Herramientas de aprendizaje supervisado

Clase 2

Manuel Benjamín

October 21, 2023

Universidad de Buenos Aires

Clasificación

```
> head(Default)
  default student   balance    income
1      No      No  729.5265 44361.625
2      No     Yes  817.1804 12106.135
3      No      No 1073.5492 31767.139
4      No      No  529.2506 35704.494
5      No      No  785.6559 38463.496
6      No     Yes  919.5885  7491.559
> summary(Default)
 default    student       balance           income     
 No :9667   No :7056   Min.   :   0.0   Min.   :  772  
 Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340  
                       Median : 823.6   Median :34553  
                       Mean   : 835.4   Mean   :33517  
                       3rd Qu.:1166.3   3rd Qu.:43808  
                       Max.   :2654.3   Max.   :73554  
```

Figure 1: Dataset Default del paquete ISLR2. Cada fila corresponde a un individuo en EEUU con su estatus de estudiante, balance de tarjeta, ingresos anuales y si dejo de pagar la tarjeta o no.

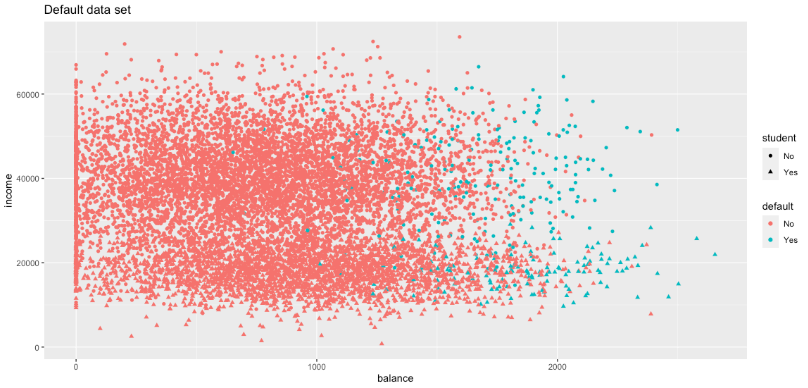


Figure 2: Visualización de todos los datos.

Objetivos

- Predecir si un individuo entrará en default o no.
- Entender la importancia de las variables.
- Como fijar el tope máximo en la tarjeta de un individuo para maximizar la ganancia esperada?

Opciones

- Vecinos cercanos.
- Regresión logística.

Consideraciones

Que tipo de error es peor? Decir que alguien va a defaultear cuando en realidad va a pagar o al revés? Que clase deberíamos codificar como clase positiva?

Vecinos cercanos

Figure 3: Regiones de clasificación según la cantidad de vecinos

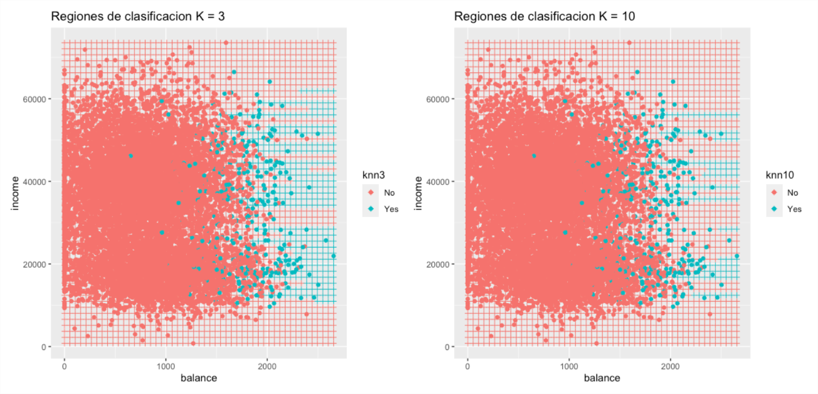


Figure 4: Regiones de clasificación segun la cantidad de vecinos con las observaciones superpuestas.

Regresión logística

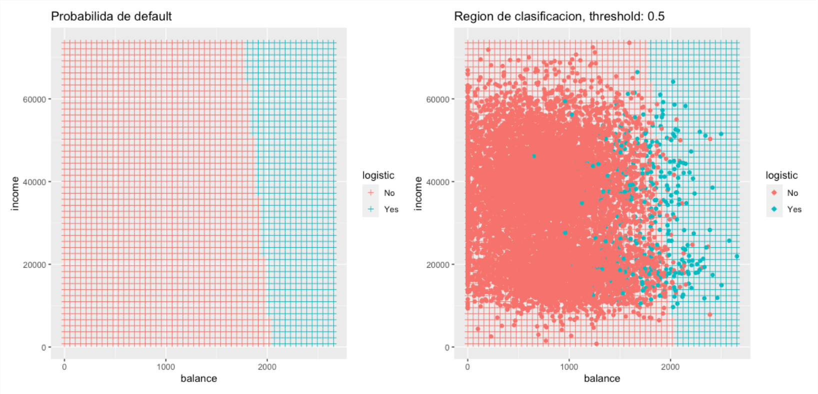


Figure 5: Región de clasificación.

Para tratar de responder cada modelo ajustado

- El ingreso de un individuo afecta la probabilidad de default?
- Dar una interpretación de las regiones obtenidas.
- Por que KNN con $K = 10$ tiene menos superficie verde que $K = 3$?
- Se puede incluir la variable Student en regresion logistica? Y en KNN?
- En KNN, ¿Por qué hay regiones predichas como ”No” a pesar de que visualmente están rodeadas de observaciones ”Si”?

Comparación de probas estimadas

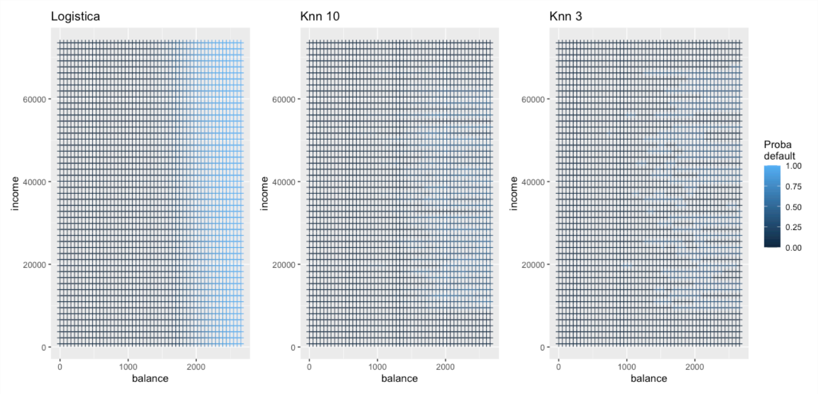


Figure 6: Probabilidad de default en función de balance e ingreso.

Regresión logística

Regresión logística

Modelo

$$
\begin{aligned}
Y|X = x &\sim \mathcal{B}_e(p(x)) \\
p(x) &= \frac{exp(\beta_0 + \beta x)}{1 + exp(\beta_0 + \beta x)} \\
&= \frac{1}{1 + exp(-\beta_0 - \beta x)}
\end{aligned}
$$

La relación entre p y x la podemos reescribir en función del logaritmo de las chances (odds)

$$
\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta x
$$

Las chances son una manera de referirse a lo verosímil de un evento y son calculadas como la probabilidad de ocurrencia sobre la probabilidad de no ocurrencia.

$$\text{odds} = \frac{p}{1-p}$$

- Las odds de sacar cara en una moneda es $\frac{1}{1}$ (1 a 1.)
- Las odds de sacar un 6 en un dado es $\frac{1}{5}$ (1 a 5.)

En un juego con finitos resultados, el numerador de las odds representa cuantas veces se espera ganar si se juegan (numerador + denominador) veces el juego.

Las odds se usan historicamente en apuestas

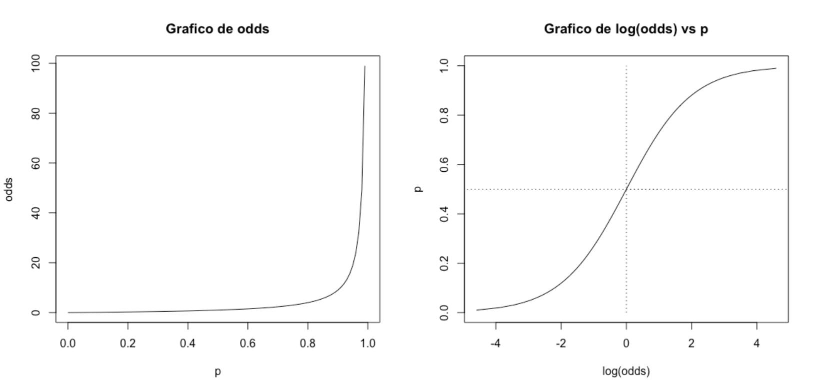


Grafico de odds
odds
p
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
Grafico de log(odds) vs p
p
log(odds)
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4

Interpretación de coeficientes del modelo

$$\text{Default} \sim \mathcal{B}_e(p(\text{balance}))$$

$$\log(\text{odds}) = -10 + 0.005 \times \text{balance}$$

- Como el signo del coeficiente default es positivo, a mayor balance, mayor log odds y por lo tanto mayor probabilidad de default.
- Si un individuo incrementa en k unidades su balance, las chances de default se multiplican por el factor $exp(0.005 \times k)$.
- Para modelos multivariados la interpretación de cada coeficiente es el incremento proporcional de los odds manteniendo el resto de las covariables fijas. (a la modelo lineal)

¿Cómo se ajustan los parámetros del modelo logístico?

Máxima la verosimilitud del modelo

$$\begin{aligned}
\log\left(\mathcal{L}(\beta)\right) &= \log\left(\Pi_{i=1}^{n} p(x_i)^{y_i}(1-p(x_i))^{1-y_i}\right) \\
&= \sum_{i=1}^{n} y_i \log(p(x_i)) + (1-y_i)\log(1-p(x_i)) \\
&= \sum_{i=1}^{n} y_i \log\left(\frac{exp(\beta_0+\beta x_i)}{1+exp(\beta_0+\beta x_i)}\right) + \\
&(1-y_i)\log\left(1-\frac{exp(\beta_0+\beta x_i)}{1+exp(\beta_0+\beta x_i)}\right)
\end{aligned}$$

O equivalentemente minimizando la deviance:

$$\text{Deviance} = -2\log\left(\mathcal{L}(\beta)\right)$$

Inferencia sobre los parametros

Si el modelo vale, tenemos estimaciones de la varianza de los coeficientes y distribuciones aproximadas para muestras grandes. Podemos hacer intervalos de confianza para los coeficientes, analisis de la deviance (Test de hipótesis como en ANOVA del modelo lineal)

Para trabajar en R

Responder las siguientes preguntas para los datos de default.

- Realizar gráficos exploratorios para entender los datos de Default.
- El balance ayuda a explicar el default?
- Ser estudiante, aumenta o disminuye la probabilidad de default?
- Existe alguna interacción entre balance e ingreso para explicar el default?
- Con el modelo ajustado

  $$\text{default} \sim \text{balance} + \text{student}$$

  Grafique la curvas de probabilidad de default en función de balance para estudiantes y no estudiantes.

Ejercicio para pensar

La fintech donde trabajas va a sacar una tarjeta de crédito. Esta tarjeta va a tener un tope máximo en el consumo que se fija para cada persona. El CEO de la empresa te encomendó hacer un estudio y determinar un mecanismo que determine el balance maximo permitod a cada persona. Este puede depender del ingreso y del estado estudiantil de la persona.

La instrucción del jefe es escueta pero directa: *“Ponga el limite de manera de ganar la mayor cantidad de plata posible.”*

Luego de juntarse con el sector de finanzas te enterás que cuando alguien paga su deuda en la tarjeta se gana el 15% del total del balance. En cambio si entra en default se pierde el total del balance si una tarjeta entra en default.

Hacete cargo del pedido del jefe. Recordá que tenés que poder explicar y justificar tu criterio de manera que otras personas lo entiendan. Si bien el objetivo primero es ganar plata, el jefe valora metodologías que el pueda entender. Sumas puntos extra si tenes graficos y podés estimar la ganancia esperada por cada tarjeta emitida.

No te olvides de dejar en claro los supuestos que vas haciendo en cada paso del proceso.

Ejercicios

- Con los datos de balance Ajustar KNN con K = 3 escalando apropiadamente las variables balance e ingreso. Comparar con los ajustes sin escalar.
- Ajustar una regresión logística con

  $$\texttt{default} \sim \texttt{balance + student + student*balance}$$

  y otra para

  $$\texttt{default} \sim \texttt{balance}$$

  pero usando solo los datos de estudiantes.
  Con ambos modelos prediga la probabilidad de default para estudiantes con distintos balances. ¿Cambian las predicciones? ¿Que está sucediendo?

Regresión de Poisson

Modelando datos de conteo

```
> head(Bikeshare)
```

	season	mnth	day	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	bikers
1	1	Jan	1	0	0	6	0	clear	0.24	0.2879	0.81	0.0000	3	13	16
2	1	Jan	1	1	0	6	0	clear	0.22	0.2727	0.80	0.0000	8	32	40
3	1	Jan	1	2	0	6	0	clear	0.22	0.2727	0.80	0.0000	5	27	32
4	1	Jan	1	3	0	6	0	clear	0.24	0.2879	0.75	0.0000	3	10	13
5	1	Jan	1	4	0	6	0	clear	0.24	0.2879	0.75	0.0000	0	1	1
6	1	Jan	1	5	0	6	0	cloudy/misty	0.24	0.2576	0.75	0.0896	0	1	1

Figure 7: Bikeshare data de Washington DC. *`bikers`* es la cantidad de usuarios del servicio de bicicleta, algunas de las otras variables *`mnth`*: mes del año, *`hr`*: hora del día, *`workingday`*: una variable indicadora de si el día es laborable ...

Tiene sentido un modelo lineal?

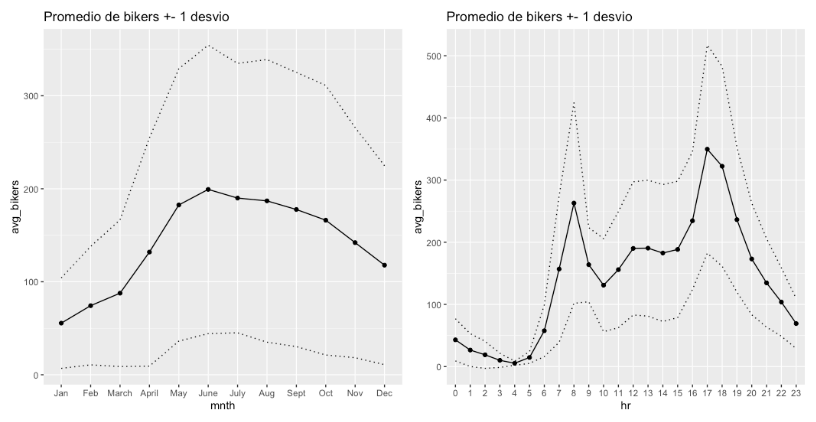


Figure 8: Promedio y desvio de bikers agrupados en distintas variables.

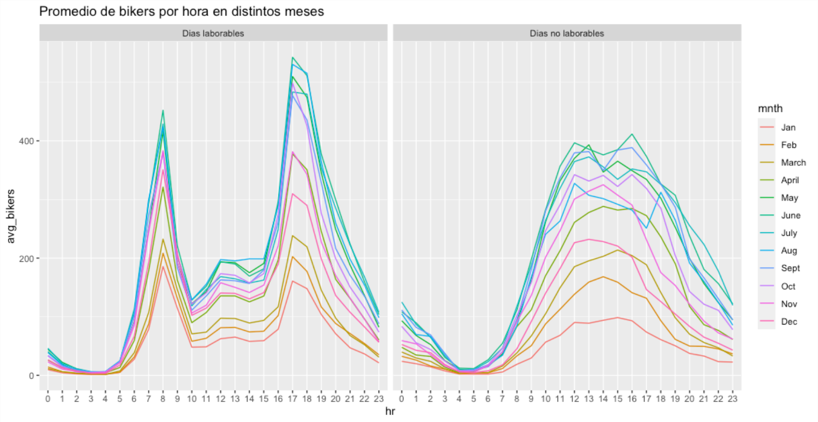


Figure 9: Mas gráficos lindos

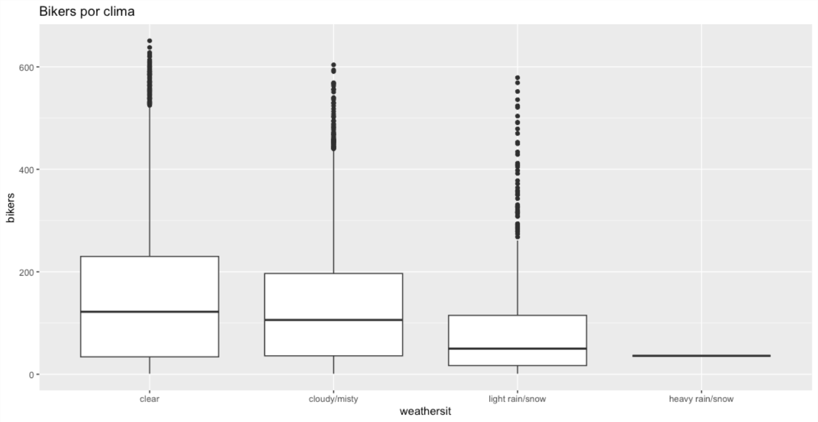
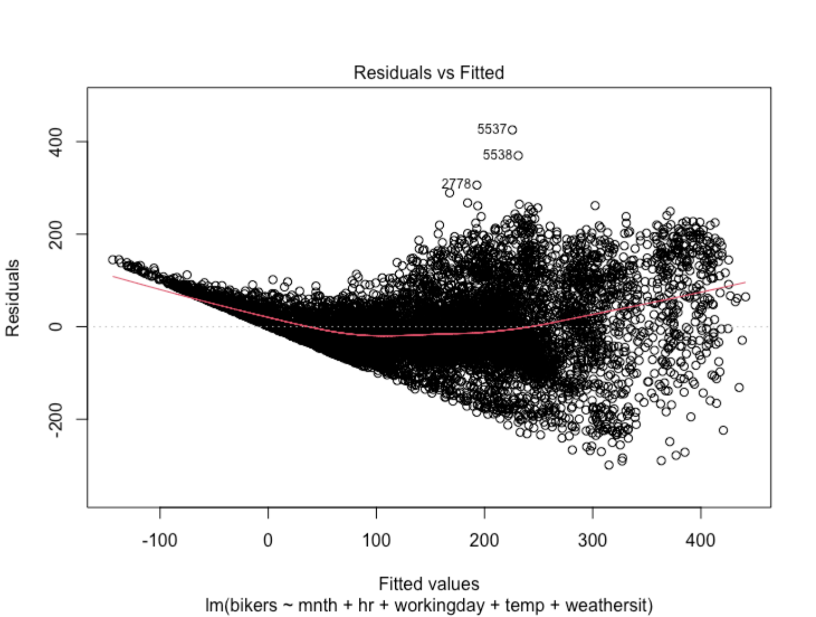


Figure 10: No tan lindo pero informativo.

¿Podemos ajustar un modelo lineal a los datos?



Problemas de ajustar una regresión lineal

- Las predicciones pueden dar cantidad de bikers negativas.
- La variabilidad de la respuesta depende de las covariables.

Hay algunas cosas que se pueden hacer para mitigar estos problema (transformaciones en la respuesta)

$$\log(Y) = \beta X + \varepsilon$$

Pero estas soluciones traen otros problemas:

- Interpretación de los coeficientes.
- Que pasa si en un horario hay cero bikers?

Variable aleatoria Poisson

Decimos que la variable Y tiene distribución de Poisson con parametro λ si

$$P(Y=k)=\frac{e^{-\lambda}\lambda^{k}}{k!}$$

La esperanza y la varianza de una Poisson coinciden

$$E(Y)=Var(Y)=\lambda.$$

$$Y|X \sim \text{Poisson}(\lambda(X))$$
$$\lambda(X) = exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)$$

Interpretación de coeficientes

Un incremento en una unidad de X_j esta asociado con un incremento de $E(Y)$ por un factor de e^{β_j}.

$$\lambda(X_1 + 1, X_2, \ldots, X_p) = e^{\beta_1}\lambda(X_1, X_2, \ldots, X_p).$$

Ajuste del modelo:

Por máxima verosimilitud!

$$\log(\mathcal{L}(\beta_0, \ldots, \beta_p)) = \sum_{i=1}^{n} y_i \log(\lambda(x_i)) - \lambda(x_i) - \log(x_i!).$$

```
Call:
glm(formula = bikers ~ workingday + temp + weathersit + mnth +
    hr, family = poisson, data = Bikeshare)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-20.7574   -3.3441   -0.6549    2.6999   21.9628

Coefficients:
                           Estimate Std. Error  z value Pr(>|z|)
(Intercept)                2.693688   0.009720  277.124  < 2e-16 ***
workingday                 0.014665   0.001955    7.502 6.27e-14 ***
temp                       0.785292   0.011475   68.434  < 2e-16 ***
weathersitcloudy/misty    -0.075231   0.002179  -34.528  < 2e-16 ***
weathersitlight rain/snow -0.575800   0.004058 -141.905  < 2e-16 ***
weathersitheavy rain/snow -0.926287   0.166782   -5.554 2.79e-08 ***
mnthFeb                    0.226046   0.006951   32.521  < 2e-16 ***
mnthMarch                  0.376437   0.006691   56.263  < 2e-16 ***
```

Figure 11: Resumen del ajusto Poisson a los datos de Bikeshare.

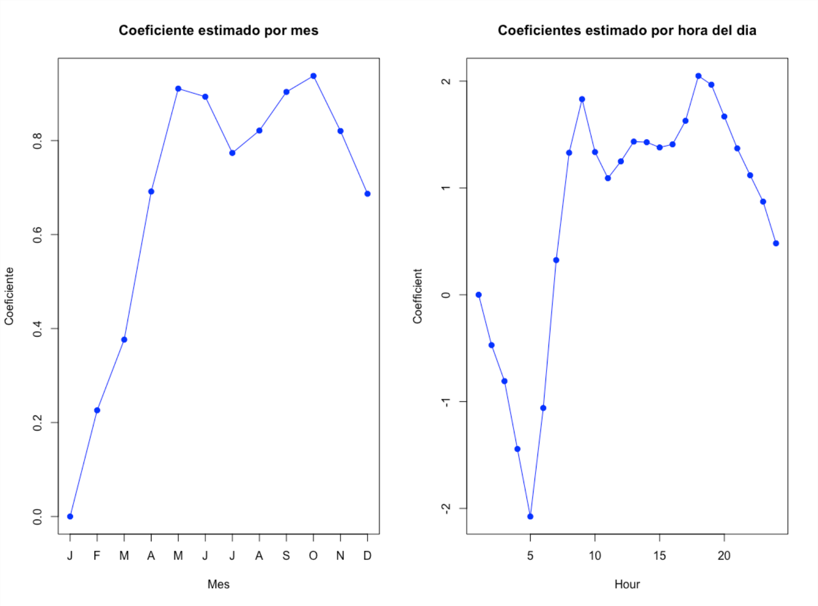
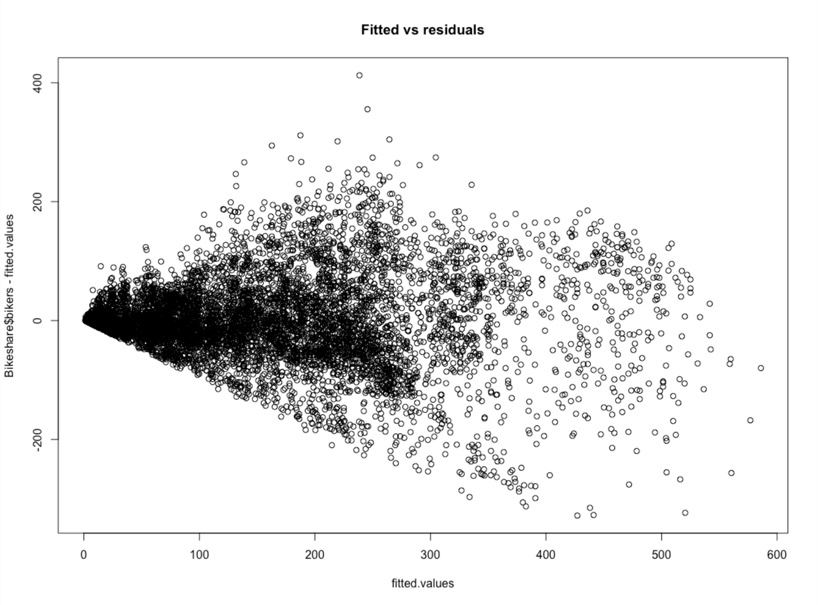


Figure 12: Graficos de los coeficientes estimados. Los valores son en diferencia del primer coeficiente que se asume cero.


Fitted vs residuals
Bikeshare$bikers - fitted.values
400
200
0
-200
0
100
200
300
400
500
600
fitted.values

Ejercicios

- Comparar las predicciones del modelo lineal y el modelo de Poisson graficando los valores predichos de cada uno. ¿Que observa?
- Predecir la cantidad total de bicicletas usadas a lo largo de un día laborable de Julio que está despejado.

Modelo lineal generalizado

Por ahora consideramos tres modelos de regresión.

- En cada uno intentamos predecir Y utilizando X. Asumimos que Y pertenece a una familia de distribuciones.
- En cada uno modelamos $E(Y|X)$.

Lineal: Normal

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \ldots, \beta_p X_p$$

Logistica: Bernoulli

$$\begin{aligned} E(Y|X) &= P(Y = 1|X_1, \ldots, X_p) \\ &= \frac{exp(\beta_0 + \beta_1 X_1 + \ldots, \beta_p X_p)}{1 + exp(\beta_0 + \beta_1 X_1 + \ldots, \beta_p X_p)} \end{aligned}$$

Poisson: Poisson

$$E(Y|X) = exp(\beta_0 + \beta_1 X_1 + \ldots, \beta_p X_p)$$

Modelo lineal generalizado

Las tres distribuciones vistas son distribuciones que pertenecen a la familia de distribuciones exponenciales. Otras distribuciones exponenciales son:

- Gamma.
- Binomial.
- Binomial negativa.
- Exponencial.
- Multinomial.

Si modelamos

$$Y|X \sim \text{ Distribución de la familia exponencial}$$

$$E(Y|X) = \eta\left(\beta_0 + \beta_1 X_1 + \ldots, \beta_p X_p\right)$$

Tenemos un modelo lineal generalizado.

Modelo lineal generalizado

- La función η se elige acorde al problema modelado.
- La interpretación de los coeficientes depende de la η elegida
- La familia exponencial garantiza algoritmos eficientes para el ajuste de máxima verosimilitud.
- Para una misma distribución podemos elegir distintas η.
- Si el modelo vale podemos hacer inferencia sobre los parámetros.