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 2.009, Tipo: A

Apellidos: _____

Nombre: _____

DNI: _____

Grupo: _____

Profesor : _____

Sección 1. Cuestiones de elección múltiple

1. Sy $P_M \boldsymbol{y}$ denota la proyección sobre un espacio vectorial M del vector $\boldsymbol{y}$,
   (a) $P_M \boldsymbol{y}$ es ortogonal a $\boldsymbol{y}$.
   (b) $(\boldsymbol{y} - P_M \boldsymbol{y})$ es ortogonal a $\boldsymbol{y}$.
   (c) $(\boldsymbol{y} - P_M \boldsymbol{y})$ es ortogonal a M.
   (d) $\boldsymbol{y}$ es ortogonal a M.
   (e) Todo falso.

2. El estimar un modelo de regresión lineal, si introducimos un nuevo regresor,
   (a) Aumentará la SSE.
   (b) En general se modificarán las estimaciones de los betas asociados a todos los demás regresores.
   (c) Los estimadores de los betas asociados a los restantes regresores, permanecerán inalterados, siempre que no haya multicolinealidad exacta.
   (d) Si el regresor añadido es irrelevante, las estimaciones de los restantes pueden resultar sesgadas.
   (e) Todo falso.

3. ¿En cuál o cuales de las siguientes regresiones de una variable Y sobre una única X te parece que sería indicado ajustar $Y = \beta_1 X$ en lugar de $Y = \beta_0 + \beta_1 X$? (En otras palabras: ¿en cuál o cuáles de las situaciones siguientes te parece que deberíamos prescindir de la columna de "unos" en la matriz de diseño?)
   (a) Y = "Consumo", X = "Renta". Objetivo: estimar la propensión al consumo con una muestra de familias.
   (b) Y = "Peso", X = "Volumen". Objetivo: estimar el peso específico (por unidad de volumen) con una muestra de objetos de una sustancia.
   (c) Y = "Peso", X = "Edad". Objetivo: estimar el "parámetro de engorde" con una muestra de animales homogéneos y alimentados con la misma dieta.
   (d) Y = "Temperatura", X = "Altura sobre el nivel del mar". Objetivo: estimar el "parámetro de enfriamiento" al progresar en altura en una ubicación dada y con situación atmosférica estable.
   (e) Todo falso.

4. El fallo de uno de los supuestos siguientes introduciría un sesgo en las estimaciones de los parámetros β. ¿Cuál?
   (a) Perturbaciones son homoscedásticas: $E(\epsilon\epsilon') = \sigma^2 I$
   (b) Modelo "escaso" (faltan regresores, sin sobrar ninguno).
   (c) Perturbaciones incorreladas.
   (d) Modelo sobreparametrizado (sobran regresores, sin faltar ninguno).
   (e) Todo falso.

5. El fallo de uno de los supuestos siguientes introduciría un sesgo en la estimación de σ^2_ϵ. ¿Cuál?
   (a) Perturbaciones no normales.
   (b) Modelo "escaso" (faltan regresores, sin sobrar ninguno).
   (c) Ausencia de multicolinealidad.
   (d) Modelo sobreparametrizado (sobran regresores, sin faltar ninguno).
   (e) Todo falso.

6. Supongamos un problema de estimación con $N = 150$ observaciones y $p = 8$ variables, sin multicolinealidad. ¿Cuál será el rango de $P_M = X(X'X)^{-1}X'$?
   (a) 142
   (b) 8
   (c) 150
   (d) Menor que 8
   (e) Todo falso.

7. Supongamos un problema de estimación con $N = 150$ observaciones y $p = 8$ variables, sin multicolinealidad. ¿Cuál será la traza de $I - P_M = I - X(X'X)^{-1}X'$?
   (a) 142
   (b) 8
   (c) 150
   (d) Menor que 142
   (e) Todo falso.

8. De entre dos modelos ajustados a una misma variable respuesta, necesariamente tendrá R^2 (no $\overline{R}^2$) más grande aquél que tenga:
   (a) Menor SSR.
   (b) Menor SSE.
   (c) Mayor SSE/SST.
   (d) Mayor SST.
   (e) Todo falso.

9. Al ajustar un modelo con $N = 100$ observaciones, $p = 6$ regresores y una restricción lineal sobre los betas, el número de grados de libertad será:
   (a) 100
   (b) 95
   (c) 94
   (d) 96
   (e) Todo falso.

10. La existencia de multicolinealidad exacta,
   (a) Hace que la dimensión del espacio sobre el que se proyecta sea inferior al número de regresores.
   (b) No permite estimar mediante MCO la totalidad de los parámetros.
   (c) No imposibilita la estimación de los parámetros cuando se emplea regresión *ridge*.
   (d) No imposibilita la estimación de los parámetros cuando se emplea regresión en componentes principales.
   (e) Todo falso.

11. Si ajustamos un modelo con y sin restricciones sobre los parámetros, el modelo con restricciones siempre proporcionará una estimación de σ^2 con más grados de libertad.
   (a) Falso.
   (b) Cierto.
   (c) No se puede contestar con la información facilitada.

12. La condición de que la matriz de diseño sea de rango completo es necesaria para garantizar que:
   (a) Exista una proyección.
   (b) Las ecuaciones normales tengan solución única.
   (c) La proyección de $\boldsymbol{y}$ sobre M, que siempre existe, sea única.
   (d) Los estimadores de los betas sean insesgados.
   (e) Todo falso.

13. El hecho de que una observación sea influyente significa que:

    (a) Su omisión alteraría de forma apreciable las estimaciones de uno o varios parámetros.

    (b) Su residuo MCO es muy grande.

    (c) Ha sido muy influída por la variable respuesta.

    (d) Su residuo studentizado es muy grande.

    (e) Todo falso.

14. El criterio C_p de Mallow tiene por expresión:

    (a) $C_p = \frac{SSE}{\hat{\sigma}^2} + p$

    (b) $C_p = \frac{SSE}{\hat{\sigma}^2} + 2p$

    (c) $C_p = \frac{SSE}{N-p} + 2\sigma^2$

    (d) $C_p = -2\log(\text{máx}(\text{Verosimilitud})) + 2p$

    (e) Todo falso.

15. De los siguientes criterios para la selección de modelos de regresión, ¿cuál sería el más permisivo a la hora de incluír nuevos regresores?

    (a) Maximización de R^2.

    (b) Maximización de $\overline{R}^2$.

    (c) Maximización de C_p.

    (d) Maximización de AIC.

    (e) No se puede responder; unas veces son más permisivos unos, otras veces otros.

Comienzo de un bloque de preguntas

16. Esta pregunta y las que siguen hasta el final del bloque hacen todas referencia a datos sobre divorcios en EE.UU. para 1920–1996. Las variables son:

Variables	Significado
`year`	Año
`divorce`	Divorcios por 1000 mujeres >15 años.
`unemployed`	Tasa desempleo
`femlab`	% mujeres pob. activa >16 años.
`military`	Militares por 1000 hab.
`marriage`	Matrimonios por 1000 > mujeres >16 años.
`birth`	Nacimientos por 1000 > mujeres 15–44 años.

Observa el siguiente código y resultados:

```
> library(faraway)
> data(divusa)
> mod <- lm(divorce ~ ., data = divusa)
> summary(mod)

Call:
lm(formula = divorce ~ ., data = divusa)

Residuals:
     Min       1Q   Median       3Q      Max
-2.90874 -0.92123 -0.09345  0.74469  3.46893

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 380.14761   99.20371   3.832 0.000274 ***
year         -0.20312    0.05333  -3.809 0.000297 ***
unemployed   -0.04933    0.05378  -0.917 0.362171
femlab        0.80793    0.11487   7.033 1.09e-09 ***
marriage      0.14977    0.02382   6.287 2.42e-08 ***
birth        -0.11695    0.01470  -7.957 2.19e-11 ***
military     -0.04276    0.01372  -3.117 0.002652 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.513 on 70 degrees of freedom
Multiple R-squared: 0.9344,        Adjusted R-squared: 0.9288
F-statistic: 166.2 on 6 and 70 DF,  p-value: < 2.2e-16
```

Es claro que el desempleo:

    (a) No es significativo a los niveles habituales del 5 %–10 %.

    (b) Aumenta la tasa de divorcios.

    (c) Disminuye la tasa de divorcios.

    (d) Todo falso.

17. Por otra parte, es claro que el numero de observaciones empleado ha sido de:

    (a) 77

    (b) 70

    (c) 70-6

    (d) 70-7

    (e) Todo falso.

18. Parece haber existido una tendencia decreciente de la tasa de divorcio en el tiempo, una vez que se considera el efecto de las demás covariables.

    (a) Cierto.

    (b) Falso.

    (c) No puedo reponder con la información facilitada.

19. Observa el siguiente estadillo:

```
> anova(mod)

Analysis of Variance Table

Response: divorce
           Df  Sum Sq Mean Sq  F value    Pr(>F)
year        1 1888.22 1888.22 825.0759 < 2.2e-16 ***
unemployed  1    0.05    0.05   0.0223  0.881843
femlab      1  169.40  169.40  74.0231 1.413e-12 ***
marriage    1   57.12   57.12  24.9587 4.141e-06 ***
birth       1  145.31  145.31  63.4934 2.090e-11 ***
military    1   22.23   22.23   9.7142  0.002652 **
Residuals  70  160.20    2.29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

    y ahora este otro:

```
> mod2 <- lm(divorce ~ birth + year + ., data = divusa)
> anova(mod2)

Analysis of Variance Table

Response: divorce
           Df  Sum Sq Mean Sq  F value    Pr(>F)
birth       1 1272.98 1272.98 556.2432 < 2.2e-16 ***
year        1  784.61  784.61 342.8458 < 2.2e-16 ***
unemployed  1   84.46   84.46  36.9069 5.771e-08 ***
femlab      1   45.58   45.58  19.9178 3.019e-05 ***
marriage    1   72.45   72.45  31.6596 3.523e-07 ***
military    1   22.23   22.23   9.7142  0.002652 **
Residuals  70  160.20    2.29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

    En el primero, `year` parecía ser la variable dando cuenta de la mayor fracción de suma de cuadrados. En el segundo, parece ser `birth`. Esto ocurre porque:

    (a) La instrucción `anova` proporciona la suma de cuadrados asociada a cada regresor cuando se introduce en el orden especificado en la regresión.

    (b) La instrucción `anova` hace uso de un generador de números aleatorios; si no se fija una semilla, los resultados varían de una ejecución a otra.

    (c) Estamos ante un experimento aleatorio: no podemos esperar resultados idénticos en dos ejecuciones.

    (d) Todo falso.

20. Observa el siguiente gráfico:

```
> plot(year, divorce, type = "l")
```

    (a) La tasa de divorcio claramente crece a lo largo del tiempo; es evidente que el modelo mod, en que `year` tenía un beta negativo, está mal especificado.

    (b) La tasa de divorcio crece a lo largo del tiempo; el coeficiente beta es negativo, pero no signficativo, de modo que no hay contradicción.

    (c) La tasa de divorcio es claramente no significativa.

    (d) El R^2 no permite extraer conclusiones del modelo.

    (e) Todo falso.

21. A la vista de los coeficientes estimados en el modelo `mod`, parece que:

    (a) La tasa de desempleo, `unemployed`, no tuvo un efecto significativo en la tasa de divorcios.

    (b) En cuando el marido se queda en paro, las mujeres americanas se buscan a otro.

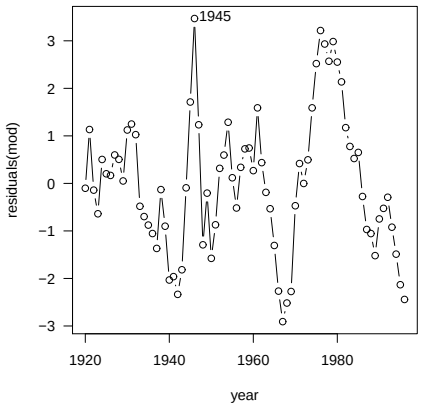
    (c) Los conyuges que se divorcian es más fácil que caigan en el paro.

    (d) Durante todo el periodo muestral, hubo prácticamente pleno empleo en EE.UU.

    (e) Todo falso.

22. Observa el siguiente gráfico de residuos MCO:

```
> plot(year, residuals(mod), type = "b")
> text(1945, 3.5, "1945", pos = 4)
```



Es evidente que:

(a) Los residuos son incorrelados.

(b) El residuo de 1945 corresponde a una observación anómala.

(c) Parece haber una pauta temporal no recogida por el modelo; el periodo de los últimos años 70 a comienzos de los 80 tuvo sistemáticamente tasas de divorcio superiores a las ajustadas por el modelo.

(d) Los residuos suman cero, por haber columna de "unos" en el modelo.

23. En el gráfico precedente,

(a) Podremos declarar al residuo de 1945 significativo si excede del cuantil de una distribución t de Student con grados de libertad adecuados, dejando a su derecha probabilidad α.

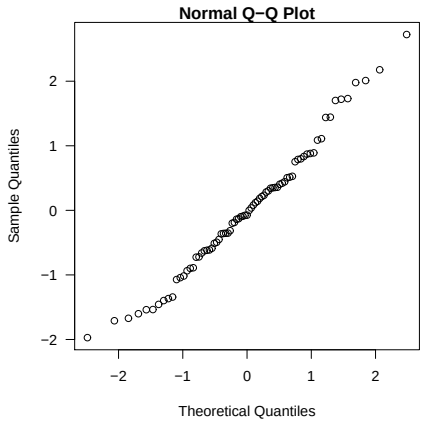
(b) Podremos declarar al residuo de 1945 significativo si excede del cuantil de una distribución máximo de k variables t de Student con grados de libertad adecuados, dejando a su derecha probabilidad α.

(c) Podremos declarar al residuo de 1945 significativo si excede del cuantil de una distribución normal, dejando a su derecha probabilidad α.

(d) Todo falso.

24. Observa el siguiente gráfico (recuerda que `rstandard` calcula residuos internamente studentizados):

```
> qqnorm(rstandard(mod))
```



(a) No parece que se incumpla flagrantemente el supuesto de normalidad de las perturbaciones.

(b) El gráfico no es afortunado; hubieran debido emplearse los residuos mínimo-cuadráticos.

(c) Es evidente la heterocedasticidad.

(d) Todo falso

25. Observa una vez más el siguiente estadillo:

```
> anova(mod2)

Analysis of Variance Table

Response: divorce
           Df  Sum Sq Mean Sq  F value    Pr(>F)
birth       1 1272.98 1272.98 556.2432 < 2.2e-16 ***
year        1  784.61  784.61 342.8458 < 2.2e-16 ***
unemployed  1   84.46   84.46  36.9069 5.771e-08 ***
femlab      1   45.58   45.58  19.9178 3.019e-05 ***
marriage    1   72.45   72.45  31.6596 3.523e-07 ***
military    1   22.23   22.23   9.7142  0.002652 **
Residuals  70  160.20    2.29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con la información en el mismo puedes deducir que el estadístico Q_h para contrastar la hipótesis de que $\beta_{\texttt{military}} = 0$, tomaría el valor:

(a) 22.23/160.20

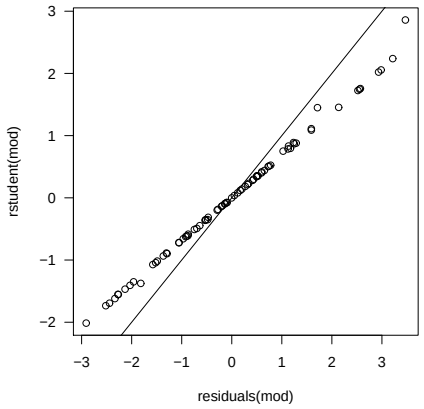
(b) 22.23/2.29

(c) (160.20 - 22.23) / 160.20

(d) Todo falso.

26. Observa el gráfico siguiente, en que se han dibujado los residuos MCO ($\hat{\epsilon}_i$, abscisas) frente a los externamente studentizados (t_i, ordenadas). Se ha superpuesto la línea que pasa por el origen con pendiente 1.

```
> plot(residuals(mod), rstudent(mod))
> lines(abline(0, 1))
```



Del gráfico se puede deducir:

(a) Que los términos diagonales p_{ii} de la matriz de diseño son aproximadamente iguales.

(b) Hay sesgo; los puntos se desvían de la línea sobre la que deberían yacer.

(c) Hay evidencia de normalidad en las perturbaciones.

(d) Los residuos borrados no diferirán mucho de los ordinarios.

(e) Todo falso.

27. Empleando un algoritmo de regresión escalonada (o "stepwise") obtenemos lo siguiente:

```
> library(MASS)
> stepAIC(mod)

Start:  AIC=70.41
divorce ~ year + unemployed + femlab + marriage + birth + military

             Df Sum of Sq     RSS     AIC
- unemployed  1     1.925 162.123  69.330
<none>                    160.197  70.410
- military    1    22.231 182.429  78.417
- year        1    33.199 193.397  82.912
- marriage    1    90.468 250.665 102.884
- femlab      1   113.214 273.411 109.572
- birth       1   144.897 305.095 118.015

Step:  AIC=69.33
divorce ~ year + femlab + marriage + birth + military

           Df Sum of Sq    RSS    AIC
<none>                  162.12  69.33
- military  1     20.96 183.08  76.69
- year      1     42.05 204.18  85.09
- marriage  1    126.64 288.77 111.78
- femlab    1    158.00 320.13 119.72
- birth     1    172.83 334.95 123.20

Call:
lm(formula = divorce ~ year + femlab + marriage + birth + military,

Coefficients:
(Intercept)         year       femlab     marriage
   405.6167      -0.2179       0.8548       0.1593
      birth     military
    -0.1101      -0.0412
```

Si decidimos emplear el AIC como criterio de selección, escogeremos como modelo:

(a) `divorce ~ year + unemployed + femlab + marriage + military`.

(b) `divorce ~ year + military + femlab + marriage + birth`.

(c) El inicial, `divorce ~ year + unemployed + femlab + marriage + birth + military`.

(d) El AIC es un diagnóstico de multicolinealidad; nada que ver con la selección de modelos.

FINAL DE UN BLOQUE DE PREGUNTAS

Sección 2. Preguntas breves

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

1. Escribe la matriz de diseño de los siguientes dos modelos ANOVA: i) Un modelo curzado con dos tratamientos de dos y tres niveles respectivamente, cuando no hay replicación. ii) Un modelo anidado $y_{iijk} = \alpha + \alpha^{A<-B} + \epsilon_{ijk}$ en que B tiene dos niveles y está anidado en A, que tiene tres, con replicación $K = 2$. (3 puntos.)

2. Enuncia y demuestra el teorema de Gauss-Markov. (5 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 para el examen de tipo A

Sección 1. Cuestiones de elección múltiple

1. Sy $P_M\boldsymbol{y}$ denota la proyección sobre un espacio vectorial M del vector $\boldsymbol{y}$,
   (a) $P_M\boldsymbol{y}$ es ortogonal a $\boldsymbol{y}$.
   (b) $(\boldsymbol{y} - P_M\boldsymbol{y})$ es ortogonal a $\boldsymbol{y}$.
   (c) $(\boldsymbol{y} - P_M\boldsymbol{y})$ es ortogonal a M.
   (d) $\boldsymbol{y}$ es ortogonal a M.
   (e) Todo falso.

2. El estimar un modelo de regresión lineal, si introducimos un nuevo regresor,
   (a) Aumentará la SSE.
   (b) En general se modificarán las estimaciones de los betas asociados a todos los demás regresores.
   (c) Los estimadores de los betas asociados a los restantes regresores, permanecerán inalterados, siempre que no haya multicolinealidad exacta.
   (d) Si el regresor añadido es irrelevante, las estimaciones de los restantes pueden resultar sesgadas.
   (e) Todo falso.

3. ¿En cuál o cuales de las siguientes regresiones de una variable Y sobre una única X te parece que sería indicado ajustar $Y = \beta_1 X$ en lugar de $Y = \beta_0 + \beta_1 X$? (En otras palabras: ¿en cuál o cuáles de las situaciones siguientes te parece que deberíamos prescindir de la columna de "unos" en la matriz de diseño?)
   (a) Y = "Consumo", X = "Renta". Objetivo: estimar la propensión al consumo con una muestra de familias.
   (b) Y = "Peso", X = "Volumen". Objetivo: estimar el peso específico (por unidad de volumen) con una muestra de objetos de una sustancia.
   (c) Y = "Peso", X = "Edad". Objetivo: estimar el "parámetro de engorde" con una muestra de animales homogéneos y alimentados con la misma dieta.
   (d) Y = "Temperatura", X = "Altura sobre el nivel del mar". Objetivo: estimar el "parámetro de enfriamiento" al progresar en altura en una ubicación dada y con situación atmosférica estable.
   (e) Todo falso.

4. El fallo de uno de los supuestos siguientes introduciría un sesgo en las estimaciones de los parámetros β. ¿Cuál?
   (a) Perturbaciones son homoscedásticas: $E(\epsilon\epsilon') = \sigma^2 I$
   (b) Modelo "escaso" (faltan regresores, sin sobrar ninguno).
   (c) Perturbaciones incorreladas.
   (d) Modelo sobreparametrizado (sobran regresores, sin faltar ninguno).
   (e) Todo falso.

5. El fallo de uno de los supuestos siguientes introduciría un sesgo en la estimación de σ^2_ϵ. ¿Cuál?

   (a) Perturbaciones no normales.

   (b) Modelo "escaso" (faltan regresores, sin sobrar ninguno).

   (c) Ausencia de multicolinealidad.

   (d) Modelo sobreparametrizado (sobran regresores, sin faltar ninguno).

   (e) Todo falso.

6. Supongamos un problema de estimación con $N = 150$ observaciones y $p = 8$ variables, sin multicolinealidad. ¿Cuál será el rango de $P_M = X(X'X)^{-1}X'$?

   (a) 142

   (b) 8

   (c) 150

   (d) Menor que 8

   (e) Todo falso.

7. Supongamos un problema de estimación con $N = 150$ observaciones y $p = 8$ variables, sin multicolinealidad. ¿Cuál será la traza de $I - P_M = I - X(X'X)^{-1}X'$?

   (a) 142

   (b) 8

   (c) 150

   (d) Menor que 142

   (e) Todo falso.

8. De entre dos modelos ajustados a una misma variable respuesta, necesariamente tendrá R^2 (no $\overline{R}^2$) más grande aquél que tenga:

   (a) Menor SSR.

   (b) Menor SSE.

   (c) Mayor SSE/SST.

   (d) Mayor SST.

   (e) Todo falso.

9. Al ajustar un modelo con $N = 100$ observaciones, $p = 6$ regresores y una restricción lineal sobre los betas, el número de grados de libertad será:

   (a) 100

   (b) 95

   (c) 94

   (d) 96

   (e) Todo falso.

10. La existencia de multicolinealidad exacta,

    (a) Hace que la dimensión del espacio sobre el que se proyecta sea inferior al número de regresores.

    (b) No permite estimar mediante MCO la totalidad de los parámetros.

    (c) No imposibilita la estimación de los parámetros cuando se emplea regresión *ridge*.

    (d) No imposibilita la estimación de los parámetros cuando se emplea regresión en componentes principales.

    (e) Todo falso.

11. Si ajustamos un modelo con y sin restricciones sobre los parámetros, el modelo con restricciones siempre proporcionará una estimación de σ^2 con más grados de libertad.

    (a) Falso.

    (b) Cierto.

    (c) No se puede contestar con la información facilitada.

12. La condición de que la matriz de diseño sea de rango completo es necesaria para garantizar que:

    (a) Exista una proyección.

    (b) Las ecuaciones normales tengan solución única.

    (c) La proyección de $\boldsymbol{y}$ sobre M, que siempre existe, sea única.

    (d) Los estimadores de los betas sean insesgados.

    (e) Todo falso.

13. El hecho de que una observación sea influyente significa que:

    (a) Su omisión alteraría de forma apreciable las estimaciones de uno o varios parámetros.

    (b) Su residuo MCO es muy grande.

    (c) Ha sido muy influída por la variable respuesta.

    (d) Su residuo studentizado es muy grande.

    (e) Todo falso.

14. El criterio C_p de Mallow tiene por expresión:

    (a) $C_p = \frac{SSE}{\hat{\sigma^2}} + p$

    (b) $C_p = \frac{SSE}{\hat{\sigma^2}} + 2p$

    (c) $C_p = \frac{SSE}{N-p} + 2\sigma^2$

    (d) $C_p = -2\log(\text{máx}(\text{Verosimilitud})) + 2p$

    (e) Todo falso.

15. De los siguientes criterios para la selección de modelos de regresión, ¿cuál sería el más permisivo a la hora de incluír nuevos regresores?

    (a) Maximización de R^2.

    (b) Maximización de $\overline{R}^2$.

    (c) Maximización de C_p.

    (d) Maximización de AIC.

    (e) No se puede responder; unas veces son más permisivos unos, otras veces otros.

Comienzo de un bloque de preguntas

16. Esta pregunta y las que siguen hasta el final del bloque hacen todas referencia a datos sobre divorcios en EE.UU. para 1920–1996. Las variables son:

Variables	Significado
`year`	Año
`divorce`	Divorcios por 1000 mujeres >15 años.
`unemployed`	Tasa desempleo
`femlab`	% mujeres pob. activa >16 años.
`military`	Militares por 1000 hab.
`marriage`	Matrimonios por 1000 > mujeres >16 años.
`birth`	Nacimientos por 1000 > mujeres 15–44 años.

Observa el siguiente código y resultados:

```
> library(faraway)
> data(divusa)
> mod <- lm(divorce ~ ., data = divusa)
> summary(mod)

Call:
lm(formula = divorce ~ ., data = divusa)

Residuals:
     Min       1Q   Median       3Q      Max
-2.90874 -0.92123 -0.09345  0.74469  3.46893

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 380.14761   99.20371   3.832 0.000274 ***
year         -0.20312    0.05333  -3.809 0.000297 ***
unemployed   -0.04933    0.05378  -0.917 0.362171
femlab        0.80793    0.11487   7.033 1.09e-09 ***
marriage      0.14977    0.02382   6.287 2.42e-08 ***
birth        -0.11695    0.01470  -7.957 2.19e-11 ***
military     -0.04276    0.01372  -3.117 0.002652 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.513 on 70 degrees of freedom
Multiple R-squared: 0.9344,        Adjusted R-squared: 0.9288
F-statistic: 166.2 on 6 and 70 DF,  p-value: < 2.2e-16
```

Es claro que el desempleo:

(a) No es significativo a los niveles habituales del 5 %–10 %.

(b) Aumenta la tasa de divorcios.

(c) Disminuye la tasa de divorcios.

(d) Todo falso.

17. Por otra parte, es claro que el numero de observaciones empleado ha sido de:

    (a) 77

    (b) 70

    (c) 70-6

    (d) 70-7

    (e) Todo falso.

18. Parece haber existido una tendencia decreciente de la tasa de divorcio en el tiempo, una vez que se considera el efecto de las demás covariables.

    (a) Cierto.

    (b) Falso.

    (c) No puedo reponder con la información facilitada.

19. Observa el siguiente estadillo:

```
> anova(mod)

Analysis of Variance Table

Response: divorce
           Df  Sum Sq Mean Sq  F value    Pr(>F)
year        1 1888.22 1888.22 825.0759 < 2.2e-16 ***
unemployed  1    0.05    0.05   0.0223  0.881843
femlab      1  169.40  169.40  74.0231 1.413e-12 ***
marriage    1   57.12   57.12  24.9587 4.141e-06 ***
birth       1  145.31  145.31  63.4934 2.090e-11 ***
military    1   22.23   22.23   9.7142  0.002652 **
Residuals  70  160.20    2.29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

    y ahora este otro:

```
> mod2 <- lm(divorce ~ birth + year + ., data = divusa)
> anova(mod2)

Analysis of Variance Table

Response: divorce
           Df  Sum Sq Mean Sq  F value    Pr(>F)
birth       1 1272.98 1272.98 556.2432 < 2.2e-16 ***
year        1  784.61  784.61 342.8458 < 2.2e-16 ***
unemployed  1   84.46   84.46  36.9069 5.771e-08 ***
femlab      1   45.58   45.58  19.9178 3.019e-05 ***
marriage    1   72.45   72.45  31.6596 3.523e-07 ***
military    1   22.23   22.23   9.7142  0.002652 **
Residuals  70  160.20    2.29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

    En el primero, `year` parecía ser la variable dando cuenta de la mayor fracción de suma de cuadrados. En el segundo, parece ser `birth`. Esto ocurre porque:

    (a) La instrucción `anova` proporciona la suma de cuadrados asociada a cada regresor cuando se introduce en el orden especificado en la regresión.

    (b) La instrucción `anova` hace uso de un generador de números aleatorios; si no se fija una semilla, los resultados varían de una ejecución a otra.

    (c) Estamos ante un experimento aleatorio: no podemos esperar resultados idénticos en dos ejecuciones.

    (d) Todo falso.

20. Observa el siguiente gráfico:

```
> plot(year, divorce, type = "l")
```

    (a) La tasa de divorcio claramente crece a lo largo del tiempo; es evidente que el modelo mod, en que `year` tenía un beta negativo, está mal especificado.

    (b) La tasa de divorcio crece a lo largo del tiempo; el coeficiente beta es negativo, pero no signficativo, de modo que no hay contradicción.

    (c) La tasa de divorcio es claramente no significativa.

    (d) El R^2 no permite extraer conclusiones del modelo.

    (e) Todo falso.

21. A la vista de los coeficientes estimados en el modelo `mod`, parece que:

    (a) La tasa de desempleo, `unemployed`, no tuvo un efecto significativo en la tasa de divorcios.

    (b) En cuando el marido se queda en paro, las mujeres americanas se buscan a otro.

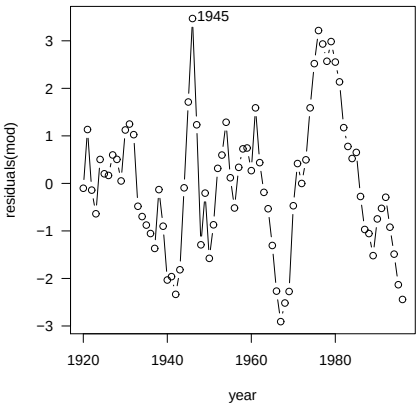
    (c) Los conyuges que se divorcian es más fácil que caigan en el paro.

    (d) Durante todo el periodo muestral, hubo prácticamente pleno empleo en EE.UU.

    (e) Todo falso.

22. Observa el siguiente gráfico de residuos MCO:

```
> plot(year, residuals(mod), type = "b")
> text(1945, 3.5, "1945", pos = 4)
```



Es evidente que:

(a) Los residuos son incorrelados.

(b) El residuo de 1945 corresponde a una observación anómala.

(c) Parece haber una pauta temporal no recogida por el modelo; el periodo de los últimos años 70 a comienzos de los 80 tuvo sistemáticamente tasas de divorcio superiores a las ajustadas por el modelo.

(d) Los residuos suman cero, por haber columna de "unos" en el modelo.

23. En el gráfico precedente,

(a) Podremos declarar al residuo de 1945 significativo si excede del cuantil de una distribución t de Student con grados de libertad adecuados, dejando a su derecha probabilidad α.

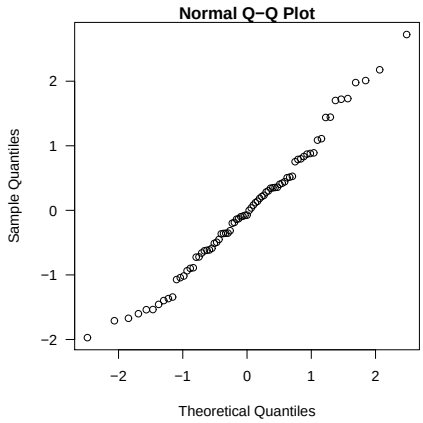
(b) Podremos declarar al residuo de 1945 significativo si excede del cuantil de una distribución máximo de k variables t de Student con grados de libertad adecuados, dejando a su derecha probabilidad α.

(c) Podremos declarar al residuo de 1945 significativo si excede del cuantil de una distribución normal, dejando a su derecha probabilidad α.

(d) Todo falso.

24. Observa el siguiente gráfico (recuerda que `rstandard` calcula residuos internamente studentizados):

```
> qqnorm(rstandard(mod))
```



(a) No parece que se incumpla flagrantemente el supuesto de normalidad de las perturbaciones.

(b) El gráfico no es afortunado; hubieran debido emplearse los residuos mínimo-cuadráticos.

(c) Es evidente la heterocedasticidad.

(d) Todo falso

25. Observa una vez más el siguiente estadillo:

```
> anova(mod2)

Analysis of Variance Table

Response: divorce
           Df  Sum Sq Mean Sq  F value    Pr(>F)
birth       1 1272.98 1272.98 556.2432 < 2.2e-16 ***
year        1  784.61  784.61 342.8458 < 2.2e-16 ***
unemployed  1   84.46   84.46  36.9069 5.771e-08 ***
femlab      1   45.58   45.58  19.9178 3.019e-05 ***
marriage    1   72.45   72.45  31.6596 3.523e-07 ***
military    1   22.23   22.23   9.7142  0.002652 **
Residuals  70  160.20    2.29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con la información en el mismo puedes deducir que el estadístico Q_h para contrastar la hipótesis de que $\beta_{\texttt{military}} = 0$, tomaría el valor:

(a) 22.23/160.20

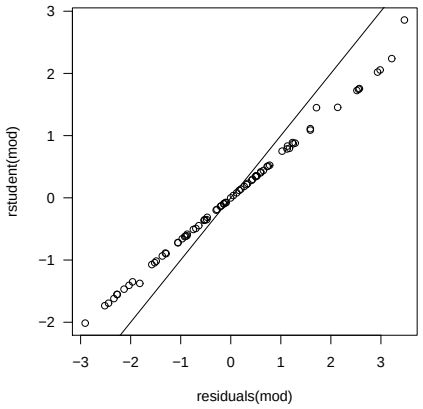
(b) 22.23/2.29

(c) (160.20 - 22.23) / 160.20

(d) Todo falso.

26. Observa el gráfico siguiente, en que se han dibujado los residuos MCO ($\hat{\epsilon}_i$, abscisas) frente a los externamente studentizados (t_i, ordenadas). Se ha superpuesto la línea que pasa por el origen con pendiente 1.

```
> plot(residuals(mod), rstudent(mod))
> lines(abline(0, 1))
```



Del gráfico se puede deducir:

(a) **Que los términos diagonales** p_{ii} de la matriz de diseño son aproximadamente iguales.

(b) Hay sesgo; los puntos se desvían de la línea sobre la que deberían yacer.

(c) Hay evidencia de normalidad en las perturbaciones.

(d) **Los residuos borrados no diferirán mucho de los ordinarios.**

(e) Todo falso.

27. Empleando un algoritmo de regresión escalonada (o "stepwise") obtenemos lo siguiente:

```
> library(MASS)
> stepAIC(mod)

Start:  AIC=70.41
divorce ~ year + unemployed + femlab + marriage + birth + military

             Df Sum of Sq     RSS     AIC
- unemployed  1     1.925 162.123  69.330
<none>                    160.197  70.410
- military    1    22.231 182.429  78.417
- year        1    33.199 193.397  82.912
- marriage    1    90.468 250.665 102.884
- femlab      1   113.214 273.411 109.572
- birth       1   144.897 305.095 118.015

Step:  AIC=69.33
divorce ~ year + femlab + marriage + birth + military

           Df Sum of Sq    RSS    AIC
<none>                  162.12  69.33
- military  1     20.96 183.08  76.69
- year      1     42.05 204.18  85.09
- marriage  1    126.64 288.77 111.78
- femlab    1    158.00 320.13 119.72
- birth     1    172.83 334.95 123.20

Call:
lm(formula = divorce ~ year + femlab + marriage + birth + military,

Coefficients:
(Intercept)         year       femlab     marriage
   405.6167      -0.2179       0.8548       0.1593
      birth     military
    -0.1101      -0.0412
```

Si decidimos emplear el AIC como criterio de selección, escogeremos como modelo:

(a) `divorce ~ year + unemployed + femlab + marriage + military`.

(b) `divorce ~ year + military + femlab + marriage + birth`.

(c) El inicial, `divorce ~ year + unemployed + femlab + marriage + birth + military`.

(d) El AIC es un diagnóstico de multicolinealidad; nada que ver con la selección de modelos.

FINAL DE UN BLOQUE DE PREGUNTAS

Sección 2. Preguntas breves

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

1. Escribe la matriz de diseño de los siguientes dos modelos ANOVA: i) Un modelo curzado con dos tratamientos de dos y tres niveles respectivamente, cuando no hay replicación. ii) Un modelo anidado $y_{iijk} = \alpha + \alpha^{A<-B} + \epsilon_{ijk}$ en que B tiene dos niveles y está anidado en A, que tiene tres, con replicación $K = 2$. (3 puntos.)

   Respuesta: Ver apuntes de clase.

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

   Respuesta: Ver apuntes de clase.