo uso de dos herramientas estadísticas:

1. Regresión por el método de los MCO.
2. Regresión Cuantílica.

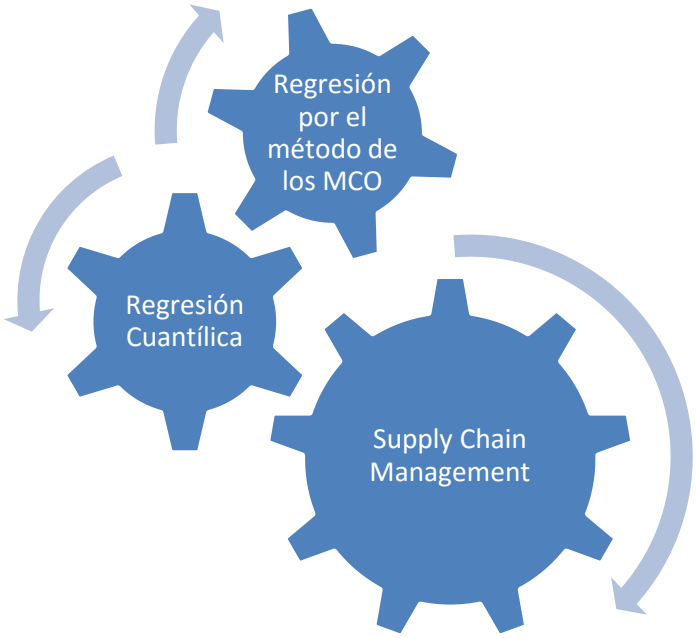


Ilustración 1 - Herramientas a utilizar - Elaboración propia

[3] En español, Gestión de la Cadena de Suministros.

3. Metodología

En una primera etapa, se realizará una descripción teórica de la regresión por el método de los Mínimos Cuadrados Ordinarios y de la Regresión Cuantílica, que servirán de punto de partida para desarrollar las etapas siguientes del proyecto.

En una segunda parte del trabajo se hará una breve descripción de la problemática a resolver y de los acercamientos previos para su resolución. Además, se presentará la información de los Costos por Carga que será utilizada como base para el trabajo.

En una tercera instancia, se realizarán estimaciones de modelos de tarifas utilizando la regresión por el método de los Mínimos Cuadrados Ordinarios y la Regresión Cuantílica, sirviéndose para ello del software R, un entorno libre para la modelización y el cálculo estadístico. Una vez obtenidos los resultados, se realizará una comparación entre los resultados con el objeto de contrastar la calidad de ambos acercamientos.

Finalmente, se extraerán las conclusiones finales pertinentes y se propondrá una metodología de trabajo para encarar posteriores estudios.

4. Objetivos del trabajo

El objetivo principal del presente trabajo es el de estimar tarifas de fletes a través del uso del método de la Regresión Cuantílica, utilizando información correspondiente a Estados Unidos de América. El objetivo final es el de elaborar un método que no solo permita abarcar el estudio en el país antes mencionado, sino que el mismo pueda ser extendido a otros países, en diferentes regiones y continentes.

Entre los objetivos particulares podemos mencionar:

- Desarrollar modelos de estimación de tarifas del Costo por Carga para movimientos de camiones en Estados Unidos de América, donde el Costo por Carga es una función de la ciudad de origen, la ciudad de destino, la distancia a recorrer, el volumen a transportar, el tipo de carga y la geografía.

- Estimar modelos de tarifas utilizando la regresión por el método de los MCO y la Regresión Cuantílica.
- Comparar y contrastar la calidad de los diferentes acercamientos mencionados en el punto anterior.

5. Límites o alcance del trabajo

El límite del presente trabajo se circunscribe al análisis de la información correspondiente a los movimientos de carga de camiones en los Estados Unidos de Norteamérica, utilizando para tal fin el software estadístico R.

6. Organización del trabajo

El presente trabajo contará con los siguientes capítulos:

- Regresión por el método de los Mínimos Cuadrados Ordinarios y Regresión Cuantílica.
- Descripción de la problemática y presentación de la información.
- Estimaciones de tarifas de fletes por ambos métodos de regresión.
- Comparación de resultados.
- Conclusiones finales.

B. Desarrollo del Proyecto

1. Análisis de regresión lineal

Muchas veces las decisiones gerenciales se basan en la relación entre dos o más variables. Por ejemplo, después de revisar la relación entre los gastos de publicidad y las ventas, un gerente de marketing podría tratar de predecir las ventas para determinado nivel de gastos de publicidad. En otro ejemplo, una empresa de electricidad podría usar la relación entre la temperatura máxima diaria y la demanda de electricidad para predecir el consumo de energía en base a las temperaturas máximas pronosticadas para el mes siguiente. A veces, un administrador confía en su intuición para juzgar cómo se relacionan dos variables. Sin embargo, si se logran obtener datos, se puede emplear un procedimiento estadístico llamado análisis de regresión para plantear una ecuación que muestre cómo dependen las variables entre sí.

El término regresión se utilizó por primera vez en el estudio de variables antropométricas, al comparar la estatura de padres e hijos, donde resultó que los hijos cuyos padres tenían una estatura muy superior al valor medio, tendían a igualarse a éste, mientras que aquellos cuyos padres eran muy bajos tendían a reducir su diferencia respecto a la estatura media. Es decir, tendían al promedio. La constatación empírica de esta propiedad se vio reforzada más tarde con la justificación teórica de ese fenómeno.

El análisis de regresión lineal incluye numerosas técnicas para el modelado, y la atención se centra en la relación entre una variable dependiente y una o más variables independientes. El término lineal se emplea para distinguirlo del resto de técnicas de regresión, que emplean modelos basados en cualquier clase de función matemática. Los modelos lineales son una explicación simplificada de la realidad, mucho más ágiles y con un soporte teórico mucho más extenso por parte de la matemática y la estadística. Es posible distinguir entre:

- Regresión lineal simple: Sólo maneja una variable independiente.
- Regresión lineal múltiple: Maneja dos o más variables independientes.

Tanto en el caso de la regresión lineal simple como en la regresión lineal múltiple, el análisis de regresión lineal puede utilizarse para explorar y cuantificar la relación entre una variable llamada dependiente (Y), y una o más variables llamadas independientes (X_1,X_2,..., X_n), así como para desarrollar una ecuación lineal con fines predictivos.

1.1. El modelo de regresión lineal

Un diagrama de dispersión[4] ofrece una idea bastante aproximada sobre el tipo de relación existente entre dos variables. Además, un diagrama de dispersión puede utilizarse como una forma de cuantificar el nivel de relación lineal existente entre dos variables: basta con observar el grado en el que la nube de puntos se ajusta a una línea recta.

Ahora bien, aunque un diagrama de dispersión permite formar una primera impresión muy rápida sobre el tipo de relación existente entre dos variables, utilizarlo como una forma de cuantificar dicha relación tiene un serio inconveniente: la relación entre dos variables no siempre es perfecta o nula. De hecho, habitualmente no es ni lo uno ni lo otro.

Podríamos encarar el análisis mediante una función matemática simple, tal como una línea recta, de la forma:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1)$$

En este caso, Y_i es la variable dependiente (también llamada variable explicada, regresando o variable endógena), X_i la variable independiente (también llamada variable explicativa, regresor o variable exógena), β_0 el punto en el que la recta corta el eje vertical, β_1 la pendiente de la recta y ε_i la perturbación aleatoria (también llamado error o residuo) que recoge todos aquellos factores de la realidad no controlables u observables, y que por lo tanto se asocian con el azar, y es la que le confiere al modelo su carácter estocástico. Es decir, la variable aleatoria ε explica la variabilidad en

[4] Un diagrama de dispersión es un tipo de diagrama matemático que utiliza las coordenadas cartesianas para mostrar los valores de dos variables para un conjunto de datos. Los datos se muestran como un conjunto de puntos, cada uno con el valor de una variable que determina la posición en el eje horizontal y el valor de la otra variable determinado por la posición en el eje vertical.

Y_i que no se puede explicar con la relación lineal entre X_i e Y_i. En general, los parámetros β_0 y β_1 son desconocidos y deben ser determinados.

En la sección siguiente describiremos todos los supuestos del modelo de regresión lineal simple y de ε_i. Uno de ellos es que la media o valor esperado de ε es cero. Una consecuencia de este supuesto es que la media, o valor esperado de Y_i, representado por $E(Y_i)$, es igual a $\beta_0+\beta_1 X_i$. En otras palabras, el valor medio de Y_i es una función lineal de X_i. La ecuación que describe la forma en que el valor medio de Y_i se relaciona con X_i se llama ecuación de regresión. La ecuación de regresión para la regresión lineal simple es la siguiente:

$$E(Y_i) = \beta_0 + \beta_1 X_i \qquad (2)$$

En una situación ideal en la que todos los puntos de un diagrama de dispersión se encuentran en una línea recta, no tendríamos que preocuparnos por encontrar la recta que mejor resume los puntos del diagrama. Simplemente uniendo los puntos entre sí obtendríamos la recta con menor ajuste a la nube de puntos. Sin embargo, en una nube de puntos más realista, es posible trazar muchas rectas diferentes. Es muy sencillo determinar que no todas ellas se ajustarán igualmente bien a la nube de puntos. El objetivo inmediato que surge es el de encontrar la recta capaz de convertirse en el mejor representante del conjunto total de puntos.

Existen diferentes procedimientos para ajustar una función simple, cada uno de los cuales intenta minimizar una medida diferente del grado de ajuste. La elección preferida ha sido, tradicionalmente, la recta que hace mínima la suma de los cuadrados de las distancias verticales entre cada punto y la recta. Esto significa que, de todas las rectas posibles, existe una y sólo una que consigue que las distancias entre cada punto y la recta sean mínimas. Esta recta se obtiene a través del método de los MCO.

Un punto interesante a resaltar es que el análisis de regresión no se puede interpretar como un procedimiento para establecer la relación causa-efecto entre variables. Sólo puede indicar cómo, o hasta qué grado, las variables están asociadas entre sí. Cualquier conclusión acerca de causa y efecto se debe basar en el juicio del o los individuos con mayores conocimientos sobre la aplicación.

1.2. Supuestos del modelo de regresión lineal

Entre los supuestos del modelo de regresión lineal, encontramos:

- Linealidad: La ecuación de regresión adopta una forma particular. En concreto, la variable dependiente es la suma de un conjunto de elementos: el origen de la recta, una combinación lineal de variables independientes o explicativas y los residuos. El incumplimiento del supuesto de linealidad suele denominarse error de especificación.
- Independencia: Los residuos son independientes entre sí, es decir, los residuos constituyen una variable aleatoria. Es frecuente encontrarse con residuos autocorrelacionados cuando se trabaja con series temporales.
- Homocedasticidad: Para cada valor de la variable independiente (o combinación de valores de las variables independientes), la varianza de los residuos es constante.
- Normalidad: Para cada valor de la variable independiente (o combinación de valores de las variables independientes), los residuos se distribuyen normalmente con media cero.
- No-colinealidad: No existe relación lineal exacta entre ninguna de las variables independientes. El incumplimiento de este supuesto da origen a la colinealidad o multicolinealidad.

Sobre el cumplimiento del primer supuesto puede obtenerse información a partir de una inspección del diagrama de dispersión: si tenemos intención de utilizar el modelo de regresión lineal, lo razonable es que la relación entre la variable dependiente y las independientes sea de tipo lineal. El quinto supuesto, no-colinealidad, no tiene sentido en regresión simple, pues es imprescindible la presencia de más de una variable independiente. El resto de los supuestos, independencia, homocedasticidad y normalidad, están estrechamente asociados al comportamiento de los residuos.

2. Método de los Mínimos Cuadrados Ordinarios (MCO)

La pregunta inminente que se plantea es ¿cómo obtener una buena estimación de los parámetros β_0 y β_1 a partir de los datos disponibles para Y_i, y para cada una de las X_i?

Uno de los procedimientos más conocidos, propuesto por Adrien-Marie Legendre (1752-1833), es el método de los MCO. Este procedimiento plantea utilizar, como estimación de los parámetros, aquella combinación de β_0 y β_1 que minimice los errores que el modelo cometerá. Es muy claro que si dispusiéramos a priori de los parámetros estimados podríamos escribir el modelo de regresión lineal simple no como la ecuación (1), sino como:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad (3)$$

La ecuación (3) se conoce como ecuación de regresión estimada o simplemente ecuación de regresión. Por lo tanto, podríamos computar el error o residuo que el modelo comente en la estimación de cada valor de la endógena comparando, de forma inmediata, el valor real de la endógena en cada observación con el valor estimado:

$$\varepsilon_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right) \qquad (4)$$

Este error dependerá, evidentemente del valor asignado a las estimaciones de los parámetros β_0 y β_1. Pues bien, el método de los MCO sugiere utilizar aquella combinación de parámetros estimados que minimice la suma al cuadrado de todos los errores cometidos para todas las observaciones disponibles:

$$\hat{\beta}_{MCO} = min\left(\sum_{i=1}^{n} (\varepsilon_i)^2\right) \qquad (5)$$

Es posible demostrar (ver apéndice 1) que los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ que minimizan la expresión (5) son:

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})\,(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \qquad (6)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad (7)$$

El método de los MCO requiere unas hipótesis previas sobre la aleatoriedad de la ecuación (1) expresadas en términos de $\varepsilon_i \sim N(0,\sigma^2)$.

3. Coeficiente de Determinación

Nos podríamos preguntar qué tan bien se ajusta a los datos la ecuación de regresión. En esta sección mostraremos que el coeficiente de determinación es una medida de la bondad de ajuste para una ecuación de regresión.

Para la i-ésima observación, la diferencia entre el valor observado de la variable dependiente, Y_i, y el valor estimado de la variable dependiente, $\hat{Y}_i$, se llama i-ésimo residual. Este valor representa el error que se comete al usar $\hat{Y}_i$ para estimar Y_i. Así, para la i-ésima observación, el residual es $Y_i - \hat{Y}_i$. La suma de los cuadrados de esos errores o residuales es la cantidad que se minimiza con el método de los MCO. Esa cantidad, se denomina suma de cuadrados debida al error y se representa por SSE[5]:

$$SSE = \sum \left(Y_i - \hat{Y}_i\right)^2 \qquad (8)$$

El valor de SSE es una medida del error que se comete al usar la ecuación de regresión para calcular los valores de la variable dependiente en la muestra.

[5] En inglés, sum of squares due to error.

Ahora bien, para la i-ésima observación, la diferencia $Y_i - \overline{Y}$ suministra una medida de error incurrido al usar $\overline{Y}$ para estimar futuros valores de la variable dependiente. A la suma correspondiente de cuadrados, llamada suma de cuadrados del total, se la representa por SST[6]:

$$SST = \sum (Y_i - \bar{Y})^2 \qquad (9)$$

Podemos imaginar que SST es una medida de lo bien que las observaciones se agrupan en torno a la recta $\overline{Y}$, y que SSE mide lo bien que las observaciones se agrupan en torno a la recta de regresión estimada.

Para medir cuánto se desvían los valores de $\hat{Y}_i$ medidos en la línea de regresión, de los valores de $\overline{Y}$, se calcula otra suma de cuadrados. A esa suma de cuadrados se la llama suma de cuadrados debida a la regresión, y se representa por SSR[7]:

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2 \qquad (10)$$

De acuerdo a lo expuesto, debemos esperar que SST, SSR y SSE estén relacionadas. En realidad, la relación entre esas tres sumas de cuadrados, es uno de los resultados más importantes de la estadística:

$$SST = SSR + SSE \qquad (11)$$

La ecuación (11) indica que la suma de cuadrados del total se puede dividir en dos componentes, la debida a la regresión y la debida al error. Se puede pensar que la SSR representa la parte explicada de la SST, y SSE la parte no explicada, también de la SST.

Ahora se puede pensar en cómo emplear las tres sumas de cuadrados, SST, SSR y SSE para suministrar una medida de la bondad de ajuste para la ecuación de regresión. Esa ecuación tendría

[6] En inglés, total sum of squares.
[7] En inglés, sum of squares due to regression.

un ajuste perfecto si cada valor de la variable dependiente Y_i estuviera sobre la línea estimada de regresión. En este caso, $Y_i - \hat{Y}_i$ sería cero para cada observación, dando como resultado $SSE=0$. Como $SST=SSR+SSE$, vemos que, para un ajuste perfecto, SSR debe ser igual a SST, y que la relación SSR/SST debe ser igual a uno. Los ajustes menos perfectos darán como resultado valores mayores de SSE. Al despejar SSE de la ecuación (11) vemos que $SSE=SST-SSR$. En consecuencia, el valor máximo de SSE (y en consecuencia el peor ajuste) se tiene cuando $SSR=0$ y $SSE=SST$.

La relación SSR/SST, que asume valores entre cero y uno, se usa para evaluar la bondad de ajuste para la ecuación de regresión. A esta relación se la conoce como coeficiente de determinación y se representa por R^2:

$$R^2 = \frac{SSR}{SST} \qquad (12)$$

Si lo expresamos como porcentaje, se puede interpretar a R^2 como el porcentaje de la suma total de cuadrados que se puede explicar aplicando la ecuación de regresión.

4. Coeficiente de Correlación

El coeficiente de correlación es una medida descriptiva de la intensidad de la asociación lineal entre dos variables, X e Y. Los valores del coeficiente de correlación siempre están entre -1 y $+1$. Un valor de $+1$ indica que las dos variables, X e Y, tienen una relación lineal positiva perfecta. Esto es, todos los puntos de datos están en una línea recta con pendiente positiva. Una valor de -1 indica que X e Y tienen una relación lineal negativa perfecta, y que todos los puntos de datos están en una recta con pendiente negativa. Los valores del coeficiente de correlación cercanos a cero indican que X e Y no tienen relación lineal.

Si ya se ha hecho un análisis de regresión y se ha calculado el coeficiente de determinación R^2, el coeficiente de correlación de la muestra se puede calcular como sigue:

$$R_{XY} = (signo\ de\ \beta_1)\sqrt{R^2} \qquad (13)$$

El signo del coeficiente de correlación es positivo si la ecuación de regresión tiene pendiente positiva y negativo si la ecuación de regresión tiene pendiente negativa.

En el caso de una relación lineal entre dos variables, el coeficiente de determinación y el coeficiente de correlación permiten tener medidas de la intensidad de una relación. El coeficiente de determinación da una medida entre 0 y 1, mientras que el coeficiente de correlación da una medida entre -1 y $+1$. Aunque el coeficiente de correlación se restringe a una relación lineal entre dos variables, el coeficiente de determinación se puede emplear en relaciones no lineales y en relaciones que tengan dos o más variables independientes. En este sentido, el coeficiente de determinación tiene una aplicabilidad más amplia.

5. Regresión Cuantílica

5.1. Introducción a la Regresión Cuantílica

Tal como lo explican Otero y Reyes (2012), es habitual que la información que se maneja dé lugar a algunos de los inconvenientes descritos al comienzo del trabajo, tales como heterocedasticidad, cambio estructural o datos atípicos, no posibilitando que la expresión $Y_i=\beta_{0_{MCO}}+\beta_{1_{MCO}}X_i+\varepsilon$ sea una buena explicación de la relación existente entre X_i e Y_i. Ante tales circunstancias, el método de Regresión Cuantílica se presenta como una buena solución.

Si bien el método de Regresión Cuantílica tiene sus inicios a finales de los años setenta de la mano de Koenker y Basset (1978), el propio Koenker (2005) afirmar que la idea básica de la que parte la Regresión Cuantílica se encuentra en los trabajos de Boskovich de la segunda mitad del siglo XVIII, acerca del estudio de la forma elíptica de la Tierra. En estos trabajos, Boskovich comenzó a utilizar la minimización del valor absoluto de los residuos para encontrar los parámetros de la función de la elíptica de la Tierra. Posteriormente, Laplace y Edgeworth investigaron sobre esta técnica, al igual que Koenker y Basset harían alrededor de un siglo después, para estimar los parámetros de la Regresión Cuantílica, tal como lo explicaremos posteriormente. Por lo tanto, las primeas ideas que pueden asociarse con la Regresión Cuantílica datan de fechas anteriores al nacimiento del método de los MCO de Adrien-Marie Legendre en 1805.

A pesar de ser un método con más de treinta años de historia, y a pesar también de las ventajas que reporta su uso bajo determinadas condiciones, resulta todavía bastante desconocido y las aplicaciones que pueden encontrarse no son muy numerosas.

5.2. Definición de cuantil

Otero y Reyes (2012) explican que, así como la regresión a través del método de los MCO se encuentra vinculada con la media, la Regresión Cuantílica se basa en el concepto de cuantil. Supongamos que disponemos de una muestra de observaciones de una variable Y con una distribución $F(.)$:

$$Y_t : t = 1,2, \dots, N \qquad (14)$$

Tendremos que el cuantil τ de la muestra, con $0 < \tau < 1$, será aquél valor b que deje una proporción τ de observaciones por debajo de b y una proporción $(1 - \tau)$ por encima. En el caso de la mediana $\tau = 0{,}5$, quedarán un 50% de los datos por debajo de $b = M_e$ y un 50% de los datos por encima. Si utilizamos el primer cuartil ($\tau = 0{,}25$) sería un 25% de los valores de Y los que quedarían por debajo de $b = Q_1$ y un 75% por encima, y de forma similar e inversa con el tercer cuartil. Los cuartiles dividen la muestra en cuatro partes, pero de igual manera podemos dividir la muestra en diez partes con los deciles, $\tau = 0{,}1; 0{,}2; \dots; 0{,}9$ o cualquier otra proporción. Los cuantiles más conocidos son la mediana, los cuartiles, los quintiles, los deciles y los percentiles.

En el cálculo de cuantiles con distribuciones de variable continua (por ejemplo, con datos agrupados) puede conseguirse fácilmente que las partes en que se divide la distribución sean exactamente iguales. Sin embargo, en las distribuciones de variable discreta (como el caso de datos aislados) debemos conformarnos con que estas partes sean aproximadamente iguales. Por desgracia, no hay consenso sobre cómo realizar esta aproximación, existiendo en la literatura científica nueve métodos diferentes, que conducen a resultados diferentes. Por ello, al calcular cualquier cuantil de datos no agrupados por medio de calculadora, software o manualmente, es básico el saber e indicar el método utilizado.

Una forma alternativa de expresar la definición de los cuantiles, que es además una primera aproximación al método de estimación de la Regresión Cuantílica, viene dada por la siguiente expresión:

$$\underset{b \in \mathbb{R}}{Min} \left[\sum_{Y_i \geq b} \tau |Y_i - b| + \sum_{Y_i < b} (1 - \tau) |Y_i - b| \right] \qquad (15)$$

Siendo τ el cuantil, Y_i los distintos valores que toman las observaciones de la muestra para la variable Y, y b el valor que minimiza la expresión, se puede demostrar fácilmente que el valor b que minimiza la expresión (15) es el de la observación que deja una proporción τ de la muestra por

debajo y una proporción $(1-\tau)$ por encima, siendo τ por tanto, un valor entre 0 y 1 correspondiente al cuantil que se quiere estimar.

5.3. Estimación cuantílica

Otero y Reyes (2012) resumen a la perfección la esencia de la Regresión Cuantílica, explicando que los objetivos que se persiguen son los mismos que en la regresión lineal por el método de los MCO, es decir, modelizar la relación entre variables. Sin embargo, por algunos motivos ya planteados (heterocedasticidad, presencia de valores atípicos, cambio estructural), el valor medio de respuesta de la variable endógena que ofrece la estimación por el método de los MCO no se siempre el más representativo. Dicho de una manera más intuitiva, al igual que la media no es siempre la medida más representativa de la distribución de una variable cuando existen en la muestra valores extremos o una elevada variabilidad, la recta de estimación obtenida por el método de los MCO, que devuelve el valor medio esperado de la variable endógena dado un valor de la exógena, tampoco es siempre la mejor expresión de la relación entre ambas variables cuando nos encontramos con un caso de heterocedasticidad, presencia de atípicos o cambio estructural.

Antes las situaciones planteadas, la Regresión Cuantílica ofrece la posibilidad de crear diferentes rectas de regresión para distintos cuantiles de la variable endógena a través de un método de estimación que se ve menos perjudicado por la presencia de tales inconvenientes. En forma genérica, la especificación del modelo de Regresión Cuantílica es la siguiente:

$$Y_i = \beta_{0_\tau} + \beta_{1_\tau} X_i + \varepsilon_{i_\tau} \qquad (16)$$

En la expresión (16), $\mathrm{Y_i}$ es la variable endógena, $\mathrm{X_i}$ la variable independiente, β_{0_τ} y β_{1_τ} los parámetros a estimar correspondientes al cuantil τ, y ε_τ la perturbación aleatoria correspondiente al cuantil τ. Tal como en el método de los MCO, en el que $\mathrm{E(Y_i|X_i)=\beta_{0_{MCO}}+\beta_{1_{MCO}}X_i}$ y por tanto $\mathrm{E(\varepsilon_i|X_i)=0}$, aquí $\mathrm{Cuant_\tau(Y_i|X_i)=\beta_{0_\tau}+\beta_{1_\tau}X_i}$ lo que implica que $\mathrm{Cuant_\tau(\varepsilon_{i_\tau}|X_i)=0}$, siendo éste el único supuesto que se hace sobre la perturbación aleatoria.

Otero y Reyes (2012) continúan exponiendo que al igual que se decía anteriormente que la mediana o los cuartiles eran casos concretos de cuantiles, ahora se tiene que la regresión mediana o la regresión cuartílica son casos concretos de la Regresión Cuantílica. Es muy importante resaltar aquí que, a diferencia de lo que ocurre en la regresión por el método de los MCO, en la que hablamos de una única recta de regresión, aquí existen tantas rectas como cuantiles estemos considerando (una recta si estimamos la regresión mediana, cuatro en el caso de la cuartílica, diez en el caso de la decílica, y así sucesivamente).

Al igual que en la regresión por el método de los MCO, en donde la media es quien minimiza la expresión (5), ahora podemos partir de la expresión (15), en la que el valor b correspondiente al cuantil τ minimiza la función. Si consideramos que el valor b en (15) es una simplificación de $\beta_{0_\tau}+\beta_{1_\tau}X_i$ cuando $X_i=1$, entonces tenemos que el problema de estimación de parámetros de Regresión Cuantílica se puede expresar de la siguiente manera:

$$\underset{\beta_{0_\tau}+\beta_{1_\tau}X_i \in \mathbb{R}}{Min}\left[\sum_{Y_i \geq \beta_{0_\tau}+\beta_{1_\tau}X_i} \tau\left|Y_i-\left(\beta_{0_\tau}+\beta_{1_\tau}X_i\right)\right| + \sum_{Y_i < \beta_{0_\tau}+\beta_{1_\tau}X_i} (1-\tau)\left|Y_i-\left(\beta_{0_\tau}+\beta_{1_\tau}X_i\right)\right|\right] \quad (17)$$

Al igual que en el método de los MCO, en donde el valor que minimizaba la expresión (5) era la media condicional de Y dada X, ahora en (17), es el cuantil condicional de Y dada X.

Otero y Reyes (2012) explican que lo que se lleva a cabo finalmente es una minimización de las desviaciones absolutas ponderadas con pesos asimétricos. Es decir, que a cada desviación correspondiente a la observación i se le da más o menos peso según el cuantil cuya recta de regresión se esté estimando. La principal ventaja que aporta el uso de las desviaciones en valor absoluto en lugar de las desviaciones al cuadrado, es el comportamiento ante la existencia de valores atípicos. Ante tal situación, la estimación que ofrece la Regresión Cuantílica prácticamente no se ve alterada por valores extremos ya que penaliza los errores de forma lineal, mientras que la regresión por el método de los MCO, al elevar los errores al cuadrado, lo que hace es darle mayor importancia precisamente a dichos valores, penalizándolos de forma cuadrática.

Respecto a la introducción de las ponderaciones asimétricas, dado que el objetivo es estimar varias rectas de regresión que pasen por distintos puntos de la distribución, la función que cumplen los pesos asimétricos es precisamente la de situar esas rectas ponderando de forma distinta los residuos positivos y los negativos. Así, por ejemplo, en la figura siguiente estaría representada la recta correspondiente al cuantil 75, que unirá los puntos del cuantil 75 condicional de Y dada X.

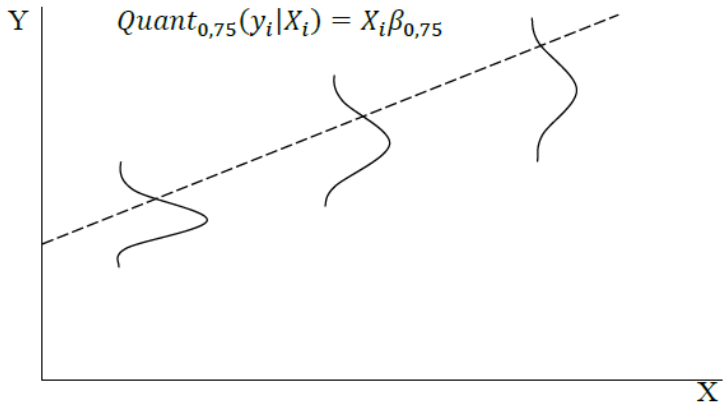


Ilustración 2 - Representación gráfica de la recta de regresión correspondiente al cuartil 75 - Otero y Reyes (2012)

Para su estimación, los residuos positivos (cuando el valor real es mayor que el estimado) se ponderan con 0,75; mientras que los negativos (cuando el valor real es menor que el estimado) se ponderan con 0,25. Otero y Reyes (2012) mencionan un interesante ejemplo en este sentido. Supongamos que el salario de una población varía según sea su nivel de formación, pero que la elasticidad y la pendiente son diferentes en el extracto de salarios altos pues su aumento es más elevado que el que corresponde por salario medio. La regresión por el método de los MCO encontrará la respuesta media en términos de pendiente y no diferenciará la existencia de un cambio en el parámetro. Por el contrario, una Regresión Cuantílica, revelará que en los cuantiles superiores el parámetro aumenta de tamaño. Podría argumentarse que una segmentación de la muestra y su posterior estimación por MCO en cada segmento nos llevaría a la misma conclusión, pero ello nos podría llevar a un sesgo de selección como señala Heckman (1979). La ventaja que aporta la Regresión Cuantílica frente a esta estrategia es que en cada cuantil intervienen todas las observaciones convenientemente ponderadas.

6. Comparación entre métodos

En función de lo expuesto, es posible construir una tabla comparativa entre el método de los MCO y la Regresión Cuantílica:

Criterio	MCO	Regresión Cuantílica
Complejidad	Simple	Compleja
Grado de Desarrollo	Ampliamente desarrollado	En desarrollo
Eficiencia ante no normalidad por presencia de valores atípicos, heterocedasticidad o cambio estructural	Ineficiente	Eficiente
Caracterización de la distribución de Y	Sólo sobre su media condicional	Muy rica, ya que permite considerar el impacto de una covariable en toda la distribución de Y, y no sólo en su media condicional.

Tabla 1 - MCO vs. Regresión Cuantílica - Elaboración Propia

7. La importancia del transporte terrestre y la estimación de la tarifa de fletes en la cadena de suministros

Cuando se habla de transporte en la cadena de suministros, se habla del movimiento de carga en todas sus formas conocidas, aérea, marítima y terrestre, a través de las cuales se trasladan materias primas, insumos y productos terminados de un punto a otro de acuerdo a una planificación de la demanda.

El presente trabajo centra su atención en el transporte de carga terrestre en los Estados Unidos de América. Este país limita al Norte con Canadá, al Este con el Océano Atlántico, al Sur con México y al Oeste con el Océano Pacífico. La superficie total de su territorio es de 9.826.630 km². Estados Unidos de América posee una infraestructura de transporte desarrollada, suficiente para soportar las necesidades de su economía. Está compuesta por una red de carreteras de 6.430.366 kilómetros, que se extiende por todo el país conectando los 50 estados que lo componen, de los cuales 75.238 kilómetros forman parte del vasto Sistema Nacional de Autopistas.

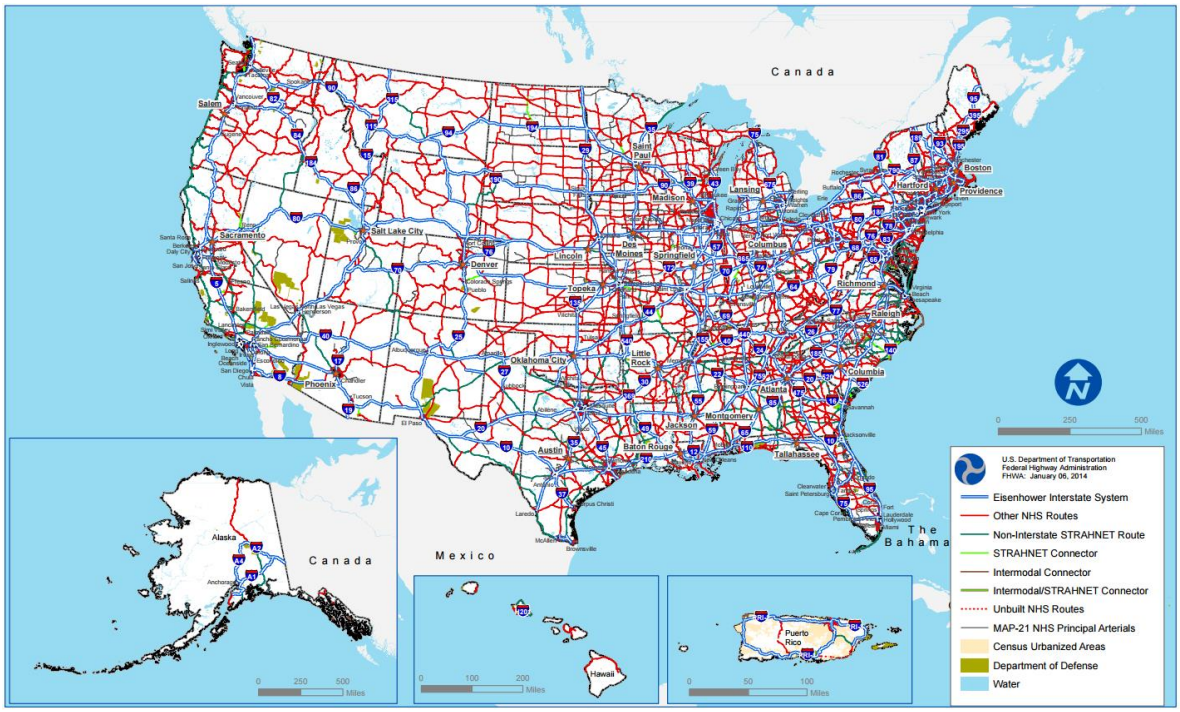
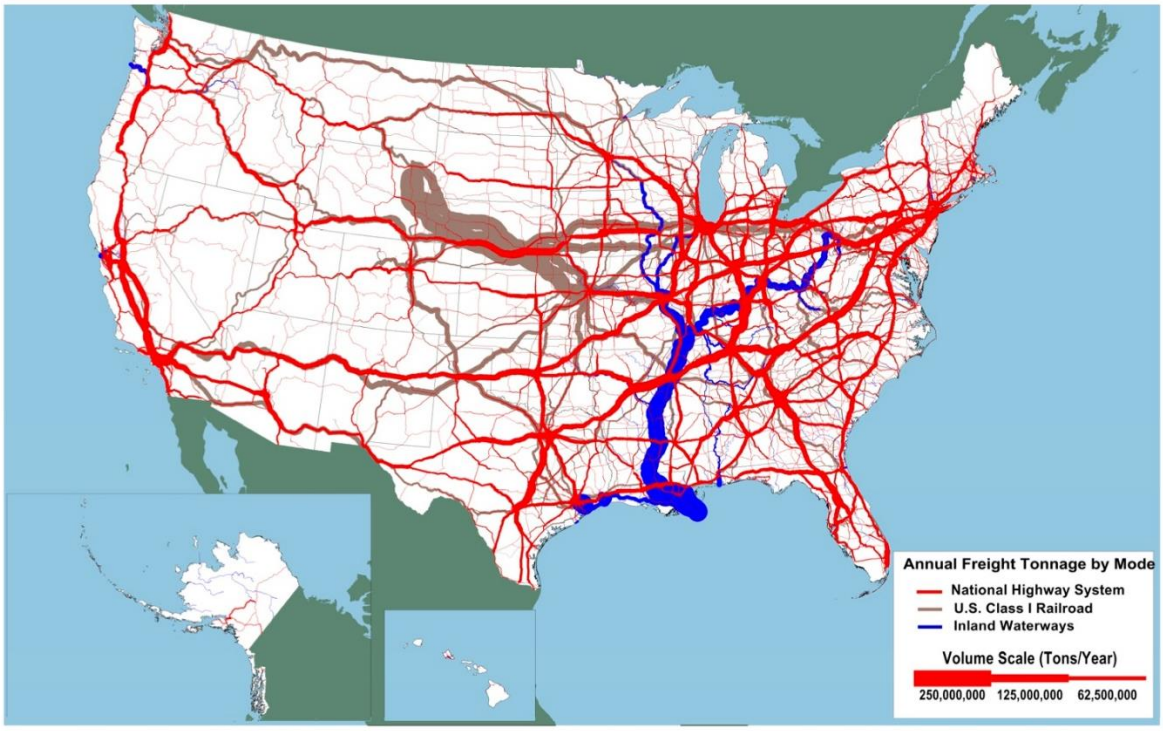


Ilustración 3 - Sistema Nacional de Autopistas - U.S. Department of Transportation - Federal Highway Administration FHWA: January 06,2014

Estados Unidos de América representa el 42% del mercado global de bienes de consumo, y por tal motivo, y para estar más cerca de sus proveedores y clientes, muchas empresas del mundo deciden invertir en este país.

En este contexto, el transporte terrestre cobra una importancia significativa, como así también la posibilidad de contar con herramientas que permitan estimar con mayor precisión la tarifa de fletes. Así, se pueden mencionar tres variables fundamentales que impactan sobre esta forma de transporte: la distancia entre el origen y el destino, la oferta de transporte y el destino final del transporte. En esta investigación, la distancia entre el origen y el destino, y el destino final del transporte, toman una relevancia significativa, como se apreciará en las próximas secciones.



Sources: Highways: U.S. Department of Transportation, Federal Highway Administration, Freight Analysis Framework, Version 3.4, 2012. Rail: Based on Surface Transportation Board, Annual Carload Waybill Sample and rail freight flow assignments done by Oak Ridge National Laboratory. Inland Waterways: U.S. Army Corps of Engineers (USACE), Annual Vessel Operating Activity and Lock Performance Monitoring System data, as processed for USACE by the Tennessee Valley Authority; and USACE, Institute for Water Resources, Waterborne Foreign Trade Data, Water flow asssignments done by Oak Ridge National Laboratory.

Ilustración 4 - Tonelaje en autopistas, ferrocarriles y vías navegables internas en los Estados Unidos de América - Fuente: U.S. Department of Transportation - Federal Highway Administration FHWA

8. Trabajo de Aplicación

8.1. Base de Datos

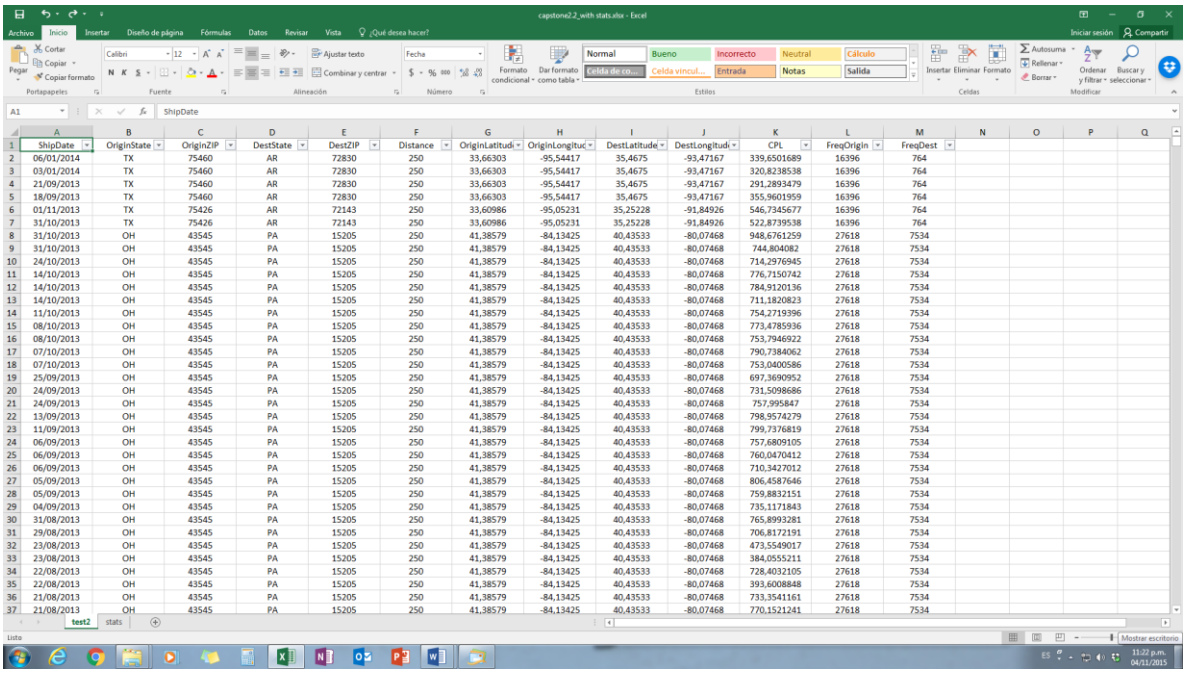
La base de datos sobre la que se sustenta el presente trabajo fue provista por el Dr. Chris Caplice, Executive Director del Center of Transportation and Logistics del Massachusetts Institute of Technology, en el marco de una investigación sobre Regresión Cuantílica durante el desarrollo del Capstone Project del Programa GCLOG (Graduate Certificate in Logistics and Supply Chain Management) 2015.

La base de datos contiene 80.926 observaciones de movimientos de fletes terrestres en los Estados Unidos de América, desarrollados entre el 01/04/2013 y el 31/03/2014, y pertenecientes a diferentes compañías de gran envergadura. Entre las variables disponibles para cada observación, podemos mencionar:

- ShipDate: Fecha del envío de la mercadería.
- OriginState: Lugar de origen, abreviado con nomenclatura ANSI[8].
- OriginZIP: Código postal de origen.
- DestState: Lugar de destino, abreviado con nomenclatura ANSI.
- DestZIP: Código postal de destino.
- Distance: Distancia recorrida en millas[9] desde el origen hasta el destino.
- OriginLatitude: Coordenada geográfica de latitud del origen.
- OriginLongitud: Coordenada geográfica de longitud del origen.
- DestLatitude: Coordenada geográfica de latitud del destino.
- DestLongitude: Coordenada geográfica de longitud del destino.
- CPL: Costo por Carga, en dólares, que es función del estado origen, el estado de destino, la distancia a recorrer, el volumen a transportar, el tipo de carga y la geografía.

[8] American National Standards Institute.
[9] La milla es una unidad de longitud que no forma parte del Sistema Métrico Decimal y que equivale a 1,609 kilómetros.

A continuación, se puede apreciar una captura de pantalla del archivo de Microsoft Excel con la base de datos provista:



ShipDate	OriginState	OriginZIP	DestState	DestZIP	Distance	OriginLatitud	OriginLongitu	DestLatitude	DestLongitud	CPL	FreqOrigin	FreqDest
06/01/2014	TX	75460	AR	72830	250	33,66303	-95,54417	35,4675	-93,47167	339,6501689	16396	764
03/01/2014	TX	75460	AR	72830	250	33,66303	-95,54417	35,4675	-93,47167	320,8238538	16396	764
21/09/2013	TX	75460	AR	72830	250	33,66303	-95,54417	35,4675	-93,47167	291,2893479	16396	764
18/09/2013	TX	75460	AR	72830	250	33,66303	-95,54417	35,4675	-93,47167	355,9601959	16396	764
01/11/2013	TX	75426	AR	72143	250	33,60986	-95,05231	35,25228	-91,84926	546,7345677	16396	764
31/10/2013	TX	75426	AR	72143	250	33,60986	-95,05231	35,25228	-91,84926	522,8739538	16396	764
31/10/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	948,6761259	27618	7534
31/10/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	744,804082	27618	7534
24/10/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	714,2976945	27618	7534
14/10/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	776,7150742	27618	7534
14/10/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	784,9120136	27618	7534
14/10/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	711,1820823	27618	7534
11/10/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	754,2719396	27618	7534
08/10/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	773,4785936	27618	7534
08/10/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	753,7946922	27618	7534
07/10/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	790,7384062	27618	7534
07/10/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	753,0400586	27618	7534
25/09/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	697,3690952	27618	7534
24/09/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	731,5098686	27618	7534
24/09/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	757,995847	27618	7534
13/09/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	798,9574279	27618	7534
11/09/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	799,7376819	27618	7534
06/09/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	757,6809105	27618	7534
06/09/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	760,0470412	27618	7534
06/09/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	710,3427012	27618	7534
05/09/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	806,4587646	27618	7534
05/09/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	759,8832151	27618	7534
04/09/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	735,1171843	27618	7534
31/08/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	765,8993281	27618	7534
29/08/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	706,81721 91	27618	7534
23/08/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	473,5549017	27618	7534
23/08/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	384,0555211	27618	7534
22/08/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	728,4032105	27618	7534
22/08/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	393,6008848	27618	7534
21/08/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	733,3541161	27618	7534
21/08/2013	OH	43545	PA	15205	250	41,38579	-84,13425	40,43533	-80,07468	770,1521241	27618	7534

Ilustración 5 - Captura de pantalla de la base de datos - Elaboración propia

El número total de datos con los que se trabajó asciende a 1.052.051, una cifra realmente contundente para poder desarrollar el estudio. Además las 80.926 observaciones implican un Costo por Carga total de USD 122.230.052, y una distancia total recorrida de 54.501.170 millas, unos 121 viajes de ida y vuelta desde la Tierra a la Luna.

En el gráfico que se muestra a continuación, podemos observar el monto total mensual de Costo por Carga y de distancia, para la base de datos bajo estudio:

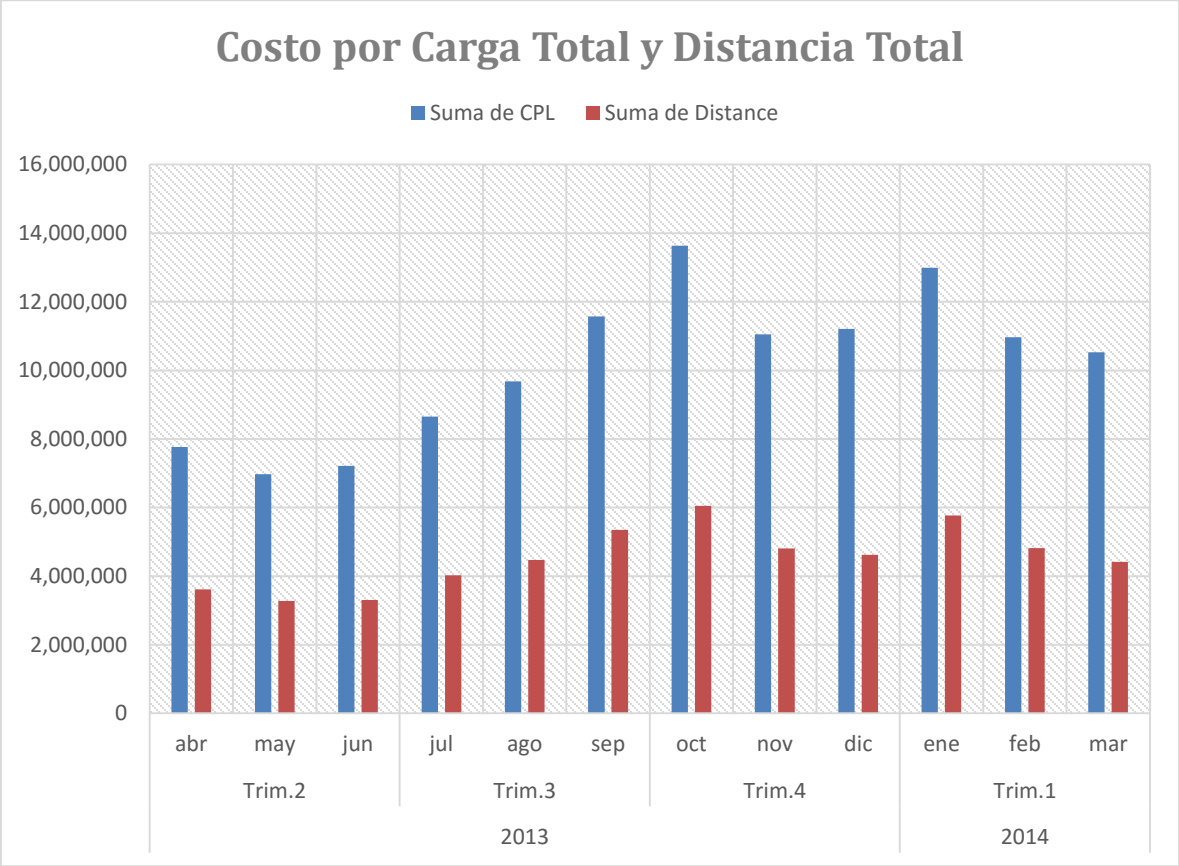


Ilustración 6 - Costo por Carga Total y distancia total como funciones del tiempo - Elaboración propia

Es interesante mencionar que, si bien las observaciones se recogieron durante el periodo de un año, las mismas no se han visto impactadas por variables macroeconómicas significativas, por lo que no es necesario realizar ningún tipo de ajuste. Esto queda claramente evidenciado en el gráfico anterior.

Con respecto a los lugares de origen y de destino, es importante conocer que Estados Unidos de América cuenta con entidades subnacionales que comparten soberanía con el gobierno federal. A continuación, se presenta una lista con los 50 estados, a los que se les anexa el Distrito de Columbia (también conocido como Washington D.C.), que es la capital del país, como un punto de partida para ciertos análisis subsiguientes.

ANSI	Estado	ANSI	Estado	ANSI	Estado
AL	Alabama	KY	Kentucky	ND	North Dakota
AK	Alaska	LA	Louisiana	OH	Ohio
AZ	Arizona	ME	Maine	OK	Oklahoma
AR	Arkansas	MD	Maryland	OR	Oregon
CA	California	MA	Massachusetts	PA	Pennsylvania
CO	Colorado	MI	Michigan	RI	Rhode Island
CT	Connecticut	MN	Minnesota	SC	South Carolina
DE	Delaware	MS	Mississippi	SD	South Dakota
DC	District of Columbia	MO	Missouri	TN	Tennessee
FL	Florida	MT	Montana	TX	Texas
GA	Georgia	NE	Nebraska	UT	Utah
HI	Hawaii	NV	Nevada	VT	Vermont
ID	Idaho	NH	New Hampshire	VA	Virginia
IL	Illinois	NJ	New Jersey	WA	Washington
IN	Indiana	NM	New Mexico	WV	West Virginia
IA	Iowa	NY	New York	WI	Wisconsin
KS	Kansas	NC	North Carolina	WY	Wyoming

Tabla 2 - Nomenclatura ANSI para los estados de los Estados Unidos de América - Elaboración Propia

De los 50 estados de los Estados Unidos de América, en la base de datos sólo 26 de ellos (Alabama, Arkansas, California, Connecticut, Florida, Georgia, Illinois, Indiana, Louisiana, Michigan, Minnesota, Missouri, Nevada, New Jersey, New York, North Carolina, North Dakota, Ohio, Oklahoma, Pennsylvania, Texas, Utah, Virginia, Washington, West Virginia, y Wisconsin) funcionan como origen (52%), mientras que 48 de ellos, a excepción de Alaska y Hawaii, funcionan como destino (96%).

A continuación, podemos apreciar gráficamente las frecuencias de origen y destino para cada uno de los estados:

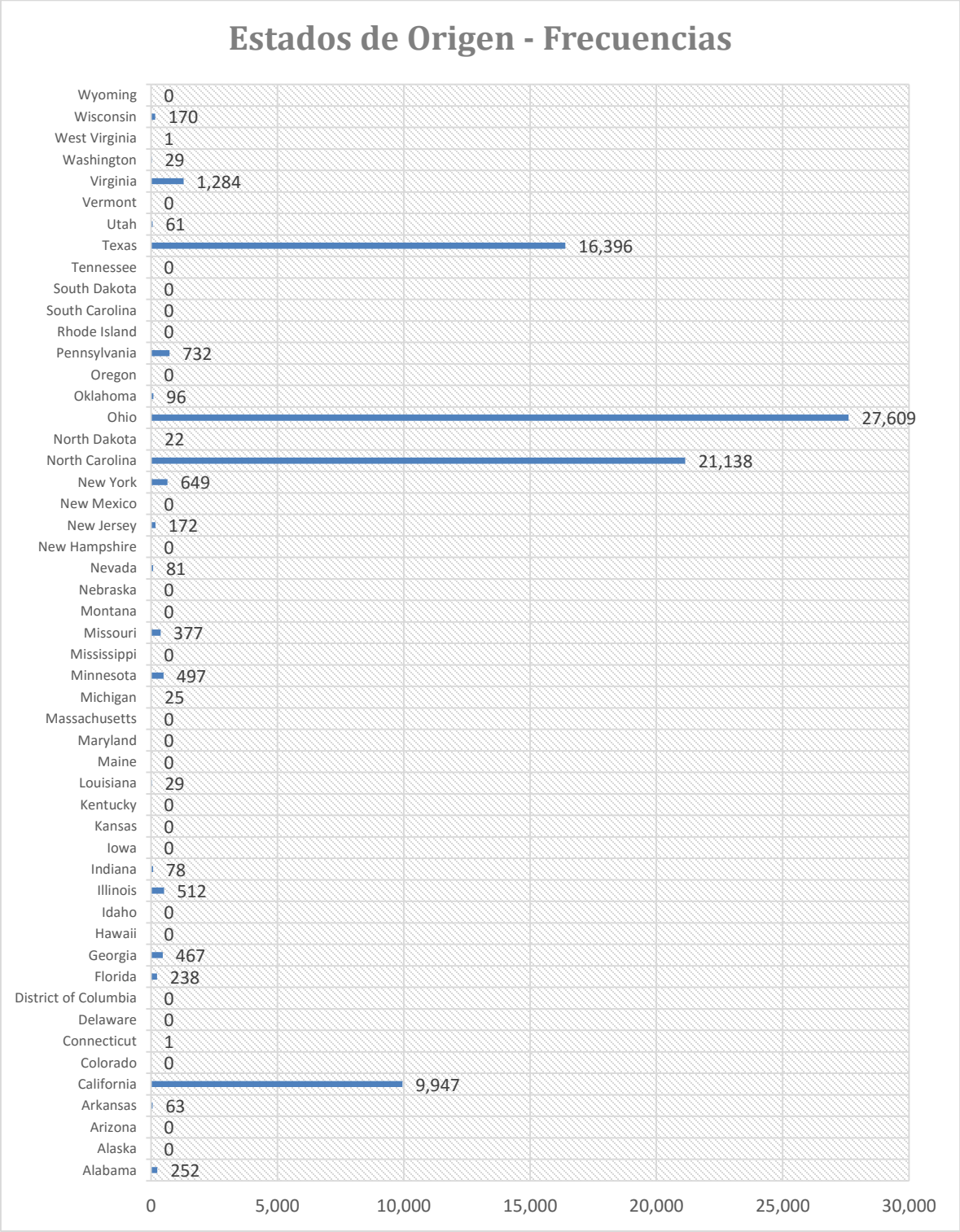


Ilustración 7 - Estados de Origen - Frecuencias - Elaboración Propia

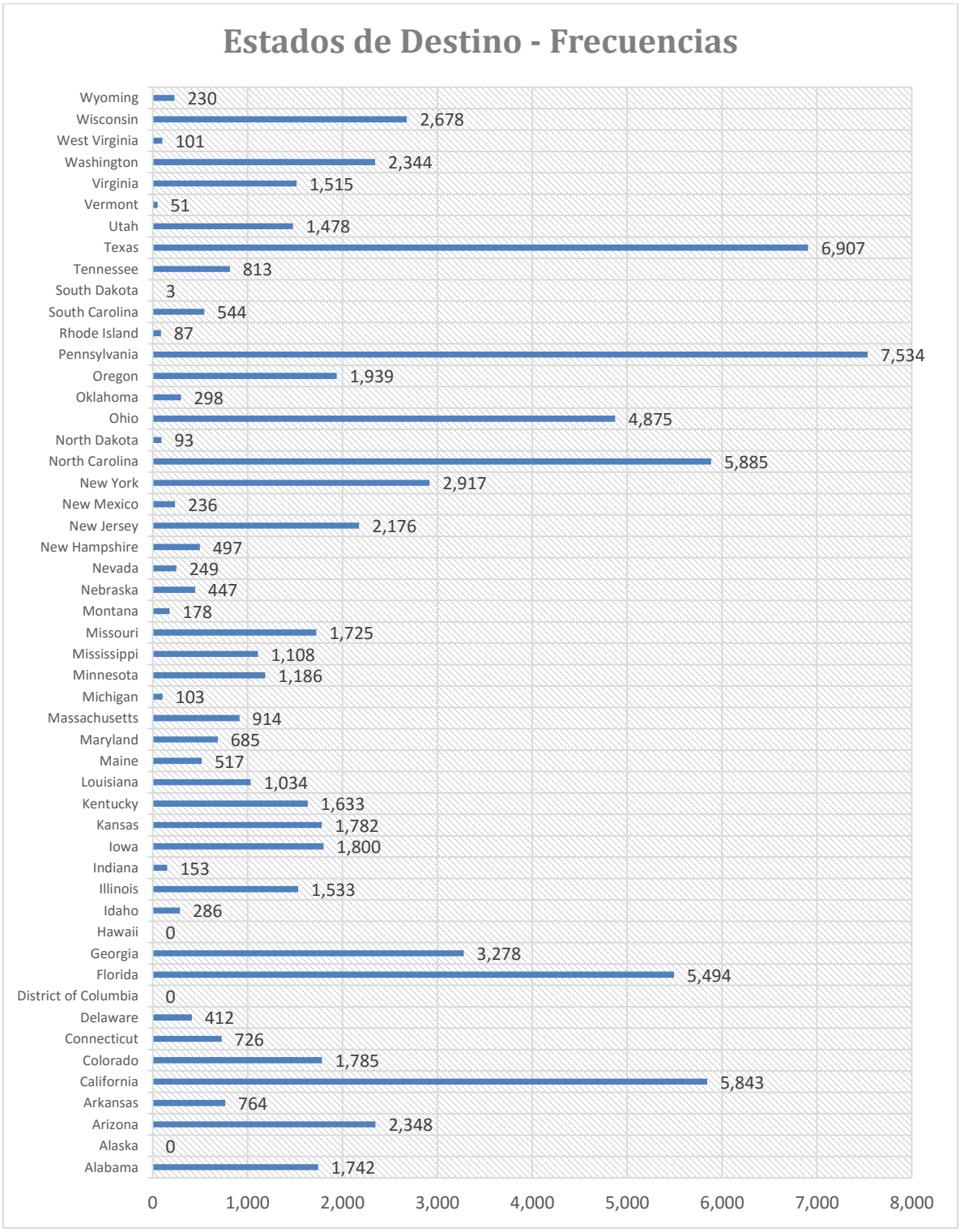


Ilustración 8 - Estados de Destino - Frecuencias - Elaboración Propia

Es interesante destacar que, de los 26 estados de origen, sólo 4 de ellos (Ohio, North Carolina, Texas y California) representan el 92,79% de las observaciones (34,12%, 26,12%, 20,12% y 12,29% respectivamente). En términos de destinos, los 4 estados con mayor participación en el número total de observaciones son Pennsylvania (9,31%), Texas (8,53%), North Carolina (7,27%) y California (7,22%), que en conjunto representan el 32,34%.

Puede resultar útil representar la información en un diagrama de Sankey[10], con los orígenes a la izquierda y los destinos a la derecha:

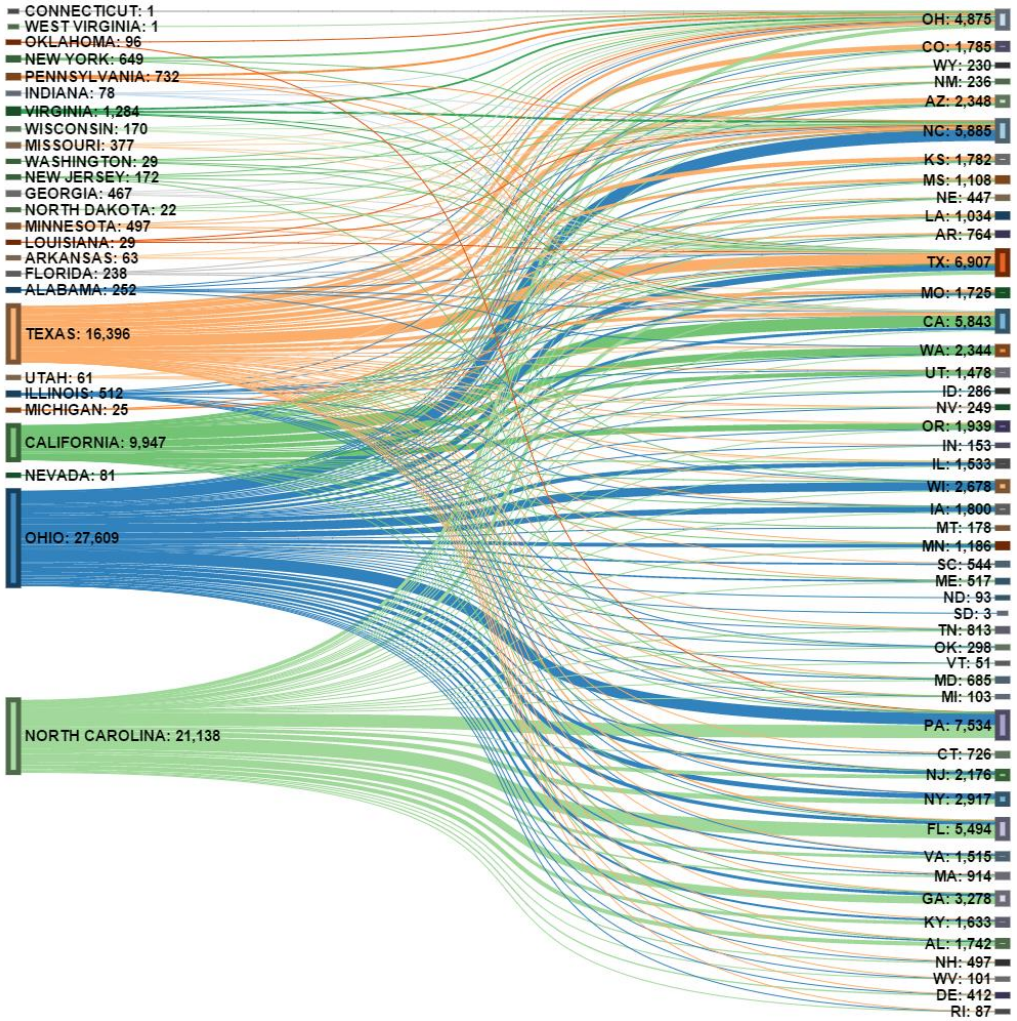


Tabla 3 - Diagrama de Sankey de Orígenes y Destinos - Elaboración Propia

[10] Tipo específico de diagrama de flujo, creado por el capitán irlandés Matthew Henry Phineas Riall Sankey (1853-1926), en el que la anchura de las flechas es proporcional a la cantidad de flujo.

8.2. Costo por Carga como una función de la distancia

A los fines del presente trabajo, analizaremos el Costo por Carga como una función de la distancia recorrida, es decir:

$$CPL = f(Distancia) \qquad (18)$$

Esta simplificación nos permitirá analizar con mayor detalle la regresión por el método de los MCO y la Regresión Cuantílica, a la vez que facilitará el desarrollo de un método estandarizado de trabajo para futuras aplicaciones.

Es posible obtener algunas medidas de tendencia central y de dispersión para la variable independiente (distancia), tal como sigue:

Medida	Variable Distancia
Media	673,47
Moda	641
Mediana	572
Mínimo	250
Máximo	3.061
Rango	2.811
Desviación Estándar	379,01
Varianza	143.647,16

Tabla 4 - Medidas de tendencia central y de dispersión para la variable distancia - Elaboración propia

8.3. Análisis en R

8.3.1. Organización de datos y mapeo

Para desarrollar el proyecto, se escribió un código en el lenguaje de programación R, utilizando el editor de texto con soporte para varios lenguajes de programación Notepad++. El código se irá presentando a lo largo del trabajo dentro de recuadros, tal como sigue:

```
##### Proyecto Final de Aplicación por Javier Martin - Tutor: C.P. Germán Tisera - ©2015 - v.1.0.000
##### Maestría en Dirección de Negocios - Cohorte 2013
##### Universidad Nacional de Córdoba - Facultad de Ciencias Económicas - Escuela de Graduados
```

Una de los primeros pasos a realizar dentro del software R, es el de organizar los datos y crear las variables que se utilizarán durante el análisis, en este caso, el Costo por Carga (CPL) y la Distancia (Distance):

```
##### Organizing the Data
table <- data.frame(read.table("capstone2.txt", header=T, sep="\t")) #reading the data
table$Distance <- as.numeric(table$Distance) #fixing the data
table$OriginLatitude <- as.numeric(as.character(gsub(",", ".", table$OriginLatitude))) #fixing the data
table$OriginLongitude <- as.numeric(as.character(gsub(",", ".", table$OriginLongitude))) #fixing the data
table$DestLatitude <- as.numeric(as.character(gsub(",", ".", table$DestLatitude))) #fixing the data
table$DestLongitude <- as.numeric(as.character(gsub(",", ".", table$DestLongitude))) #fixing the data
table$CPL <- as.numeric(as.character(gsub(",", ".", table$CPL))) #fixing the data
table$FreqOrigin <- as.numeric(table$FreqOrigin) #fixing the data
table$FreqDest <- as.numeric(table$FreqDest) #fixing the data
Distance <- table$Distance
CPL <- table$CPL
```

Con el objeto de profundizar el análisis y comprender la verdadera dimensión de la base de datos, se realizó un mapeo de todas las observaciones, utilizando para ello los paquetes ggmap y ggplot, y se anexaron dos nuevas variables a la base datos:

- FreqOrigin: Indica, para una determinada observación, el número de observaciones totales con el mismo estado de origen.
- FreqDest: Indica, para una determinada observación, el número de observaciones totales con el mismo estado de destino.

```
##### Mapping the Data
#Installing and executing the required packages
install.packages("ggmap")
install.packages("ggplot2")
library(ggmap)
library(ggplot2)
#Creating paths
aux.lat <- order(c(seq_along(table$OriginLongitude), seq_along(table$DestLongitude)))
lat <- unlist(c(table$OriginLongitude, table$DestLongitude))[aux.lat]
aux.long <- order(c(seq_along(table$OriginLatitude), seq_along(table$DestLatitude)))
long <- unlist(c(table$OriginLatitude,table$DestLatitude))[aux.long]
paths <- data.frame(lat, long)
#Plotting the paths
pdf("Truckload Movements.pdf", width=10, height=10)
par(mar=c(5,3,2,2)+0.1)
ggmap(get_googlemap(center="us", zoom=4, maptype="roadmap"), extent="device", legend="none")+
geom_point(data=table, aes(x=table$OriginLongitude, y=table$OriginLatitude, size=table$FreqOrigin), colour="blue",
alpha=0.05)+
geom_point(data=table, aes(x=table$DestLongitude, y=table$DestLatitude, size=table$FreqDest), colour="red",
alpha=0.05)+
geom_path(data=paths, aes(x=lat, y=long), colour="black", alpha=0.05)+
theme(legend.position="none")+
ggtitle("Truckload Movements")+
scale_size(range=c(1,15))
dev.off()
```

El mapa que se aprecia a continuación es el resultado de ejecutar el código especificado anteriormente:

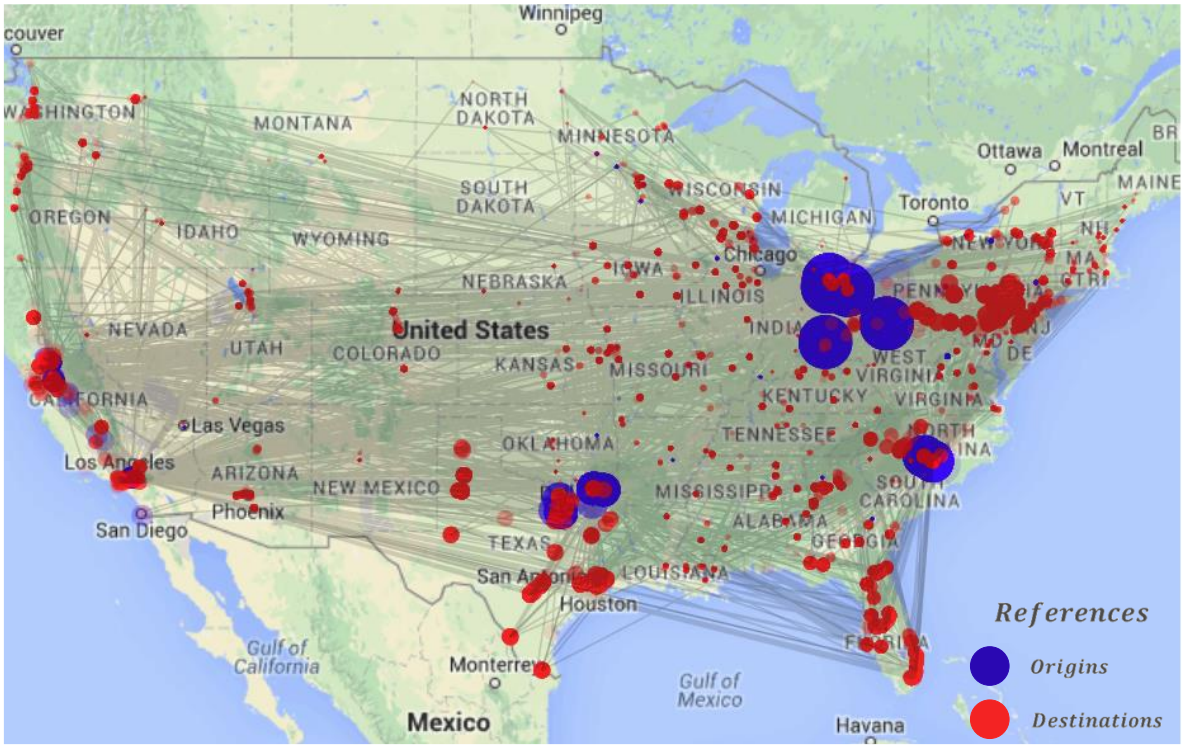


Ilustración 9 - Mapeo de Movimiento de Camiones - Elaboración propia

En el mapeo de movimiento de camiones se pueden observar en azul los orígenes y en rojo los destinos. El mapeo refuerza nuevamente lo previamente mencionado de que el 92,79% de las observaciones se corresponden con tan sólo 4 estados de origen, uno de ellos ubicados en la costa este (North Carolina), otro en el medio-oeste (Ohio), otro en el sudoeste (Texas), y otro en el Oeste (California). Es interesante resaltar este punto ya que 4 de las 5 grandes regiones de los Estados Unidos están incluidas en el porcentaje antes mencionado, e incluso la quinta región (noreste) también tiene participación, pero en menor proporción.

Además, al igual que el diagrama de Sankey, el mapeo de movimiento de camiones refleja la extrema complejidad de la base de datos, y la consecuente enorme necesidad de contar con diferentes mecanismos de estimación de tarifas de fletes.

8.3.2. Análisis generales sobre los datos

Se llevaron adelante algunos análisis generales sobre los datos en pos de profundizar el entendimiento sobre la base de datos, a través del siguiente código:

```
##### Analizing the Data
#Scatterplot
plot(Distance, CPL, main="CPL vs. Distance", type="p", xlab="Distance", ylab="CPL", col="ivory3")
#Frequencies
freq.Origin <- data.frame((table(table$OriginState)))
colnames(freq.Origin) <- c("OriginState","Frequency")
freq.Dest <- data.frame((table(table$DestState)))
colnames(freq.Dest) <- c("DestinationState","Frequency")
freq.Origin
freq.Dest
#Distance and CPL
summary(Distance)
summary(CPL)
#Q-Q Plot
qqnorm(CPL, col="ivory3"); qqline(CPL)
#Histogram With Fitted Density Curve for CPL
pdf("Histogram With Fitted Density Curve for CPL.pdf", width=10, height=10)
CPL.hist <- hist(CPL, freq=F, col="grey", ylim=c(0,0.0008), main="Histogram With Fitted Density Curve for CPL")
lines(density(CPL), col="orange", lwd=2)
abline(v=mean(CPL), col="red", lwd=2)
abline(v=median(CPL), col="blue", lwd=2)
legend(6000, 0.0001, legend=c("Mean CPL", "Median CPL", "Density Curve"), lwd=c(2,2,2), col=c("red","blue", "orange"),
bty="n")
text(6000, 0.0004, "The CPL distribution is positively skewed")
dev.off()
```

En primer lugar, se diseñó un gráfico de dispersión en el cual se aprecia que la mayor parte de las observaciones se sitúa en valores de distancia que varían entre las 250 a 1500 millas. Además, se observa una cierta tendencia de crecimiento del Costo por Carga a medida que aumenta la distancia, algo totalmente previsible y lógico.

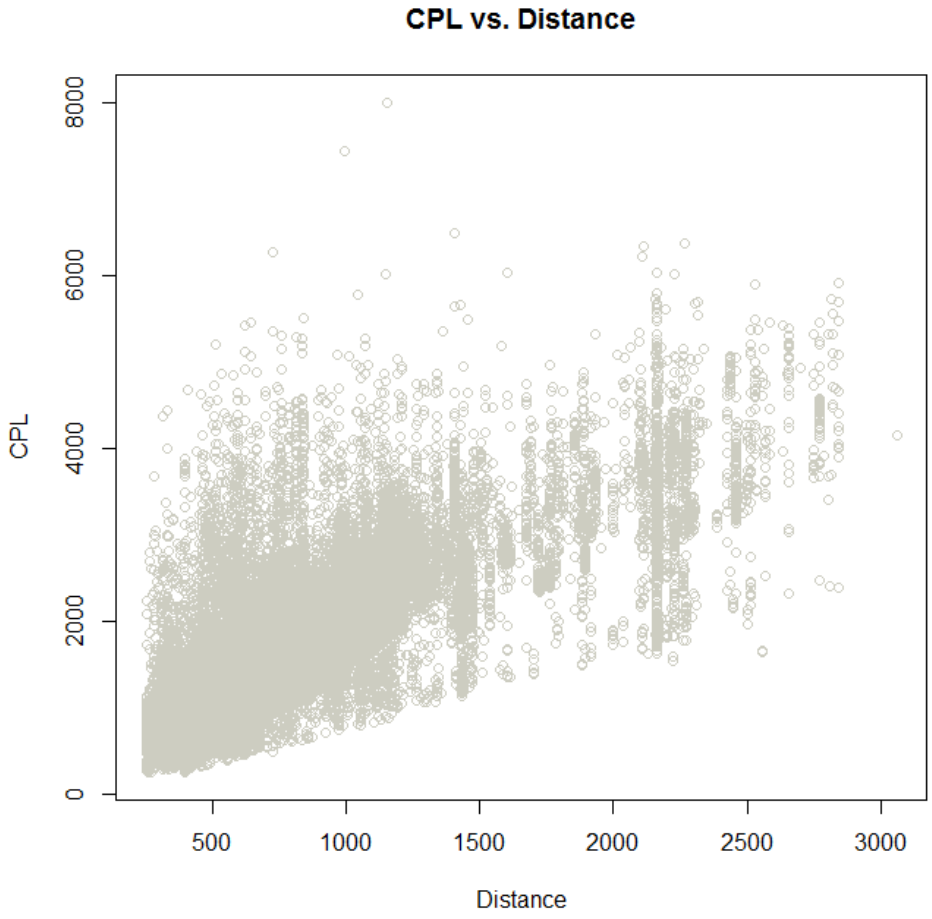


Ilustración 10 - Diagrama de dispersión CPL vs. Distance - Elaboración propia

Luego de verificar las frecuencias de orígenes y destinos, se realizó un análisis general sobre las dos variables seleccionadas, en pos de determinar algunas medidas importantes. Los resultados obtenidos fueron:

```
summary(Distance)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
250.0   421.0   572.0   673.5   815.0  3061.0
summary(CPL)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
254.5   929.9  1343.0  1510.0  1923.0  8002.0
```

Ilustración 11 - Medidas importantes para las variables bajo estudio

De la lectura de la ilustración anterior, se puede extraer que los montos del Costo por Carga varían entre los USD 254,50 y los USD 8.002, y poseen una media de USD 1.510.

El gráfico Q-Q[11] normal que se muestra a continuación nos permite comparar la distribución de la variable Costo por Carga con la distribución normal, y permite inferir que la primera no se encuentra normalmente distribuida y presenta colas pesadas a la derecha.

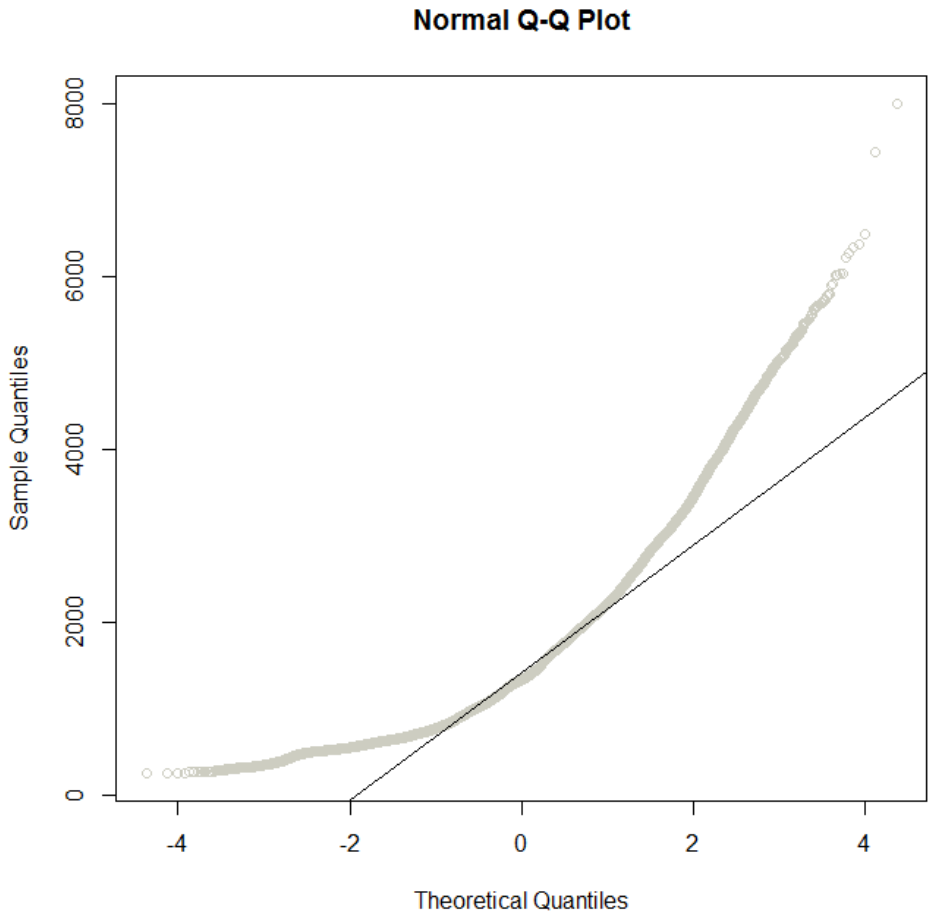


Ilustración 12 - Gráfico Q-Q normal - Elaboración propia

En términos de asimetría, la distribución del Costo por Carga es asimétrica hacia la derecha o positiva, tal cual se aprecia en el siguiente gráfico:

[11] Método gráfico para el diagnóstico de diferencias entre la distribución de probabilidad de una población de la que se ha extraído una muestra aleatoria, y una distribución usada para la comparación.

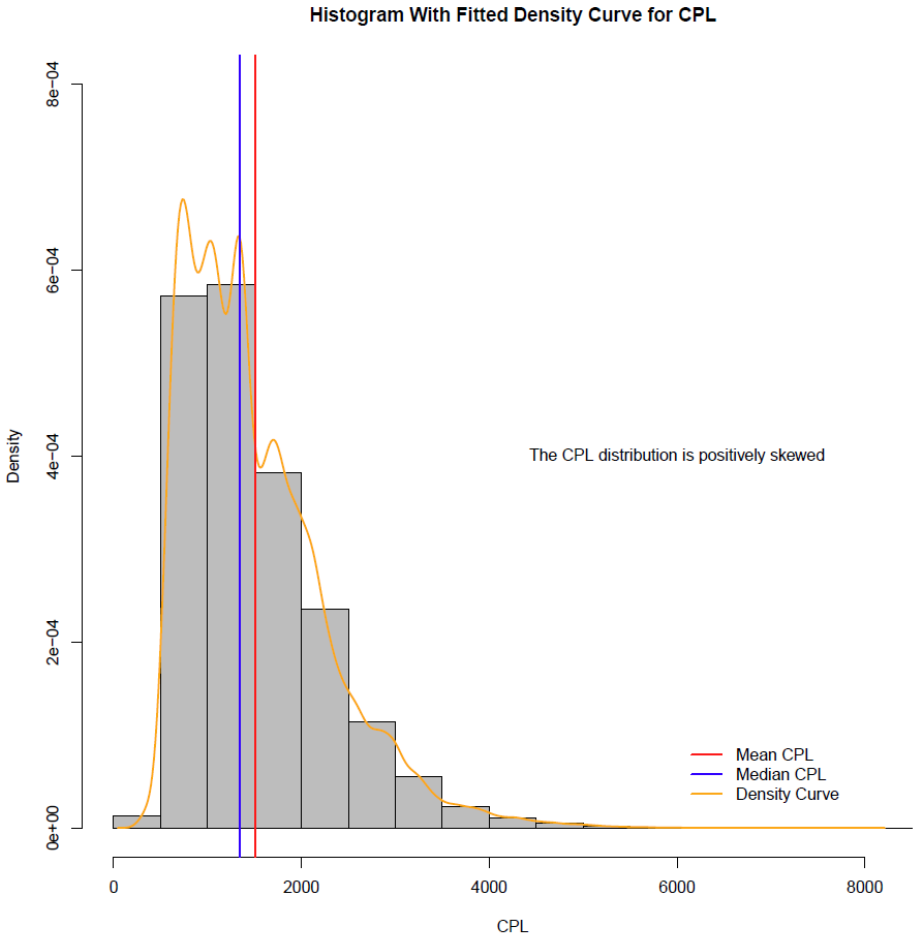


Ilustración 13 - Histograma con curva de densidad para el Costo por Carga - Elaboración propia

Algunas variables que pueden explicar este comportamiento de la variable Costo por Carga son:

- Round Trip vs. One Way: Algunos viajes son con retorno (round trip), y otros sin retorno (one way). Sin lugar a dudas estos últimos encarecen la tarifa.
- Contratos: Al momento de contratar los servicios de transporte se celebran contratos que contemplan seguros sobre la carga, accidentes y daños a terceros, con el objeto de que exista un respaldo ante cualquier eventualidad.

- Nivel de servicio: Hace referencia al nivel de servicio ofrecido por la empresa de transporte, el cual suele medirse periódicamente y suele impactar directamente sobre la tarifa que se cobra.
- Polinomio de reajustabilidad: con el fin de evitar subjetividades y no pagar "ineficiencias" en la tarifa de transporte, se suele generar un polinomio que involucra al menos tres aspectos de los costos de transporte que influyan en la tarifa final, por ejemplo: el petróleo, los neumáticos y el dólar.
- Imagen y seguridad: El transporte es siempre la cara visible final con los clientes, estos deben cumplir estándares de buena imagen, limpieza de los equipos y sobre todo seguridad, en aspectos tales como horas de conducción, control de velocidad vía GPS y equipos de carga y descarga cuando lo amerite. Sin lugar a dudas, la imagen de la empresa se ve reflejada en la tarifa final que se paga por un transporte.

8.3.3. Regresión por el método de los MCO

Luego de haber organizado y analizado los datos, se llevó adelante un análisis de regresión, utilizando el método de los MCO.

```
###### OLS Regression
#Running OLS
ols <- lm(CPL~Distance)
summary(ols)
#Plotting the results
plot(Distance, CPL, main="CPL vs. Distance", type="p", xlab="Distance", ylab="CPL", col="ivory3")
abline(ols, col="red", lwd=1) #adding the regression line to the scatterplot
points(Distance, predict(ols), col="red") #adding predicted values to the graph
ci <- predict(ols, interval="confidence") #adding confidence intervals
lines(Distance, ci[,2], lty=2, lwd=1, col="red")
lines(Distance, ci[,3], lty=2, lwd=1, col="red")
legend("bottomright", legend=c("OLS Line", "Lower and Upper Limits"), lty=c(1,2), col=c("red","red"), bty="n")
```

Los resultados pueden apreciarse en a continuación:

```
Call:
lm(formula = CPL ~ Distance)

Residuals:
    Min      1Q  Median      3Q     Max
-2934.9  -289.3   -55.8   219.7  5715.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.137e+02  3.242e+00   127.6   <2e-16 ***
Distance    1.628e+00  4.196e-03   388.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 452.4 on 80924 degrees of freedom
Multiple R-squared:  0.6505,    Adjusted R-squared:  0.6505
F-statistic: 1.506e+05 on 1 and 80924 DF,  p-value: < 2.2e-16
```

Ilustración 14 - Resultados de la regresión por el método de los MCO - Elaboración propia

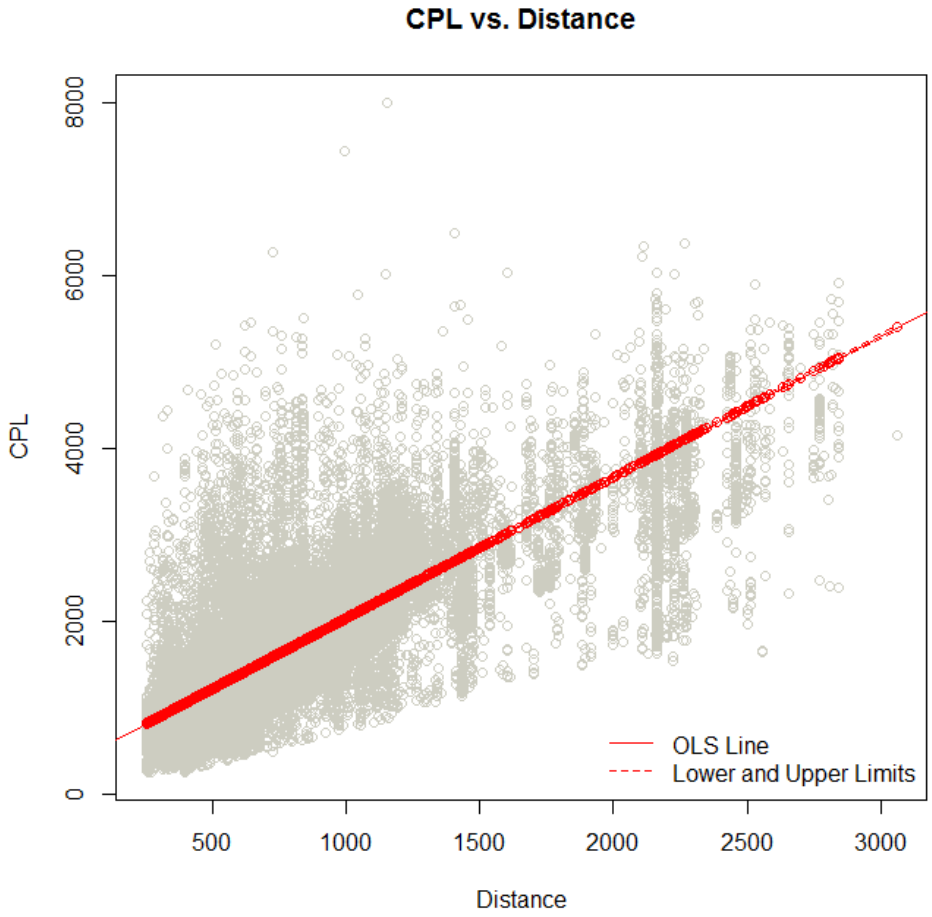


Ilustración 15 - Regresión por el método de los MCO - Elaboración propia

La recta de regresión que se obtiene por el método de los MCO es:

$$CPL_{MCO} = 413{,}70 + 1{,}63 \cdot Distance \qquad (19)$$

Podría inferirse a partir de la ecuación anterior, que el costo fijo de transporte promedio para todas las observaciones de la base de datos es de USD 413,70, e incluye costos de seguros, amortizaciones, salarios de los conductores, entre otros. De la misma manera, podría decirse que, por cada milla recorrida, el Costo por Carga se incrementa en USD 1,63. Este valor de pendiente

explica costos variables como el combustible, el mantenimiento de los camiones, peajes, entre otros.

8.3.4. Coeficientes de correlación y determinación

En pos de determinar los coeficientes de correlación y de determinación, se construyó el siguiente código:

```
###### Coefficients of Determination and Correlation
# Coefficient of Correlation
R <- cor(CPL,Distance)# Coefficient of Determination
summary(ols)$r.squared
# ANOVA
anova <- aov(CPL~Distance)
# Sum of Squares due to Regression
SSR <- sum((predict(anova) - mean(CPL))^2)
# Sum of Squares due to Error
SSE <- sum(anova$residuals^2)
# Total Su
SST <- SSR + SSE
# Coefficient of Determination
R² <- SSR/SST
# Results
a. <- SSR
b. <- SSE
c. <- SST
d. <- R²
Results <- data.frame(a.,b.,c.,d.)
colnames(Results) <- c("SSR","SSE","SST", "R² ")
rownames (Results)  <-  c ("Results")
```

Los resultados se resumen a continuación:

```
> Results
                SSR         SSE         SST        R²
Results 30823684399 16560727125 47384411524 0.6505026
```

Ilustración 16 - SSR, SSE, SST y coeficiente de determinación - Elaboración propia

Se observa que 65,05% de la variación puede ser explicada por el modelo de regresión por el método de los MCO, algo que podría definirse como relativamente aceptable, a pesar de la clara no normalidad de la distribución.

```
> R
[1] 0.8065374
```

Ilustración 17 - Coeficiente de correlación - Elaboración propia

En términos del coeficiente de correlación, tenemos que las variables Costo por Carga y Distancia se encuentran correlacionadas positivamente, algo que, como mencionamos, es totalmente esperable.

8.3.5. Regresión Cuantílica

Para poder realizar la Regresión Cuantílica, se utiliza el paquete del software R llamado quantreg, versión 5.19, creado por Roger Koenker, uno de los padres de la Regresión Cuantílica. El código desarrollado es el siguiente:

```
##### Quantile Regression
#Installing and executing the package
install.packages("quantreg")
library(quantreg)
#Running QR and plotting the results
plot(Distance, CPL, main="CPL vs. Distance", type="p", xlab="Distance", ylab="CPL", col="ivory3")
abline(rq(CPL~Distance, tau=0.5), lwd=1, col="blue") #adding the median regression line
abline(lm(CPL~Distance), lwd=1, col="red") #adding the insipid OLS regression line
taus <- c(0.05,0.10,0.25,0.75,0.90,0.95) #the other six common quantile points are chosen (in addition to 0.5)
for(i in 1:length(taus)){
abline(rq(CPL~Distance, tau=taus[i]), lwd=1, col="black")
}
legend("bottomright", legend=c("OLS Line","Median Line","QR Lines"), lty=c(1,1,1), col=c("red","blue","black"), bty="n")
#Plotting the slope and intercept of the estimated linear quantile regression of the data as a function of tau
plot(summary(rq(CPL~Distance, tau=1:99/100)))
```

Se eligieron seis cuantiles representativos (y muy comunes en el campo de la investigación), además del cuantil 0,5 (que proporciona la regresión mediana). Estos cuantiles representativos son: 0,05; 0,10; 0.25; 0,75; 0,90; 0,95. Los resultados obtenidos se aprecian en el gráfico a continuación:

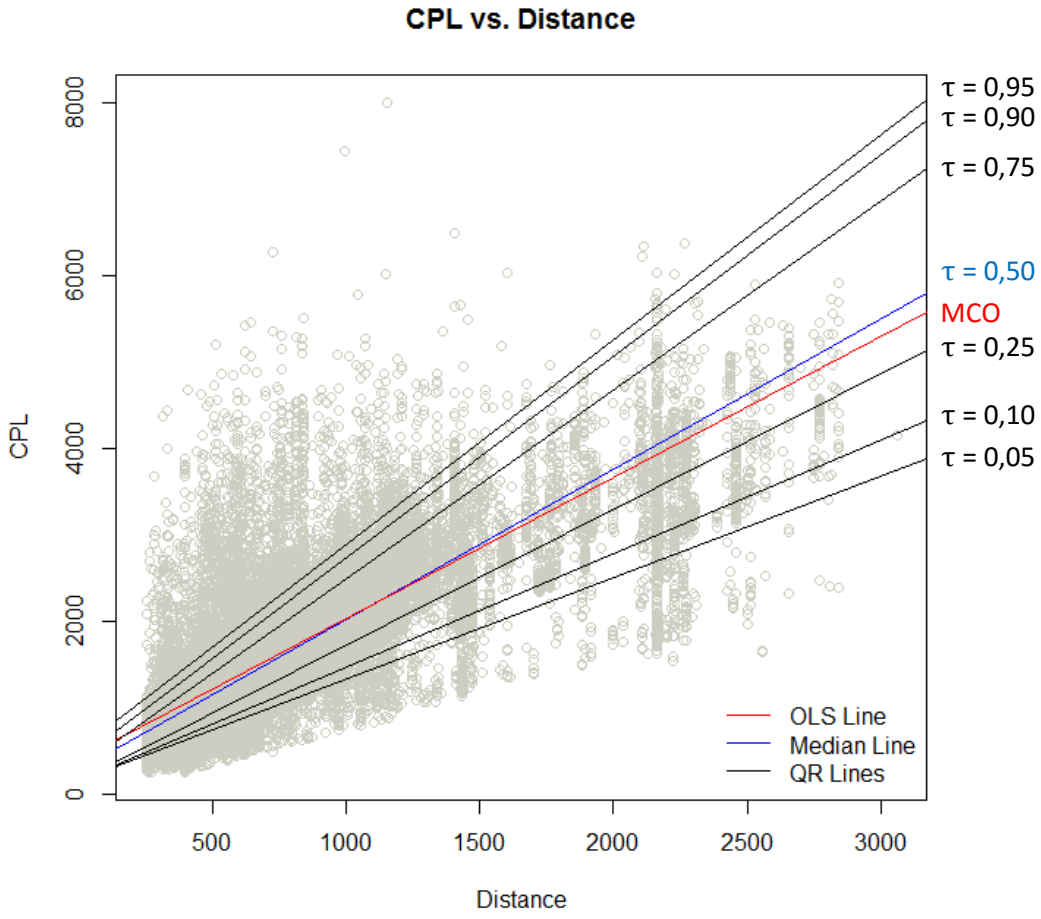


Ilustración 18 - Regresión Cuantílica - Elaboración propia

En la imagen anterior pueden apreciarse ocho rectas de regresión:

- OLS Line: En rojo, la recta de regresión del método de los MCO.
- Median Line: En azul, la recta de regresión mediana, un caso especial de Regresión Cuantílica cuando τ es igual a 0,5.
- QR Lines: En negro, las seis rectas de Regresión Cuantílica para los valores de τ representativos anteriormente mencionados (0,05 la recta con menor pendiente, y 0,95 la recta con mayor pendiente). Si se hubieran elegido más cuantiles, hubiéramos obtenido tantas rectas de regresión como cuantiles hubiésemos seleccionado.

Es interesante destacar la cercanía entre las rectas de regresión por el método de los MCO y la recta de regresión mediana, en torno a las 1.000 millas de distancia.

Sin lugar a dudas, el gráfico más representativo de la Regresión Cuantílica es aquél en el que se muestran los valores de ordenada al origen y de pendiente derivados de las rectas de Regresión Cuantílica, como una función de los cuantiles de la variable endógena.

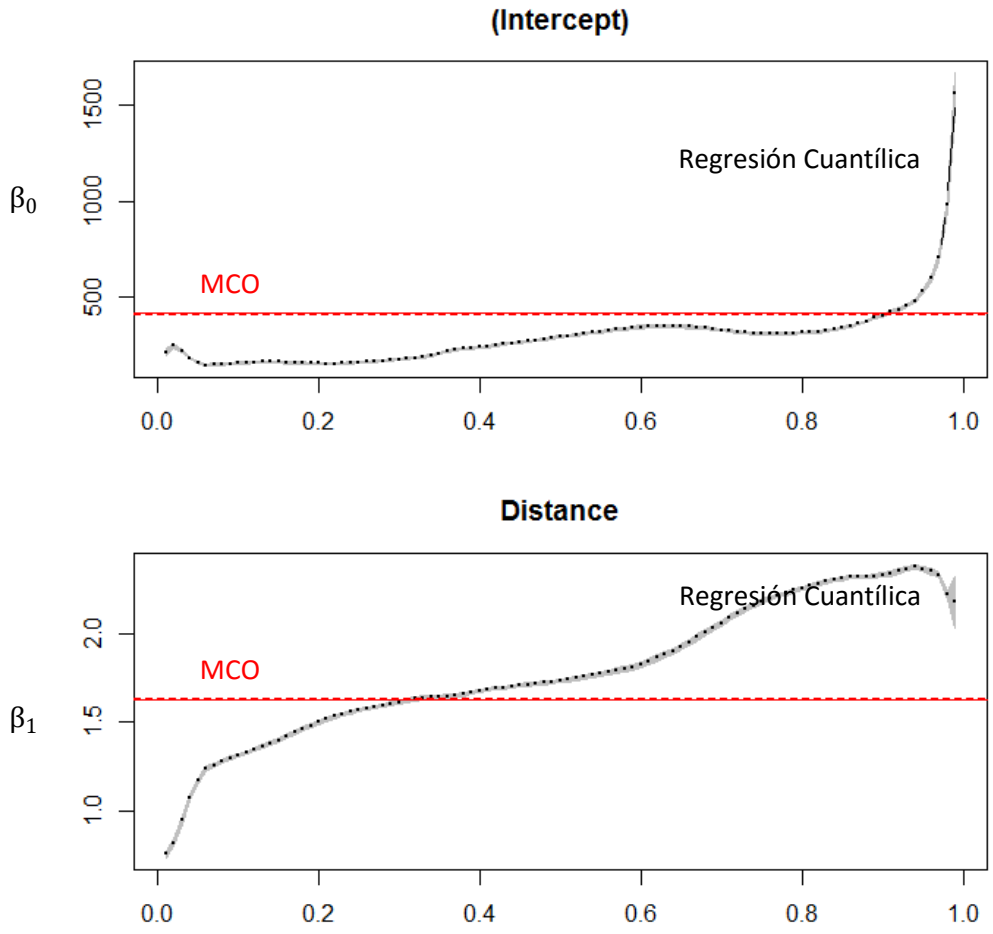


Ilustración 19 - Valores de ordenada al origen y de pendiente derivados de las rectas de Regresión Cuantílica, como una función del cuantil - Elaboración propia

Observando el gráfico de los valores de ordenada al origen de las rectas de Regresión Cuantílica, se observa que existen diferencias estadísticamente significativas entre ambos métodos, para prácticamente casi todo el espectro de cuantiles.

Si observamos el gráfico correspondiente a valores de pendiente, se aprecia que, para valores bajos del Costo por Carga, el método de los MCO sobreestima el impacto de la distancia en la determinación del Costo por Carga, mientras que subestima completamente el impacto para valores elevados. Esto es una consecuencia directa del hecho que, a medida que se incrementa la distancia y aumenta el Costo por Carga, los costos fijos del transporte se licuan y comienzan a emerger como determinantes los costos variables mencionados anteriormente.

En términos de costos fijos, el método de los MCO sobreestima los montos prácticamente para todos los valores de τ, y la Regresión Cuantílica muestra una marcada estabilidad, sólo creciendo en los cuantiles muy superiores de la variable endógena. Esta última situación permite realizar algunas interpretaciones:

- Las empresas analizadas poseen una estructura de costos similar.
- Los viajes de grandes distancias, que coinciden en la mayoría de los casos con valores muy altos del Costo por Carga, sufren el impacto de los altos costos de seguro y de las compensaciones de los conductores por las condiciones demandantes de tales travesías.

Queda demostrado por medio de este análisis que la Regresión Cuantílica ofrece una descripción más amplia y sistémica de la distribución de la variable endógena que se desea estudiar, ofreciendo mejores bases para estimar tarifas de fletes.

8.3.6. Estimación de la tarifa de fletes

Con el objetivo de estimar la tarifa de fletes a través de la recta de regresión obtenida por el método de los MCO y a través del uso de la Regresión Cuantílica, se creó el siguiente código:

```
##### CPL Forecasting Function
CPL.forecasting <- function(dist, tau) {
#Function to determine the CPL given a Distance (using OLS)
CPL.pred.ols <- function(dist){
ols$coefficients[[1]]+ols$coefficients[[2]]*dist
}
OLS.Prediction <- CPL.pred.ols(dist)
#Function to determine the CPL given a Distance (using Quantile Regression)
CPL.pred.rq <- function(dist){
rq <- rq(CPL~Distance, tau=tau)
rq$coefficients[[1]]+rq$coefficients[[2]]*dist
}
rq.Prediction <- CPL.pred.rq(dist)
#Slopes
slope.ols <- round(ols$coefficients[[2]], 2)
rq <- rq(CPL~Distance, tau=tau)
slope.rq <- round(rq$coefficients[[2]], 2)
#Intercepts
intercept.ols <- round(ols$coefficients[[1]], 2)
intercept.rq <- round(rq$coefficients[[1]], 2)
#Results
a <- c(round(OLS.Prediction,2), slope.ols, intercept.ols)
b <- c(round(rq.Prediction,2), slope.rq, intercept.rq)
Output <- data.frame(a,b)
rownames(Output) <- c("Prediction","Slope","Intercept")
colnames(Output) <- c("OLS","RQ")
return(Output)
}
```

Con este código es factible seleccionar un valor para la variable independiente y para τ, y obtener el valor predicho por la recta de regresión por el método de los MCO y el predicho por la recta de Regresión Cuantílica para el valor del cuantil seleccionado, incluso para valores de distancia fuera del espectro contenido en la base de datos. Por ejemplo, si elegimos un valor de distancia de 200 millas y un valor de cuantil de 0,25 arribamos a los siguientes resultados:

```
                OLS      RQ
Prediction 739.41 473.03
Slope        1.63   1.57
Intercept  413.74 159.99
```

Ilustración 20 - Ejemplo de estimación de tarifas de fletes - Elaboración propia

La moda para la variable distancia es de 641 millas, por lo que podemos calcular los valores predichos de Costo por Carga, ordenada al origen y pendiente para la recta de regresión obtenida por el método de los MCO y para las siete rectas Regresión Cuantílica de los cuantiles representativos mencionados en la sección anterior:

Concepto	CPL Prom. Real	MCO	Regresión Cuantílica						
			τ = 0,05	τ = 0,10	τ = 0,25	τ = 0,50	τ = 0,75	τ = 0,90	τ = 0,95
CPL	1323,44	1457,52	910,00	998,78	1163,29	1403,25	1706,81	1898,89	2046,74
Pendiente	-	1,63	1,17	1,31	1,57	1,73	2,18	2,33	2,36
Ordenada al Origen	-	413,74	158,87	156,72	159,99	293,69	310,11	405,69	531,00

Tabla 5 - Comparación entre el método de los MCO y la Regresión Cuantílica para la estimación de tarifas de fletes - Elaboración propia

Los resultados presentados en la tabla anterior permiten efectuar los siguientes comentarios:

- La estimación del Costo por Carga por el método de los MCO y para la regresión mediana (Regresión Cuantílica con τ igual a 0,5) son muy similares. A pesar de esta similitud, el valor de la ordenada al origen es bastante mayor para el caso del MCO, situación que se ve compensada por la mayor pendiente presentada por la regresión mediana.
- Tomando un simple promedio entre los valores de Costo por Carga obtenidos a través de la Regresión Cuantílica, obtenemos un valor de cercano a los USD 1.447, más cercano al valor real que el provisto por el método de los MCO.

8.3.7. Animación para apreciar todas las rectas de regresión obtenidas a través de la Regresión Cuantílica

Con el objeto de dinamizar los resultados obtenidos por medio de la Regresión Cuantílica, se creó el siguiente código para crear una secuencia animada con 100 rectas de regresión:

```
##### Animation
#Installing Animation Package
install.packages("animation")
library(animation)
#Creating PNG files
png(file="rq%03d.png", width=1920, heigh=1080)
for (i in seq(0, 1, 0.01)){
par(mar=c(5,5,5,5)+0.1)
plot(Distance, CPL, main="CPL vs. Distance", type="p", xlab="Distance (miles)", ylab="CPL ($)", col="ivory3", cex.main=2,
cex.axis=2, cex.lab=2)
ols <- lm(CPL~Distance)
rq <- rq(CPL~Distance, tau=i)
text(2500, 1000, paste("tau =", i , "\n ß1_OLS =", round(ols$coefficients[[2]], 2), "\n ß1_RQ =", round(rq$coefficients[[2]],
2)), cex=3)
abline(lm(CPL~Distance), lwd=1, col="red")
abline(rq(CPL~Distance, tau=i), lwd=1, col="blue")
legend("bottomright", legend=c("OLS Line","QR Lines"), lty=c(1,1), col=c("red","blue"), bty="n", cex=2)
}
dev.off()
#Converting PNG files to one GIF using ImageMagick (ImageMagick must be installed)
ani.options(convert="C:/Program Files/ImageMagick-6.9.0-Q16/convert.exe", interval=0.15)
im.convert("rq*.png", output="rq.gif")
#Removing PNG Files
file.remove(list.files(pattern=".png"))
```

A continuación, se ofrece una breve muestra de los resultados (el gráfico se lee de izquierda a derecha y de arriba hacia abajo):

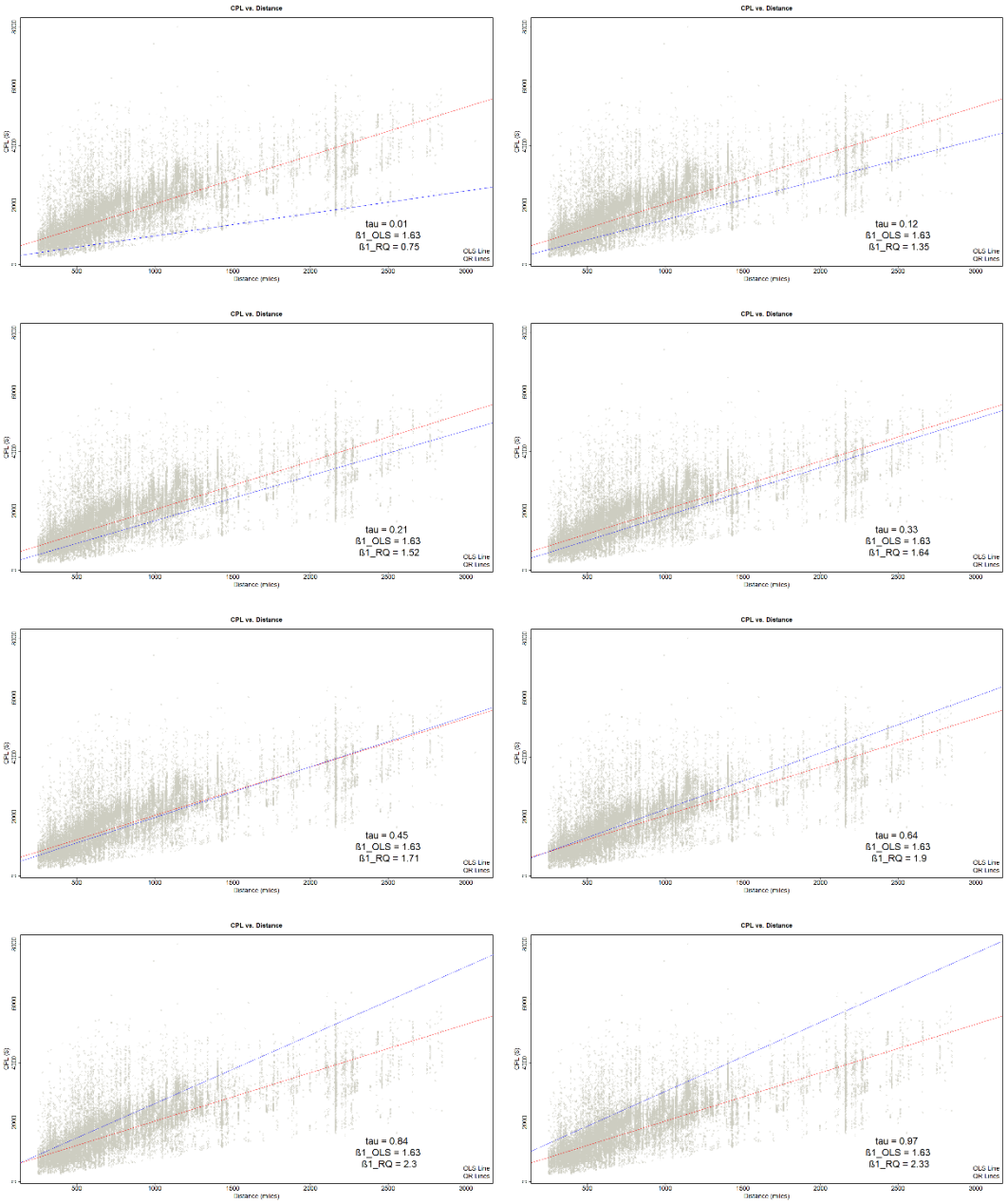


Ilustración 21 - Secuencia de rectas de regresión derivadas de la Regresión Cuantílica - Elaboración propia

8.3.8. Comparación entre la regresión por el método de los MCO y la Regresión Cuantílica

Con el objeto de comparar ambos métodos, se escribió un código que por motivos de extensión no se ubica dentro del cuerpo principal del presente trabajo, sino en el apéndice 2. Para proceder con en el análisis, se agruparon las observaciones de distancias en rangos de 250 millas y se calcularon las predicciones de Costo por Carga utilizando la recta de regresión sugerida por el método de los MCO y la recta de regresión sugerida por la Regresión Cuantílica para tres cuantiles: 0,25; 0,50; 0,75. Luego, se calcularon los errores relativos derivados de comparar cada una de las predicciones con los valores reales de Costo por Carga, para cada una de las observaciones. Los resultados se graficaron en diagramas de caja[12], indicando con rojo los resultados para el método de los MCO y en azul para la regresión cunatílica.

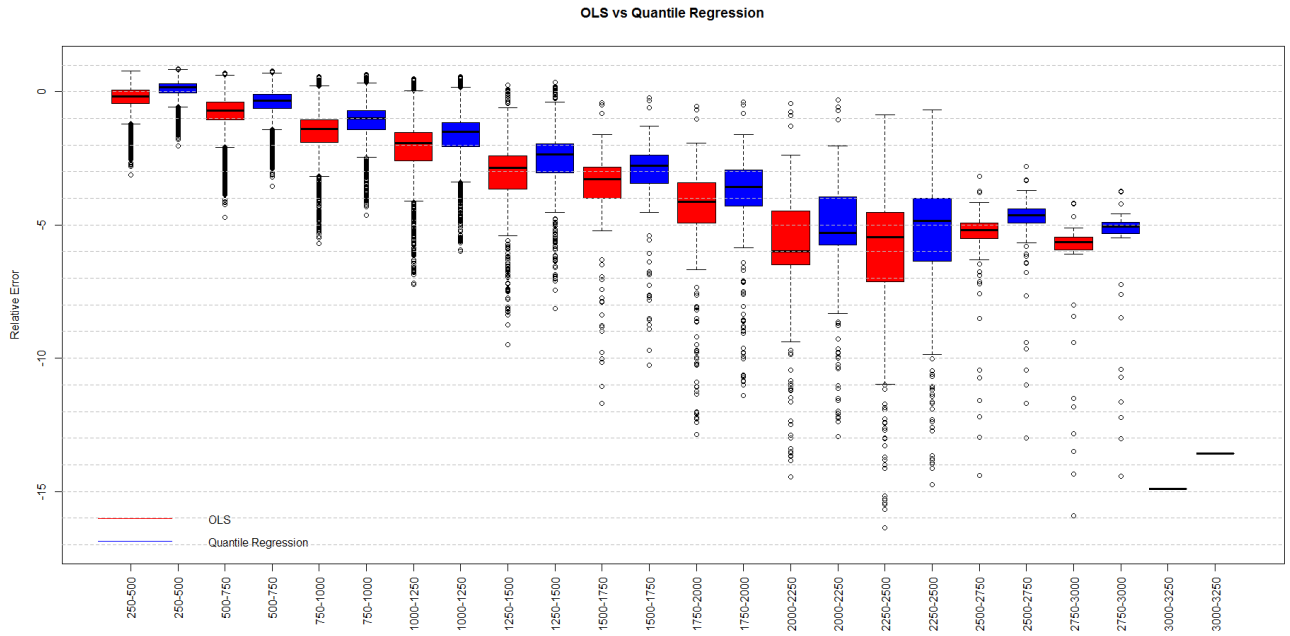


Ilustración 22 - Comparación entre métodos para τ = 0,25 - Elaboración propia

[12] Gráfico que suministra información sobre los valores mínimo y máximo, los cuartiles Q1, Q2 o mediana y Q3, y sobre la existencia de valores atípicos y la simetría de la distribución.

Utilizando un τ igual a 0,25 se aprecia que la Regresión Cuantílica ofrece estimaciones más robustas para todos los rangos de distancias.

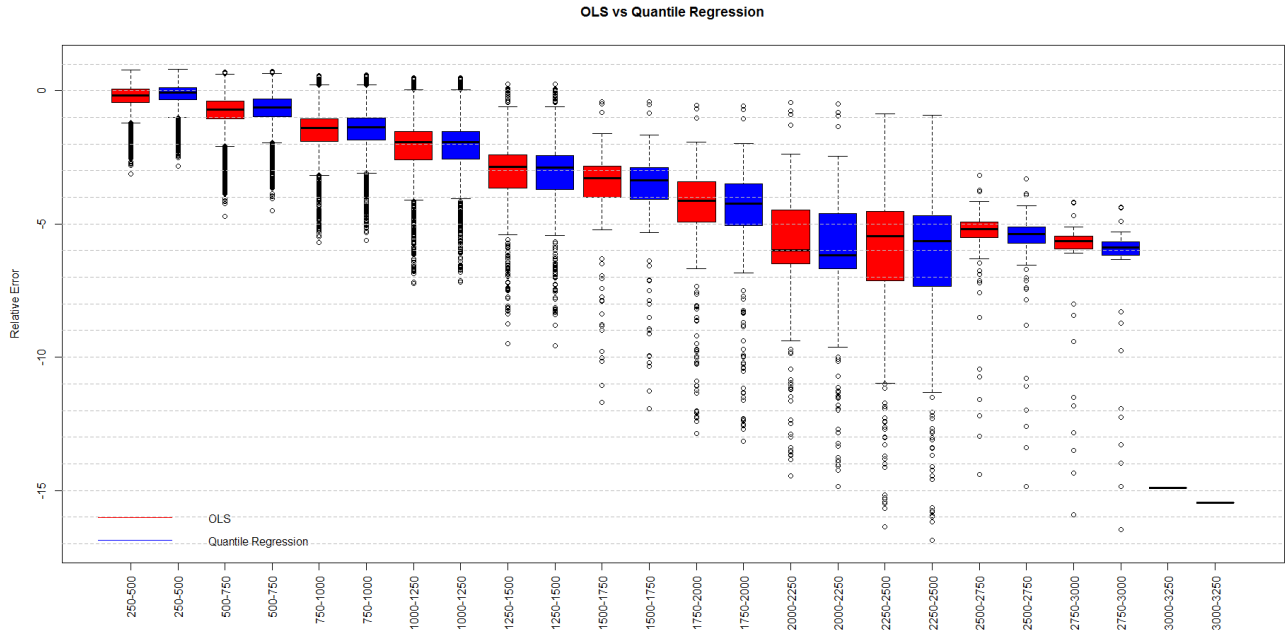


Ilustración 23 - Comparación entre métodos para τ = 0,50 - Elaboración propia

Para el caso de τ igual a 0,50 se aprecia que la Regresión Cuantílica ofrece estimaciones más robustas hasta el rango de distancias que corre desde las 1.000 a las 1.250 millas inclusive, situación que cambia al movernos hacia rangos de distancias superiores, aunque en todos los casos con un comportamiento muy similar al obtenido a través del método de los MCO.

Diferente es la situación cuando se utiliza un valor de τ igual a 0,75 para realizar las estimaciones a través de la Regresión Cuantílica, ya que en todos los rangos de distancias el método de los MCO ofrece estimaciones más robustas. Esto puede apreciarse en el gráfico que se ubica a continuación.

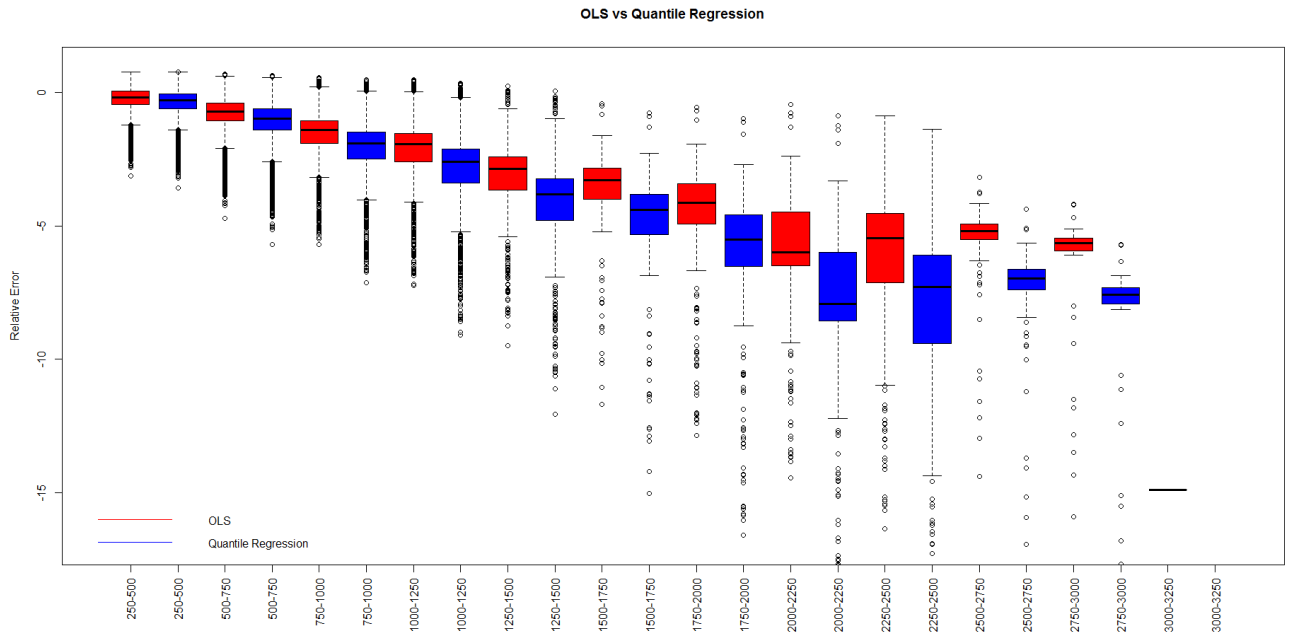


Ilustración 24 - Comparación entre métodos para τ = 0,75 - Elaboración propia

C. Cierre del Proyecto

1. Sobre ambos métodos de regresión

A partir de los resultados obtenidos a lo largo del presente trabajo, es posible remarcar que la Regresión Cuantílica otorga una imagen más completa del efecto de la variable independiente sobre la variable dependiente y permite modelar cualquier número de cuantiles condicionales sin riesgo de sesgo en la estimación de los parámetros del modelo. Además, en el caso de la Regresión Cuantílica existe robustez en la estimación contra la posible presencia de heterocedasticidad y observaciones extremas, dado que los parámetros de la misma se estiman minimizando la suma de los valores absolutos ponderados de los residuos, en lugar de la suma de cuadrados de residuos como es en el caso del método de los MCO.

Respecto a las desventajas de la Regresión Cuantílica, comparada con la regresión por el método de los MCO, podemos mencionar el alto grado de trabajo computacional que se requiere para desarrollar un estudio de características similares a las del presente trabajo, no sólo en términos de calidad de información, sino también en términos de flexibilidad. Otros puntos interesantes a mencionar son que la función objetivo de la Regresión Cuantílica no es diferenciable en el origen, y por consiguiente no puede darse una solución cerrada, y el escaso desarrollo teórico que existe sobre la misma en algunos ámbitos.

2. Metodología estándar de trabajo

Con el objeto de poder extender el presente estudio otras bases de datos pertenecientes a otros países en diferentes regiones o continentes, es necesario organizar el estudio en una serie de pasos, cada uno de los cuales apalancará los resultados positivos finales. La propuesta es como sigue:

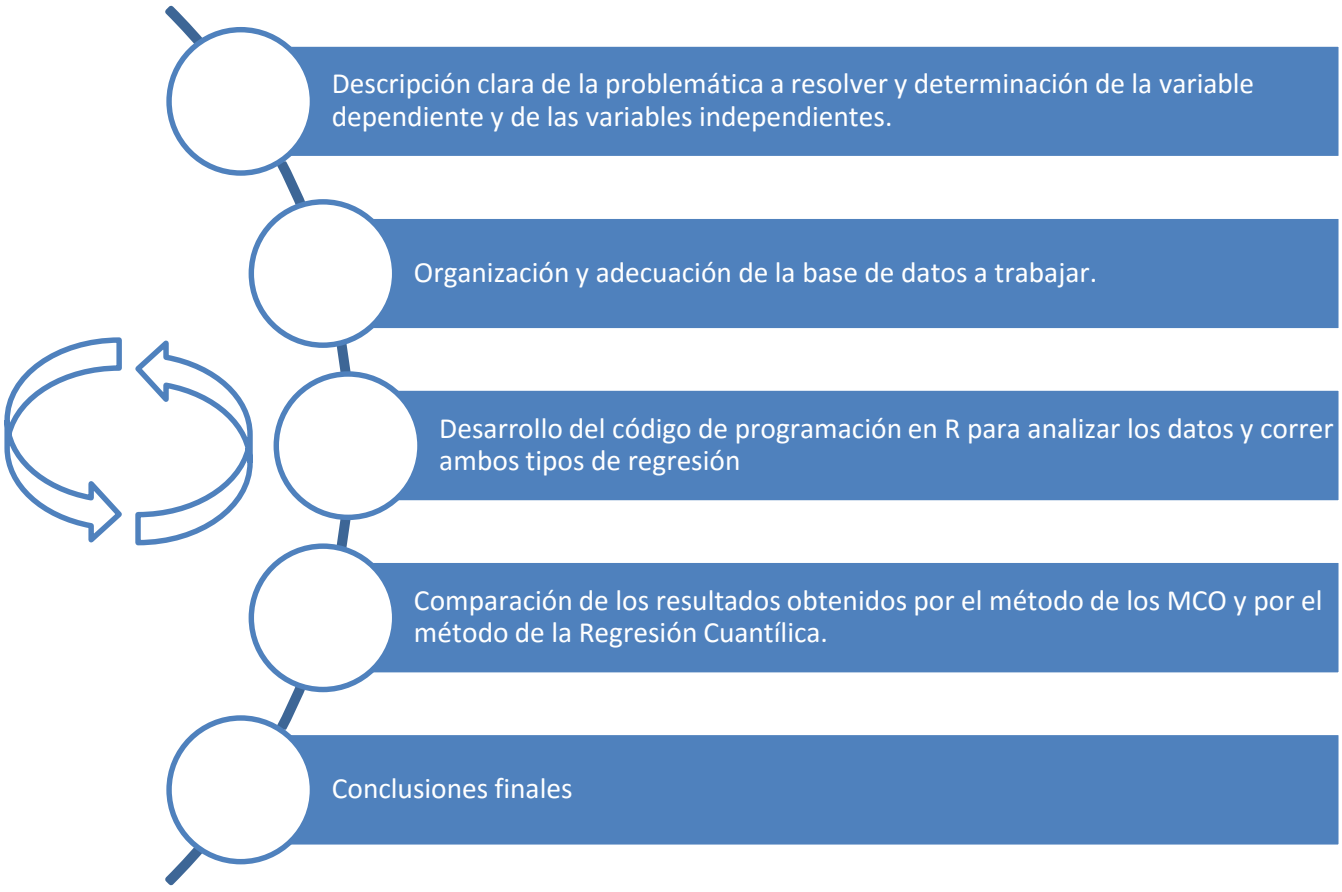


Ilustración 25 - Metodología de trabajo estandarizada - Elaboración propia

Esta metodología estándar de trabajo también puede hacerse extensiva a otros campos de la Gestión de la Cadena de Suministros, incluso del tipo humanitario o ambiental. En este sentido, el presente trabajo es una invitación a encarar este tipo de estudios.

3. Palabras finales

A lo largo del presente trabajo se ha logrado combinar dos campos de investigación en pleno desarrollo, es decir, el Supply Chain Management y la Regresión Cuantílica. Esta situación es totalmente novedosa, no sólo en el marco del Trabajo Final de Aplicación de la Maestría en Dirección de Negocios de la Facultad de Ciencias Económicas de la Universidad Nacional de Córdoba, sino también en ambos ámbitos de investigación a nivel mundial. En este sentido, el presente informe puede servir como punto de partida elemental para futuras investigaciones referidas a la estimación de tarifas de fletes, no sólo para los Estados Unidos de América, desde donde proviene la base de datos que se utilizó durante la investigación, sino también para cualquier otro país, independientemente de su grado de desarrollo en materia de transporte terrestre o de la complejidad de su cadena de suministros.

Por otra parte, la Regresión Cuantílica brinda tal flexibilidad que puede ser utilizada en otros campos del Supply Chain Management, no sólo el del transporte. A tal fin, se puede tomar como punto de partida la metodología estándar propuesta en el presente trabajo.

D. Fuentes

Bibliografía

Anderson, D., Sweeney, D., & Williams, T. (2005). *Estadística para administración y economía.* Thomson.

Chen, F., & Chalhoub-Deville, M. (2014). Principles of Quantile Regression and an Aplication. *SAGE*, 63-87.

Chopra, S., & Meindl, P. (2013). *Supply Chain Management - Strategy, Planning and Operation.* Pearson.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press.

Koenker, R. (2012). Quantile Regression in R: A Vignette.

Koenker, R., & Bassett, G. (1978). Regression Quantiles. *Econometrica, 46*(1), 33-50.

Koenker, R., & Hallock, K. (2001). Quantile Regression. *Journal of Economic Perspectives*, 143-156.

Mahía, R., & de Arce, R. (2011). Breve Apunte de la Estimación de los Parámetros MCO y Máxima Verosimilitud.

Otero, J., & Reyes, B. (2012). *Regresión Cuantílica: Estimación y Contrastes.* Madrid: Instituto Lawrence R. Klein.

SPPS. (s.f.). El Análisis de Regresión Lineal: El Procedimiento de Regresión Lineal. SPPS.

Venables, W., Smith, D., & R Core Team. (2014). An Introduction to R.

Sitios Web

- edX: https://www.edx.org/
- R Project: http://www.r-project.org

Software

- Microsoft Office: https://products.office.com
- Notepad++: https://notepad-plus-plus.org
- R Software: http://www.r-project.org
- SankeyMATIC: http://sankeymatic.com

Apéndices

Apéndice 1: Deducción con cálculo infinitesimal de las fórmulas de cuadrados mínimos

El método de los MCO es un procedimiento para determinar los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ que minimizan la suma de los residuales elevados al cuadrado. Esta suma se expresa como sigue:

$$\sum (Y_i - \hat{Y}_i)^2$$

Al sustituir $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ obtenemos:

$$\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

El objetivo será el de minimizar esta última expresión. Para hacerlo, debemos obtener derivadas parciales respecto a $\hat{\beta}_0$ y $\hat{\beta}_1$, igualarlas a cero y resolver las ecuaciones. Así, obtenemos:

$$\frac{\partial \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\partial \hat{\beta}_0} = -2 \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\frac{\partial \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\partial \hat{\beta}_1} = -2 \sum X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

Dividiendo la penúltima ecuación entre dos y sumando individualmente cada término, se obtiene:

$$-\sum Y_i + \sum \hat{\beta}_0 + \sum \hat{\beta}_1 X_i = 0$$

Pasamos $\sum Y_i$ al otro lado del signo igual y sustituimos $\sum \hat{\beta}_0 = n\hat{\beta}_0$ para obtener:

$$n\hat{\beta}_0 + \left(\sum X_i\right)\hat{\beta}_1 = \sum Y_i$$

Multiplicando ambos términos por X_i tenemos la siguiente expresión:

$$\left(\sum X_i\right)\hat{\beta}_0 + \left(\sum X_i^2\right)\hat{\beta}_1 = \sum X_i Y_i$$

Las últimas dos expresiones se conocen como ecuaciones normales. De la primera de ellas podemos despejar $\hat{\beta}_0$ y obtener:

$$\hat{\beta}_0 = \frac{\sum Y_i}{n} - \hat{\beta}_1 \frac{\sum X_i}{n}$$

Reemplazando esta expresión, en la segunda ecuación normal, tenemos:

$$\frac{\sum X_i \sum Y_i}{n} - \hat{\beta}_1 \frac{(\sum X_i)^2}{n} + \left(\sum X_i^2\right)\hat{\beta}_1 = \sum X_i Y_i$$

Podemos acomodar esta última expresión y obtener:

$$\hat{\beta}_1 = \frac{\sum X_i Y_i - (\sum X_i \sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2/n} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2}$$

Como $\bar{Y} = \sum Y_i/n$ y $\bar{X} = \sum X_i/n$, podemos inferir que:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Las últimas dos expresiones son las fórmulas que propusimos en la sección 2.

Apéndice 2: Código para diagramas de caja de errores relativos

```
##### Box Plots of Errors
#Creating the variables
df.capstone <- data.frame(table$Distance, table$CPL)
colnames(df.capstone) <- c("Distance","CPL")
quantile.regression <- rq(CPL~Distance, tau=0.25)
quantile.regression.coefficients <- data.frame(quantile.regression$coefficients)
#Ranges of distances
dist.250.500 <- subset(df.capstone, Distance>=250 & Distance<500); dist.500.750 <- subset(df.capstone, Distance>=500
& Distance<750)
dist.750.1000 <- subset(df.capstone, Distance>=750 & Distance<1000); dist.1000.1250 <- subset(df.capstone,
Distance>=1000 & Distance<1250)
dist.1250.1500 <- subset(df.capstone, Distance>=1250 & Distance<1500); dist.1500.1750 <- subset(df.capstone,
Distance>=1500 & Distance<1750)
dist.1750.2000 <- subset(df.capstone, Distance>=1750 & Distance<2000); dist.2000.2250 <- subset(df.capstone,
Distance>=2000 & Distance<2250)
dist.2250.2500 <- subset(df.capstone, Distance>=2250 & Distance<2500); dist.2500.2750 <- subset(df.capstone,
Distance>=2500 & Distance<2750)
dist.2750.3000 <- subset(df.capstone, Distance>=2750 & Distance<3000); dist.3000.3250 <- subset(df.capstone,
Distance>=3000 & Distance<3250)
#Ranges of OLS
ols.250.500 <- data.frame(ols$coefficients[[1]]+ols$coefficients[[2]]*dist.250.500$Distance); colnames(ols.250.500) <-
"CPL"; ols.500.750 <- data.frame(ols$coefficients[[1]]+ols$coefficients[[2]]*dist.500.750$Distance);
colnames(ols.500.750) <- "CPL"
ols.750.1000 <- data.frame(ols$coefficients[[1]]+ols$coefficients[[2]]*dist.750.1000$Distance); colnames(ols.750.1000)
<- "CPL" ; ols.1000.1250 <- data.frame(ols$coefficients[[1]]+ols$coefficients[[2]]*dist.1000.1250$Distance);
colnames(ols.1000.1250) <- "CPL"
ols.1250.1500 <- data.frame(ols$coefficients[[1]]+ols$coefficients[[2]]*dist.1250.1500$Distance);
colnames(ols.1250.1500) <- "CPL"; ols.1500.1750 <-
data.frame(ols$coefficients[[1]]+ols$coefficients[[2]]*dist.1500.1750$Distance); colnames(ols.1500.1750) <- "CPL"
ols.1750.2000 <- data.frame(ols$coefficients[[1]]+ols$coefficients[[2]]*dist.1750.2000$Distance);
colnames(ols.1750.2000) <- "CPL"; ols.2000.2250 <-
data.frame(ols$coefficients[[1]]+ols$coefficients[[2]]*dist.2000.2250$Distance); colnames(ols.2000.2250) <- "CPL"
```

```
ols.2250.2500 <- data.frame(ols$coefficients[[1]]+ols$coefficients[[2]]*dist.2250.2500$Distance);
colnames(ols.2250.2500) <- "CPL"; ols.2500.2750 <-
data.frame(ols$coefficients[[1]]+ols$coefficients[[2]]*dist.2500.2750$Distance); colnames(ols.2500.2750) <- "CPL"
ols.2750.3000 <- data.frame(ols$coefficients[[1]]+ols$coefficients[[2]]*dist.2750.3000$Distance);
colnames(ols.2750.3000) <- "CPL"; ols.3000.3250 <-
data.frame(ols$coefficients[[1]]+ols$coefficients[[2]]*dist.3000.3250$Distance); colnames(ols.3000.3250) <- "CPL"
#Ranges of Quantile Regression
rq.250.500 <-
data.frame(quantile.regression$coefficients[[1]]+quantile.regression$coefficients[[2]]*dist.250.500$Distance);
colnames(rq.250.500) <- "CPL"; rq.500.750 <-
data.frame(quantile.regression$coefficients[[1]]+quantile.regression$coefficients[[2]]*dist.500.750$Distance);
colnames(rq.500.750) <- "CPL"
rq.750.1000 <-
data.frame(quantile.regression$coefficients[[1]]+quantile.regression$coefficients[[2]]*dist.750.1000$Distance);
colnames(ols.750.1000) <- "CPL"; rq.1000.1250 <-
data.frame(quantile.regression$coefficients[[1]]+quantile.regression$coefficients[[2]]*dist.1000.1250$Distance);
colnames(rq.1000.1250) <- "CPL"
rq.1250.1500 <-
data.frame(quantile.regression$coefficients[[1]]+quantile.regression$coefficients[[2]]*dist.1250.1500$Distance);
colnames(rq.1250.1500) <- "CPL"; rq.1500.1750 <-
data.frame(quantile.regression$coefficients[[1]]+quantile.regression$coefficients[[2]]*dist.1500.1750$Distance);
colnames(rq.1500.1750) <- "CPL"
rq.1750.2000 <-
data.frame(quantile.regression$coefficients[[1]]+quantile.regression$coefficients[[2]]*dist.1750.2000$Distance);
colnames(rq.1750.2000) <- "CPL"; rq.2000.2250 <-
data.frame(quantile.regression$coefficients[[1]]+quantile.regression$coefficients[[2]]*dist.2000.2250$Distance);
colnames(rq.2000.2250) <- "CPL"
rq.2250.2500 <-
data.frame(quantile.regression$coefficients[[1]]+quantile.regression$coefficients[[2]]*dist.2250.2500$Distance);
colnames(rq.2250.2500) <- "CPL"; rq.2500.2750 <-
data.frame(quantile.regression$coefficients[[1]]+quantile.regression$coefficients[[2]]*dist.2500.2750$Distance);
colnames(rq.2500.2750) <- "CPL"
rq.2750.3000 <-
data.frame(quantile.regression$coefficients[[1]]+quantile.regression$coefficients[[2]]*dist.2750.3000$Distance);
colnames(rq.2750.3000) <- "CPL"; rq.3000.3250 <-
data.frame(quantile.regression$coefficients[[1]]+quantile.regression$coefficients[[2]]*dist.3000.3250$Distance);
colnames(rq.3000.3250) <- "CPL"
```

```
#Errors for OLS
error.ols.250.500 <- (df.capstone$CPL-ols.250.500)/df.capstone$CPL; colnames(error.ols.250.500) <- "Error";
error.ols.500.750 <- (df.capstone$CPL-ols.500.750)/df.capstone$CPL; colnames(error.ols.500.750) <- "Error"
error.ols.750.1000 <- (df.capstone$CPL-ols.750.1000)/df.capstone$CPL; colnames(error.ols.750.1000) <- "Error";
error.ols.1000.1250 <- (df.capstone$CPL-ols.1000.1250)/df.capstone$CPL; colnames(error.ols.1000.1250) <- "Error"
error.ols.1250.1500 <- (df.capstone$CPL-ols.1250.1500)/df.capstone$CPL; colnames(error.ols.1250.1500) <- "Error";
error.ols.1500.1750 <- (df.capstone$CPL-ols.1500.1750)/df.capstone$CPL; colnames(error.ols.1500.1750) <- "Error"
error.ols.1750.2000 <- (df.capstone$CPL-ols.1750.2000)/df.capstone$CPL; colnames(error.ols.1750.2000) <- "Error";
error.ols.2000.2250 <- (df.capstone$CPL-ols.2000.2250)/df.capstone$CPL; colnames(error.ols.2000.2250) <- "Error"
error.ols.2250.2500 <- (df.capstone$CPL-ols.2250.2500)/df.capstone$CPL; colnames(error.ols.2250.2500) <- "Error";
error.ols.2500.2750 <- (df.capstone$CPL-ols.2500.2750)/df.capstone$CPL; colnames(error.ols.2500.2750) <- "Error"
error.ols.2750.3000 <- (df.capstone$CPL-ols.2750.3000)/df.capstone$CPL; colnames(error.ols.2750.3000) <- "Error";
error.ols.3000.3250 <- (df.capstone$CPL-ols.3000.3250)/df.capstone$CPL; colnames(error.ols.3000.3250) <- "Error"
#Errors for Quantile Regression
error.rq.250.500 <- (df.capstone$CPL-rq.250.500)/df.capstone$CPL;colnames(error.rq.250.500) <- "Error";
error.rq.500.750 <- (df.capstone$CPL-rq.500.750)/df.capstone$CPL;colnames(error.rq.500.750) <- "Error"
error.rq.750.1000 <- (df.capstone$CPL-rq.750.1000)/df.capstone$CPL;colnames(error.rq.750.1000) <- "Error";
error.rq.1000.1250 <- (df.capstone$CPL-rq.1000.1250)/df.capstone$CPL;colnames(error.rq.1000.1250) <- "Error"
error.rq.1250.1500 <- (df.capstone$CPL-rq.1250.1500)/df.capstone$CPL;colnames(error.rq.1250.1500) <- "Error";
error.rq.1500.1750 <- (df.capstone$CPL-rq.1500.1750)/df.capstone$CPL;colnames(error.rq.1500.1750) <- "Error"
error.rq.1750.2000 <- (df.capstone$CPL-rq.1750.2000)/df.capstone$CPL;colnames(error.rq.1750.2000) <- "Error";
error.rq.2000.2250 <- (df.capstone$CPL-rq.2000.2250)/df.capstone$CPL;colnames(error.rq.2000.2250) <- "Error"
error.rq.2250.2500 <- (df.capstone$CPL-rq.2250.2500)/df.capstone$CPL;colnames(error.rq.2250.2500) <- "Error";
error.rq.2500.2750 <- (df.capstone$CPL-rq.2500.2750)/df.capstone$CPL;colnames(error.rq.2500.2750) <- "Error"
error.rq.2750.3000 <- (df.capstone$CPL-rq.2750.3000)/df.capstone$CPL;colnames(error.rq.2750.3000) <- "Error";
error.rq.3000.3250 <- (df.capstone$CPL-rq.3000.3250)/df.capstone$CPL;colnames(error.rq.3000.3250) <- "Error"
#Data frames for both regression methods
df.error.ols.summary <-
c(error.ols.250.500,error.ols.500.750,error.ols.750.1000,error.ols.1000.1250,error.ols.1250.1500,error.ols.1500.1750,err
or.ols.1750.2000,error.ols.2000.2250,error.ols.2250.2500,error.ols.2500.2750,error.ols.2750.3000,error.ols.3000.3250)
df.error.rq.summary <- c(error.rq.250.500,error.rq.500.750,
        error.rq.750.1000,error.rq.1000.1250,error.rq.1250.1500,error.rq.1500.1750,error.rq.1750.2000,error.rq.2000.
2250,error.rq.2250.2500,error.rq.2500.2750,error.rq.2750.3000,error.rq.3000.3250)
#Summary data frame
df.error.capstone.summary <-
c(error.ols.250.500,error.rq.250.500,error.ols.500.750,error.rq.500.750,error.ols.750.1000,error.rq.750.1000,error.ols.10
00.1250,error.rq.1000.1250,error.ols.1250.1500,error.rq.1250.1500,error.ols.1500.1750,error.rq.1500.1750,error.ols.175
```

```
0.2000,error.rq.1750.2000,error.ols.2000.2250,error.rq.2000.2250,error.ols.2250.2500,error.rq.2250.2500,error.ols.2500
.2750,error.rq.2500.2750,error.ols.2750.3000,error.rq.2750.3000,error.ols.3000.3250,error.rq.3000.3250)
#Box plots (without interleaving)
pdf("OLS vs. Quantile Regression.jpg", width=20, height=10)
par(mar=c(7,5,5,5))
boxplot(c(df.error.ols.summary,df.error.rq.summary), names=c("250-500","500-750","750-1000","1000-1250","1250-
1500","1500-1750","1750-2000","2000-2250","2250-2500","2500-2750","2750-3000","3000-3250","250-500","500-
750","750-1000","1000-1250","1250-1500","1500-1750","1750-2000","2000-2250","2250-2500","2500-2750","2750-
3000","3000-3250"),
col=c("red","red","red","red","red","red","red","red","red","red","red","red","red","blue","blue","blue","blue","blue","blue","bl
ue","blue","blue","blue","blue","blue"), main="OLS vs Quantile Regression", ylab="Relative Error", ylim=c(-17,1), las=3)
legend("bottomleft", legend=c("OLS", "Quantile Regression"), lty=c(1,1), col=c("red","blue"), bty="n")
abline(h=1:-20, lwd=1, lty=2, col="gray")
dev.off()
#Box plots (with interleaving)
pdf("OLS vs. Quantile Regression.pdf", width=20, height=10)
par(mar=c(7,5,5,5))
boxplot(df.error.capstone.summary, names=c("250-500","250-500","500-750","500-750","750-1000","750-1000","1000-
1250","1000-1250","1250-1500","1250-1500","1500-1750","1500-1750","1750-2000","1750-2000","2000-2250","2000-
2250","2250-2500","2250-2500","2500-2750","2500-2750","2750-3000","2750-3000","3000-3250","3000-3250"),
col=c("red","blue","red","blue","red","blue","red","blue","red","blue","red","blue","red","blue","red","blue","red","blue","red","blue",
"red","blue","red","blue","red","blue"), main="OLS vs Quantile Regression", ylab="Relative Error", ylim=c(-17,1), las=3)
legend("bottomleft", legend=c("OLS", "Quantile Regression"), lty=c(1,1), col=c("red","blue"), bty="n")
abline(h=1:-20, lwd=1, lty=2, col="gray")
dev.off()
```

Índice de Palabras