Estadística (M) - Segundo cuatrimestre 2018

Práctica 6 - Regresión lineal

Se recomienda empezar por los ejercicios marcados sin †.

A) *Propiedades del estimador de mínimos cuadrados*

Supongamos que tenemos una muestra $(Y_1, \mathbf{x}_1), \ldots, (Y_n, \mathbf{x}_n)$, $\mathbf{x}_i \in \mathbb{R}^p$ (fijos) y que existe un $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ tal que

$$Y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_i, \qquad i = 1, \ldots, n \tag{1}$$

donde ε_i son errores independientes con media 0 y varianza σ^2. Sea $\mathbf{X} \in \mathbb{R}^{n \times p}$ la matriz compuesta por las filas $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (en ese orden) y sea $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. Notar que si el modelo tiene intercept, $\mathbf{X}$ tendrá una primera columna compuesta por unos. En esta sección asumimos que la matriz de diseño $\mathbf{X}$ tiene rango completo. Llamamos $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ al estimador de mínimos cuadrados y $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)^{\mathrm{T}}$ el vector estimado de respuestas.

1. (a) Si el modelo tiene intercept, probar que $\sum_{i=1}^n \widehat{Y}_i = \sum_{i=1}^n Y_i$.

   (b) Probar que $\sum_{i=1}^n \widehat{Y}_i(Y_i - \widehat{Y}_i) = 0$ (incluso si no hay intercept).

   Sugerencia: las cuentas se simplifican mucho trabajando con álgebra lineal. Probar y usar que $\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$ donde $\mathbf{H} \in \mathbb{R}^{n \times n}$ es un proyector (llamado matriz *Hat*). Para la parte (a), probar y usar que el vector de unos $\mathbf{1}_n$ pertenece a la imagen del operador $\mathbf{H}$.

2. (a) En el modelo (1), los errores son homocedásticos, es decir, su varianza es siempre la misma. Estos errores pueden ser estimados por los residuos $r_i = Y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$. ¿Es cierto que estos residuos también son homocedásticos?

   (b) Si el modelo tiene intercept, calcular la suma $\sum_{i=1}^n r_i$. (Sugerencia: usar el ejercicio 1).

3. Supongamos que con la matriz de diseño $\mathbf{X}$ y vector de respuestas $\mathbf{Y}$ obtenemos, por el método de mínimos cuadrados, el vector estimado de respuestas $\widehat{\mathbf{Y}}_1$. Supongamos ahora que, en vez de observar $\mathbf{X}$, hubiéramos observado la matriz de diseño $\mathbf{W}$, donde la columna j de $\mathbf{W}$ es igual a la columna j de la matriz $\mathbf{X}$ multiplicada por una constante $c_j \neq 0$. Con la matriz $\mathbf{W}$ y el vector de respuestas $\mathbf{Y}$, se obtiene el vector estimado de respuestas $\widehat{\mathbf{Y}}_2$. Probar que $\widehat{\mathbf{Y}}_1 = \widehat{\mathbf{Y}}_2$. Deducir que el vector estimado de respuestas es invariante respecto a la escala con la que se miden las variables explicativas.

   Sugerencia: escribir en la forma más explícita posible la relación entre las matrices $\mathbf{X}$ y $\mathbf{W}$.

† 4. **(Coeficiente de correlación múltiple)**

   (a) Probar que si el modelo (1) tiene intercept y ninguna otra covariable (o sea, $\mathbf{X}$ es un vector columna de unos), entonces $\widehat{Y}_i = \overline{Y}$ (el promedio de las Y_i).

   (b) Probar que si el modelo (1) tiene intercept, entonces

$$\sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2.$$

   Esto también se puede escribir como $\mathrm{SCT} = \mathrm{SCE} + \mathrm{SCR}$ donde

   - $\mathrm{SCT} = \sum_{i=1}^n (Y_i - \overline{Y})^2$ es la suma de cuadrados totales, y refleja la variabilidad total de la respuesta $\mathbf{Y}$.

- SCE $= \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ es la suma de cuadrados de los errores y refleja la variabilidad de $\mathbf{Y}$ que no fue explicada por el modelo de regresión.
- SCR $= \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$ es la suma de cuadrados de la regresión y refleja la variabilidad de $\mathbf{Y}$ explicada por el modelo de regresión (en comparación con lo que obtendríamos si sólamente consideráramos el intercept).

El **coeficiente de correlación múltiple** R se define como la raíz cuadrada de

$$R^2 = \frac{\text{SCR}}{\text{SCT}}.$$

Sugerencia: usar el ejercicio 1.

(c) Probar que si hay intercept, $0 \leq R^2 \leq 1$. En este caso, demostrar también que $R^2 = 0$ si todos los coeficientes de las covariables que no son intercept fueron estimados como cero. Notar que $R^2 = 1$ en el caso en que $\widehat{Y}_i = Y_i$ para todo $i = 1, \ldots, n$ (modelo saturado).

B) *Inferencia para el vector de regresión e interpretación de salidas del* `R`

1. En el puerto de la Ciudad de Grand Lakes, en Canadá, se quiere ver cómo influye el peso de un cargamento en el tiempo necesario para descargarlo. Para eso, se registra el peso y el tiempo de descarga para 30 cargamentos y se plantea el modelo lineal

$$\text{Tiempo} = \alpha + \text{Peso} * \beta + \varepsilon_i \tag{2}$$

donde se asume que los errores ε_i son independientes y tienen distribución normal con media 0. Los datos se encuentran en el archivo `glakes.csv`. A partir de la siguiente salida de `R`, contestar las preguntas que figuran abajo.

```
> glakes.fit = lm(Tiempo ~ Peso)
> summary(glakes.fit)

Call:
lm(formula = Tiempo ~ Peso)

Residuals:
    Min      1Q  Median      3Q     Max
-23.882  -6.397  -1.261   5.931  21.850

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.344707   2.642633   4.671 6.32e-05 ***
Peso         0.006518   0.000531  12.275 5.22e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.7 on 29 degrees of freedom
Multiple R-squared:  0.8386,Adjusted R-squared:  0.833
F-statistic: 150.7 on 1 and 29 DF,  p-value: 5.218e-13
```

(a) ¿Cuáles son los coeficientes estimados para α y β usando mínimos cuadrados?

(b) ¿Hay evidencia suficiente a nivel 0.01 para decir que $\beta \neq 0$? ¿Cuál es el estadístico del test correspondiente y cuál es su distribución bajo H_0? ¿Cuánto vale este estadístico para estos datos? Hallar el $p-$valor. Interpretar la conclusión de este test.

(c) Considerar las hipótesis $H_0 : \alpha = 10$ vs. $H_1 : \alpha > 10$. ¿Hay evidencia suficiente para rechazar H_0 a nivel 0.05? Hallar el $p-$valor.

(d) Hallar un intervalo de confianza para α de nivel 0.95.

† (e) ¿Cuánto vale el R^2 en este caso y cómo interpretaría esto?

(f) Llega un nuevo cargamento con peso 1000. ¿Cómo estimaría el tiempo que se tarda en descargarlo?

2. En este conjunto de datos, queremos explicar la deuda (variable `Balance`) en tarjeta de crédito de 400 clientes en función de varias características de cada uno. Para cada cliente, tenemos las variables `Income` (sueldo anual), `Rating` (rating crediticio), `Limit` (lmite de crédito), `Cards` (número de tarjetas), `Age` (Edad), `Education` (número de años de educación), `GenderFemale` (vale 1 si el cliente es mujer, 0 si no), `StudentYes` (vale 1 si el cliente es estudiante, 0 si no), `MarriedYes` (vale 1 si el cliente está casado, 0 si no). Se asume válido un modelo lineal con intercept, variable respuesta `Balance` y las variables explicativas ya descriptas. Los datos se encuentran en el archivo `credit.txt`. Se obtiene la siguiente salida de `R`:

```
> credit.fit = lm(Balance ~ Income + Rating + Limit +  Cards +
+                   Age + Education + Gender + Student + Married)
> summary(credit.fit)

Call:
lm(formula = Balance ~ Income + Rating + Limit + Cards + Age +
    Education + Gender + Student + Married)

Residuals:
    Min      1Q  Median      3Q     Max
-171.66  -75.32  -11.29   54.42  309.98

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -468.40374   34.35512 -13.634  < 2e-16 ***
Income         -7.80200    0.23395 -33.349  < 2e-16 ***
Rating          1.10227    0.48923   2.253   0.0248 *
Limit           0.19308    0.03268   5.909 7.52e-09 ***
Cards          17.92327    4.33228   4.137 4.31e-05 ***
Age            -0.63468    0.29325  -2.164   0.0310 *
Education      -1.11503    1.59592  -0.699   0.4852
GenderFemale  -10.40665    9.90410  -1.051   0.2940
StudentYes    426.46919   16.67770  25.571  < 2e-16 ***
MarriedYes     -7.01910   10.27803  -0.683   0.4951
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 98.72 on 390 degrees of freedom
Multiple R-squared:  0.9549,Adjusted R-squared:  0.9539
F-statistic: 918.2 on 9 and 390 DF,  p-value: < 2.2e-16
```

(a) ¿Cuáles parecen ser las variables más relevantes en este modelo?

(b) Sea β_{Age} el coeficiente correspondiente a la variable `Age`. Mirando la salida, hallar el $p-$valor del test con hipótesis $H_0 : \beta_{Age} = 0$ vs. $H_1 : \beta_{Age} \neq 0$. ¿Cuál sería el $p - valor$ del test con hipótesis $H_0 : \beta_{Age} = 0$ vs. $H_1 : \beta_{Age} < 0$.

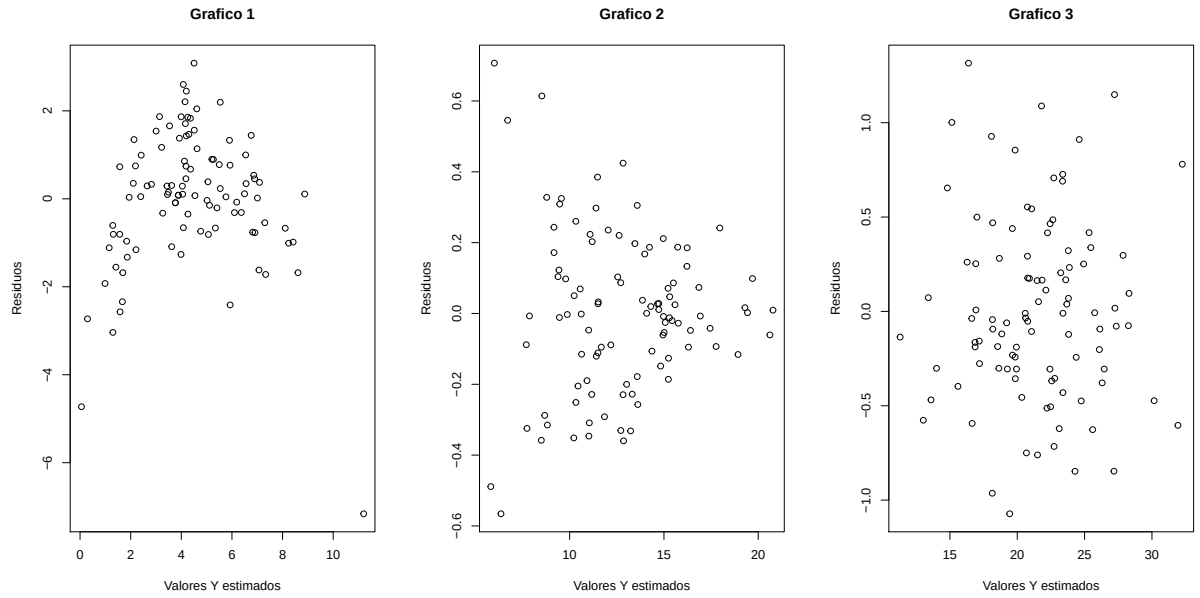
(c) Supongamos que los errores ε_i del modelo tienen distribución $N(0, \sigma^2)$. Estimar σ^2 en este caso.

(d) Supongamos que dos clientes comparten las mismas características (en términos de las variables explicativas consideradas), salvo que un cliente tiene tres años más que el otro. Sea B_V el valor esperado de la deuda del cliente más viejo y B_J el valor esperado de la deuda del paciente joven. Hallar $B_J - B_V$ y estimarlo en este caso.

(e) Hallar un intervalo de confianza de nivel 0.9 para $\beta_{Education}$ (el coeficiente correspondiente a la variable `Education`).

C) *Diagnóstico*

1. Supongamos que tenemos una muestra $(Y_1, x_1), \ldots, (Y_n, x_n)$ con $n = 100$ y ajustamos un modelo $Y_i = \alpha + \beta x_i + \varepsilon_i$ para $i = 1, \ldots, 100$ y ε_i son independientes con media 0 y varianza σ^2. Consideramos los siguientes tres gráficos de vector estimado de respuestas vs. residuos:



Supongamos que para nuestro conjunto de datos observamos el gráfico 1. ¿Le parece válido el modelo? ¿Hay alguna hipótesis que parezca no cumplirse? Si le parece que hay un problema, ¿se le ocurre alguna forma de modificar el modelo para que resulte válido? Responder las mismas preguntas suponiendo que se observó el gráfico 2 y finalmente responderlas suponiendo que se observó el gráfico 3.

2. (Para hacer con el R) Hacer el gráfico de vector estimado de respuestas vs. residuos en el para los datos `glakes.csv`. ¿Le parece que el modelo planteado en el ejercicio 1 de la parte B) es válido? Considerar ahora el modelo

$$\log(\text{Tiempo}) = \alpha + \beta * \text{Peso}^{0.25} + \varepsilon_i \tag{3}$$

donde ε_i son independientes y distribuídas como $N(0, \sigma^2)$. Hacer nuevamente el gráfico. ¿Le parece válido el modelo ahora?

3. (Para hacer con el R)

   (a) ¿Es razonable asumir que los errores en el modelo (2) son normales? Responder lo mismo para el modelo (3).

   (b) Para el modelo del conjunto de datos de `credit.txt`, ¿se puede asumir normalidad de los errores?

   Sugerencia: investigar y usar el test de Shapiro-Wilk (comando `shapiro.test` en `R`) aplicado a los residuos.

D) <u>*Selección de modelos*</u>

† 1. (Para hacer con el R) Consideremos el conjunto de datos `credit.txt`. Separar las observaciones en dos grupos : *Entrenamiento* y *Testeo*. Poner en el grupo *Entrenamiento* aproximadamente 2/3 del total de observaciones (elegidas aleatoriamente) y las observaciones restantes ponerlas en el grupo *Testeo*. Dado un cierto modelo, obtener los coeficientes estimados que correspondan únicamente usando las observaciones del grupo *Entrenamiento*. Con dichas estimaciones, predecir cada una de las observaciones del grupo *Testeo*, obteniendo así los valores estimados $\widehat{Y}_1, \ldots, \widehat{Y}_k$, donde k es el tamaño del grupo *Testeo*. Finalmente, calcular $W = \sum_{i=1}^{k}(Y_i - \widehat{Y}_i)^2$, donde en dicha suma sólamente intervienen las observaciones del grupo *Testeo*. En este ejercicio, si se tienen varios modelos, se elige el que tiene menor W.

   Consideremos todos los modelos lineales (siempre con intercept) cuyas variables explicativas son un subconjunto de $\{\texttt{Income}, \texttt{Cards}, \texttt{Age}\}$ (son 8 modelos en total). Calcular el W de cada uno de ellos y decidir cuál de estos 8 modelos preferiría.

† 2. **(Leave-one-out Cross-Validation)** (Para hacer con el R) Para el conjunto `glakes.csv`, se proponen los siguientes modelos:

   (a) $\text{Tiempo} = \beta * \text{Peso} + \varepsilon_i$

   (b) $\text{Tiempo} = \alpha + \beta * \text{Peso} + \varepsilon_i$

   (c) $\log(\text{Tiempo}) = \alpha + \beta * \text{Peso}^{0.25} + \varepsilon_i$

   (d) $\log(\text{Tiempo}) = \alpha + \beta * \text{Peso}^{0.25} + \gamma * \text{Peso}^{0.5} + \varepsilon_i$

   Consideremos el siguiente procedimiento como modo de "evaluar" cada modelo:

   - Fijado un $i \in \{1, \ldots, n\}$, obtener los coeficientes estimados que correspondan usando todas las observaciones salvo la $i-$ésima.
   - Con los coeficientes obtenidos, predecir el valor de la $i-$ésima observacion, llamemos $\widehat{Y}_i$ a dicha predicción. Luego, obtener $r_i = Y_i - \widehat{Y}_i$.
   - Hacer variar el i de 1 a n (donde n es la cantidad de observaciones) y para cada i, seguir los pasos anteriores, obteniendo de esta forma los residuos $r_1, \ldots, r_n$
   - Obtener $W = \sum_{i=1}^{n} r_i^2$.

   Si se tienen varios modelos, se elige el que tiene menor W. Aplicando este procedimiento a cada uno de los cuatro modelos propuestos, decidir cuál elegiría. **Importante:** Para los modelos (c) y (d), las predicciones se obtienen haciendo $\widehat{Y}_i = \exp\{\widehat{\alpha} + \widehat{\beta} * \text{Peso}_i^{0.25}\}$ y $\widehat{Y}_i = \exp\{\widehat{\alpha} + \widehat{\beta} * \text{Peso}_i^{0.25} + \widehat{\gamma} * \text{Peso}_i^{0.5}\}$ respectivamente.