



RUN: 21.266.659-9⁵

Segunda prueba escrita Parte escrita

Instrucciones

1. Hoy corresponde responder las preguntas escritas, para lo cual dispone de 70 minutos.
2. Esta prueba es individual. Identifíquese solo con su RUN, en el espacio provisto para ello.
3. Apague su teléfono celular y guárdelo en su mochila. Está prohibido sacarlo antes de salir de la sala.
4. Sus respuestas se entregan en este enunciado, utilizando solo el espacio otorgado para ello. Use un lápiz adecuado y letra clara.
5. Si tiene alguna duda sobre el enunciado, levante la mano y el profesor le entregará un papel para que la haga por escrito. No hable durante el desarrollo de la prueba.
6. La calificación se calcula teniendo en consideración tanto la parte práctica (41 puntos) como la de esta parte escrita (36 puntos). Los criterios de evaluación de las siguientes preguntas son los siguientes:

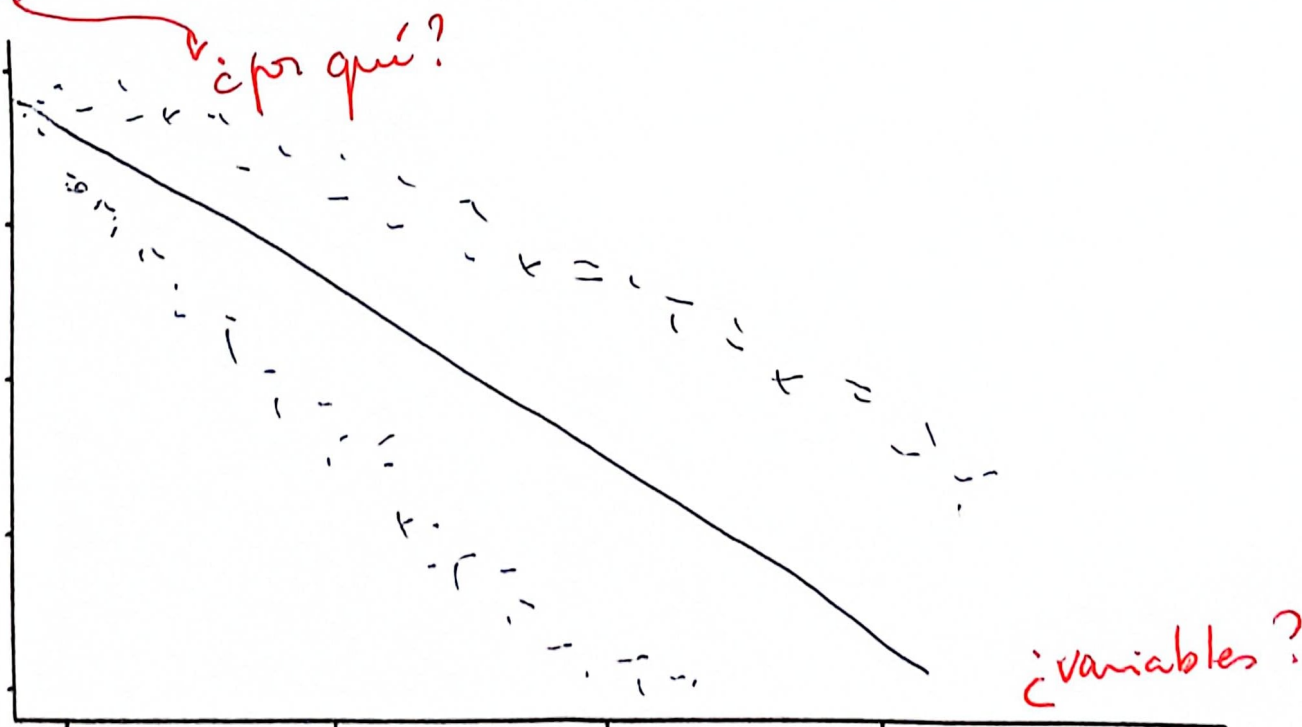
| Problema | Categoría | Nivel de logro | Puntos |
|----------|------------------------|--|--------|
| P3 | Estudio | Describe un estudio pertinente que es interesante y novedoso (no visto anteriormente en lecturas dadas o en clases), que se da naturalmente en el contexto solicitado, y argumenta sólidamente que su análisis requiere del modelo indicado y que es posible obtener los datos requeridos para su construcción. | 6 |
| | Ejemplo visual | Dibuja, con orden y sin borrones, un gráfico pertinente (de dispersión), cuidando completar las etiquetas de ambos ejes, en que se aprecia un ejemplo de los datos que podría observar en el estudio propuesto, con valores verosímiles, que ejemplifican la correlación y problemática indicadas en el enunciado. | 6 |
| P4 | Cálculo de métricas | Calcula correctamente los tres valores solicitados (exactitud, sensibilidad, especificidad), mostrando con claridad, y sin borrones, el procedimiento seguido y qué información ha usado para cada métrica. | 5 |
| | Curva ROC | Dibuja, con orden y sin borrones, un gráfico donde se aprecia con claridad la curva ROC que resulta de los datos entregados y los calculados correctamente en la pregunta anterior, cuidando completar las etiquetas de ambos ejes. | 6 |
| | Evaluación del modelo | Utilizando argumentos sólidos y completos, de forma clara y sin ideas espurias, emite un juicio correcto de la confiabilidad del modelo obtenido por el enólogo, que se basa en los resultados que se aprecian en el enunciado, y explica el tipo correcto de regresión utilizada para su construcción. | 6 |
| P5 | Estudio | Explica con argumentos sólidos, de forma clara y sin ideas espurias, cómo puede utilizar una técnica de remuestreo pertinente (bootstrapping) para identificar si existe sesgo en la media estimada (si la media observada se encuentra en el centro del histograma de la distribución bootstrap, entonces no hay evidencia de sesgo; pero si la media observada se encuentra hacia algún extremo, eso podría ser evidencia de sesgo). | 5 |
| General | Ortografía y redacción | Utiliza entre el 70% y 100% del espacio provisto para responder (no más), escribiendo con buena ortografía y redacción (≤5 errores), usando vocabulario propio de la disciplina y el contexto de las preguntas que responde. | 2 |
| TOTAL | | | 36 |

Pregunta 3.

a) Proponga un ejemplo novedoso (no mencionado en clase ni que aparezca en las lecturas dadas) de un estudio relación con la obligación de votar en todas las elecciones que requiera utilizar un modelo de regresión lineal simple. Justifique que es plausible obtener los datos para realizar el estudio que propone.

Se desea mejorar la cantidad de personas que van a votar, para esto se busca revisar si hay una relación entre la distancia media entre el local de votación y la residencia de sus votantes, con el porcentaje de asistencia a la votación por local. Para esto se requerirían los datos domiciliarios de los votantes asignados a un local de votación, para obtener la distancia entre ellos, estos datos se pueden obtener de encuestas censarias o de información recolectada por las mismas votaciones, para los porcentajes de asistencia similarmente debe existir un registro accesible con la información necesaria, ya que al votar se anota en un libro usando la cédula de identidad.

b) En el espacio provisto, dibuje cómo se verían un conjunto de datos en el estudio que propone si las variables estudiadas exhibieran un coeficiente de correlación de Pearson de aproximadamente $-0,3$ y presentarían problemas de linealidad que dificultarían su modelado.



Pregunta 4. Un enólogo de la región del Maule está estudiando la calidad de la producción sus cepas de Merlot en esta temporada. Basándose en estudios previos [Aeberhard et al. (1994) Pattern Recognition 27(8); Jackson et al. (1978) Journal of the Science of Food and Agriculture 29(8); Jover et al. (2004) Food Quality and preference 15(5); entre otros] construyó y evaluó un modelo de regresión logística usando el siguiente código (que no está completo):

```
set.seed(29)
i.entrenamiento <- sample(1:nrow(datos), 100) # 50% de los datos
datos.entrenamiento <- datos[i.entrenamiento, ]
datos.prueba <- datos[-i.entrenamiento, ]

modelo.rlog <- glm(quality ~ chlorides + sulphates + alcohol,
                  family = binomial,
                  data = datos.entrenamiento)

cat("\nCook:\n")
cd <- cooks.distance(modelo.rlog)
print(cd[cd > 1])

cat("\nDurbin-Watson:\n")
print(durbinWatsonTest(modelo.rlog))

cat("\nVIF:\n")
print(round(vif(modelo.rlog), 1))

muestra_rendimiento(modelo = modelo.rlog,
                    datos = datos.prueba,
                    umbrales = c(0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8)
                    )
```

77
22
99

Obteniendo los siguientes resultados:

Cook:
named numeric(0) ← Ninguno Mayor A 1.

Durbin-Watson:
lag Autocorrelation D-W Statistic p-value
1 0.1104422 1.760642 0.196 → Mayor a 0,05, residuos ind.
Alternative hypothesis: rho != 0

VIF:
chlorides sulphates alcohol vif Menor A 10.
1.3 1.3 1.1

Umbral: 0.2
Observado
Predicho Bueno Malo
Bueno 41 46
Malo 12 1
Exactitud: 0.42, sensibilidad: 0.77, especificidad: 0.02

Umbral: 0.3
Observado
Predicho Bueno Malo
Bueno 31 46
Malo 22 1
Exactitud: 0.42, sensibilidad: 0.58, especificidad: 0.02

VP + VN
n

Umbral: 0.4

| | Observado | |
|----------|-----------|------|
| Predicho | Bueno | Malo |
| Bueno | 26 | 46 |
| Malo | 27 | 1 |

VP
VP+FN

Exactitud: 0.27, sensibilidad: 7777, especificidad: 0.02

Umbral: 0.5

| | Observado | |
|----------|-----------|------|
| Predicho | Bueno | Malo |
| Bueno | 16 | 44 |
| Malo | 37 | 3 |

Exactitud: 0.19, sensibilidad: 0.3, especificidad: 0.06

Umbral: 0.6

| | Observado | |
|----------|-----------|------|
| Predicho | Bueno | Malo |
| Bueno | 15 | 37 |
| Malo | 38 | 10 |

VN
FN+VN

Exactitud: 0.25, sensibilidad: 0.28, especificidad: 7777

Umbral: 0.7

| | Observado | |
|----------|-----------|------|
| Predicho | Bueno | Malo |
| Bueno | 7 | 31 |
| Malo | 46 | 16 |

Exactitud: 0.23, sensibilidad: 0.13, especificidad: 0.34

Umbral: 0.8

| | Observado | |
|----------|-----------|------|
| Predicho | Bueno | Malo |
| Bueno | 2 | 20 |
| Malo | 51 | 27 |

Exactitud: 0.29, sensibilidad: 0.04, especificidad: 0.57

Se le pide:

a) Calcule los valores que faltan en la salida a pantalla del script (marcados con 7777).

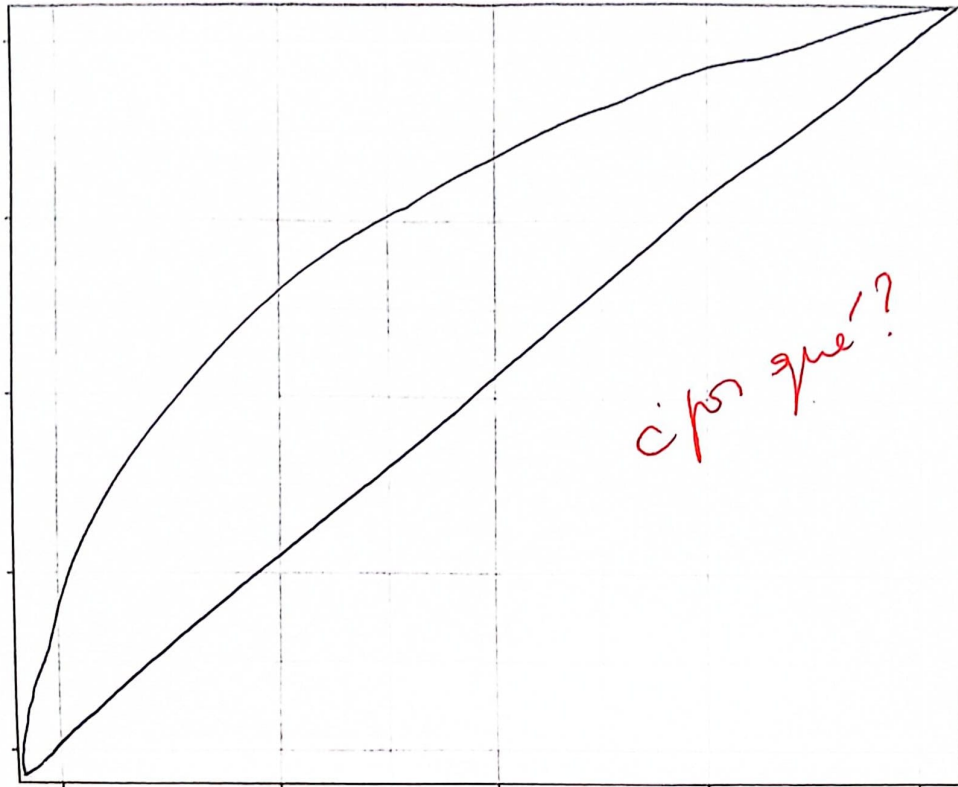
Realice sus cálculos en este espacio

1) Exactitud: $\frac{31+1}{100} = 0,32$ ✓

2) Sensibilidad: $\frac{26}{26+27} = \frac{26}{53}$ ✓

3) Especificidad: $\frac{10}{37+10} = \frac{10}{47}$ ✓

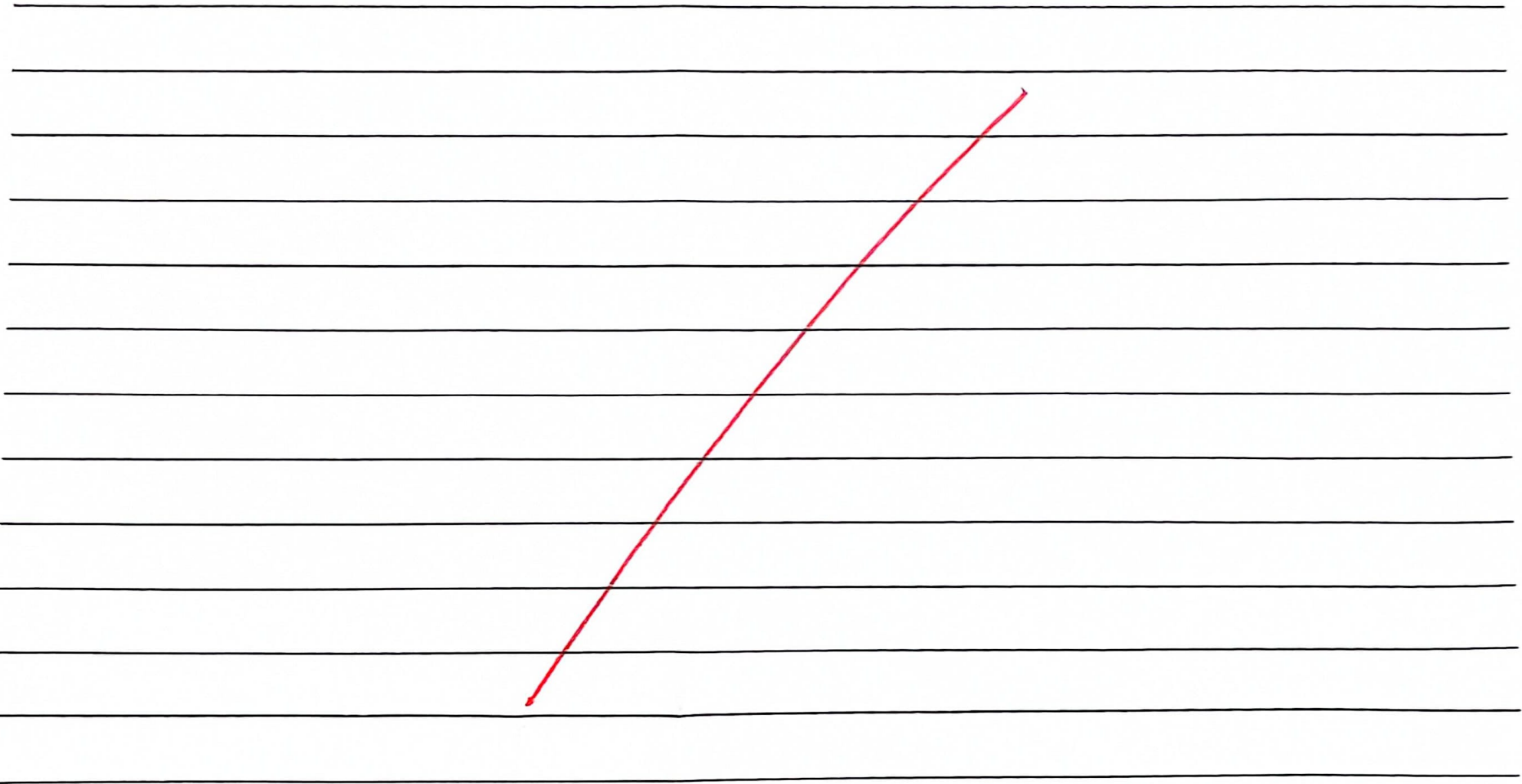
b) Construya una curva ROC que presente la calidad predictiva del modelo conseguido por el enólogo.



c) ¿Qué tipo de regresión ha utilizado el enólogo para construir su modelo de regresión logística? ¿Qué puede decir de la confiabilidad del modelo que obtuvo el enólogo?. Justifique su respuesta.

Uso una Regresión logística Binomial, el Modelo presenta buenas distancias de Cook, ninguna sobre 1, la prueba de Durbin-Watson No puede rechazar la hipótesis de que los Residuos son independientes y los Nif son menores a 10, todos son menores que 1, así que con esas datos no tengo evidencia para decir que el Modelo es poco confiable, aunque se podrían realizar más pruebas.

Pregunta 5. Considere que se ha medido la distancia diaria que caminan al interior del campus una muestra aleatoria de 40 estudiantes de primer año de Ingeniería durante una semana de clases normal, encontrando una media de 3 km y una desviación estándar de 1 km. ¿Cómo se puede utilizar remuestreo para detectar sesgos en la estimación de la media?



Buena suerte.