eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

INSTRUCCIONES

1. Salvo que se indique lo contrario, las preguntas bien contestadas valen un punto. Puede haber más de una respuesta correcta, y para obtener puntuación has de señalarlas todas. Preguntas que no estén bien contestadas puntúan -0.5 veces su valor.

2. Intento medir conocimientos y no agudeza visual. Inevitablemente, en un examen de este tipo hay que prestar mucha atención. Cada curso hay personas que echan a perder una nota potencialmente buena por responder temeraria o atolondradamente.

¡Por favor, fíjate bien en todos los detalles!

3. Te ayudará proceder por exclusión de absurdos. Si una pregunta te parece ambigua, anota brevemente la razón al margen y no la contestes.

4. Al final, hay una Sección con unas pocas preguntas breves, que requieren cálculos no superiores a dos líneas: respóndelas directamente en el papel aparte que se te proporciona.

5. El tiempo previsto para el examen es de 1h 30'.

¡No pases la página hasta que se indique!

Estadística: Modelos Lineales

Final Enero 2.008, Tipo: A

Apellidos: _____

Nombre: _____

DNI: _____

Grupo: _____

Profesor : _____

Sección 1. Cuestiones de elección múltiple

1. Para evitar la mezcla de efectos y el tomar el valor de un estimador por lo que no es, si estamos interesados en la repercusión sobre Y de la variable X_4, haremos bien en limitarnos a estimar $Y = \beta_0 + \beta_4 X_4 + \epsilon$. La inclusión de variables adicionales es totalmente desaconsejable, salvo que sean ortogonales a X_4.
   (a) Cierto
   (b) Falso
2. La distancia de Cook proporciona:
   (a) Una medida global de la influencia de una observación sobre el conjunto de los β's.
   (b) Una medida individualizada de la influencia de cada observación sobre cada uno de los β's.
   (c) Una medida de bondad de ajuste alternativa a la R^2.
   (d) Todo falso.
3. La omisión de regresores que hubieran debido incluirse en una regresión tiene como consecuencia:
   (a) Sesgos en los estimadores de los $\hat{\beta}_i$.
   (b) En general, un sesgo a la alza de $\hat{\sigma}^2$.
   (c) En general, mayores problemas de multicolinealidad.
   (d) Una mayor pérdida de grados de libertad.
   (e) Todo falso.
4. El supuesto de normalidad de las perturbaciones es un requisito para que se verifique:
   (a) El teorema que garantiza que el estimador MCO puede ser mejorado en términos de ECM por otros estimadores (como el *ridge*).
   (b) El teorema de Gauss-Markov.
   (c) Que los estimadores MCO son insesgados.
   (d) Que los estimadores MCO son óptimos en el sentido mínimo-cuadrático.
   (e) Todo falso.

5. El modelo de regresión lineal permite cuando se verifican los supuestos necesarios:
   (a) Establecer relaciones de causalidad desde una (o varias) variables X (regresores) hacia una variable Y (respuesta).
   (b) Decidir si una proyección es lineal.
   (c) Hacer predicciones sobre los valores de los regresores.
   (d) Contrastar hipótesis acerca de la existencia (o no) de relación lineal entre los regresores y la respuesta.
6. Si estuviéramos radicalmente en contra de introducir ningún sesgo en las estimaciones de los β's, no haríamos en ningún caso:
   (a) Regresión ridge.
   (b) Regresión en componentes principales.
   (c) Regresión stepwise.
   (d) Regresión *all subsets* o sobre todos los subconjuntos posibles de regresores.
   (e) Todo falso.
7. Indica cuál o cuales de los modelos que siguen *no* son reconducibles a un modelo lineal que pueda estimarse con la teoría estudiada en el curso:
   (a) $Y = X_1\beta_1 + X_2\beta_2 + \epsilon$
   (b) $Y = X_1^{\beta_1} + X_2^{\beta_2} + \epsilon$
   (c) $Y = X_1^{\beta_1} X_2^{\beta_2} \epsilon$
   (d) $\log(Y) = X_1\beta_1 + \log(X_2)\beta_2 + \epsilon$
   (e) Todo falso.

Comienzo de un bloque de preguntas

Las preguntas hasta el siguiente trazo horizontal hacen referencia a los datos que siguen, referentes a un experimento realizado para investigar la posibilidad de provocar lluvia. El procedimiento consiste en lanzar desde un avión en mitad de las nubes bengalas de yoduro de plata, que supuestamente podrían favorecer la precipitación del vapor de agua en la nube. Los datos resumidos pueden verse a continuación:

```
> library(MASS)
> nubes <- dget("nubes.dge")
> str(nubes)

'data.frame':        216 obs. of  7 variables:
 $ PERIOD: int  1 2 3 4 5 6 7 8 9 10 ...
 $ SEEDED: Factor w/ 2 levels "S","U": 1 2 1 2 1 2 2 1 2 1 ...
 $ SEASON: Factor w/ 4 levels "AUTUMN","SPRING",..: 1 1 4 4 4 4 4 4 2 2 ...
 $ TE    : num  1.69 0.74 0.81 1.44 2.48 0.84 0.37 0.37 1.33 3.38 ...
 $ NC    : num  1.65 1.09 2.39 2.96 4.16 2.76 1.08 0.26 2.53 2.76 ...
 $ SC    : num  1.8 0.79 0.36 1.27 2.16 0.87 0.85 0.47 1.08 3.1 ...
 $ NWC   : num  3.33 1.59 2.06 4.05 6 4.17 3.45 0.9 3.65 5.06 ...
```

```
> nubes[1:3, ]

  PERIOD SEEDED SEASON   TE   NC   SC  NWC
1      1      S AUTUMN 1.69 1.65 1.80 3.33
2      2      U AUTUMN 0.74 1.09 0.79 1.59
3      3      S WINTER 0.81 2.39 0.36 2.06

> attach(nubes)

        The following object(s) are masked from nubes ( position 3 ) :

         NC NWC PERIOD SC SEASON SEEDED TE
```

Los significados de las variables se recogen a continuación:

Variable	Tipo	Significado
PERIOD	Numérica	Periodo del experimento
SEEDED	Cualitativa	U = "unseeded", nube no tratada.
		S = "seeded", nube tratada.
SEASON	Cualitativa	Estación: AUTUMN=Otoño, SPRING=Primavera,
		SUMMER=Verano, WINTER=Invierno.
TE	Numérica	Precipitación área tratada.
NC	Numérica	Precipitación área contigua Norte.
SC	Numérica	Precipitación área contigua Sur.
NWC	Numérica	Precipitación área contigua Noroeste.

Ajustamos un modelo con la variable `TE` como respuesta y `SEEDED` como único regresor y obtenemos lo siguiente:

```
> mod1 <- lm(TE ~ SEEDED, data = nubes)
> summary(mod1)

Call:
lm(formula = TE ~ SEEDED, data = nubes)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6631 -0.9638 -0.3959  0.5919  5.3541

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5759     0.1324  11.907   <2e-16 ***
SEEDEDU       0.1271     0.1872   0.679    0.498
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.375 on 214 degrees of freedom
Multiple R-Squared: 0.002151,        Adjusted R-squared: -0.002512
F-statistic: 0.4613 on 1 and 214 DF,  p-value: 0.4977
```

8. ¿Cuál es tu conclusión preliminar a la vista de estos resultados?
   (a) La variable SEEDED no parece tener nada que ver con la precipitación.
   (b) La variable SEEDEDU es claramente no significativa, pero el otro coeficiente no reportado, SEEDEDS, podría serlo.
   (c) Para contrastar la hipótesis de interés —influencia de la inseminación en la precipitación— habría que hacer un contraste Q_h.
   (d) Todo falso.
9. Parece ser que...
   (a) A pesar de ser `SEEDEDU` no significativa ($\alpha = 0{,}05$), la regresión en su conjunto es significativa a dicho nivel: esto avala la creencia de que el `SEEDEDS` si es significativa.
   (b) La regresión en su conjunto no es significativa, como resultaba de esperar a la vista de los resultados precedentes.
   (c) Se ha producido un error, quizá por presencia de fuerte multicolinealidad. De otro modo, $\overline{R}^2$ (`Adjusted R-square`) no podría tomar el valor que toma.
   (d) Todo falso.

Procedemos a continuación a ajustar un modelo que incluya todas las variables disponibles:

```
> mod2 <- lm(TE ~ ., data = nubes)
> summary(mod2)

Call:
lm(formula = TE ~ ., data = nubes)

Residuals:
     Min       1Q   Median       3Q      Max
-2.03368 -0.44510 -0.08008  0.41227  2.35100

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1620997  0.1650393  -0.982    0.327
PERIOD        0.0010206  0.0008839   1.155    0.250
SEEDEDU      -0.0840900  0.1013870  -0.829    0.408
SEASONSPRING  0.1156252  0.1411977   0.819    0.414
SEASONSUMMER  0.1856097  0.1565883   1.185    0.237
SEASONWINTER -0.2092056  0.1490409  -1.404    0.162
NC            0.0732731  0.0691289   1.060    0.290
SC            0.6999560  0.0822030   8.515 3.41e-15 ***
NWC           0.3501814  0.0675418   5.185 5.13e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7379 on 207 degrees of freedom
Multiple R-Squared: 0.7222,        Adjusted R-squared: 0.7115
F-statistic: 67.27 on 8 and 207 DF,  p-value: < 2.2e-16
```

10. La interpretación del coeficiente de la variable `SEASONWINTER` es...

    (a) Que la precipitación por periodo en las experiencias realizadas en invierno fue, $-0{,}2092$ menor que en otoño (AUTUMN), *una vez tenida en cuenta la influencia de las restantes variables.*

    (b) Que la precipitación por periodo en las experiencias realizadas en invierno fue, en promedio, $-0{,}2092$ menor que en las experiencias restantes.

    (c) Que la precipitación por periodo en las experiencias realizadas en invierno fue, en promedio, $-0{,}2092$ menor que en otoño (AUTUMN), que es aquí el nivel de referencia.

    (d) Todo falso.

11. La $\overline{R}^2$ (`Adjusted R-square`) es menor que la R^2. Ello es indicativo:

    (a) De nada: ocurre siempre.

    (b) De la baja calidad del ajuste.

    (c) De la presencia de multicolinealidad.

    (d) Todo falso.

12. El estadístico Q_h para contrastar la significación conjunta de la regresión toma un valor 67.27, con un p-value `<2.2e-16`. Ello quiere decir que:

    (a) Obtener una regresión como la obtenida, si realmente los regresores no ayudaran a predecir la variable respuesta, tiene una probabilidad ridiculamente baja. Hay que concluir que los regresores en su conjunto ayudaan a predecir la respuesta.

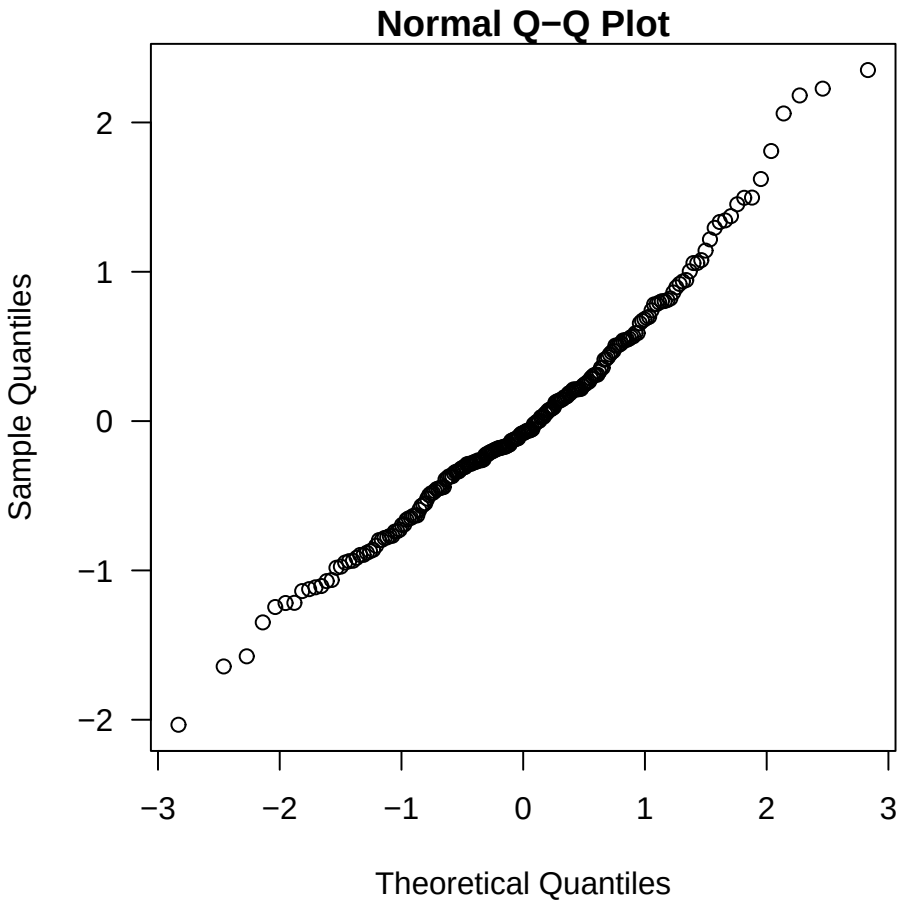
    (b) Que la regresión no es significativa, y resultados similares podrían haberse obtenido por puro azar.

    (c) Que la pendiente de la recta de regresión es casi cero.

    (d) Todo falso.

13. Realizando un qqplot de los residuos, obtenemos:

```
> qqnorm(mod2$residuals)
```



El resultado sugiere que:

(a) Los residuos siguen una distribución no muy separada de la normal.

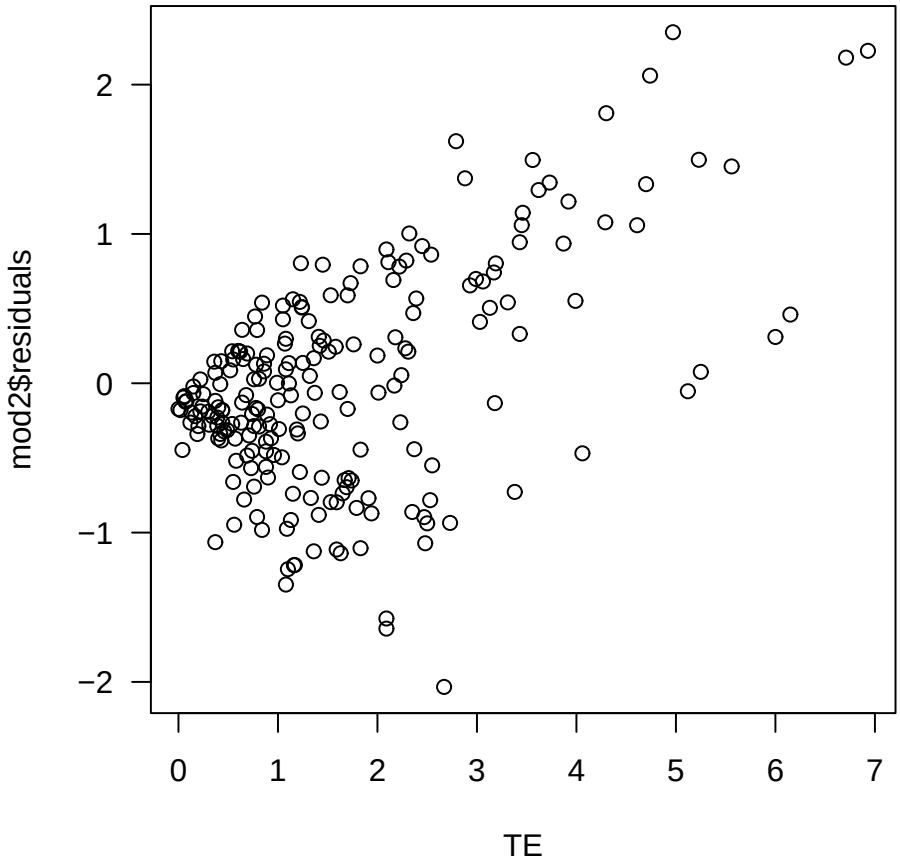
(b) El supuesto de linealidad de la regresión es correcto.

(c) La R^2 es cercana a uno, dado que los puntos se sitúan aproximadamente sobre la bisectriz del primer cuadrante (cuya pendiente es 1).

(d) Todo falso.

14. Dibujando ahora los residuos frente a los valores de la variable respuesta `TE` (= precipitación en áreas tratadas)

```
> plot(TE, mod2$residuals)
```



vemos claramente que:

(a) Las observaciones con elevada `TE` han sido sistemáticamente mal ajustadas (el modelo predice valores *por debajo de los observados*).

(b) Las observaciones con elevada `TE` han sido sistemáticamente mal ajustadas (el modelo predice valores *por encima de los observados*).

(c) Claramente, hace falta un término que recoja una tendencia cuadrática.

(d) La varianza de los residuos es claramente creciente.

(e) Todo falso.

15. Examinemos ahora los residuos borrados. Una de las muchas manera de obtenerlos en R consiste en obtener primero la diagonal de la matriz de proyección $X(X'X)^{-1}X'$,

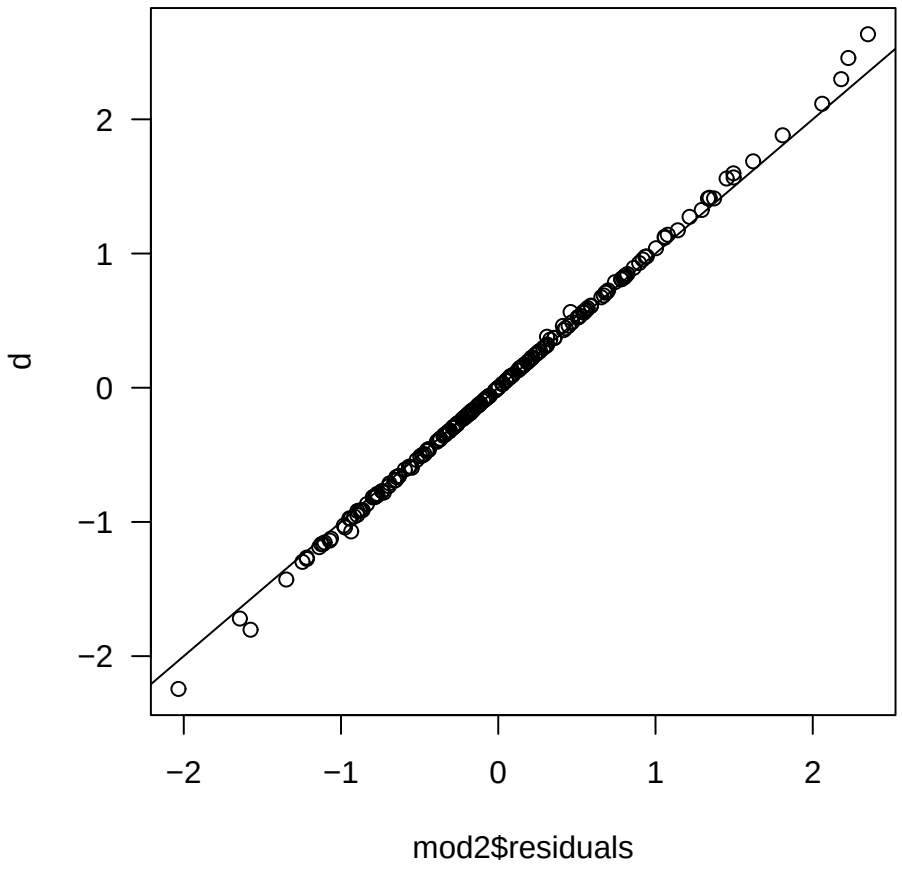
```
> diagP <- lm.influence(mod2)$hat
```

y a continuación calcular los residuos borrados así:

```
> d <- mod2$residuals/(1 - diagP)
```

Si dibujamos residuos borrados frente a residuos ordinarios,

```
> plot(mod2$residuals, d)
> abline(a = 0, b = 1)
```



aquéllas observaciones en que los puntos se separan más de la bisectriz del primer cuadrante ($\Longrightarrow$ residuo borrado muy diferente del residuo ordinario) sugieren:

(a) Que estamos ante una observación relativamente influyente.

(b) Que estamos ante una observación mal ajustada.

(c) Que estamos ante un fallo del supuesto de normalidad, y en presencia de una distribución de colas "gordas".

(d) Todo falso.

Podríamos ayudarnos de un criterio como AIC para seleccionar un modelo. Supón que hacemos:

```
> mod3 <- stepAIC(mod2, trace = FALSE)
> summary(mod3)

Call:
lm(formula = TE ~ SEASON + SC + NWC, data = nubes)

Residuals:
     Min       1Q   Median       3Q      Max
-2.04651 -0.39348 -0.05817  0.43377  2.47024

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.12309    0.13211  -0.932    0.353
SEASONSPRING  0.10853    0.14109   0.769    0.443
SEASONSUMMER  0.19236    0.15453   1.245    0.215
SEASONWINTER -0.20785    0.14823  -1.402    0.162
SC            0.72693    0.08057   9.022  < 2e-16 ***
NWC           0.39569    0.04667   8.478 4.05e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7379 on 210 degrees of freedom
Multiple R-Squared: 0.7182,        Adjusted R-squared: 0.7115
F-statistic:   107 on 5 and 210 DF,  p-value: < 2.2e-16
```

y al mejor de los modelos seleccionados le añadimos la variable SEEDED:

```
> mod4 <- lm(TE ~ SEASON + SC + NWC + SEEDED, data = nubes)
> summary(mod4)

Call:
lm(formula = TE ~ SEASON + SC + NWC + SEEDED, data = nubes)

Residuals:
     Min       1Q   Median       3Q      Max
-2.09632 -0.40753 -0.06112  0.43374  2.44094

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.09053    0.13911  -0.651    0.516
SEASONSPRING  0.10880    0.14124   0.770    0.442
SEASONSUMMER  0.19539    0.15474   1.263    0.208
SEASONWINTER -0.21032    0.14842  -1.417    0.158
SC            0.72490    0.08070   8.983  < 2e-16 ***
NWC           0.39931    0.04696   8.502 3.54e-15 ***
SEEDEDU      -0.07634    0.10120  -0.754    0.451
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7387 on 209 degrees of freedom
Multiple R-Squared: 0.719,        Adjusted R-squared: 0.7109
F-statistic: 89.11 on 6 and 209 DF,  p-value: < 2.2e-16
```

16. Tu conclusión a la vista de los resultados es:
    (a) El tratamiento de nubes experimentado parece ineficaz para incrementar la precipitación.
    (b) El tratamiento de nubes experimentado parece eficaz para incrementar la precipitación.
    (c) Todo falso.

FINAL DE UN BLOQUE DE PREGUNTAS

---

17. A menudo seleccionamos un modelo de regresión minimizando o maximizando un criterio. ¿Cuál de entre los siguientes es máximamente favorable a la inclusión de nuevos regresores?
    (a) C_p de Mallows.
    (b) AIC de Akaike.
    (c) $\overline{R}^2$
    (d) R^2
    (e) Todo falso.
18. Un modelo ANOVA con un único tratamiento de cinco niveles y con replicación 3, tendrá:
    (a) 15 grados de libertad.
    (b) 12 grados de libertad.
    (c) 10 grados de libertad.
    (d) 13 grados de libertad.
    (e) Todo falso.
19. Supongamos un modelo con 20 regresores. Si para contrastar la hipótesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{20} = 0$ seguimos el criterio de rechazarla siempre que algún $\hat{\beta}$ tiene un p-value asociado menor que 0.05, la probabilidad de rechazar indebidamente H_0 será:
    (a) Mayor que 0.05.
    (b) Exactamente igual a 0.05.
    (c) Menor que 0.05.
    (d) Quizá tan grande como $1 - (20 \times 0{,}05)$, pero no más.
    (e) Todo falso.

20. Cuando imponemos un restricción lineal sobre los parámetros:
    (a) Podemos estar incrementando el sesgo de los $\hat{\beta}_i$.
    (b) Podemos estar incrementando las varianzas de los $\hat{\beta}_i$.
    (c) Necesariamente disminuye $\hat{\sigma}^2$.
    (d) Nunca disminuye SSE.
    (e) Todo falso.

Sección 2. Preguntas breves

Responde a *una sola* pregunta de entre las dos siguientes:

1. Enuncia y demuestra el teorema de Gauss-Markov. (5 puntos.)

2. En general, si estimamos el modelo
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
y a continuación el modelo ampliado
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$
los estimadores de los parámetros $\boldsymbol{\beta}$ no son iguales en ambos modelos. No obstante, los estimadores de $\boldsymbol{\beta}$ *si* son los mismos en el caso particular de que $\boldsymbol{X}$ y $\boldsymbol{Z}$ tengan todas sus columnas mutuamente ortogonales. Demuéstralo. (3 puntos.)

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

INSTRUCCIONES

1. Salvo que se indique lo contrario, las preguntas bien contestadas valen un punto. Puede haber más de una respuesta correcta, y para obtener puntuación has de señalarlas todas. Preguntas que no estén bien contestadas puntúan -0.5 veces su valor.

2. Intento medir conocimientos y no agudeza visual. Inevitablemente, en un examen de este tipo hay que prestar mucha atención. Cada curso hay personas que echan a perder una nota potencialmente buena por responder temeraria o atolondradamente.

¡Por favor, fíjate bien en todos los detalles!

3. Te ayudará proceder por exclusión de absurdos. Si una pregunta te parece ambigua, anota brevemente la razón al margen y no la contestes.

4. Al final, hay una Sección con unas pocas preguntas breves, que requieren cálculos no superiores a dos líneas: respóndelas directamente en el papel aparte que se te proporciona.

5. El tiempo previsto para el examen es de 1h 30'.

¡No pases la página hasta que se indique!

Respuestas al examen de tipo A

Sección 1. Cuestiones de elección múltiple

1. Para evitar la mezcla de efectos y el tomar el valor de un estimador por lo que no es, si estamos interesados en la repercusión sobre Y de la variable X_4, haremos bien en limitarnos a estimar $Y = \beta_0 + \beta_4 X_4 + \epsilon$. La inclusión de variables adicionales es totalmente desaconsejable, salvo que sean ortogonales a X_4.
   (a) Cierto
   (b) Falso
2. La distancia de Cook proporciona:
   (a) Una medida global de la influencia de una observación sobre el conjunto de los β's.
   (b) Una medida individualizada de la influencia de cada observación sobre cada uno de los β's.
   (c) Una medida de bondad de ajuste alternativa a la R^2.
   (d) Todo falso.
3. La omisión de regresores que hubieran debido incluirse en una regresión tiene como consecuencia:
   (a) Sesgos en los estimadores de los $\hat{\beta}_i$.
   (b) En general, un sesgo a la alza de $\hat{\sigma}^2$.
   (c) En general, mayores problemas de multicolinealidad.
   (d) Una mayor pérdida de grados de libertad.
   (e) Todo falso.
4. El supuesto de normalidad de las perturbaciones es un requisito para que se verifique:
   (a) El teorema que garantiza que el estimador MCO puede ser mejorado en términos de ECM por otros estimadores (como el *ridge*).
   (b) El teorema de Gauss-Markov.
   (c) Que los estimadores MCO son insesgados.
   (d) Que los estimadores MCO son óptimos en el sentido mínimo-cuadrático.
   (e) Todo falso.
5. El modelo de regresión lineal permite cuando se verifican los supuestos necesarios:
   (a) Establecer relaciones de causalidad desde una (o varias) variables X (regresores) hacia una variable Y (respuesta).
   (b) Decidir si una proyección es lineal.
   (c) Hacer predicciones sobre los valores de los regresores.
   (d) Contrastar hipótesis acerca de la existencia (o no) de relación lineal entre los regresores y la respuesta.

6. Si estuviéramos radicalmente en contra de introducir ningún sesgo en las estimaciones de los β's, no haríamos en ningún caso:

   (a) Regresión ridge.

   (b) Regresión en componentes principales.

   (c) Regresión stepwise.

   (d) Regresión *all subsets* o sobre todos los subconjuntos posibles de regresores.

   (e) Todo falso.

7. Indica cuál o cuales de los modelos que siguen *no* son reconducibles a un modelo lineal que pueda estimarse con la teoría estudiada en el curso:

   (a) $Y = X_1\beta_1 + X_2\beta_2 + \epsilon$

   (b) $Y = X_1^{\beta_1} + X_2^{\beta_2} + \epsilon$

   (c) $Y = X_1^{\beta_1} X_2^{\beta_2} \epsilon$

   (d) $\log(Y) = X_1\beta_1 + \log(X_2)\beta_2 + \epsilon$

   (e) Todo falso.

Comienzo de un bloque de preguntas

Las preguntas hasta el siguiente trazo horizontal hacen referencia a los datos que siguen, referentes a un experimento realizado para investigar la posibilidad de provocar lluvia. El procedimiento consiste en lanzar desde un avión en mitad de las nubes bengalas de yoduro de plata, que supuestamente podrían favorecer la precipitación del vapor de agua en la nube. Los datos resumidos pueden verse a continuación:

```
> library(MASS)
> nubes <- dget("nubes.dge")
> str(nubes)

'data.frame':        216 obs. of  7 variables:
 $ PERIOD: int  1 2 3 4 5 6 7 8 9 10 ...
 $ SEEDED: Factor w/ 2 levels "S","U": 1 2 1 2 1 2 2 1 2 1 ...
 $ SEASON: Factor w/ 4 levels "AUTUMN","SPRING",..: 1 1 4 4 4 4 4 4 2 2 ...
 $ TE    : num  1.69 0.74 0.81 1.44 2.48 0.84 0.37 0.37 1.33 3.38 ...
 $ NC    : num  1.65 1.09 2.39 2.96 4.16 2.76 1.08 0.26 2.53 2.76 ...
 $ SC    : num  1.8 0.79 0.36 1.27 2.16 0.87 0.85 0.47 1.08 3.1 ...
 $ NWC   : num  3.33 1.59 2.06 4.05 6 4.17 3.45 0.9 3.65 5.06 ...

> nubes[1:3, ]

  PERIOD SEEDED SEASON   TE   NC   SC  NWC
1      1      S AUTUMN 1.69 1.65 1.80 3.33
2      2      U AUTUMN 0.74 1.09 0.79 1.59
3      3      S WINTER 0.81 2.39 0.36 2.06

> attach(nubes)
```

```
The following object(s) are masked from nubes ( position 3 ) :

 NC NWC PERIOD SC SEASON SEEDED TE
```

Los significados de las variables se recogen a continuación:

Variable	Tipo	Significado
PERIOD	Numérica	Periodo del experimento
SEEDED	Cualitativa	U = "unseeded", nube no tratada.
		S = "seeded", nube tratada.
SEASON	Cualitativa	Estación: AUTUMN=Otoño, SPRING=Primavera,
		SUMMER=Verano, WINTER=Invierno.
TE	Numérica	Precipitación área tratada.
NC	Numérica	Precipitación área contigua Norte.
SC	Numérica	Precipitación área contigua Sur.
NWC	Numérica	Precipitación área contigua Noroeste.

Ajustamos un modelo con la variable `TE` como respuesta y `SEEDED` como único regresor y obtenemos lo siguiente:

```
> mod1 <- lm(TE ~ SEEDED, data = nubes)
> summary(mod1)

Call:
lm(formula = TE ~ SEEDED, data = nubes)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6631 -0.9638 -0.3959  0.5919  5.3541

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5759     0.1324  11.907   <2e-16 ***
SEEDEDU       0.1271     0.1872   0.679    0.498
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.375 on 214 degrees of freedom
Multiple R-Squared: 0.002151,        Adjusted R-squared: -0.002512
F-statistic: 0.4613 on 1 and 214 DF,  p-value: 0.4977
```

8. ¿Cuál es tu conclusión preliminar a la vista de estos resultados?

   (a) La variable SEEDED no parece tener nada que ver con la precipitación.

   (b) La variable SEEDEDU es claramente no significativa, pero el otro coeficiente no reportado, SEEDEDS, podría serlo.

   (c) Para contrastar la hipótesis de interés —influencia de la inseminación en la precipitación— habría que hacer un contraste Q_h.

   (d) Todo falso.

9. Parece ser que...

   (a) A pesar de ser `SEEDEDU` no significativa ($\alpha = 0{,}05$), la regresión en su conjunto es significativa a dicho nivel: esto avala la creencia de que el `SEEDEDS` si es significativa.

   (b) La regresión en su conjunto no es significativa, como resultaba de esperar a la vista de los resultados precedentes.

   (c) Se ha producido un error, quizá por presencia de fuerte multicolinealidad. De otro modo, $\overline{R}^2$ (`Adjusted R-square`) no podría tomar el valor que toma.

   (d) Todo falso.

Procedemos a continuación a ajustar un modelo que incluya todas las variables disponibles:

```
> mod2 <- lm(TE ~ ., data = nubes)
> summary(mod2)

Call:
lm(formula = TE ~ ., data = nubes)

Residuals:
     Min       1Q   Median       3Q      Max
-2.03368 -0.44510 -0.08008  0.41227  2.35100

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1620997  0.1650393  -0.982    0.327
PERIOD        0.0010206  0.0008839   1.155    0.250
SEEDEDU      -0.0840900  0.1013870  -0.829    0.408
SEASONSPRING  0.1156252  0.1411977   0.819    0.414
SEASONSUMMER  0.1856097  0.1565883   1.185    0.237
SEASONWINTER -0.2092056  0.1490409  -1.404    0.162
NC            0.0732731  0.0691289   1.060    0.290
SC            0.6999560  0.0822030   8.515 3.41e-15 ***
NWC           0.3501814  0.0675418   5.185 5.13e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7379 on 207 degrees of freedom
Multiple R-Squared: 0.7222,        Adjusted R-squared: 0.7115
F-statistic: 67.27 on 8 and 207 DF,  p-value: < 2.2e-16
```

10. La interpretación del coeficiente de la variable `SEASONWINTER` es...

    (a) Que la precipitación por periodo en las experiencias realizadas en invierno fue, $-0{,}2092$ menor que en otoño (AUTUMN), *una vez tenida en cuenta la influencia de las restantes variables.*

    (b) Que la precipitación por periodo en las experiencias realizadas en invierno fue, en promedio, $-0{,}2092$ menor que en las experiencias restantes.

    (c) Que la precipitación por periodo en las experiencias realizadas en invierno fue, en promedio, $-0{,}2092$ menor que en otoño (AUTUMN), que es aquí el nivel de referencia.

    (d) Todo falso.

11. La $\overline{R}^2$ (`Adjusted R-square`) es menor que la R^2. Ello es indicativo:

    (a) De nada: ocurre siempre.

    (b) De la baja calidad del ajuste.

    (c) De la presencia de multicolinealidad.

    (d) Todo falso.

12. El estadístico Q_h para contrastar la significación conjunta de la regresión toma un valor 67.27, con un p-value `<2.2e-16`. Ello quiere decir que:

    (a) Obtener una regresión como la obtenida, si realmente los regresores no ayudaran a predecir la variable respuesta, tiene una probabilidad ridiculamente baja. Hay que concluir que los regresores en su conjunto ayudaan a predecir la respuesta.

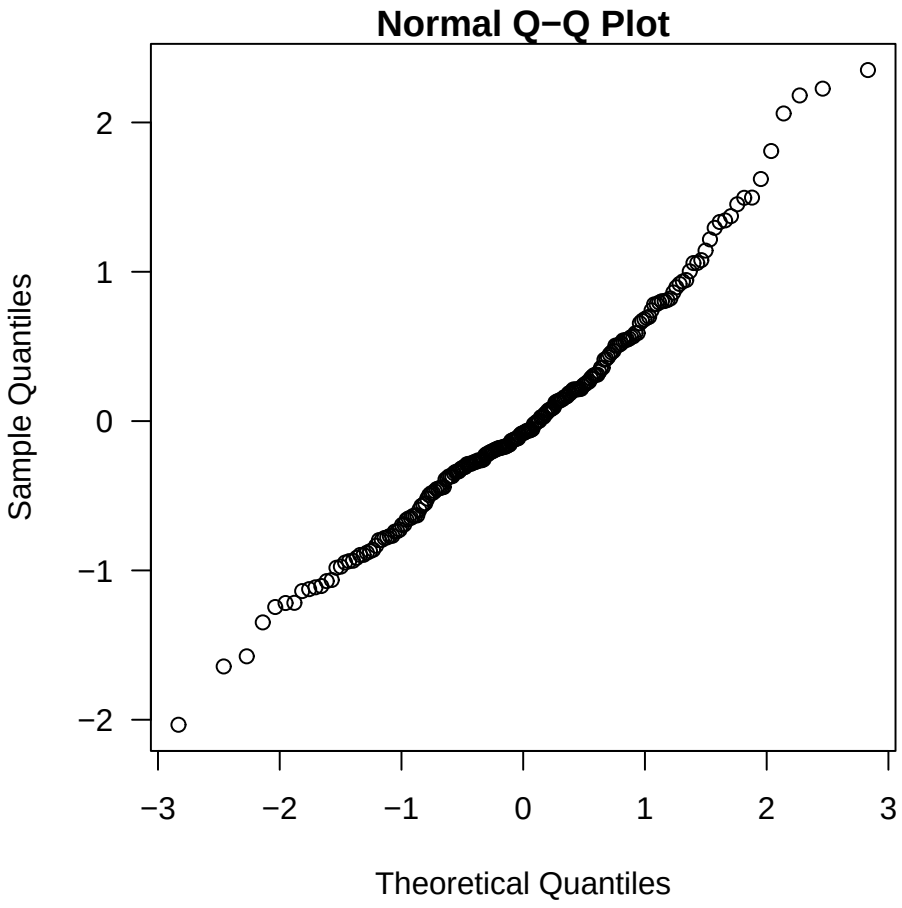
    (b) Que la regresión no es significativa, y resultados similares podrían haberse obtenido por puro azar.

    (c) Que la pendiente de la recta de regresión es casi cero.

    (d) Todo falso.

13. Realizando un qqplot de los residuos, obtenemos:

```
> qqnorm(mod2$residuals)
```



El resultado sugiere que:

(a) Los residuos siguen una distribución no muy separada de la normal.

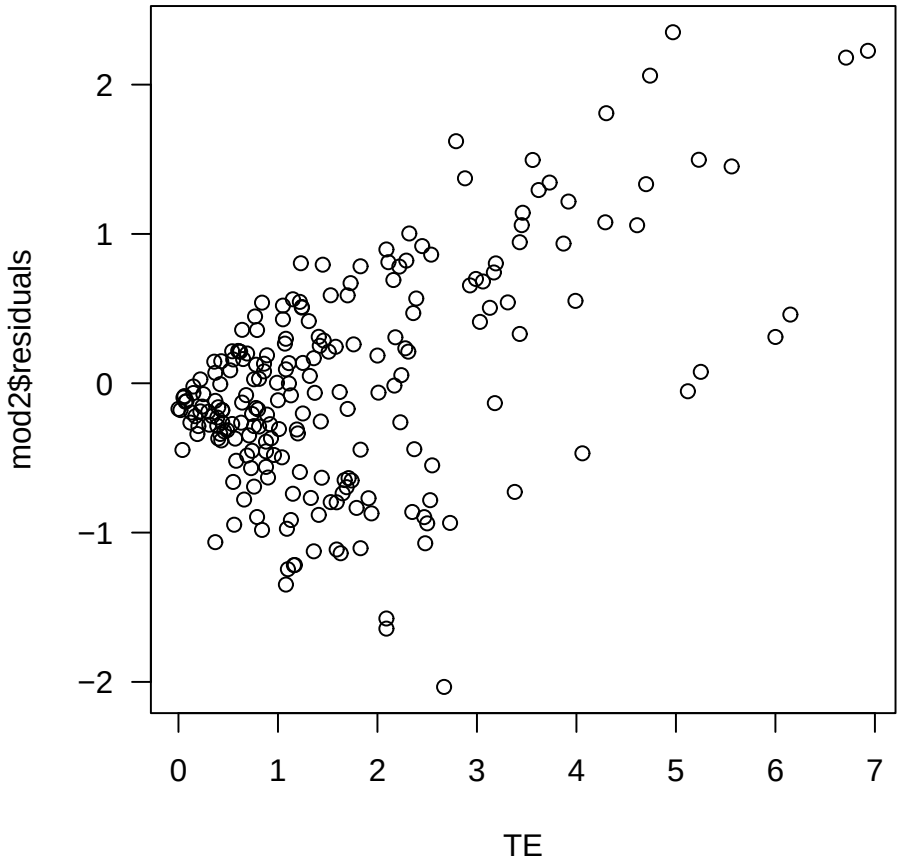
(b) El supuesto de linealidad de la regresión es correcto.

(c) La R^2 es cercana a uno, dado que los puntos se sitúan aproximadamente sobre la bisectriz del primer cuadrante (cuya pendiente es 1).

(d) Todo falso.

14. Dibujando ahora los residuos frente a los valores de la variable respuesta `TE` (= precipitación en áreas tratadas)

```
> plot(TE, mod2$residuals)
```



vemos claramente que:

(a) Las observaciones con elevada `TE` han sido sistemáticamente mal ajustadas (el modelo predice valores *por debajo de los observados*).

(b) Las observaciones con elevada `TE` han sido sistemáticamente mal ajustadas (el modelo predice valores *por encima de los observados*).

(c) Claramente, hace falta un término que recoja una tendencia cuadrática.

(d) La varianza de los residuos es claramente creciente.

(e) Todo falso.

15. Examinemos ahora los residuos borrados. Una de las muchas manera de obtenerlos en R consiste en obtener primero la diagonal de la matriz de proyección $X(X'X)^{-1}X'$,

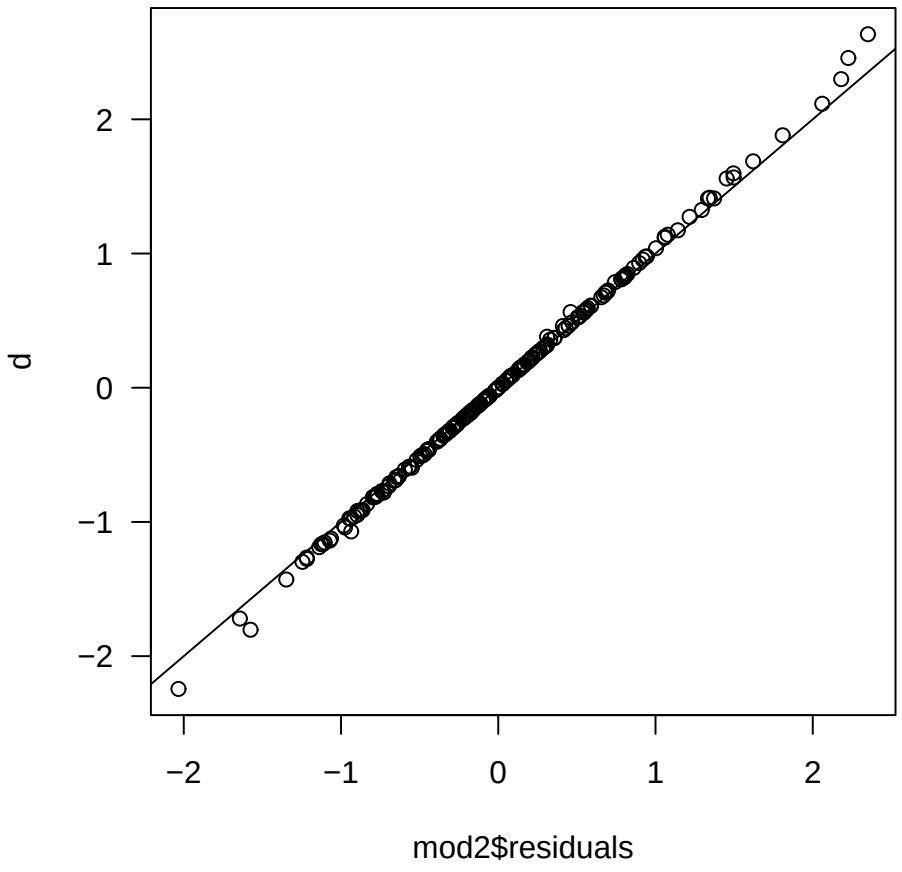
```
> diagP <- lm.influence(mod2)$hat
```

y a continuación calcular los residuos borrados así:

```
> d <- mod2$residuals/(1 - diagP)
```

Si dibujamos residuos borrados frente a residuos ordinarios,

```
> plot(mod2$residuals, d)
> abline(a = 0, b = 1)
```



aquéllas observaciones en que los puntos se separan más de la bisectriz del primer cuadrante ($\Longrightarrow$ residuo borrado muy diferente del residuo ordinario) sugieren:

(a) Que estamos ante una observación relativamente influyente.

(b) Que estamos ante una observación mal ajustada.

(c) Que estamos ante un fallo del supuesto de normalidad, y en presencia de una distribución de colas "gordas".

(d) Todo falso.

Podríamos ayudarnos de un criterio como AIC para seleccionar un modelo. Supón que hacemos:

```
> mod3 <- stepAIC(mod2, trace = FALSE)
> summary(mod3)

Call:
lm(formula = TE ~ SEASON + SC + NWC, data = nubes)

Residuals:
     Min       1Q   Median       3Q      Max
-2.04651 -0.39348 -0.05817  0.43377  2.47024

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.12309    0.13211  -0.932    0.353
SEASONSPRING  0.10853    0.14109   0.769    0.443
SEASONSUMMER  0.19236    0.15453   1.245    0.215
SEASONWINTER -0.20785    0.14823  -1.402    0.162
SC            0.72693    0.08057   9.022  < 2e-16 ***
NWC           0.39569    0.04667   8.478 4.05e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7379 on 210 degrees of freedom
Multiple R-Squared: 0.7182,        Adjusted R-squared: 0.7115
F-statistic:   107 on 5 and 210 DF,  p-value: < 2.2e-16
```

y al mejor de los modelos seleccionados le añadimos la variable SEEDED:

```
> mod4 <- lm(TE ~ SEASON + SC + NWC + SEEDED, data = nubes)
> summary(mod4)

Call:
lm(formula = TE ~ SEASON + SC + NWC + SEEDED, data = nubes)

Residuals:
     Min       1Q   Median       3Q      Max
-2.09632 -0.40753 -0.06112  0.43374  2.44094

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.09053    0.13911  -0.651    0.516
SEASONSPRING  0.10880    0.14124   0.770    0.442
SEASONSUMMER  0.19539    0.15474   1.263    0.208
SEASONWINTER -0.21032    0.14842  -1.417    0.158
SC            0.72490    0.08070   8.983  < 2e-16 ***
NWC           0.39931    0.04696   8.502 3.54e-15 ***
SEEDEDU      -0.07634    0.10120  -0.754    0.451
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7387 on 209 degrees of freedom
Multiple R-Squared: 0.719,        Adjusted R-squared: 0.7109
F-statistic: 89.11 on 6 and 209 DF,  p-value: < 2.2e-16
```

16. Tu conclusión a la vista de los resultados es:

    (a) El tratamiento de nubes experimentado parece ineficaz para incrementar la precipitación.

    (b) El tratamiento de nubes experimentado parece eficaz para incrementar la precipitación.

    (c) Todo falso.

FINAL DE UN BLOQUE DE PREGUNTAS

---

17. A menudo seleccionamos un modelo de regresión minimizando o maximizando un criterio. ¿Cuál de entre los siguientes es máximamente favorable a la inclusión de nuevos regresores?

    (a) C_p de Mallows.

    (b) AIC de Akaike.

    (c) $\overline{R}^2$

    (d) R^2

    (e) Todo falso.

18. Un modelo ANOVA con un único tratamiento de cinco niveles y con replicación 3, tendrá:

    (a) 15 grados de libertad.

    (b) 12 grados de libertad.

    (c) 10 grados de libertad.

    (d) 13 grados de libertad.

    (e) Todo falso.

19. Supongamos un modelo con 20 regresores. Si para contrastar la hipótesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{20} = 0$ seguimos el criterio de rechazarla siempre que algún $\hat{\beta}$ tiene un p-value asociado menor que 0.05, la probabilidad de rechazar indebidamente H_0 será:

    (a) Mayor que 0.05.

    (b) Exactamente igual a 0.05.

    (c) Menor que 0.05.

    (d) Quizá tan grande como $1 - (20 \times 0{,}05)$, pero no más.

    (e) Todo falso.

20. Cuando imponemos un restricción lineal sobre los parámetros:
    (a) Podemos estar incrementando el sesgo de los $\hat{\beta}_i$.
    (b) Podemos estar incrementando las varianzas de los $\hat{\beta}_i$.
    (c) Necesariamente disminuye $\hat{\sigma}^2$.
    (d) Nunca disminuye SSE.
    (e) Todo falso.

Sección 2. Preguntas breves

Responde a *una sola* pregunta de entre las dos siguientes:

1. Enuncia y demuestra el teorema de Gauss-Markov. (5 puntos.)

   Respuesta: Ver apuntes de clase.

2. En general, si estimamos el modelo
   $$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
   y a continuación el modelo ampliado
   $$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$
   los estimadores de los parámetros $\boldsymbol{\beta}$ no son iguales en ambos modelos. No obstante, los estimadores de $\boldsymbol{\beta}$ *si* son los mismos en el caso particular de que $\boldsymbol{X}$ y $\boldsymbol{Z}$ tengan todas sus columnas mutuamente ortogonales. Demuéstralo. (3 puntos.)

   Respuesta: Ver apuntes de clase.