Estadística: Modelos Lineales

Final Enero 2.007, Tipo: A

Apellidos y Nombre: _____

DNI: _____

Grupo: _____

Profesor : _____

Sección 1. 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!

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

3. 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.

Sección 2. Cuestiones de elección múltiple

1. Dado un modelo de regresión lineal $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, las estimaciones $\hat{\boldsymbol{\beta}}$ de los parámetros permiten construir $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ tal que:
   (a) $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ es la proyección de $\boldsymbol{y}$ sobre el subespacio generado por las columnas de $\boldsymbol{X}$.
   (b) $||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2$ es máximo
   (c) $\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$
   (d) $\boldsymbol{y}$ tiene correlación 1 con $\boldsymbol{X}\hat{\boldsymbol{\beta}}$.

2. En un modelo ANOVA con dos tratamientos,
$$y_{ijk} = \alpha + \alpha_i^A + \alpha_j^B + \epsilon_{ijk}$$
con replicación y las restricciones habituales,
$$\sum_i \alpha_i^A = \sum_j \alpha_j^B = 0$$
se verifica que:
   (a) El efecto estimado para un nivel i cualquiera del tratamiento A es el mismo sea cual fuere el nivel j del tratamiento B presente.
   (b) Los residuos son independientes unos de otros.
   (c) Las interacciones resultan estimadas por $(\hat{\alpha}_i^A \hat{\alpha}_j^B)$.
   (d) Todo falso.

3. Para facilitar al analista la selección de modelos de regresión hay una porción de estrategias o algoritmos de búsqueda. ¿Cuál de ellos, si es que alguno, garantiza hallar *el mejor* de los modelos, de acuerdo con el criterio que hayamos especificado?
   (a) La estrategia *all subsets* o búsqueda entre todos los subconjuntos de regresores.
   (b) La regresión escalonada hacia delante.
   (c) La regresión esclonada hacia atrás.
   (d) La regresión escalonada mixta.
   (e) Todo falso.

4. Un valor propio cercano a cero de la matriz $X'X$ sería indicio de:
   (a) Multicolinealidad aproximada.
   (b) No linealidad del diseño.
   (c) R^2 muy pequeña.
   (d) Varianza de la perturbación muy grande.

5. Los factores de incremento de varianza (o VIF's, *variance inflation factors*):
   (a) Son estadísticos que permiten diagnosticar multicolinealidad.
   (b) Están relacionados con los residuos internamente studentizados.
   (c) Están relacionados con los residuos externamente studentizados.
   (d) Están relacionados con la R^2 de la regresión.
   (e) Toman valores entre 0 y 1.

6. El estimador *ridge:*
   (a) Es un estimador no lineal: por eso puede lograr un mejor ECM (error cuadrático medio) que el estimador MCO.
   (b) No puede mejorar al MCO en términos de varianza, como quedó suficientemente demostrado en el teorema de Gauss-Markov.
   (c) Es particularmente interesante en situaciones de fuerte multicolinealidad, aunque puede mejorar el ECM en todos los casos.

7. Si tenemos 10 sucesos, $E_1, \ldots, E_{10}$, independientes o no, cada uno de los cuales se verifica con probabilidad $0{,}99$, la probabilidad Prob$\{\cap_{i=1}^{1} 0 E_i\}$ de que *los diez simultáneamente* se verifiquen puede acotarse mediante la desigualdad de Bonferroni. Efectuado el cálculo resulta que
   (a) $1 \geq$ Prob$\{\cap_{i=1}^{1} 0 E_i\} > 0{,}99$.
   (b) Prob$\{\cap_{i=1}^{1} 0 E_i\} \geq 0{,}90$.
   (c) $0{,}90 >$ Prob$\{\cap_{i=1}^{1} 0 E_i\} \geq 0{,}10$.
   (d) Prob$\{\cap_{i=1}^{1} 0 E_i\} \leq 0{,}10$.

8. Cuando en un modelo que incluye la columna de "unos" tomamos una variable y la reescalamos (expresándola, por ejemplo, en unidades que son mil veces mayores que las primitivas, de modo que la variable correspondiente toma valores mil veces más pequeños), se verifica que:
   (a) El ajuste es idéntico al primitivo; el coeficiente estimado de la variable reescalada será ahora mil veces más pequeño.
   (b) El ajuste es idéntico al primitivo; el coeficiente estimado de la variable reescalada será ahora mil veces más grande.
   (c) R^2 se multiplica por 1000^2.
   (d) SSE y SSR no varían, pero SST puede hacerlo.

9. El estimador MCO verifica, con los supuestos habituales, que:
   (a) Es de mínima varianza entre todos los estimadores lineales.
   (b) Es de mínima varianza entre todos los estimadores insesgados.
   (c) Es de mínimo error cuadrático medio entre todos los estimadores lineales.
   (d) Es de mínimo error cuadrático medio entre todos los estimadores lineales e insesgados.

10. La validación cruzada *(cross-validation*) consiste en utilizar alternativamente las observaciones para estimar y validar u modelo. La versión más extrema *(leave one out)* reserva por turno una sóla observación para validar y emplea las $N-1$ restantes para estimar el modelo. Esto requiere en principio estimar el modelo N veces, lo que es muy costoso. En el caso concreto del modelo de regresión lineal, sin embargo, el cálculo es mucho más económico. Basta estimar el modelo una única vez. Los errores de ajuste de cualquier observación si la dejamos fuera y estimamos mediante las restantes coinciden con:
    (a) Los residuos borrados.
    (b) Los residuos internamente *studentizados*
    (c) Los residuos externamente *studentizados*
    (d) Todo falso.

11. ¿Cuales de las siguientes propiedades posee *necesariamente* la matriz de covarianzas de los residuos de un modelo de regresión lineal ordinaria?
    (a) Diagonal.
    (b) Idempotente.
    (c) Simétrica.
    (d) De rango completo.

12. Se conoce como distancia de Cook:
    (a) El mayor número de millas náuticas que el capitán Cook recorrió durante un periodo de 24 horas consecutivas en sus exploraciones por los mares australes.
    (b) Una medida de influencia de la observación i-ésima sobre *uno* cualquiera de los β_i.
    (c) Una medida de influencia de la observación i-ésima sobre *el conjunto* de los β_j.
    (d) El denominador de los residuos borrados.
    (e) Todo falso.

13. ¿En cuál o cuales de las siguientes situaciones resultaría manifiestamente inadecuada la inclusión en el modelo de una columna de "unos"?

    (a) Regresión de Y = "Consumo en el periodo t" sobre X = "Renta en el periodo t" para una muestra de N familias, con objeto de estimar la propensión marginal al consumo (supuesta constante).

    (b) Regresión de Y = "Peso de un objeto" sobre X = "Volumen del mismo objeto", para una muestra de N objetos de la misma materia con el fín de estimar el peso específico de dicha materia.

    (c) Regresión de Y = "Velocidad de sedimentación de una solución acuosa" sobre X = "Temperatura en ºC" para una muestra de N soluciones de la misma materia y concentración.

14. Cuando se omite en un modelo de regresión lineal un regresor que hubiera debido aparecer:

    (a) *Siempre* se sesgan los estimadores de los β's correspondientes a los regresores incluidos.

    (b) En general, se sesgan los estimadores de los β's correspondientes a los regresores incluidos; pero podría ocurrir que no fuera así si el regresor omitido es ortogonal a todos los presentes.

    (c) *Nunca* se sesga el estimador de σ^2, varianza de la perturbación. Sólamente se pierden grados de libertad.

    (d) El estimador de σ^2, varianza de la perturbación, continúa siendo insesgado si el regresor omitido es ortogonal a los incluidos.

15. Cuando se ajusta un modelo $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ y la columna de "unos" está entre los regresores, la R^2 ordinaria (no corregida) puede interpretarse como:

    (a) El ángulo que forma $\boldsymbol{y}$ con el subespacio que generan las columnas de $\boldsymbol{X}$.

    (b) El coseno del ángulo que forma $\boldsymbol{y}$ con el subespacio que generan las columnas de $\boldsymbol{X}$.

    (c) El coseno al cuadrado del ángulo que forma $\boldsymbol{y}$ con el subespacio que generan las columnas de $\boldsymbol{X}$.

    (d) El nivel de significación de los $\boldsymbol{\beta}$ estimados.

16. Cuando se incluyen regresores irrelevantes en un modelo de regresión:

    (a) Las estimaciones de cualquiera de los β's pueden resultar sesgadas.

    (b) La estimación de σ^2 resultará sesgada por exceso.

    (c) La estimación de σ^2 resultará sesgada por defecto.

    (d) La estimación de σ^2 tendrá menos grados de libertad que los que hubiera tenido de ajustarse el modelo correcto.

COMIENZO DE UN BLOQUE DE PREGUNTAS

Las preguntas hasta el siguiente trazo horizontal hacen referencia a los datos que siguen, relacionando para diferentes Estados USA los resultados en el SATM (un examen de Matemáticas) y las siguientes otras variables:

`region`	Región USA	
	ENC	East North Central
	ESC	East South Central
	MA	Mid-Atlantic
	NE	New England
	PAC	Pacific
	SA	South Atlantic
	WNC	West North Central
	WSC	West South Central
`SATV`	Resultado examen aptitud verbal.	
`pop`	Población en miles.	
`percent`	Porcentaje de graduados tomando el SATM.	
`dollars`	Gasto en 1000's$ por alumno.	
`pay`	Salario en 1000's$ por profesor.	

Ajustamos un modelo de regresión del modo que sigue:

```
> mod1 <- lm(SATM ~ . , data=States)
> summary(mod1)

Call:
lm(formula = SATM ~ ., data = States)

Residuals:
     Min       1Q   Median       3Q      Max
-15.5931  -3.2600  -0.5519   3.6316  24.7618

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.598e+01  4.878e+01   1.968   0.0566 .
regionESC   -1.039e+01  5.919e+00  -1.756   0.0873 .
regionMA    -6.204e+00  9.084e+00  -0.683   0.4989
regionMTN   -2.136e+00  5.002e+00  -0.427   0.6718
regionNE    -9.434e+00  7.757e+00  -1.216   0.2316
regionPAC   -5.672e-01  5.590e+00  -0.101   0.9197
regionSA    -1.231e+01  5.868e+00  -2.098   0.0428 *
regionWNC    4.708e+00  5.730e+00   0.822   0.4165
regionWSC   -1.247e+01  5.831e+00  -2.138   0.0392 *
pop          4.910e-04  2.999e-04   1.637   0.1100
SATV         9.082e-01  1.027e-01   8.839 1.18e-10 ***
percent     -1.855e-01  1.730e-01  -1.072   0.2907
dollars      1.602e+00  2.334e+00   0.687   0.4966
pay         -1.560e-01  5.652e-01  -0.276   0.7841
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05

Residual standard error: 7.704 on 37 degrees of freedom
Multiple R-Squared: 0.9632,  Adjusted R-squared: 0.9503
F-statistic:  74.6 on 13 and 37 DF,  p-value: < 2.2e-16
```

Adicionalmente, calculamos la suma de cuadrados de los residuos,

```
> sum(residuals(mod1)^2)
[1] 2195.949
```

Haciendo uso de la información precedente, responde a las siguientes cuestiones:

17. En este caso, el coeficiente R^2 ordinario es mayor que el $\overline{R}^2$ (o corregido).
    (a) Es pura casualidad, podría ser al revés.
    (b) Esto puede ocurrir en presencia de fuerte multicolinealidad.
    (c) Esto indica claramente que hay *outliers*
    (d) Este es necesariamente el caso.
18. Del estadillo anterior deducimos:
    (a) Se ha empleado una muestra con 13 observaciones.
    (b) Se ha empleado una muestra con 37 observaciones.
    (c) Se ha empleado una muestra con 51 observaciones.
    (d) Se ha empleado una muestra con 50 observaciones.
19. La suma total de cuadrados (SST) será:
    (a) 2086.810
    (b) 2310.796
    (c) 2115.138
    (d) 2279.847
    (e) Nada de lo anterior.
20. La variable `pay` tiene asociado un $\hat{\beta} = -1{,}560 \times 10^{-1}$. Ello es evidencia de que cuanto más se paga a los vagos de los profesores, más a la ligera se lo toman y peores son los resultados de los alumnos en el SATM.
    (a) Cierto
    (b) Falso
21. "Como cabría esperar, los resultados en el examen verbal (SATV) y matemático, tienen poco que ver."
    (a) Cierto: parece lógico que así sea. El SATV reflejaría un perfil del estudiante "de letras" y el SATM el de uno "de ciencias", perfiles que como se sabe son bastante antagónicos.
    (b) Falso, el resultado en el SATV acontece que es muy buen predictor del resultado en SATM. Los alumnos "buenos" parecen serlo en todo, y los "malos" también son malos en todo.
22. El coeficiente 0.1100 en la línea `pop` y bajo la columna `Pr(>|t|)` significa:
    (a) Que en el 11 % de los Estados, la variable `pop` influyó en el resultado del SATM.
    (b) Que el valor estimado del parámetro es apenas mayor que la décima parte de su desviación típica.
    (c) Que si afirmáramos que la variable `pop` influye en el resultado de SATM, estaríamos diciendo la verdad el 11 % de las veces.
    (d) Que si el verdadero coeficiente fuera cero, y todos los supuestos habituales se cumplieran, habría probabilidad 0.1100 de estimar un $|\hat{\beta}| \geq 0{,}000491$ para la variable `pop`.
    (e) Todo es falso.
23. El coeficiente $\hat{\beta} = -1{,}247 \times 10^{1}$ estimado para la variable `regionWSC` significa (prescindiendo de si es o no significativo) que los estudiantes en la región WSC, en igualdad de todo lo demás, obtienen una nota en el SATM:
    (a) Inferior en 12.47 % a la obtenida por los estudiantes de cualquier otra región.
    (b) Todo falso
    (c) Inferior en 12.47 a la obtenida por los estudiantes en la región de referencia, que en este caso es ENC.
    (d) Inferior en 12.47 a la obtenida por los estudiantes de otras regiones.

24. "Hay evidencia concluyente de que por cada 1 % adicional de estudiantes que se presentan al SATM, la nota desciende 0.1855. Ello es lógico: en los Estados en que sólo se presentan los mejores, la nota será más alta que en los Estados en que se presentan muchos (al igual que sucede con el porcentaje de aprobados en acceso a la Universidad, mayor para los institutos/colegios que presentan a pocos y buenos alumnos, que para los que dejan presentarse a todos)."

    (a) Todo falso.

    (b) Lo que la estimación sugiere es justamente lo contrario: allá donde se presenta un mayor porcentaje al examen, también las notas son superiores: de la cantidad nace la calidad.

    (c) Los datos corroboran el párrafo entrecomillado, dado el signo del $\hat{\beta}$.

    (d) El argumento anterior puede tener lógica, pero no resulta avalado por los datos: el $\hat{\beta}$ correspondiente no es significativo.

25. Como hay muchos parámetros no significativos de entre los que recogen el efecto Región, podemos plantearnos un modelo sin ellos:

```
> mod2 <- lm(SATM ~ . - region , data=States)
> summary(mod2)

Call:
lm(formula = SATM ~ . - region, data = States)

Residuals:
    Min      1Q  Median      3Q     Max
-14.749  -5.448  -0.649   3.873  34.560

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.2975427 42.4162502   0.856   0.3967
pop          0.0005321  0.0002716   1.959   0.0563
SATV         1.0030845  0.0867660  11.561 4.57e-15
percent     -0.2584571  0.1289731  -2.004   0.0511
dollars      1.5690544  1.9961347   0.786   0.4360
pay          0.3090260  0.5009215   0.617   0.5404
---

Residual standard error:
  8.754 on 45 degrees of freedom
Multiple R-Squared: 0.9423,
Adjusted R-squared: 0.9359
F-statistic: 147 on 5 and 45 DF, p-value: < 2.2e-16

> sum(residuals(mod2)^2)
[1] 3448.105
```

    Para contrastar la hipótesis de que el efecto Región está efectivamente ausente, podríamos con la información proporcionada calcular el estadístico Q_h, que en el caso presente tomaría el valor aproximado:

    (a) -1.844

    (b) 4.123

    (c) 3.751

    (d) 2.637

26. El valor de Q_h seleccionado en la cuestión anterior debería, para valorar si es significativo o no, compararse con los cuantiles de una distribución

    (a) t de Student con 8 grados de libertad.

    (b) t de Student con 37 grados de libertad.

    (c) t de Student con (37-8) grados de libertad.

    (d) $\mathcal{F}$ de Snedecor con 8 y 38 grados de libertad.

    (e) Todo falso.

27. Si hiciéramos lo correcto, comprobaríamos que el efecto Region es significativo (a los niveles habituales del 5 %) y no debe ser eliminado del modelo. Podemos no obstante tratar ahora de modificar `mod1` eliminando algunos otros de los regresores cuyos coeficientes no son significativos.

```
> mod3 <- lm(SATM ~ region+SATV)
> summary(mod3)

Call:
lm(formula = SATM ~ region + SATV)

Residuals:
     Min       1Q   Median       3Q      Max
-16.2787  -3.7087  -0.2828   5.0038  21.0492

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  79.9707    31.0761   2.573 0.013785 *
regionESC   -12.5729     5.4782  -2.295 0.026919 *
regionMA     -7.4787     6.1297  -1.220 0.229410
regionMTN    -5.7553     4.4878  -1.282 0.206892
regionNE    -18.0306     4.9264  -3.660 0.000713 ***
regionPAC    -4.1508     5.1623  -0.804 0.425995
regionSA    -17.4405     4.8938  -3.564 0.000945 ***
regionWNC     2.2544     5.3692   0.420 0.676772
regionWSC   -13.6344     5.2368  -2.604 0.012789 *
SATV          0.9508     0.0685  13.880  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05

Residual standard error:
7.759 on 41 degrees of freedom
Multiple R-Squared: 0.9587,
Adjusted R-squared: 0.9496
F-statistic: 105.7 on 9 and 41 DF, p-value: < 2.2e-16

> sum(residuals(mod3)^2)
[1] 2468.156
```

En `mod3`, los regresores cuyos nombres comienzan por `region`),

(a) Son claramente colineales: un síntoma evidente de ello es que siendo el efecto Region claramente significativo, haya muchos parámetros estimados no significativos.

(b) Lejos de ser colineales, se trata de columnas de ceros y unos, mutuamente ortogonales.

(c) Todo falso.

28. *Suponiendo* que el modelo `mod1` es el más parametrizado entre los razonables, una estimación insesgada de σ^2, varianza de la perturbación, sería:

(a) 2195.949 / 37 = 59.35

(b) 2195.949 / 13 = 168.92

(c) 2195.949 / 51 = 43.058

(d) Todo falso.

29. Haciendo uso de la información anterior y de la suma de cuadrados de residuos de los modelos `mod2` y `mod3` puedes calcular fácilmente sus estadísticos C_p que resultan ser respectivamente:

(a) 123.345 y 213.234

(b) 12.134 y 23.432

(c) 32.491 y 43.456

(d) 70.098 y 61.586

30. Sobre la base del cálculo en la cuestión anterior, resultaría que:

(a) Todo falso.

(b) La diferencia entre `mod2` y `mod3` no es significativa.

(c) Es preferible el modelo `mod2`.

(d) Es preferible el modelo `mod3`.

31. Examinemos algunas cuestiones acerca de los residuos del modelo `mod1`. Si calculamos la media aritmética de dichos residuos mediante

```
> mean(residuals(mod1))
[1] 6.731407e-17
```

obtenemos un valor indistinguible de cero ($6{,}731407 \times 10^{-17}$). Esto indica:

(a) Todo falso.

(b) Que el modelo ajusta bien.

(c) No indica nada: ocurre siempre que hay columna de "unos".

(d) Que la estimación de la varianza es insesgada.

32. Si ahora obtenemos los residuos ordinarios y recordamos que el número de observaciones y grados de libertad en el modelo `mod1`, llegaremos a la conclusión de que podemos comparar el mayor de ellos (en valor absoluto) con

(a) La distribución del máximo de 51 variables t de Student con 37 grados de libertad.

(b) La distribución del máximo de 51 variables t de Student con 36 grados de libertad.

(c) La distribución t de Student con 37 grados de libertad.

(d) La distribución t de Student con 36 fgrados de libertad.

(e) Todo falso: los residuos MCO no tienen la misma distribución, son heterocedásticos.

FIN DEL BLOQUE DE PREGUNTAS

33. Una observación con residuo MCO muy grande:
    (a) Contribuye de modo importante a engrosar SSE.
    (b) Siempre tendrá también un residuo borrado grande.
    (c) Tendrá gran influencia sobre al menos uno de los β's estimados.
    (d) Corresponde a una observación que siempre tendremos interés en desechar.
34. Si para estimar un modelo con p regresores (excluida la columna de "unos") empleamos el método de componentes principales y construimos el estimador $\hat{\boldsymbol{\beta}}_{\boldsymbol{CP}}$ haciendo uso de todas (las p) componentes principales,
    (a) Todo falso.
    (b) $\hat{\boldsymbol{\beta}}_{\boldsymbol{CP}}$ será idéntico al $\hat{\boldsymbol{\beta}}_{\boldsymbol{MCO}}$ o estimador mínimo cuadrático ordinario.
    (c) El $\hat{\boldsymbol{\beta}}_{\boldsymbol{CP}}$ obtenido será sesgado.
    (d) Es imposible hacer lo que se propone: las componentes principales son linealmente dependientes unas de otras, y hemos de excluir al menos una.
35. 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.
36. La coexistencia de un R^2 muy elevado y unos t-ratios en su totalidad no significativos —salvo, quizá, el correspondiente a la columna de "unos"— es un síntoma indicativo de que:
    (a) La especificación lineal es inadecuada, y se hace preciso probar una regresión no lineal.
    (b) El número de observaciones es claramente insuficiente.
    (c) Existen *outliers*.
    (d) Existe multicolinealidad.
37. Si se hace regresión en componentes principales y se toman tantas componentes principales como regresores hay (excluida, en su caso, la columna de "unos"),
    (a) Se obtiene un estimador insesgado pero no lineal.
    (b) Se soluciona radicalmente el problema de la multicolinealidad, porque al ser la componentes principales ortogonales por construcción, no puede existir multicolinealidad ni ninguno de sus perniciosos efectos.
    (c) Se obtiene un estimador no lineal y sesgado.
    (d) Se obtiene una solución idéntica a la que obtendríamos mediante el estimador *ridge* haciendo $k = p$.
38. El supuesto de que las perturbaciones $\boldsymbol{\epsilon}$ del modelo $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ son incorreladas es imprescindible para poder demostrar:
    (a) Que los estimadores de los parámetros $\boldsymbol{\beta}$ son insesgados.
    (b) Que los estimadores de los parámetros $\boldsymbol{\beta}$ son lineales.
    (c) Que la matriz de diseño es de rango completo.
    (d) Todo falso.

Sección 3. Preguntas breves

39. En general, si estimamos el modelo

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

y a continuación el modelo ampliado

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\epsilon}$$

los estimadores de los parámetros $\boldsymbol{\beta}$ no son iguales en ambos modelos, salvo en el caso particular de que las matrices de regresores $\boldsymbol{X}$ y $\boldsymbol{Z}$ tengan todas sus columnas mutuamente ortogonales. Demuéstralo.

40. Enuncia y demuestra el teorema de Gauss-Markov.

Respuestas al examen de tipo A

Sección 1. 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!

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

3. 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.

Sección 2. Cuestiones de elección múltiple

1. Dado un modelo de regresión lineal $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, las estimaciones $\hat{\boldsymbol{\beta}}$ de los parámetros permiten construir $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ tal que:

   (a) **$\boldsymbol{X}\hat{\boldsymbol{\beta}}$ es la proyección de $\boldsymbol{y}$ sobre el subespacio generado por las columnas de $\boldsymbol{X}$.**

   (b) $||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2$ es máximo

   (c) $\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$

   (d) $\boldsymbol{y}$ tiene correlación 1 con $\boldsymbol{X}\hat{\boldsymbol{\beta}}$.

2. En un modelo ANOVA con dos tratamientos,

$$y_{ijk} = \alpha + \alpha_i^A + \alpha_j^B + \epsilon_{ijk}$$

   con replicación y las restricciones habituales,

$$\sum_i \alpha_i^A = \sum_j \alpha_j^B = 0$$

   se verifica que:

   (a) **El efecto estimado para un nivel i cualquiera del tratamiento A es el mismo sea cual fuere el nivel j del tratamiento B presente.**

   (b) Los residuos son independientes unos de otros.

   (c) Las interacciones resultan estimadas por $(\hat{\alpha}_i^A \hat{\alpha}_j^B)$.

   (d) Todo falso.

3. Para facilitar al analista la selección de modelos de regresión hay una porción de estrategias o algoritmos de búsqueda. ¿Cuál de ellos, si es que alguno, garantiza hallar *el mejor* de los modelos, de acuerdo con el criterio que hayamos especificado?

   (a) **La estrategia *all subsets* o búsqueda entre todos los subconjuntos de regresores.**

   (b) La regresión escalonada hacia delante.

   (c) La regresión esclonada hacia atrás.

   (d) La regresión escalonada mixta.

   (e) Todo falso.

4. Un valor propio cercano a cero de la matriz $X'X$ sería indicio de:

   (a) **Multicolinealidad aproximada.**

   (b) No linealidad del diseño.

   (c) R^2 muy pequeña.

   (d) Varianza de la perturbación muy grande.

5. Los factores de incremento de varianza (o VIF's, *variance inflation factors*):

   (a) Son estadísticos que permiten diagnosticar multicolinealidad.

   (b) Están relacionados con los residuos internamente studentizados.

   (c) Están relacionados con los residuos externamente studentizados.

   (d) Están relacionados con la R^2 de la regresión.

   (e) Toman valores entre 0 y 1.

6. El estimador *ridge:*

   (a) Es un estimador no lineal: por eso puede lograr un mejor ECM (error cuadrático medio) que el estimador MCO.

   (b) No puede mejorar al MCO en términos de varianza, como quedó suficientemente demostrado en el teorema de Gauss-Markov.

   (c) Es particularmente interesante en situaciones de fuerte multicolinealidad, aunque puede mejorar el ECM en todos los casos.

7. Si tenemos 10 sucesos, $E_1, \ldots, E_{10}$, independientes o no, cada uno de los cuales se verifica con probabilidad 0,99, la probabilidad Prob$\{\cap_{i=1}^{1} 0E_i\}$ de que *los diez simultáneamente* se verifiquen puede acotarse mediante la desigualdad de Bonferroni. Efectuado el cálculo resulta que

   (a) $1 \geq$ Prob$\{\cap_{i=1}^{1} 0E_i\} > 0{,}99$.

   (b) **Prob**$\{\cap_{i=1}^{1} 0E_i\} \geq 0{,}90$.

   (c) $0{,}90 >$ Prob$\{\cap_{i=1}^{1} 0E_i\} \geq 0{,}10$.

   (d) Prob$\{\cap_{i=1}^{1} 0E_i\} \leq 0{,}10$.

8. Cuando en un modelo que incluye la columna de "unos" tomamos una variable y la reescalamos (expresándola, por ejemplo, en unidades que son mil veces mayores que las primitivas, de modo que la variable correspondiente toma valores mil veces más pequeños), se verifica que:

   (a) El ajuste es idéntico al primitivo; el coeficiente estimado de la variable reescalada será ahora mil veces más pequeño.

   (b) El ajuste es idéntico al primitivo; el coeficiente estimado de la variable reescalada será ahora mil veces más grande.

   (c) R^2 se multiplica por 1000^2.

   (d) SSE y SSR no varían, pero SST puede hacerlo.

9. El estimador MCO verifica, con los supuestos habituales, que:

   (a) Es de mínima varianza entre todos los estimadores lineales.

   (b) Es de mínima varianza entre todos los estimadores insesgados.

   (c) Es de mínimo error cuadrático medio entre todos los estimadores lineales.

   (d) Es de mínimo error cuadrático medio entre todos los estimadores lineales e insesgados.

10. La validación cruzada *(cross-validation*) consiste en utilizar alternativamente las observaciones para estimar y validar u modelo. La versión más extrema *(leave one out)* reserva por turno una sóla observación para validar y emplea las $N-1$ restantes para estimar el modelo. Esto requiere en principio estimar el modelo N veces, lo que es muy costoso. En el caso concreto del modelo de regresión lineal, sin embargo, el cálculo es mucho más económico. Basta estimar el modelo una única vez. Los errores de ajuste de cualquier observación si la dejamos fuera y estimamos mediante las restantes coinciden con:

    (a) Los residuos borrados.

    (b) Los residuos internamente *studentizados*

    (c) Los residuos externamente *studentizados*

    (d) Todo falso.

11. ¿Cuales de las siguientes propiedades posee *necesariamente* la matriz de covarianzas de los residuos de un modelo de regresión lineal ordinaria?

    (a) Diagonal.

    (b) Idempotente.

    (c) Simétrica.

    (d) De rango completo.

12. Se conoce como distancia de Cook:

    (a) El mayor número de millas náuticas que el capitán Cook recorrió durante un periodo de 24 horas consecutivas en sus exploraciones por los mares australes.

    (b) Una medida de influencia de la observación i-ésima sobre *uno* cualquiera de los β_i.

    (c) Una medida de influencia de la observación i-ésima sobre *el conjunto* de los β_j.

    (d) El denominador de los residuos borrados.

    (e) Todo falso.

13. ¿En cuál o cuales de las siguientes situaciones resultaría manifiestamente inadecuada la inclusión en el modelo de una columna de "unos"?

    (a) Regresión de Y = "Consumo en el periodo t" sobre X = "Renta en el periodo t" para una muestra de N familias, con objeto de estimar la propensión marginal al consumo (supuesta constante).

    (b) Regresión de Y = "Peso de un objeto" sobre X = "Volumen del mismo objeto", para una muestra de N objetos de la misma materia con el fín de estimar el peso específico de dicha materia.

    (c) Regresión de Y = "Velocidad de sedimentación de una solución acuosa" sobre X = "Temperatura en ºC" para una muestra de N soluciones de la misma materia y concentración.

14. Cuando se omite en un modelo de regresión lineal un regresor que hubiera debido aparecer:

    (a) *Siempre* se sesgan los estimadores de los β's correspondientes a los regresores incluidos.

    (b) En general, se sesgan los estimadores de los β's correspondientes a los regresores incluidos; pero podría ocurrir que no fuera así si el regresor omitido es ortogonal a todos los presentes.

    (c) *Nunca* se sesga el estimador de σ^2, varianza de la perturbación. Sólamente se pierden grados de libertad.

    (d) El estimador de σ^2, varianza de la perturbación, continúa siendo insesgado si el regresor omitido es ortogonal a los incluidos.

15. Cuando se ajusta un modelo $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ y la columna de "unos" está entre los regresores, la R^2 ordinaria (no corregida) puede interpretarse como:

    (a) El ángulo que forma $\boldsymbol{y}$ con el subespacio que generan las columnas de $\boldsymbol{X}$.

    (b) El coseno del ángulo que forma $\boldsymbol{y}$ con el subespacio que generan las columnas de $\boldsymbol{X}$.

    (c) El coseno al cuadrado del ángulo que forma $\boldsymbol{y}$ con el subespacio que generan las columnas de $\boldsymbol{X}$.

    (d) El nivel de significación de los $\boldsymbol{\beta}$ estimados.

16. Cuando se incluyen regresores irrelevantes en un modelo de regresión:

    (a) Las estimaciones de cualquiera de los β's pueden resultar sesgadas.

    (b) La estimación de σ^2 resultará sesgada por exceso.

    (c) La estimación de σ^2 resultará sesgada por defecto.

    (d) La estimación de σ^2 tendrá menos grados de libertad que los que hubiera tenido de ajustarse el modelo correcto.

COMIENZO DE UN BLOQUE DE PREGUNTAS

Las preguntas hasta el siguiente trazo horizontal hacen referencia a los datos que siguen, relacionando para diferentes Estados USA los resultados en el SATM (un examen de Matemáticas) y las siguientes otras variables:

`region`	Región USA	
	ENC	East North Central
	ESC	East South Central
	MA	Mid-Atlantic
	NE	New England
	PAC	Pacific
	SA	South Atlantic
	WNC	West North Central
	WSC	West South Central
`SATV`	Resultado examen aptitud verbal.	
`pop`	Población en miles.	
`percent`	Porcentaje de graduados tomando el SATM.	
`dollars`	Gasto en 1000's$ por alumno.	
`pay`	Salario en 1000's$ por profesor.	

Ajustamos un modelo de regresión del modo que sigue:

```
> mod1 <- lm(SATM ~ . , data=States)
> summary(mod1)

Call:
lm(formula = SATM ~ ., data = States)

Residuals:
     Min       1Q   Median       3Q      Max
-15.5931  -3.2600  -0.5519   3.6316  24.7618

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.598e+01  4.878e+01   1.968   0.0566 .
regionESC   -1.039e+01  5.919e+00  -1.756   0.0873 .
regionMA    -6.204e+00  9.084e+00  -0.683   0.4989
regionMTN   -2.136e+00  5.002e+00  -0.427   0.6718
regionNE    -9.434e+00  7.757e+00  -1.216   0.2316
regionPAC   -5.672e-01  5.590e+00  -0.101   0.9197
regionSA    -1.231e+01  5.868e+00  -2.098   0.0428 *
regionWNC    4.708e+00  5.730e+00   0.822   0.4165
regionWSC   -1.247e+01  5.831e+00  -2.138   0.0392 *
pop          4.910e-04  2.999e-04   1.637   0.1100
SATV         9.082e-01  1.027e-01   8.839 1.18e-10 ***
percent     -1.855e-01  1.730e-01  -1.072   0.2907
dollars      1.602e+00  2.334e+00   0.687   0.4966
pay         -1.560e-01  5.652e-01  -0.276   0.7841
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05

Residual standard error: 7.704 on 37 degrees of freedom
Multiple R-Squared: 0.9632,  Adjusted R-squared: 0.9503
F-statistic:  74.6 on 13 and 37 DF,  p-value: < 2.2e-16
```

Adicionalmente, calculamos la suma de cuadrados de los residuos,

```
> sum(residuals(mod1)^2)
[1] 2195.949
```

Haciendo uso de la información precedente, responde a las siguientes cuestiones:

17. En este caso, el coeficiente R^2 ordinario es mayor que el $\overline{R}^2$ (o corregido).
    (a) Es pura casualidad, podría ser al revés.
    (b) Esto puede ocurrir en presencia de fuerte multicolinealidad.
    (c) Esto indica claramente que hay *outliers*
    (d) Este es necesariamente el caso.

18. Del estadillo anterior deducimos:
    (a) Se ha empleado una muestra con 13 observaciones.
    (b) Se ha empleado una muestra con 37 observaciones.
    (c) Se ha empleado una muestra con 51 observaciones.
    (d) Se ha empleado una muestra con 50 observaciones.

19. La suma total de cuadrados (SST) será:
    (a) 2086.810
    (b) 2310.796
    (c) 2115.138
    (d) 2279.847
    (e) Nada de lo anterior.

20. La variable `pay` tiene asociado un $\hat{\beta} = -1{,}560 \times 10^{-1}$. Ello es evidencia de que cuanto más se paga a los vagos de los profesores, más a la ligera se lo toman y peores son los resultados de los alumnos en el SATM.
    (a) Cierto
    (b) Falso

21. "Como cabría esperar, los resultados en el examen verbal (SATV) y matemático, tienen poco que ver."
    (a) Cierto: parece lógico que así sea. El SATV reflejaría un perfil del estudiante "de letras" y el SATM el de uno "de ciencias", perfiles que como se sabe son bastante antagónicos.
    (b) Falso, el resultado en el SATV acontece que es muy buen predictor del resultado en SATM. Los alumnos "buenos" parecen serlo en todo, y los "malos" también son malos en todo.

22. El coeficiente 0.1100 en la línea `pop` y bajo la columna `Pr(>|t|)` significa:
    (a) Que en el 11 % de los Estados, la variable `pop` influyó en el resultado del SATM.
    (b) Que el valor estimado del parámetro es apenas mayor que la décima parte de su desviación típica.
    (c) Que si afirmáramos que la variable `pop` influye en el resultado de SATM, estaríamos diciendo la verdad el 11 % de las veces.
    (d) Que si el verdadero coeficiente fuera cero, y todos los supuestos habituales se cumplieran, habría probabilidad 0.1100 de estimar un $|\hat{\beta}| \geq 0{,}000491$ para la variable `pop`.
    (e) Todo es falso.

23. El coeficiente $\hat{\beta} = -1{,}247 \times 10^{1}$ estimado para la variable `regionWSC` significa (prescindiendo de si es o no significativo) que los estudiantes en la región WSC, en igualdad de todo lo demás, obtienen una nota en el SATM:
    (a) Inferior en 12.47 % a la obtenida por los estudiantes de cualquier otra región.
    (b) Todo falso
    (c) Inferior en 12.47 a la obtenida por los estudiantes en la región de referencia, que en este caso es ENC.
    (d) Inferior en 12.47 a la obtenida por los estudiantes de otras regiones.

24. "Hay evidencia concluyente de que por cada 1 % adicional de estudiantes que se presentan al SATM, la nota desciende 0.1855. Ello es lógico: en los Estados en que sólo se presentan los mejores, la nota será más alta que en los Estados en que se presentan muchos (al igual que sucede con el porcentaje de aprobados en acceso a la Universidad, mayor para los institutos/colegios que presentan a pocos y buenos alumnos, que para los que dejan presentarse a todos)."

    (a) Todo falso.

    (b) Lo que la estimación sugiere es justamente lo contrario: allá donde se presenta un mayor porcentaje al examen, también las notas son superiores: de la cantidad nace la calidad.

    (c) Los datos corroboran el párrafo entrecomillado, dado el signo del $\hat{\beta}$.

    (d) El argumento anterior puede tener lógica, pero no resulta avalado por los datos: el $\hat{\beta}$ correspondiente no es significativo.

25. Como hay muchos parámetros no significativos de entre los que recogen el efecto Región, podemos plantearnos un modelo sin ellos:

```
> mod2 <- lm(SATM ~ . - region , data=States)
> summary(mod2)

Call:
lm(formula = SATM ~ . - region, data = States)

Residuals:
    Min      1Q  Median      3Q     Max
-14.749  -5.448  -0.649   3.873  34.560

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.2975427 42.4162502   0.856   0.3967
pop          0.0005321  0.0002716   1.959   0.0563
SATV         1.0030845  0.0867660  11.561 4.57e-15
percent     -0.2584571  0.1289731  -2.004   0.0511
dollars      1.5690544  1.9961347   0.786   0.4360
pay          0.3090260  0.5009215   0.617   0.5404
---

Residual standard error:
  8.754 on 45 degrees of freedom
Multiple R-Squared: 0.9423,
Adjusted R-squared: 0.9359
F-statistic: 147 on 5 and 45 DF, p-value: < 2.2e-16

> sum(residuals(mod2)^2)
[1] 3448.105
```

Para contrastar la hipótesis de que el efecto Región está efectivamente ausente, podríamos con la información proporcionada calcular el estadístico Q_h, que en el caso presente tomaría el valor aproximado:

(a) -1.844

(b) 4.123

(c) 3.751

(d) 2.637

26. El valor de Q_h seleccionado en la cuestión anterior debería, para valorar si es significativo o no, compararse con los cuantiles de una distribución

    (a) t de Student con 8 grados de libertad.

    (b) t de Student con 37 grados de libertad.

    (c) t de Student con (37-8) grados de libertad.

    (d) $\mathcal{F}$ de Snedecor con 8 y 38 grados de libertad.

    (e) Todo falso.

27. Si hiciéramos lo correcto, comprobaríamos que el efecto Region es significativo (a los niveles habituales del 5 %) y no debe ser eliminado del modelo. Podemos no obstante tratar ahora de modificar `mod1` eliminando algunos otros de los regresores cuyos coeficientes no son significativos.

```
> mod3 <- lm(SATM ~ region+SATV)
> summary(mod3)

Call:
lm(formula = SATM ~ region + SATV)

Residuals:
     Min       1Q   Median       3Q      Max
-16.2787  -3.7087  -0.2828   5.0038  21.0492

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  79.9707    31.0761   2.573 0.013785 *
regionESC   -12.5729     5.4782  -2.295 0.026919 *
regionMA     -7.4787     6.1297  -1.220 0.229410
regionMTN    -5.7553     4.4878  -1.282 0.206892
regionNE    -18.0306     4.9264  -3.660 0.000713 ***
regionPAC    -4.1508     5.1623  -0.804 0.425995
regionSA    -17.4405     4.8938  -3.564 0.000945 ***
regionWNC     2.2544     5.3692   0.420 0.676772
regionWSC   -13.6344     5.2368  -2.604 0.012789 *
SATV          0.9508     0.0685  13.880  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05

Residual standard error:
7.759 on 41 degrees of freedom
Multiple R-Squared: 0.9587,
Adjusted R-squared: 0.9496
F-statistic: 105.7 on 9 and 41 DF, p-value: < 2.2e-16

> sum(residuals(mod3)^2)
[1] 2468.156
```

En `mod3`, los regresores cuyos nombres comienzan por `region`),

(a) Son claramente colineales: un síntoma evidente de ello es que siendo el efecto Region claramente significativo, haya muchos parámetros estimados no significativos.

(b) Lejos de ser colineales, se trata de columnas de ceros y unos, mutuamente ortogonales.

(c) Todo falso.

28. *Suponiendo* que el modelo `mod1` es el más parametrizado entre los razonables, una estimación insesgada de σ^2, varianza de la perturbación, sería:

(a) 2195.949 / 37 = 59.35

(b) 2195.949 / 13 = 168.92

(c) 2195.949 / 51 = 43.058

(d) Todo falso.

29. Haciendo uso de la información anterior y de la suma de cuadrados de residuos de los modelos `mod2` y `mod3` puedes calcular fácilmente sus estadísticos C_p que resultan ser respectivamente:

(a) 123.345 y 213.234

(b) 12.134 y 23.432

(c) 32.491 y 43.456

(d) 70.098 y 61.586

30. Sobre la base del cálculo en la cuestión anterior, resultaría que:

(a) Todo falso.

(b) La diferencia entre `mod2` y `mod3` no es significativa.

(c) Es preferible el modelo `mod2`.

(d) Es preferible el modelo `mod3`.

31. Examinemos algunas cuestiones acerca de los residuos del modelo `mod1`. Si calculamos la media aritmética de dichos residuos mediante

```
> mean(residuals(mod1))
[1] 6.731407e-17
```

obtenemos un valor indistinguible de cero ($6{,}731407 \times 10^{-17}$). Esto indica:

(a) Todo falso.

(b) Que el modelo ajusta bien.

(c) No indica nada: ocurre siempre que hay columna de "unos".

(d) Que la estimación de la varianza es insesgada.

32. Si ahora obtenemos los residuos ordinarios y recordamos que el número de observaciones y grados de libertad en el modelo `mod1`, llegaremos a la conclusión de que podemos comparar el mayor de ellos (en valor absoluto) con

(a) La distribución del máximo de 51 variables t de Student con 37 grados de libertad.

(b) La distribución del máximo de 51 variables t de Student con 36 grados de libertad.

(c) La distribución t de Student con 37 grados de libertad.

(d) La distribución t de Student con 36 fgrados de libertad.

(e) Todo falso: los residuos MCO no tienen la misma distribución, son heterocedásticos.

FIN DEL BLOQUE DE PREGUNTAS

33. Una observación con residuo MCO muy grande:
    (a) Contribuye de modo importante a engrosar SSE.
    (b) Siempre tendrá también un residuo borrado grande.
    (c) Tendrá gran influencia sobre al menos uno de los β's estimados.
    (d) Corresponde a una observación que siempre tendremos interés en desechar.

34. Si para estimar un modelo con p regresores (excluida la columna de "unos") empleamos el método de componentes principales y construimos el estimador $\hat{\boldsymbol{\beta}}_{\boldsymbol{CP}}$ haciendo uso de todas (las p) componentes principales,
    (a) Todo falso.
    (b) $\hat{\boldsymbol{\beta}}_{\boldsymbol{CP}}$ será idéntico al $\hat{\boldsymbol{\beta}}_{\boldsymbol{MCO}}$ o estimador mínimo cuadrático ordinario.
    (c) El $\hat{\boldsymbol{\beta}}_{\boldsymbol{CP}}$ obtenido será sesgado.
    (d) Es imposible hacer lo que se propone: las componentes principales son linealmente dependientes unas de otras, y hemos de excluir al menos una.

35. 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.

36. La coexistencia de un R^2 muy elevado y unos t-ratios en su totalidad no significativos —salvo, quizá, el correspondiente a la columna de "unos"— es un síntoma indicativo de que:
    (a) La especificación lineal es inadecuada, y se hace preciso probar una regresión no lineal.
    (b) El número de observaciones es claramente insuficiente.
    (c) Existen *outliers*.
    (d) Existe multicolinealidad.

37. Si se hace regresión en componentes principales y se toman tantas componentes principales como regresores hay (excluida, en su caso, la columna de "unos"),
    (a) Se obtiene un estimador insesgado pero no lineal.
    (b) Se soluciona radicalmente el problema de la multicolinealidad, porque al ser la componentes principales ortogonales por construcción, no puede existir multicolinealidad ni ninguno de sus perniciosos efectos.
    (c) Se obtiene un estimador no lineal y sesgado.
    (d) Se obtiene una solución idéntica a la que obtendríamos mediante el estimador *ridge* haciendo $k = p$.

38. El supuesto de que las perturbaciones $\boldsymbol{\epsilon}$ del modelo $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ son incorreladas es imprescindible para poder demostrar:
    (a) Que los estimadores de los parámetros $\boldsymbol{\beta}$ son insesgados.
    (b) Que los estimadores de los parámetros $\boldsymbol{\beta}$ son lineales.
    (c) Que la matriz de diseño es de rango completo.
    (d) Todo falso.

Sección 3. Preguntas breves

39. En general, si estimamos el modelo
$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$
y a continuación el modelo ampliado
$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\epsilon}$$
los estimadores de los parámetros $\boldsymbol{\beta}$ no son iguales en ambos modelos, salvo en el caso particular de que las matrices de regresores $\boldsymbol{X}$ y $\boldsymbol{Z}$ tengan todas sus columnas mutuamente ortogonales. Demuéstralo.

Respuesta: Ver apuntes de clase.

40. Enuncia y demuestra el teorema de Gauss-Markov.

Respuesta: Ver apuntes de clase.