Introducción a la Estadística y Ciencia de Datos - Segundo cuatrimestre 2021

Práctica 6 - Regresión lineal

Propiedades

1. Sea $(X, Y) \in \mathbb{R}^2$ un vector aleatorio.

   a) Probar que la función $g(X) = aX + b$ que minimiza $ECM = E[(Y - g(X))^2]$ es la recta de regresión, dada por

   $$g(X) = \frac{cov(X, Y)}{var(X)}(X - E(X)) + E(Y)\,.$$

   b) Hallar cuánto vale el ECM óptimo.

2. Sea $Y_i = \theta_0 + \theta_1 x_i + \epsilon_i$, $1 \leq i \leq n$, con x_i fijas, $\mathbb{E}(\epsilon_i) = 0$, $\mathbb{V}ar(\epsilon_i) = \sigma^2$ y $\epsilon_1, \ldots, \epsilon_n$ independientes.

   a) Hallar el estimador de mínimos cuadrados (EMC) de (θ_0, θ_1).

   b) Probar que si $\epsilon_i \sim N(0, \sigma^2)$, el EMC y el EMV de (θ_0, θ_1) coinciden. ¿Cuánto vale el estimador de máxima verosimilitud de σ^2? ¿Es el que usamos habitualmente?

   c) Estimar los parámetros que minimizan el ECM poblacional, hallados en el ejercicio 1. Puede ser útil recordar el ejercicio 5 de la práctica 4 para estimar la covarianza. ¿Coincide con el EMC de (θ_0, θ_1)?

3. Sean $(Y_1, \mathbf{x}_1), \ldots, (Y_n, \mathbf{x}_n)$, $\mathbf{x}_i \in \mathbb{R}^p$ (fijos) que satisfacen el modelo

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

   para algún un $\boldsymbol{\beta}_0 \in \mathbb{R}^p$, donde ε_i son errores independientes con media 0 y varianza σ^2. Llamemos $\mathbf{X} \in \mathbb{R}^{n \times p}$ a la matriz compuesta por las filas $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (en ese orden) y sea $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$. Notar que si el modelo tiene intercept, $\mathbf{X}$ tendrá una primera columna compuesta por unos. Asumimos que la matriz de diseño $\mathbf{X}$ tiene rango completo. Llamamos $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ al estimador de mínimos cuadrados y $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)^T$ el vector predicho de respuestas.

   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). ¿Cómo interpreta esto desde el punto de vista geométrico?

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

   Sug: Escribir al vector de residuos $\mathbf{r}$ en términos de la matriz de proyección $\mathbf{P}$ y usar las propiedades del cálculo de la matriz de covarianza.

d) Si el modelo tiene intercept, calcular la suma $\sum_{i=1}^{n} r_i$.

e) 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).

Estimación e inferencia

4. 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 en toneladas y el tiempo de descarga para 30 cargamentos y se plantea el modelo lineal

$$\log(\text{Tiempo}) = \theta_0 + \text{Peso}^{0.25} \cdot \theta_1 + \varepsilon_i \,, \tag{2}$$

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) Construya las variables transformadas $LT = \log(\text{Tiempo})$ y $W = \text{Peso}^{0.25}$ a partir de los datos. Ajuste para las variables transformadas un modelo lineal con intercept θ_0 y pendiente θ_1. A partir de la salida de `R`, se piden los siguientes items.

b) ¿Cuánto valen las estimaciones de θ_0 y θ_1 usando el método de mínimos cuadrados?

c) ¿Cuánto vale la estimación de σ^2, la varianza de los errores?

d) ¿Hay evidencia suficiente a nivel 0.01 para decir que $\theta_1 \neq 0$? ¿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.

e) A los investigadores les interesa determinar si la intercept es 10 o mayor a 10. Para resolver esto, considerar las hipótesis $H_0 : \theta_0 = 10$ vs. $H_1 : \theta_0 > 10$. ¿Hay evidencia suficiente para rechazar H_0 a nivel 0.05? Hallar el $p-$valor.

f) Hallar un intervalo de confianza para θ_0 de nivel 0.95. ¿Cuánto vale el coeficiente de regresión múltiple R^2 en este caso y cómo interpretaría este valor en términos de las variables transformadas? Calcular la correlación muestral entre la variable LT observada y los valores predichos por el modelo (el comando `cor` del R puede ser de utilidad). Compare el cuadrado de dicha correlación con el coeficiente de regresión múltiple R^2.

g) Una empresa envía habitualmente cargamentos de peso 625 toneladas y está interesada en obtener una estimación de la media del logaritmo del tiempo que se tarda en descargar cada uno de ellos. ¿Cómo estimaría este valor? Calcular un intervalo de confianza de nivel 0.95 para esta cantidad.

h) Llega un nuevo cargamento con peso 625 toneladas. ¿Cuál es su predicción del logaritmo del tiempo que se tardará en descargarlo? Calcular un intervalo de predicción de nivel 0.95. Compárelo con el intervalo del ítem anterior.

i) Para el cargamento del ítem anterior ¿cuál es su predicción del tiempo que se tardará en descargarlo? Calcular un intervalo de predicción de nivel 0.95.

5. A partir del conjunto de datos `credit.txt`, se busca explicar la deuda (variable `Balance`) en tarjeta de crédito de 400 clientes en función de varias características de cada uno. Para cada cliente, se midieron las variables `Income` (sueldo anual), `Rating` (rating crediticio), `Limit` (límite 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 para la variable respuesta `Balance` y las variables explicativas ya descriptas. Tras realizar un ajuste de dicho modelo y asumiendo que los errores ε_i tienen distribución $\mathcal{N}(0, \sigma^2)$, responda las siguientes preguntas.

   a) ¿Cuál es el modelo propuesto? ¿Cuál es el modelo ajustado?

   b) ¿Cuál es el valor predicho para la segunda observación? ¿Cuánto vale el residuo para la segunda observación? ¿Cuánto vale la suma de los residuos? ¿Cuánto vale la correlación muestral entre la variable `Education` y los residuos? ¿Cuánto vale la correlación muestral entre la variable `Limit` y los residuos? ¿Cuánto vale la correlación muestral entre los residuos y cada una de las demás variables explicativas incluidas en el modelo ajustado?

   c) Dé el valor estimado de σ^2.

   d) ¿Cuáles parecen ser las variables estadísticamente más relevantes en este modelo?

   e) 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. $H_1 : \beta_{Age} < 0$.

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

   g) 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 cliente más joven. Hallar $B_J - B_V$ y estimarlo en este caso. Hallar un intervalo de confianza de nivel 0.95 para esta diferencia.

6. (Para hacer en clase) Para los datos considerados en el ejercicio 7 de la Práctica 5 Complemento, es decir, los datos de cáncer de páncreas, plantee un modelo lineal para comparar las medias del logaritmo de `LYVE1` de los tres grupos determinados por los tres niveles de la variable `diagnosis` a nivel simultáneo 0.05.

   Sug: Explorar el comando `factor` y también comparar las estimaciones de los coeficientes con los promedios muestrales de la variable logaritmo de `LYVE1` en cada grupo.

Selección de Modelos

7. 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 solamente 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.

8. Para el conjunto `glakes.csv`, se proponen los siguientes modelos:

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

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

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

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

   Consideremos el siguiente procedimiento para 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 $\widehat{Y}_i = \exp\{\widehat{\alpha} + \widehat{\beta} \times \text{Peso}_i^{0.25}\}$ y $\widehat{Y}_i = \exp\{\widehat{\alpha} + \widehat{\beta} \times \text{Peso}_i^{0.25} + \widehat{\gamma} \times \text{Peso}_i^{0.5}\}$ respectivamente.

9. Los datos del archivo `cemento.txt` fueron tomados en un estudio experimental para relacionar el calor generado (Y) al fraguar 14 muestras de cemento con distinta composición. Las variables explicativas son los pesos (medidos en porcentajes del peso de cada muestra de cemento) de 5 componentes del cemento.

   a) Calcular la matriz de correlación de todas las variables comprendidas en el problema, incluyendo a la variable Y. Inspeccionando esta matriz, determinar cuáles parecen ser las variables que pueden contribuir significativamente a explicar la variación de Y.

   b) Usar Y como variable dependiente y todas las covariables y una intercept para realizar un ajuste lineal. Calcular el estimador de mínimos cuadrados de los parámetros y para cada uno de ellos testear la hipótesis de que sea 0. ¿Cuáles son significativamente distintos de 0? ¿Es la regresión significativa? ¿Observa alguna contradicción con el

resultado obtenido en los tests individuales anteriores? ¿Vale la pena hacer un nuevo intento para seleccionar qué variables entran en la regresión?

c) Calcular la suma de las 5 covariables. ¿Qué observa? ¿Cómo se justifica este parecido entre los totales? A partir de este resultado, ¿puede justificar ésto que eliminemos del modelo la intercept?

d) Realizar un nuevo ajuste lineal usando las 5 variables independientes y eliminando la intercept. ¿Cuáles coeficientes son significativamente distintos de 0?

e) Plantear un nuevo modelo en el que intervengan aquellas variables que contribuyen significativamente y estimar los parámetros por mínimos cuadrados.

f) ¿Cuál modelo, entre los 3 planteados, eligiría finalmente? Seleccionarlo efectuando el procedimiento de *validación cruzada* propuesto en el ejercicio 8.

Experimentos Numéricos

10. En este ejercicio se crearán datos simulados y se ajustará un modelo de regresión lineal simple de la forma

$$Y = \beta_0 + \beta_1 X + \varepsilon\,.$$

a) Utilizando el comando `rnorm`, crear un vector x que contenga $n = 100$ observaciones provenientes de una distribución $\mathcal{N}(0, 1)$.

b) Utilizando el comando `rnorm`, crear un vector $epsilon$ que contenga $n = 100$ observaciones provenientes de una distribución $\mathcal{N}(0, \sigma^2)$ con $\sigma^2 = 0.025$.

c) Usando x y $epsilon$, generar un vector acorde al modelo:

$$y = -1 + 0.5x + epsilon$$

¿Cuál es la longitud del vector y? ¿Cuáles son los verdaderos valores de β_0 y β_1 en el modelo?

d) Realizar un scatterplot de x vs. y y observar la relación entre x e y.

e) Ajustar un modelo lineal para predecir y en función de x utilizando el método de cuadrados mínimos. Comparar los valores exactos de β_0 y β_1 con sus estimaciones. ¿Son estadísticamente significativos? Hallar una estimación de σ^2 y compárela con su valor real.

f) Graficar la recta de cuadrados mínimos sobre el gráfico realizado en (d). En otro color graficar la recta verdadera $Y = -1 + 0.5X$.

g) Ajustar un modelo polinomial que prediga y usando x y x^2 y superponga al gráfico anterior en un nuevo color. ¿Encuentra alguna evidencia de que el término cuadrático mejora el ajuste del modelo?

Sug: Explorar el comando `I(x^2)` para formular el modelo dentro del comando `lm`.

h) Repetir los ítems (a) a (g) modificando los datos generados de manera que haya más ruido en los datos. Una forma de hacerlo es aumentando el valor de la varianza de la distribución normal usada para generar el término del error $epsilon$. Considere multiplicar por 10 el valor de σ^2 dado en b). ¿Qué pasa si lo multiplica por 100?

11. *a*) Generar el siguiente modelo: crear dos vectores de datos de tamaño 100 x_1 y x_2 a partir de una distribución uniforme en el intervalo $(0, 1)$ y a partir de ellos generar el vector $y = 2 + 2x_1 + 0.3x_2 + \epsilon$, con ϵ que contenga 100 observaciones provenientes de una distribución $\mathcal{N}(0, \sigma^2)$ con $\sigma^2 = 1$. ¿Cuáles son los verdaderos coeficientes de regresión?

b) Realizar tres scatterplots con el comando `pairs` para graficar todas las variables de a pares. ¿Se identifica alguna estructura?

c) Utilizando los datos generados, ajustar un modelo lineal para predecir y en función de x_1 y x_2, utilizando el método de cuadrados mínimos y comparar los valores verdaderos de $\boldsymbol{\beta}$ con sus valores estimados. ¿Son significativos? ¿Qué valor toma el coeficiente de regresión múltiple R^2? ¿Y el test de F que conclusión da?

d) Repetir los incisos (a) y (b), pero variando el desvío del error. En primera instancia utilice $\sigma = 0.5$ y en segundo término con $\sigma = 0.25$. Compare entre sí y con los gráficos, resultados y conclusiones anteriores.

12. *a*) Generar $n = 50$ vectores (x_1, x_2) con distribución conjunta $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ siendo $\boldsymbol{\mu} = (1, 2)^t$ y $\boldsymbol{\Sigma}$ la matriz con 1 en la diagonal y 0.5 fuera de la diagonal. Explorar el comando `mvrnorm` de la librería `MASS`. ¿Las componentes de una $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con el $\boldsymbol{\Sigma}$ dado más arriba, son independientes?

b) Generar un vector de respuestas y siguiendo el siguiente modelo

$$y_i = 1 + 2x_{i1} + 4x_{i2} + \epsilon_i \, , 1 \leq i \leq n$$

donde ϵ_i tiene distribución $N(0, 1)$.

c) Utilizando los datos generados, ajustar un modelo lineal para predecir y en función de x_1 y x_2, utilizando el método de cuadrados mínimos. Guardar los tres coeficientes estimados en un vector. Comparar los valores verdaderos de $(\beta_0, \beta_1, \beta_2)$ con sus valores estimados.

d) Utilizando lo visto en clase teórica respecto de la matriz de diseño X, estime las tres correlaciones entre $\widehat{\beta}_0$, $\widehat{\beta}_1$ y $\widehat{\beta}_2$. Según esto, ¿los estimadores de los coeficientes son variables aleatorias independientes?

e) Repetir los incisos $(12a)$, $(12b)$ y $(12c)$ un número $Nrep = 1000$ veces, guardando los coeficientes estimados en tres vectores de longitud $Nrep$.

f) Graficar scatterplots de los tres vectores de coeficientes estimados usando el comando `pairs`. ¿Parecen correlacionados entre sí? Estime la correlación entre ellos usando la correlación muestral propuesta en el ejercicio 5 de la Práctica 4 (sobre los datos de `MustangPrice`). Lo obtenido, ¿se condice con el gráfico realizado?

g) Repita los items $(12a)$ a $(12f)$ pero modificando las siguientes cuestiones en la simulación: $\boldsymbol{\mu} = (5, 2)^t$ y $\boldsymbol{\Sigma}$ la matriz con 1 en la diagonal y 0 fuera de la diagonal, es decir, la matriz identidad. Responda a las preguntas anteriores y compare sus respuestas con lo hecho en el ítem (f).

h) (opcional) Finalmente, puede repetir $(12a)$ a $(12f)$ pero modificando las siguientes cuestiones en la simulación: tome $\boldsymbol{\mu} = (0, 2)^t$ y $\boldsymbol{\Sigma}$ la matriz con 1 en la diagonal y 0 fuera de la diagonal. Responda a las preguntas anteriores y compare sus respuestas con lo hecho en el ítem g). ¿Qué cambia?