Práctica 2 - Regresión lineal (Parte 2)

Ejercicios teóricos (Optativos)

.

---

1. Sean $Y_i = \beta_o + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n$, donde los ϵ_i son independientes y $N(0, \sigma^2)$. Deducir el estadístico F para testear $H_o : \beta_o = 0$.

---

2. Dado que $\bar{x} = 0$, derive el test F para testear $H_o : \beta_o = \beta_1$ en el ejercicio anterior.

---

3. Sean $U_1, \ldots, U_n$ v.a. i.i.d. con distribución $N(\mu_1, \sigma^2)$ y $V_1, \ldots, V_n$ v.a. i.i.d. con distribución $N(\mu_2, \sigma^2)$ e independientes de las anteriores. Derive un test F para $H_o : \mu_1 = \mu_2$

---

4. Una serie de $n+1$ observaciones independientes $Y_i, i = 1, \ldots, n+1$ son tomadas de una población con distribución normal con varianza desconocida σ^2. Después de las n primeras observaciones se sospecha que ha habido un repentino cambio en la media de la distribución. Derive un test para testear la hipótesis de que la observación $(n+1)$–ésima tiene la misma media que las n anteriores.

---

5. Supongamos que queremos comparar la media de k poblaciones, para lo cual se toman muestras aleatorias independientes entre sí de tamaño J de cada una de la poblaciones. Sea Y_{ij} la j-ésima observación de la i-ésima población, $i = 1, \ldots, k$, $j = 1, \ldots, J$ y supongamos que $Y_{ij} \sim N(\mu_i, \sigma^2)$.

1. Supongamos que se plantea el modelo

$$Y_{ij} = \mu_i + \epsilon_{ij} \,,$$

   con $\epsilon_{ij} \sim N(0, \sigma^2)$.

2. Deduzca los estimadores de mínimos cuadrados de los parámetros y un test de nivel α para la hipótesis de que las medias poblacionales son iguales.

3. Deduzca intervalos de confianza para $\mu_i - \mu_j$, $1 \leq i < j \leq k$ de nivel global $1 - \alpha$.

Ejercicios prácticos

---

6. Los datos en el archivo `peak.txt` son datos simulados que corresponden al caudal de agua (Y) en distintas cuencas de ríos después de episodios de tormenta. Las variables independientes son:
X_1= área de la cuenca
X_2= área impermeable al agua
X_3= pendiente promedio del terreno
X_4= máxima longitud de los afluentes de la cuenca
X_5= índice de absorción del agua (0= absorción completa, 100= no absorción)
X_6= capacidad de depósito del suelo
X_7= velocidad de infiltración del agua en el suelo
X_8= cantidad de lluvia caída
X_9= tiempo durante el cual la cantidad de lluvia excedió 0.25 pulgada por hora

1. 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 a explicar la variación de Y. Si tuviera que usar una sola variable, ¿cuál usaría?

2. Calcular la matriz de correlación de $ln(Y)$ con el ln de cada una de las variables independientes. ¿Cómo cambian las correlaciones y sus conclusiones acerca de cuáles serían las variables que contribuyen significativamente a la variación de $ln(Y)$?

3. Use $ln(Y)$ como variable dependiente y los logaritmos de las variables independientes 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 que es 0. ¿ Cuáles son significativamente distintos de 0? ¿Cuáles son las variables que eliminaría en primera instancia para simplificar el modelo? ¿Es la regresión significativa?

4. Eliminar del modelo las variables que resultan menos interesantes y estimar nuevamente los parámetros. ¿Son todos los parámetros significativos con un nivel de 0.05? ¿Es la regresión significativa? Si no es asi, continúar eliminando aquellas variables menos importantes en cada paso y estimar los parámetros. Pare cuando todas las variables sean significativas . ¿Le parece que $\beta_o = 0$ tiene sentido en este ejemplo?

---

7. *(Para entregar)*

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.

1. 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.

2. 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 es 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?

3. 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?

4. Realizar un nuevo ajuste lineal usando las 5 variables independientes y eliminando la intercept. ¿Cómo se comparan estos resultados con los obtenidos anteriormente? ¿Cuáles son significativamente distintos de 0?

5. Plantear un nuevo modelo en el que intervengan aquellas variables que contribuyen significativamente y estimar los parámetros por mínimos cuadrados. ¿Qué modelo eligiría finalmente?

6. A partir de la estimación del error cuadrático medio, determinar si de todos los modelos planteados en el ejercicio, el sellecionado en el ítem anterior parece ser el mejor.

---

8. Con los datos del archivo `Vapor.txt`

1. Calcular, para cada punto del diseño, el intervalo de confianza de nivel 0.95 para la respuesta.
2. Calcular, para cada punto del diseño, el intervalo de predicción de nivel 0.95 para la respuesta.
3. Realizar en un mismo gráfico los pares de puntos (x,y), la recta de mínimos cuadrados y los límites de los intervalos obtenidos en a) y b) para cada punto del diseño.
4. Calcular, para cada punto del diseño, el intervalo de confianza para la respuesta de manera que el nivel global de los 25 intervalos obtenidos sea 0.95.
5. Graficar los pares de puntos (x,y), la recta de mínimos cuadrados y las curvas que se obtendrían si se unieran los límites superiores por un lado y los inferiores por otro, de los intervalos de confianza computados en d). Superponga la banda de confianza de nivel total 0.95 para la recta ajustada. ¿Cómo se interpretan estas curvas?

---

9. (*Opcional*)
Supongamos que queremos comparar k rectas de regresión dadas por

$$Y = \alpha_i + \beta_i x + \epsilon \quad i = 1, \ldots, k,$$

donde $E(\epsilon) = 0$ y $Var(\epsilon) = \sigma^2$. Para ello tomamos n_i pares (x_{ij}, y_{ij}), $j = 1, \ldots, n_i$ correspondientes a la i-ésima recta, $i = 1, \ldots, k$, de manera que

$$Y_{ij} = \alpha_i + \beta_i x_{ij} + \epsilon_{ij}$$

donde los ϵ_{ij} son independientes y con distribución $N(0, \sigma^2)$.

1. Encontrar una expresión matricial adecuada para plantear este problema.
2. Hallar los estimadores de mínimos cuadrados de los parámetros.
3. Supongamos que se desea testear la hipótesis de que las k rectas son paralelas. Expresar las hipótesis nula y alternativa para este problema y deducir un test de nivel α para decidir entre H_o y H_1.
4. Si al realizar el test planteado en c) se rechazara la hipótesis de que las rectas son paralelas, ¿tendría sentido tratar de identificar aquellos β_i que son diferentes? ¿Intervalos de confianza para qué combinación de los parámetros serían adecuados para detectar los β_i que difieren ? ¿Cuántos intervalos debe plantear? Deduzca los intervalos de confianza de manera tal que tengan un nivel global $1 - \alpha$. ¿Qué posibilidades tiene?

obs.	$x1j$	$y1j$	$x2j$	$y2j$
1	1.86	4.53	5.19	7.48
2	−1.65	−10.88	3.13	6.25
3	4.26	0.93	3.67	6.73
4	9.43	16.35	−0.15	1.09
5	3.09	−0.13	2.39	7.56
6	11.19	6.2	9.07	12.15
7	5.12	−0.93	6.44	8.32
8	5.04	3.76	2.3	4.92
9	0.69	−3.1	3.27	3.53
10	10.88	5.8	3.33	2.64
11	−1.16	0.81	6.19	8.6
12	0.96	−5.5	5.3	9.69
13	−4.77	−6.05	3.59	6.06
14	5.78	9.36	0.61	7.78
15	10.58	6.41	4.28	8.93
16	8.57	10.46	7.32	13.29
17	3.67	−0.36	7.63	11.43
18	6.25	1.59	6.75	10.37
19	9.6	11.59	6.78	7.93
20	6.57	7.29	1.77	5.63