**Pronóstico de ventas para Walmart**

Modelos Lineales Generalizados

13 de diciembre de 2017

Ricardo Lastra 160167

Federico Riveroll 105898

Daniel Sharp 138176

**Pronóstico de ventas para Walmart**

1. Introducción

En 2013, Walmart abrió una competencia en Kaggle que consistía en pronosticar las ventas de 45 de sus tiendas en Estados Unidos. Los datos consisten en variables macroeconómicas y ambientales para cada una de las tiendas, junto con los descuentos que ellos han aplicado en las tiendas. Los datos presentados en la competencia abarcan desde la semana del 5 de febrero del 2010 hasta la semana del 1 de noviembre de 2012. Las ventas están agrupadas en datos semanales.

Para este análisis, se modificaron los datos con el objetivo de simplificarlos para su uso en análisis bayesiano. Esta simplificación consistió en tomar una muestra de 10 tiendas únicamente para desarrollar el análisis. Además, se agruparon las ventas de los departamentos en cada tienda, por lo cual nuestro análisis consistirá en la predicción de las ventas de cada tienda. Como variables adicionales de las tiendas tenemos Tipo y Tamaño, donde la primera hace referencia al tamaño y número de servicios que se ofrecen en dicha tienda y la segunda es el área de la tienda en pies cuadrados. Las tiendas del tipo A tienen un área promedio de 187,000 pies cuadrados, las del tipo B de 104,000 pies cuadrados y finalmente, las tiendas del tipo C, que se implementaron en 2010, tienen un área promedio de 42,000 pies cuadrados. Adicionalmente realizamos un estudio sobre las diferencias en el tamaño (tipo) de tienda, para poder inferir el rendimiento por metro cuadrado ya que consideramos un parámetro importante desde el punto de vista del cliente. Como se mencionó, también hay variables macroeconómicas y ambientales, tales como; temperatura, el precio de la gasolina, el índice de precios del consumidor, la tasa de desempleo y una variable binaria que define si hubo día feriado en esa semana o no. Esta última variable nos hace pensar en el supuesto de que los días feriados impactan en las ventas de sus sucursales, por lo que habrá que contemplar esto en el análisis.

1. Objetivos

* Determinar si hay diferencia significativa en el comportamiento de las ventas de las diferentes sucursales. Es decir, si hay sucursales que difieran en rentabilidad (por rentabilidad tomamos ventas por metro cuadrado).
* Determinar la relación que tienen las covariables sobre las ventas de las tiendas. Es importante diferenciar entre las variables exógenas (ajenas al negocio) y las endógenas (controlables por el negocio).
* Desarrollar un modelo para realizar predicciones de las ventas en las 4 semanas siguientes y evaluar su desempeño.

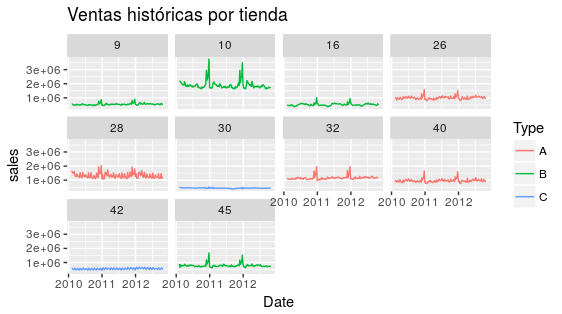
1. Análisis descriptivo de la información

Como se mencionó en la introducción, los datos incluyen tanto variables endógenas de la empresa y exógenas. La lista completa de las variables y su análisis descriptivo es la siguiente:

*Tabla 1: Análisis descriptivo*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Tipo** | **Media** | **Desviación Estándar** | **Mínimo** | **Máximo** |
| Número de Tienda | Numérica | 27.8 | 12.17 | 9 | 45 |
| Fecha | Fecha o ‘Timestamp’ | N/A | N/A | 05-02-2010 | 26-10-2012 |
| Temperatura | Numérica | 59.83 | 19.52 | 5.54 | 99.22 |
| Precio de Gasolina | Numérica | 3.398 | 0.46 | 2.514 | 4.468 |
| Descuento 1-5 | Numérica | 1,054.08 | 4,172.85 | 0 | 141,630.6 |
| Índice de Precios al Consumidor | Numérica | 166.5 | 36.7 | 126.1 | 227.2 |
| Tasa de Desempleo | Numérica | 7.995 | 2.19 | 4.125 | 14.313 |
| Día Feriado | Booleana | 0.06993 | 0.255 | 0 | 1 |
| Ventas | Numérica | 920,075 | 455,806.8 | 368,600 | 3,749,058 |
| Tipo | Categórica | 1.8 | 0.75 | 1 | 3 |
| Tamaño (pies cuadrados) | Numérica | 122,735 | 57,427.13 | 34,690 | 206,302 |

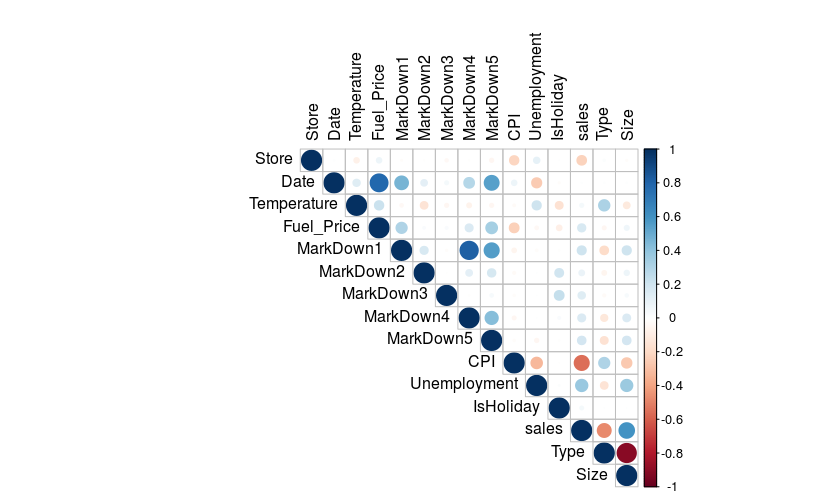
A continuación se muestra un análisis exploratorio de las variables en el dataset. El objetivo es obtener una perspectiva de que tipo de relaciones hay entre ellas. El primer análisis es de las ventas contra el tiempo, para identificar tendencia o estacionalidad en los datos para las 10 tiendas diferentes objeto de análisis.



*Figura 1: Ventas históricas por tienda*

En la Figura 1 se puede observar que las tiendas tienen variaciones en los montos de venta, donde la tienda 10 tiene los montos de venta más altos. Por otro lado, las tiendas 9, 16, 30 y 42 tienen los menores montos de ventas. Se puede ver que hay picos en las ventas para todas las tiendas en las mismas fechas. Analizando los datos encontramos que los picos fuertes en la ventas son en la semana navideña y en la semana en que se lleva a cabo ‘Black Friday’. La intensidad del efecto de estas fechas en las ventas depende de cada tienda. El tipo de tienda no parece tener un efecto claro en las ventas. Únicamente parece que las tiendas tipo 3 son las de menores ventas, que tiene sentido puesto que son las tiendas más pequeñas. Sin embargo, hay más variabilidad en las tiendas tipo 1 y 2.

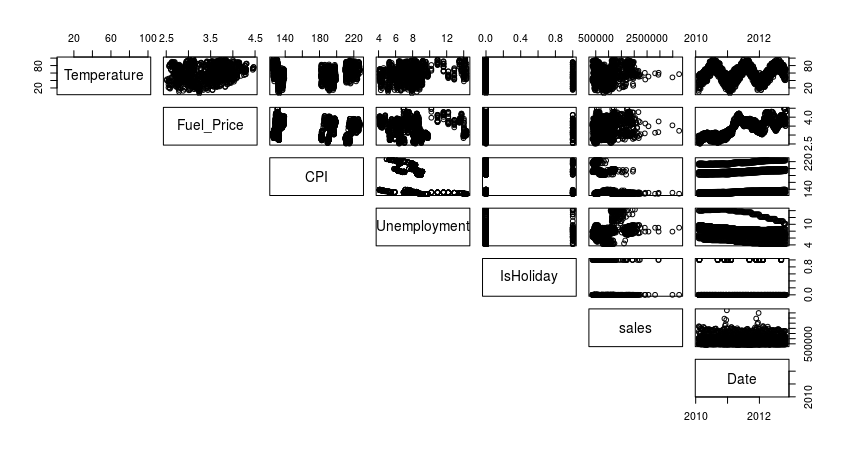
Para realizar una primera exploración a la interacción de las variables, graficamos un correlograma de nuestros datos. Para hacer esto se transformó la variable de fecha a numérica, con valores del 1 al 143.



*Figura 2: Correlograma*

Después de observar la Figura 2 podemos decir que las principales correlaciones positivas son entre la fecha y el precio de la gasolina; la fecha y los descuentos 1 y 5. Hay también fuerte correlación entre el descuento 1 y 4 y 5. Las ventas están altamente correlacionadas con el tamaño de la tienda, lo cual también es de esperar. Por otro lado, se puede observar que las principales correlaciones negativas son entre la fecha y la tasa de desempleo, y entre las ventas y el índice de precios al consumidor. Hay correlación negativa entre el tipo de tienda y el tamaño, pero esto sucede únicamente porque las tiendas más pequeñas son del tipo 3 y las más grandes del tipo 1. Es un fenómeno por como están construidas las categorías.

En la siguiente gráfica se pueden observar las variables macroeconómicas y ambientales, su interacción entre ellas y ante las ventas.



*Figura 3: Exploración de variables exógenas*

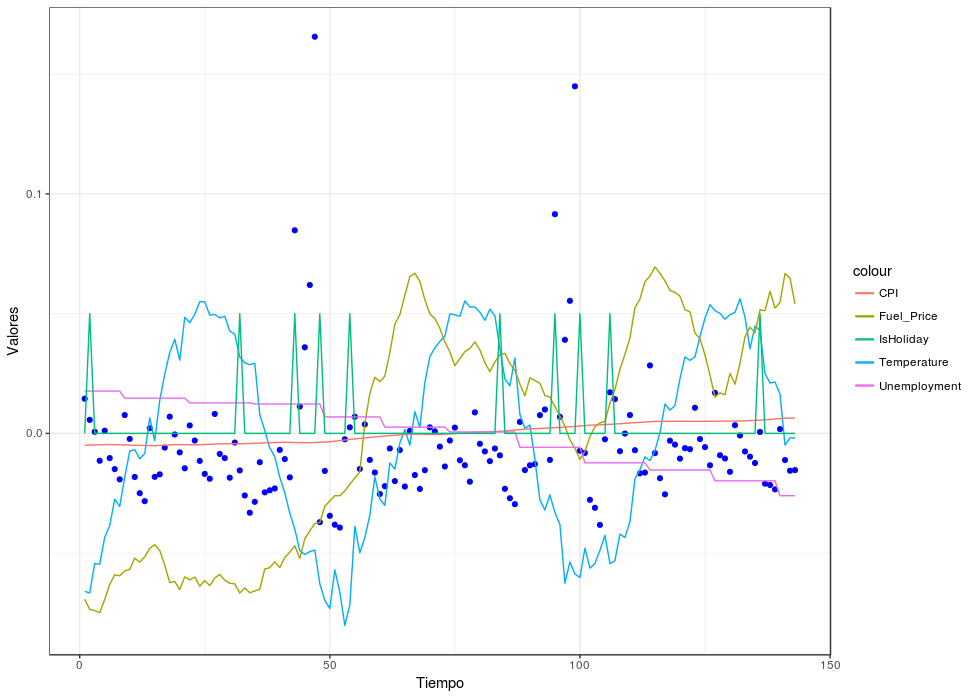
De la Figura 3 también se aprecia que las variables tienen alta estacionalidad, esto es visible en todas las variables mostradas. Por otro lado, no parece haber una clara relación entre las covariables y las ventas de las tiendas. A continuación se realiza el mismo análisis para las variables accionables del negocio, es decir, los descuentos aplicados:



*Figura 4: Exploración de variables accionables*

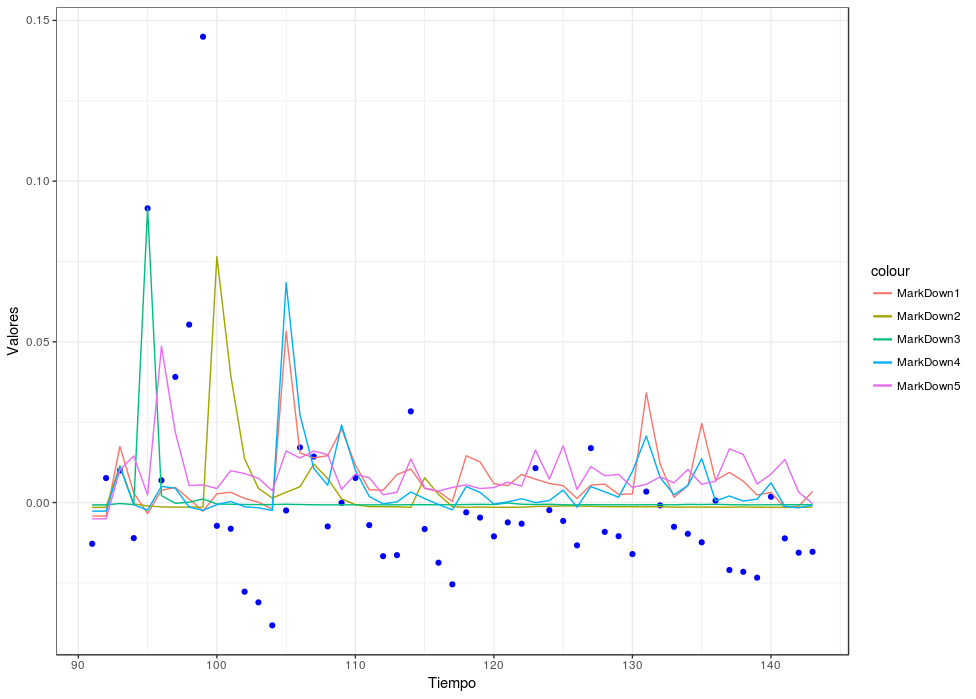
La dinámica de los descuentos se implementó por primera vez en la segunda semana de noviembre del 2011 en todas las tiendas, es por esto que las variables tienen valor de cero antes de esta fecha. El efecto de los descuentos sobre las ventas no es especialmente claro, aunque en algunos casos, como con el MarkDown 3 y el 4 se alcanza a observar una relación casi lineal entre las variables, donde a mayor descuento parece haber un valor más alto para las ventas.

En la Figura 5 podemos ver el diagrama sobre el tiempo de variables exógenas (las que no dependen de la empresa). Las variables están escaladas con el fin de poder observar mejor el efecto y las ventas se muestran en los puntos azules.



*Figura 5: Variables exógenas en el tiempo*

Podemos ver que es complicado tener una referencia visual sobre las variables exógenas, de antemano sabemos que la temperatura (-0.21) y si es día feriado (0.18) son las correlaciones más fuertes de las variables que no dependen del cliente. Ahora veamos las endógenas (los descuentos) en la Figura 6. De igual forma, los datos están escalados.



*Figura 6: Variables endógenas en el tiempo*

Aquí observamos que la variable de mayor relación es “Markdown3”, la cual tiene 0.309 de correlación con las ventas.

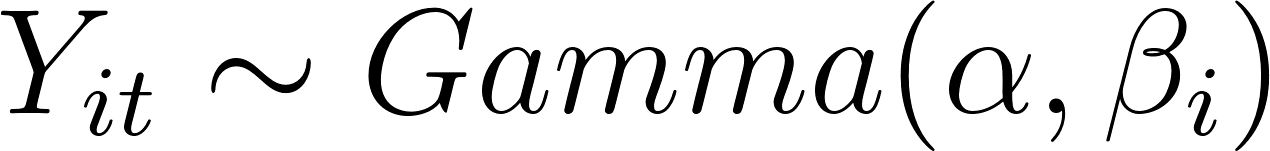
4. Análisis

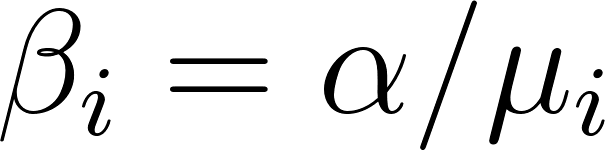
Los diferentes objetivos del análisis requieren también la utilización de diferentes modelos. A continuación se presenta, para cada objetivo, los modelos utilizados, junto con métricas relevantes de comparación entre los modelos.

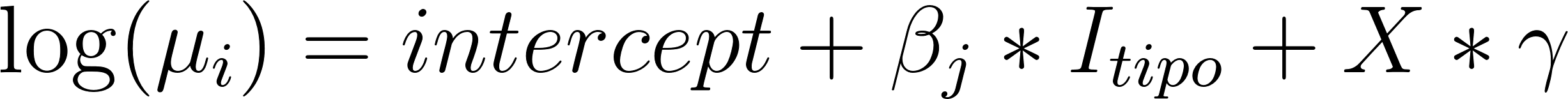
**Evaluación de diferencia en las ventas de las tiendas**

Para evaluar las diferencias en el comportamiento de las ventas de las 10 tiendas elegidas para la muestra se utilizó un modelo de efectos fijos. La idea es identificar si el tipo de tienda tiene efecto sobre el nivel de ventas por pie cuadrado de las tiendas, y en caso de que sí, cómo lo afecta. Es de esperar que las tiendas más grandes tengan mayores ventas, por lo que se utilizará la métrica de Ventas/Tamaño para identificar cómo varían los ingresos por pie cuadrado de tienda según la clasificación que Walmart haya asignado. Para este inciso los tipos de tienda se etiquetaron como 1, 2 y 3 para A, B y C, respectivamente. Se eligió un modelo de efectos fijos pues se espera que, dado que los servicios que ofrece cada tipo de tienda son muy diferentes, estos sean heterogéneos.

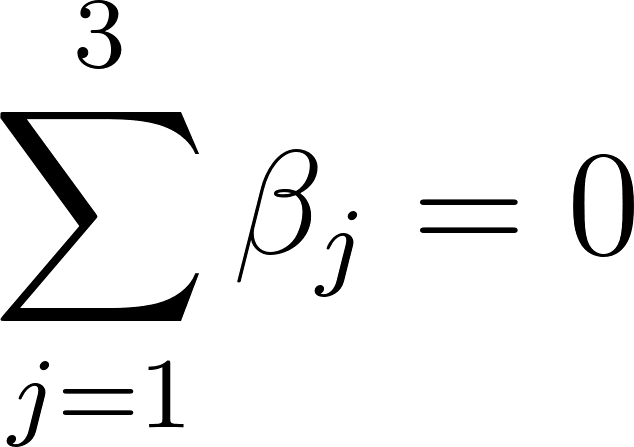
Modelo de efectos fijos:

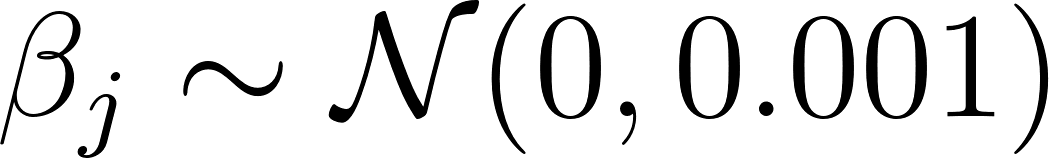
[](about:blank)

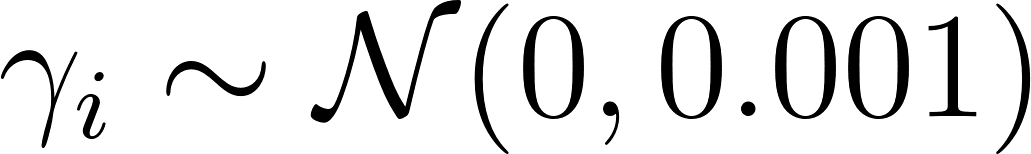
[](about:blank) y

[](about:blank)

Donde:

[](about:blank)

[](about:blank)

[](about:blank)

[](about:blank)

Siendo:

[](about:blank)

Se utilizó una distribución Gamma con función liga logarítmica en un modelo de efectos fijos para los coeficientes betas. Se definió que las betas provienen, cada una, de una distribución normal con media 0 y precisión 0.0001, con el objetivo de que sean vagos. Los coeficientes de interés son las betas. La inclusión de las demás variables es con el objetivo de que ‘absorban’ el efecto que tienen sobre las ventas. De este análisis se encontró que existe diferencia significativa en las ventas por el tipo de tienda, como se muestra en la siguiente gráfica:

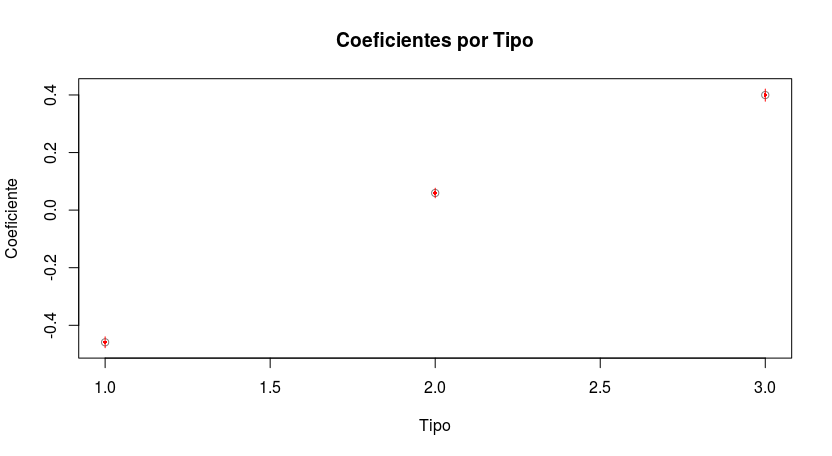
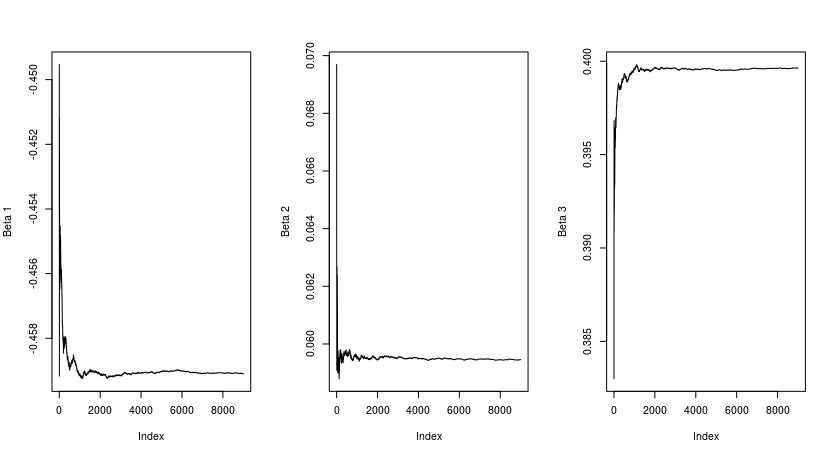


Figura 7: Coeficientes por tipo de tienda.

La siguiente gráfica contiene la convergencia de los coeficientes:



*Figura 8: Convergencia de los coeficientes Beta.*

Con este modelo se alcanzó una correcta convergencia para cada uno de los coeficientes de las tiendas, por lo que se decidió utilizar este modelo para resolver la pregunta.

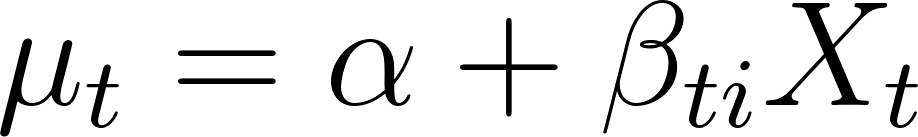
**Predicción de ventas y análisis del efecto de las covariables**

Para el objetivo de predicción se decidió utilizar dos modelos dinámicos para modelar las ventas de las sucursales y generar predicciones por las siguientes 4 semanas. Para esto se eliminaron las últimas cuatro variables respuestas de cada tienda. Estos modelos se muestran a continuación:

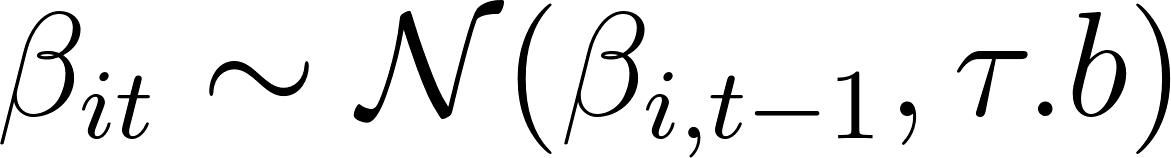
Modelo de predicción 1: Betas dinámicas

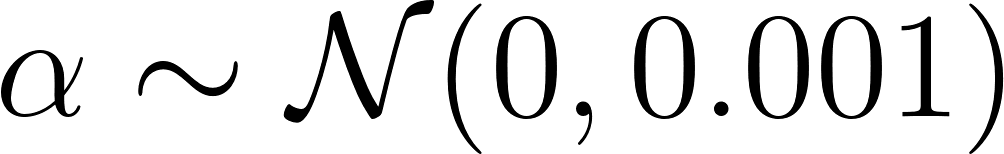
Observación:

[](about:blank)

[](about:blank)

Evolución:

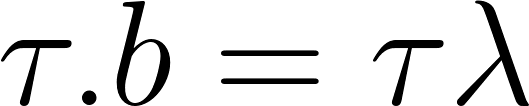
[](about:blank)

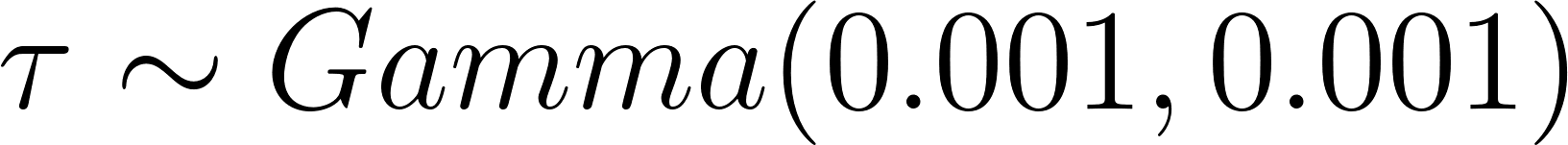
[](about:blank)

Y con:

[](about:blank)

Hiper Parámetros:

[](about:blank)

[](about:blank)

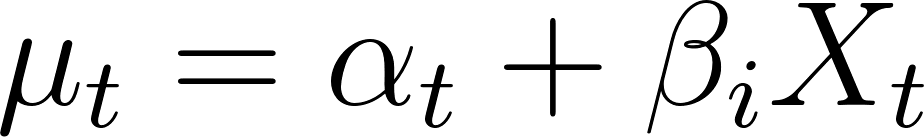
[](about:blank)

Modelo de predicción 2: Alphas dinámicas

(*\*El cambio fué hacer la alpha dinámica en lugar de las betas*)

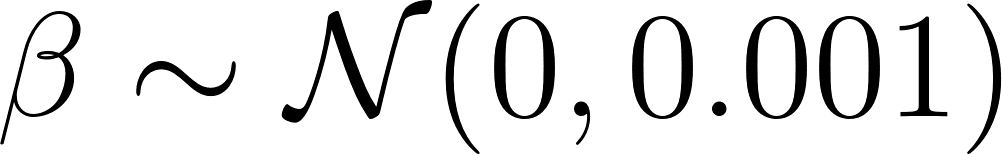
Observación:

[](about:blank)

[](about:blank)

Evolución:

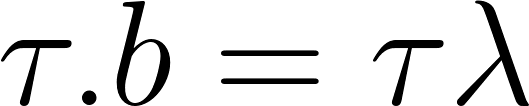
[](about:blank)

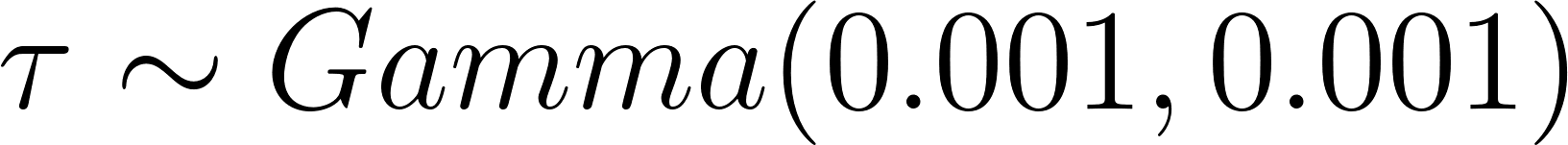
[](about:blank)

Y con:

[](about:blank)

Hiper Parámetros:

[](about:blank)

[](about:blank)

[](about:blank)

En ambos modelos tenemos el hiper parámetro lambda, conocido como el factor de suavizamiento, que controla en qué proporción es más grande la precisión de las distribuciones de nuestros parámetros en relación con la precisión de la verosimilitud. El efecto que tiene este hiper parámetro es que entre más alto es su valor, el modelo ajusta en menor proporción a los datos históricos y tiene intervalos de predicción menores, es decir, se suaviza. Por otro lado, entre más pequeño es lambda, más ajusta a los datos históricos y por lo tanto, mayores son los intervalos de predicción. Ambos modelos se ejecutaron sobre los datos de la tienda 9 (elegida arbitrariamente) y el efecto se muestra para los dos modelos en las gráficas de la figura 9:

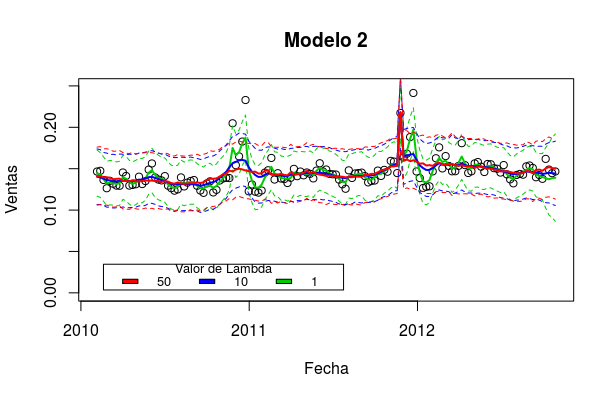
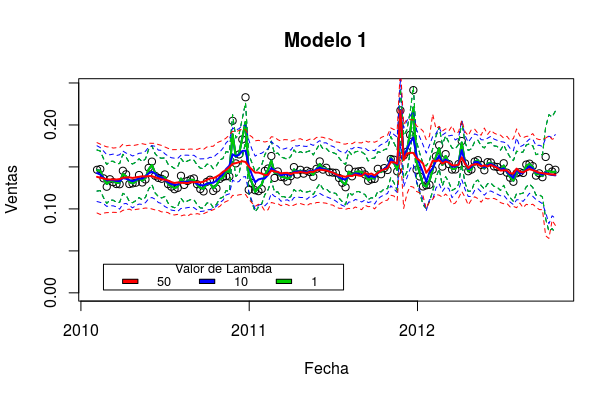


Figura 9: Modelos aplicados a la tienda 9

El efecto del cambio en el valor de lambda es notorio en ambos modelos, donde la lambda igual a 1 tiene intervalos menores en los datos observados pero estos se extienden significativamente para valores no observados. Esto es más notorio en el Modelo 1, que por su configuración, tiene intervalos de predicción más amplios que el Modelo 2. Para poder hacer una comparación de los modelos se utilizaron métricas como el DIC y las pseudo-R2. Estos se muestran para cada modelo y para cada tienda, con cada valor de lambda, como se muestra en la figura 10:

DIC

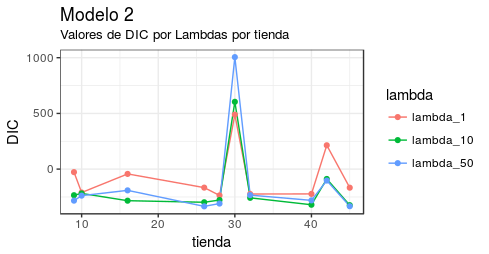
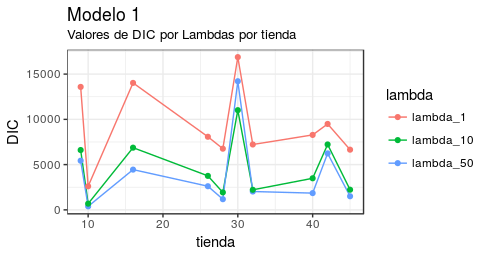


Figura 10: DIC en tiendas con diferentes lambdas

De las gráficas anteriores podemos ver que el Modelo 1 obtuvo valores positivos de DIC para la mayoría de las tiendas, mientras que por otro lado, el Modelo 2 obtuvo DICs negativas para casi todas las tiendas. Esto se explica por la configuración del modelo; el Modelo 1 ajusta 1431 coeficientes en la regresión (10 betas por observación - 143 en total - más un alpha), lo cual penaliza el DIC que se calcula. Por otro lado, el Modelo 2 calcula 153 coeficientes (10 betas + 143 alphas), lo que explica los menores valores de DIC. En este sentido es mejor el Modelo 2 pues el Modelo 1, al tener tantos coeficientes, sobre ajusta a los datos, lo cual es perjudicial para el caso de predicción. Los valores de lambda más altos aseguran que ajusten menos a los datos, es por esto que para el Modelo 1, el DIC menor lo tengan los modelos con lambda igual a 50. Por otro lado, en el Modelo 2 sucede algo interesante y diferente a lo que sucede en el Modelo 1. Mientras que en el Modelo 1, los modelos con lambda 1 tuvieron DICs consistentemente mayores que los otros - por el sobre ajuste ya mencionado -, en el Modelo 2 no hubo una tendencia clara donde un valor de lambda tuviera un DIC consistentemente mejor que los otros. Incluso, el valor más consistente a lo largo de los modelos parece haber sido el de lambda igual a 1, que por la construcción del modelo no sobre ajusta demasiado a los datos. Los modelos parecen haber tenido problemas con la tienda 30, pues el valor de DIC aumenta significativamente con respecto a los demás.

R2

Para robustecer el análisis, se exploraron los resultados de los mismos modelos pero en términos de las pseudo-R2 calculadas. Las gráficas de estos se muestran a continuación:

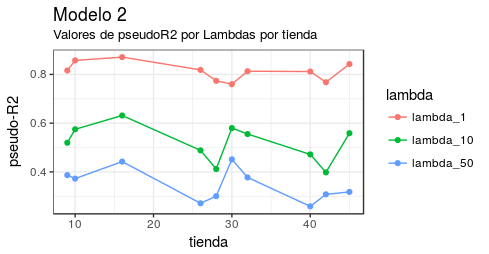
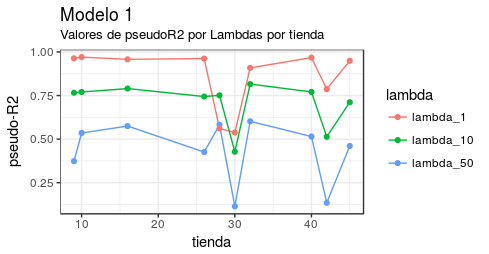


Figura 11: Pseudo-R2 por cada valor lambda definido por tienda.

Soportando el análisis anterior, en las gráficas anteriores se puede observar que los modelos con lambdas más pequeñas obtuvieron R2 mayores, implicando un fuerte ajuste a los datos. De nuevo, el modelo con lambda igual a uno parece ser el más consistente (menos variable) en todas las tiendas, además de que es el que tiene valores mayores de R2.

Por lo mencionado anteriormente, se decidió utilizar el Modelo 2, que no sobre ajusta a los datos y por lo tanto tiene valores DIC significativamente menores a los que produce el Modelo 1. Además, se decidió utilizar con un valor de lambda igual a 10, pues obtuvo valores del DIC relativamente consistentes en las tiendas (con excepción de la tienda 30, que parece un outlier). Además, presenta pseudo-R2 medias, que son menores que lambda 1, pero porque sobre ajusta menos a los datos anteriores; parece ser un modelo estable con buen ajuste de los datos, pero no tal que le impida generar predicciones eficientes.

5. Interpretación de los modelos

**Diferencia en el comportamiento de las tiendas**

Del primer objetivo, que consistía en encontrar si había diferencia entre los diferentes tipos de las tiendas explicados anteriormente, se encontró lo siguiente:

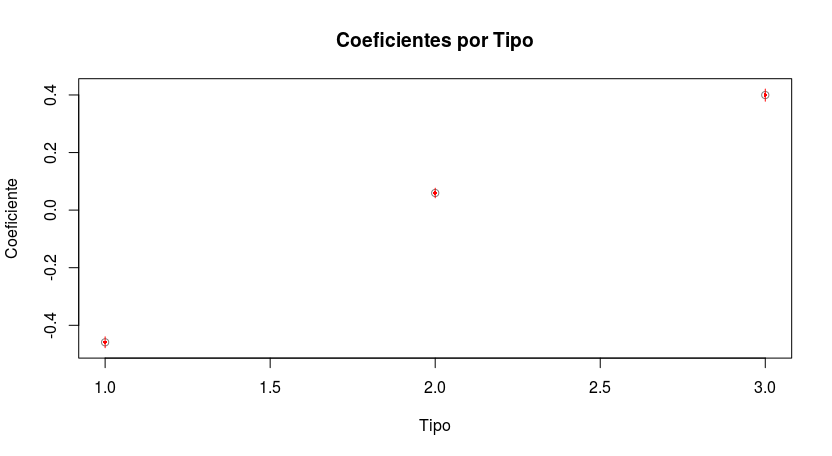


Figura 12: Coeficientes por tipo de tienda

En la imagen anterior se demuestra que hay diferencia significativa entre los tres tipos de tienda. Aplicando la siguiente transformación a los coeficientes podemos encontrar el porcentaje de cambio que tiene cada tienda en sus ventas respecto al caso medio.

Transformación: (e^(Beta) - 1)\*100

Los coeficientes transformados se encuentran en la siguiente tabla:

|  |  |  |
| --- | --- | --- |
| **Beta 1** | **Beta 2** | **Beta 3** |
| -36.7% | 6.1% | 49% |

De estos, podemos observar que las tiendas del tipo 3 tienen ventas cerca del 50% mayores que el caso medio. Las tiendas del tipo 1 tienen los ingresos por pie cuadrado más bajos, con un valor ~35% menor que el caso medio. Finalmente, las tiendas del tipo 2 están muy cercanas al caso medio. Todos los coeficientes resultaron significativos. De este análisis podemos concluir que las tiendas del tipo 3 obtienen los mayores ingresos por pie cuadrado, a pesar de tener las menores ventas. Esto puede ser muy relevante para Walmart pues puede implicar un cambio a su estrategia para la apertura de más tiendas de este tipo.

**Predicción de ventas**

Para realizar las predicciones se utilizó el Modelo 2 con lambda igual a 10, como se definió al comparar los modelos candidatos. Dado el objetivo de que el modelo pudiera predecir las siguientes 4 semanas, eliminamos las últimas 4 observaciones del dataset. El modelo produjo las siguientes predicciones:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Figura 13: Predicción de las 10 tiendas seleccionadas de las ventas de las próximas 4 semanas

Las predicciones se realizaron sobre las últimas cuatro observaciones, y como se puede observar, quedaron dentro del intervalo de predicción para todas las tiendas. También se puede observar cómo las predicciones obedecen a la tendencia general de las ventas de la tienda en particular, y en los casos de alta variabilidad en los datos observados, también generan intervalos de predicción más altos.

La interpretación de los coeficientes de regresión se muestran a continuación:

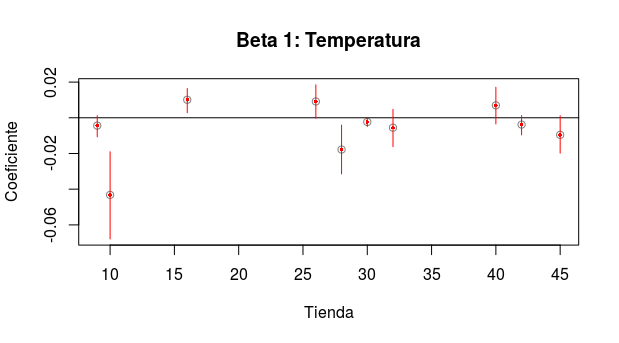


Figura 14: Coeficientes de temperatura por tienda.

En la figura 14 se observa que los coeficientes de temperatura fueron no significativos en algunas de las tiendas (9, 32, 40, 42 y 45). En las que sí fue significativo hay diferencias en que fue positivo para algunas tiendas y negativo para otras. Una hipótesis es que la variabilidad de este coeficiente posiblemente dependa altamente en la región de la tienda, donde regiones con temperaturas más extremas posiblemente si afecten las ventas. Por otro lado, tiendas en regiones más templadas se ven menos afectadas.

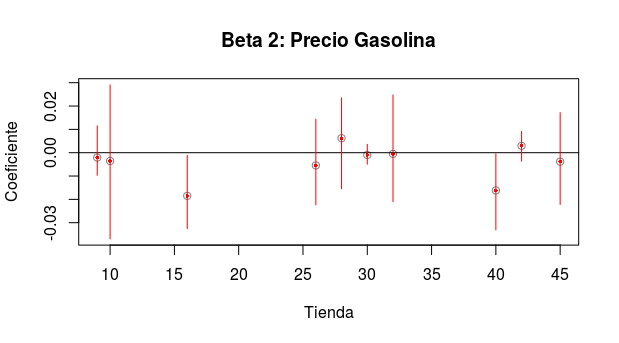


Figura 15: Coeficientes del precio de gasolina por tienda.

Vemos que este coeficiente es no significativo para la mayoría de los casos. Únicamente es significativo en la tienda 16 y en la 40 y tiene coeficiente negativo, lo cual implica una relación negativa, es decir, que a mayores precios de gasolina menores son las ventas.

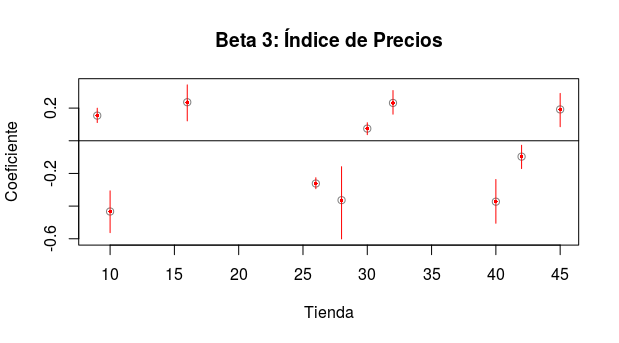


Figura 16: Coeficientes de CPI por tienda.

El coeficiente de CPI fue el que más efecto tuvo en las ventas de las tiendas, tal como se observa en la Figura 16, pues es significativo en ocho tiendas de las 10 seleccionadas. Interesantemente, no tuvo efectos consistentes en todas las tiendas. En cinco tiene coeficiente positivo y en las otras cinco es negativo. De nuevo, esto se puede deber a diferencias regionales, donde el CPI varíe más o menos intensamente. También podría ser consecuencia del nivel socioeconómico que atienden las tiendas, donde las tiendas de mayor nivel socioeconómico tienen una demanda menos elástica y por lo tanto el volumen de compras no varía a pesar de incrementos en el precio, que serían las tiendas con coeficiente positivo. Las tiendas con coeficiente negativo podrían ser las que atienden a niveles socioeconómicos más bajos, con mayor elasticidad, y que se ven más perjudicados por los incrementos por inflación.

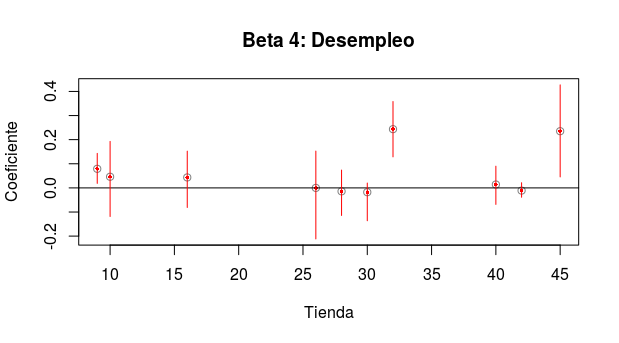


Figura 16: Coeficientes del Desempleo por tienda

De la Figura 16 podemos observar que el coeficiente para la tasa de desempleo es positivo para tres de las diez tiendas, lo que implica que a mayor desempleo, mayores las ventas, pues tienen una relación positiva. Este es interesante pues es poco intuitivo, uno esperaría quizás el resultado opuesto. Una hipótesis sobre este comportamiento es que cuando hay un mayor número de personas desempleadas, tienden a tener más tiempo libre, y por lo tanto, algunas lo utilizan para ir de compras. Esta podía ser la explicación para el fenómeno en las tiendas 32 y 45.

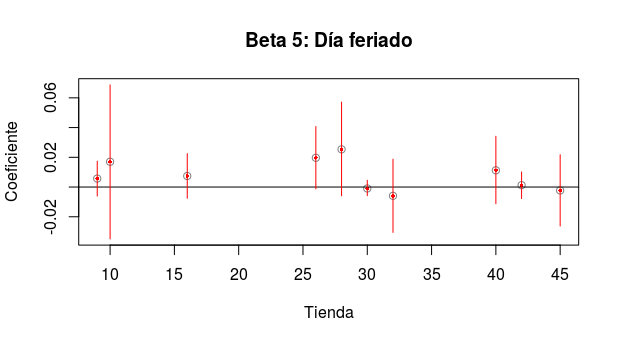


Figura 17: Coeficiente de día feriado por tienda

En la figura anterior observamos que no hay coeficientes significativos para la variable de día feriado, por lo que podemos decir que no hay evidencia que demuestre que el día feriado tiene relación positiva o negativa con las ventas para alguna de las tiendas.

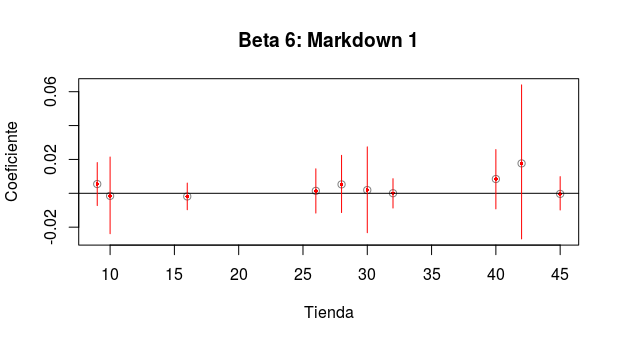


Figura 18: Coeficientes de Markdown 1 por tienda.

La variable Markdown 1, que es el ‘primer’ descuento aplicado a los productos en las tiendas parece no tener relación significativa con las ventas ya que los coeficientes pasan por el 0 y sus intervalos son muy cercanos a este.

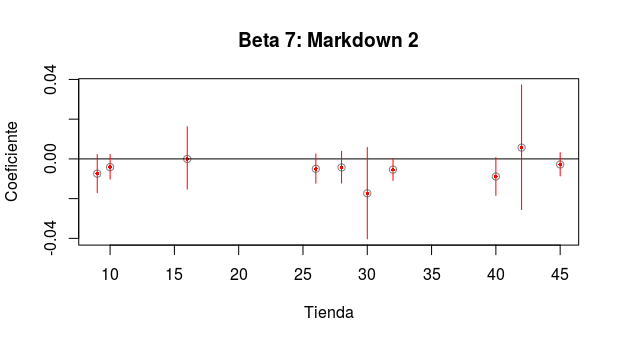


Figura 19: Coeficientes de Markdown 2 por tienda.

Para este coeficiente sucede el mismo fenómeno que en el anterior. El segundo descuento aplicado a los productos no parece tener relación con las ventas de las tiendas. No hay mucha diferencia entre las tiendas excepto porque la tienda 42 tiene una alta variabilidad en sus resultados con respecto a las demás.

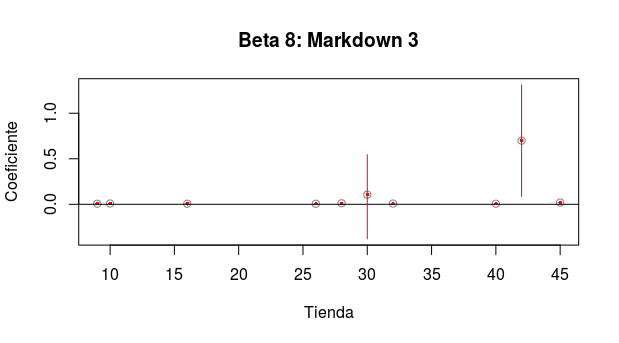


Figura 20: Coeficientes de Markdown 3 por tienda.

Como se aprecia en la figura 20, el tercer descuento aplicado a los productos únicamente obtuvo efecto significativo en la tienda 42, donde a mayor descuento se obtuvieron mayor ventas. Los demás coeficientes están sobre la línea del 0, con un efecto prácticamente nulo.

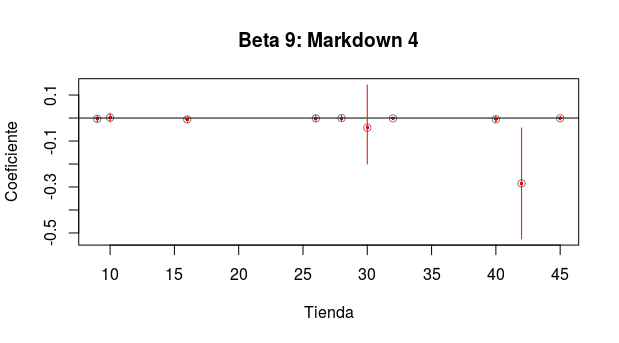


Figura 21: Coeficientes de Markdown 4 por tienda.

El descuento 4 no parece tener relación alguna con las ventas excepto por la tienda 42, esto se podría deber a que los descuentos llegan a un punto tan alto que, aunque quizás se vendan más unidades, el bajo precio perjudica las ventas totales. Sin embargo, es un punto que se debería de analizar más a fondo con mayor información.

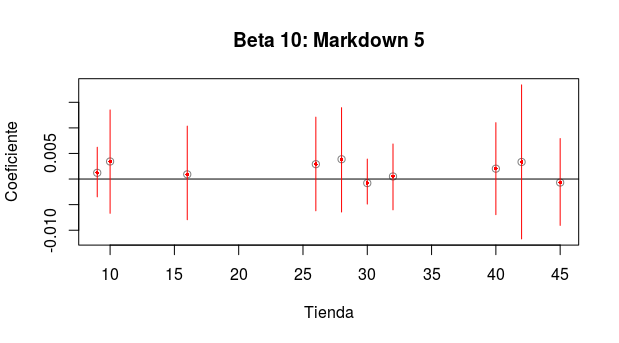


Figura 22: Coeficientes de Markdown 5 por tienda.

El coeficiente Beta 5, referente al quinto descuento aplicado a los productos también resultó ser no significativo para las diez tiendas de estudio al igual que la variable Markdown 1, ya que los coeficientes pasan por el 0 y sus intervalos son muy cercanos a este.

Predicciones para las últimas cuatro semanas con el modelo elegido

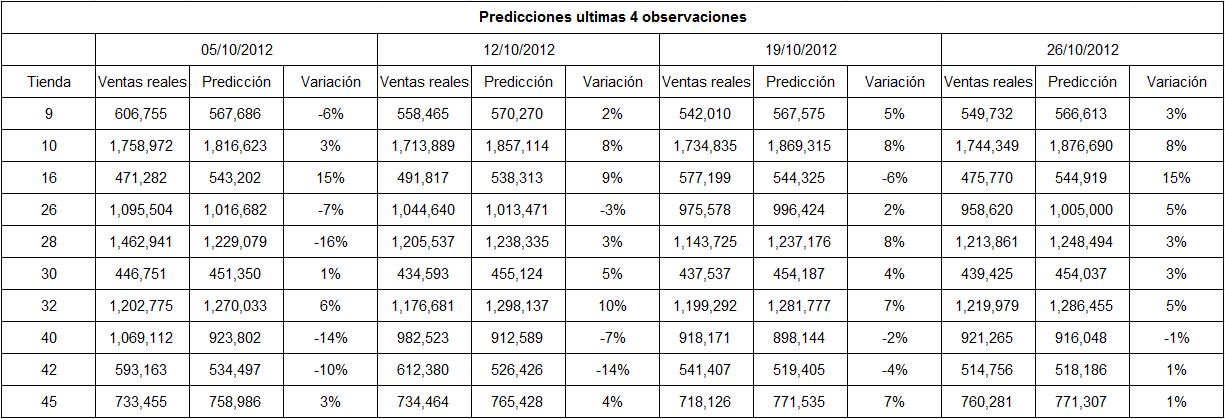


Figura 23: Predicciones puntuales ante valor real

Tomando el valor puntual de las predicciones y comparándolo contra el valor real de la muestra, se observa un promedio del 2% de variación de forma general, también se observa que las predicciones en un tiempo mayor tienden a ser por encima del valor real y en pocas ocasiones por debajo del mismo. Cumpliendo con uno de los objetivos principales, entonces tenemos un modelo seleccionado **“**Modelo 2 con lambda igual a 10**”** el cual nos permite generar muy buenas predicciones con un desempeño destacable debido a la baja variación de los valores reales comparados contra las predicciones.

5. Conclusiones

Del análisis se encontró una diferencia significativa entre las rentabilidades (definida como las ventas/pies cuadrados, dado que no tenemos datos de costos) por tipo de tienda, donde las tiendas más pequeñas - tipo C - resultaron tener una rentabilidad significativamente mayor a las otras, a pesar de tener ventas menores. Para este punto sería interesante hacer un análisis más profundo de las causas para este fenómeno y si podrían extenderse algunas características a los otros tipos de tienda. De igual manera, es interesante que esta iniciativa relativamente reciente parece haber sido efectiva. Para esto sugerimos que se tome la siguiente decisión: que se implemente el desarrollo estratégico de tiendas tipo C a través de la apertura de por lo menos 1 tienda mas por cada 10 de las existentes del tipo A y B, con la finalidad de incrementar la proporción de tiendas, pudiendo definir este plan de forma anual durante los próximos 3 años, dentro de esta estrategia sugerimos hacer evaluaciones anuales del comportamiento de cada tienda, así como del comportamiento de las demás y así evitar un comportamiento negativo. Un factor de alta importancia es que efectivamente las tiendas de tipo C sean más rentables.

Por el lado de las predicciones de ventas, el modelo final produjo estimaciones puntuales para las últimas 4 semanas que estuvieron en un intervalo dentro del ±10% con respecto al valor real y con un promedio de variación global del 2%. La apertura de los intervalos de predicción depende directamente de las ventas históricas de cada tienda. Sin embargo, como se pudo observar en la interpretación de los coeficientes, hay algunos que para ciertas tiendas resultaron no significativos. Esto es explicable por el hecho de que hay particularidades para cada tienda, posiblemente por la región en la que se encuentran, que puede cambiar la relación que tienen sus ventas con las covariables. Esto se demostró en la interpretación del coeficiente del índice de precios al consumidor, que resultó positivo para algunas tiendas y negativo para otras. El principal aprendizaje en este punto es que es difícil diseñar un modelo que pueda ser aplicable a diferentes tiendas y probablemente sería mejor diseñar un modelo para cada tienda, que considere sus características. Otro camino podría ser buscar agrupar a las tiendas más similares respecto a su comportamiento espacial, de ventas y de promociones, en caso de poder obtener más variables, podríamos tener grupos hasta por tipo de productos o tipo de cliente y con esto podríamos desarrollar un modelo para cada grupo, y sea con las variables actuales o con variables adicionales. A partir de estas predicciones sugerimos se tomen las siguientes iniciativas: se deben redefinir las políticas de descuentos, ya que hay muy pocas tiendas donde el descuento está relacionado directamente a las ventas, es necesario evaluar si los descuentos maximizan o no las ventas y a su vez si maximizan o no la ganancia. Si se desea un ROI en las nuevas tiendas óptimo se deberá considerar el CPI de la zona y se deben de reforzar políticas de estructuras de precios que ayuden a incrementar la dependencia de esta variable en las ventas. Se recomienda un análisis y desarrollo puntual en el tratamiento de esta variable y su impacto en el negocio.

La estadística bayesiana nos permite la combinación de diferentes fuentes de información, agregar subjetividad a un procedimiento estadístico, la perspectiva de los parámetros como variables aleatorias, la combinación de múltiples fuentes que interactúen con información de manera jerárquica o la inferencia en series de tiempo con pocos datos. En este curso enfocamos el poder del ajuste de modelos regresivos a datos desde éste enfoque con la ayuda de los paquetes JAGS y BUGS, los cuales permiten definir verosimilitud, distribución previa y realizan muestreo con el sampleador de Gibbs y entregan una distribución posterior de los parámetros (y de predicciones) para distribuciones que no pueden ser conjugadas.

En el contexto de nuestro problema, tuvimos oportunidad de realizar con datos absolutamente reales varias implementaciones vistas en clase como estimación e interpretación de parámetros, modelos para análisis de efectos fijos y modelos de tiempo dinámicos que se actualizan con los parámetros del ‘momento’ anterior, por nombrar algunas.

6. Referencias

Los datos utilizados fueron obtenidos de la competencia de reclutamiento que llevó a cabo Walmart en Kaggle, este se publicó en 2013. Se puede acceder a los datos a través de la siguiente liga: *https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting*

Es importante mencionar que para fines del proyecto se modificaron los datos con el propósito de simplificarlos para el análisis bayesiano.

A lo largo del proyecto se consultaron las notas de clase, especialmente el capítulo 5 que habla de los Modelos Dinámicos, dado que tratamos con datos con dependencia temporal con parametros con tendencia Dinámica. También se consultaron los apartados referentes a los modelos de efectos fijos

Se consultó el libro Bayesian Data Analysis de Andrew Gelman et al, para los temas de modelos dinámicos y modelos de efectos fijos. Los detalles del libro son: *Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. (2002). Bayesian Data Analysis, 2a. edición. Chapman & Hall: Boca Raton.*

7. Anexos

Listado de modelos en JAGS

Modelo de efectos fijos

*model*

*{*

*#Likelihood*

*for (i in 1:n) {*

*y[i] ~ dgamma(alpha,betast[i])*

*betast[i] <- alpha/mu[i]*

*log(mu[i]) <- intercept + beta[T[i]] + t(x[i,]) %\*% gama*

*}*

*#Priors*

*intercept ~ dnorm(0.0001,0.0001)*

*for (j in 1:3) { beta[j] ~ dnorm(0.001, 0.001) }*

*alpha ~ dgamma(1,1)*

*for (j in 1:11) {*

*gama[j] ~ dnorm(0, 0.001)*

*}*

*#Prediction*

*for (i in 1:n) {*

*yf[i] ~ dgamma(alpha,betast[i])*

*}*

*intercept.adj <- intercept+mean(beta[])*

*for (j in 1:3) {*

*beta.adj[j] <- beta[j]-mean(beta[])*

*}*

*}*

Modelo de predicción 1

*model*

*{*

*#Likelihood*

*for (i in 1:n) {*

*y[i] ~ dnorm(mu[i],tau)*

*mu[i]<-alpha+beta[1,i]\*x1[i]+beta[2,i]\*x2[i]+beta[3,i]\*x3[i]+beta[4,i]\*x4[i]+beta[5,i]\*x5[i]+beta[6,i]\*x6[i]+beta[7,i]\*x7[i]+beta[8,i]\*x8[i]+beta[9,i]\*x9[i]+beta[10,i]\*x10[i]*

*}*

*#State eq.*

*for (i in 2:n) {*

*for (j in 1:10) {*

*beta[j,i] ~ dnorm(beta[j,i-1],tau.b[j])*

*}*

*}*

*#Priors*

*alpha ~ dnorm(0,0.001)*

*for (j in 1:10) { beta[j,1] ~ dnorm(0,0.001) }*

*tau ~ dgamma(0.001,0.001)*

*for (j in 1:10){*

*tau.b[j]<- lam\*tau*

*}*

*lam<-1*

*#Prediction 1*

*for (i in 1:n) { yf[i] ~ dnorm(mu[i],tau)*

*}*

Modelo de predicción 2

*model*

*{*

*#Likelihood*

*for (i in 1:n) {*

*y[i] ~ dnorm(mu[i],tau)*

*mu[i]<-alpha[i]+beta[1]\*x1[i]+beta[2]\*x2[i]+beta[3]\*x3[i]+beta[4]\*x4[i]+beta[5]\*x5[i]+beta[6]\*x6[i]+beta[7]\*x7[i]+beta[8]\*x8[i]+beta[9]\*x9[i]+beta[10]\*x10[i]*

*}*

*#State eq.*

*for (i in 2:n) {*

*alpha[i] ~ dnorm(alpha[i-1],tau.b)*

*}*

*#Priors*

*alpha[1] ~ dnorm(0,0.001)*

*for (j in 1:10) { beta[j] ~ dnorm(0,0.001) }*

*tau ~ dgamma(0.001,0.001)*

*tau.b<- lam\*tau*

*lam<-1*

*#Prediction 1*

*for (i in 1:n) { yf[i] ~ dnorm(mu[i],tau) }*

*}*