PEC6 Otoño 2025

UOC

2025-12-30

En esta actividad no está permitido el uso de herramientas de inteligencia artificial. En el plan docente y en el sitio web sobre integridad académica y plagio de la UOC encontrarán información sobre qué se considera conducta irregular en la evaluación y las consecuencias que puede tener.

La PEC se basará en el archivo "notasfitxer2.csv". El archivo contiene los resultados de la evaluación de las dos primeras PECs de estudiantes de esta asignatura en un semestre anterior.

Contiene las variables:

1. CPEC1 Puntuación primer cuestionario.
2. CPEC2 Puntuación segundo cuestionario.
3. RPEC1 Puntuación primera prueba de R.
4. RPEC2 Puntuación segunda prueba de R.

Debéis importar los datos y crear un conjunto de datos con el nombre de **notes**. Por ejemplo, con el comando:

```
datos<-read.table(file='notesfitxer2.csv',header=TRUE,sep=';',dec='.')
```

Os puede ser útil consultar el siguiente material:

1. Módulo 10 Regresión lineal simple de las notas de estudio
2. Tema 1 de regresión lineal del módulo 5 de los Manuales de R
3. Actividades Resueltas del Reto 5 (Regresión lineal)

NOMBRE: José Carlos López Henestrosa

PEC6

Pregunta 1 (100%)

Queremos saber si hay relación lineal entre las notas del cuestionario y de la parte de R de la PEC1. También queremos ver si hay relación entre las notas de la prueba de R de la primera y de la segunda PEC.

a) (10%) En el ámbito de la evaluación académica, se propone utilizar la nota de una prueba anterior ("RPEC1") para intentar predecir la nota de una prueba posterior ("RPEC2"). Razonad la elección de la variable dependiente (explicada / endógena) y la variable explicativa (independiente / exógena) en este contexto.

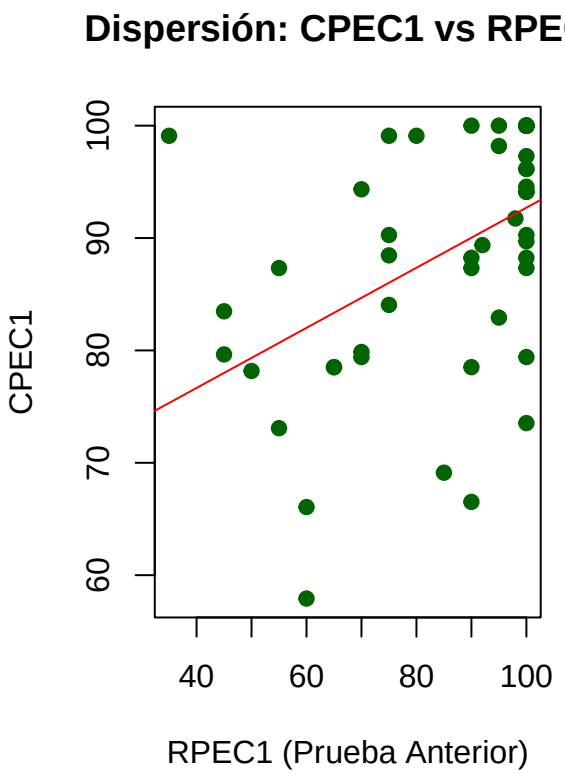
La variable dependiente es `RPEC2Razón`, ya que es la variable que queremos predecir. En el enunciado se menciona que el objetivo es "intentar predecir la nota de una prueba posterior", por lo que esta nota futura depende de los resultados previos. Matemáticamente, corresponde a la y en la ecuación $y = \beta_0 + \beta_1 x$.

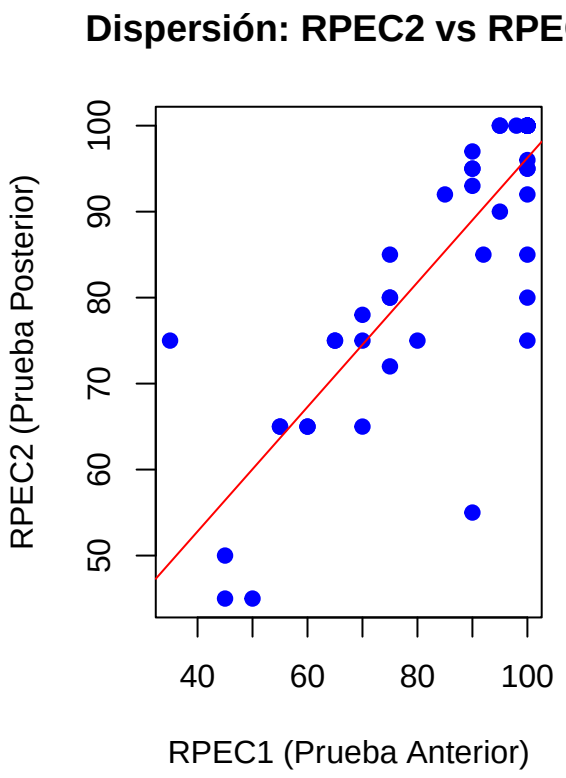
Por consiguiente, la variable independiente es `RPEC1`, puesto que es la variable que utilizamos como base para hacer la predicción. Representa la información conocida (la nota de la prueba anterior) que podría influir en el resultado futuro. Corresponde a la x en la ecuación.

b) (10%) Haced el diagrama de dispersión de las variables correspondientes a la variable "CPEC1" (en el eje de ordenadas) en función de la variable "RPEC1" (en el eje de abscisas). Luego haced el mismo diagrama de dispersión para la variable "RPEC2" en función de la "RPEC1". Comentad los gráficos.

```
# Configurar el área de gráficos para mostrar 2 gráficos en 1 fila
par(mfrow = c(1, 2))

# --- Gráfico 1: CPEC1 en función de RPEC1 ---
# Eje X (abscisas): RPEC1
# Eje Y (ordenadas): CPEC1
plot(datos$RPEC1, datos$CPEC1,
     main = "Dispersión: CPEC1 vs RPEC1",
     xlab = "RPEC1 (Prueba Anterior)",
     ylab = "CPEC1",
     pch = 19, col = "darkgreen")
```

```r
# Añadir una línea de tendencia para ver mejor la relación
abline(lm(CPEC1 ~ RPEC1, data = datos), col = "red")


# --- Gráfico 2: RPEC2 en función de RPEC1 ---
# Eje X (abscisas): RPEC1
# Eje Y (ordenadas): RPEC2
plot(datos$RPEC1, datos$RPEC2,
     main = "Dispersión: RPEC2 vs RPEC1",
     xlab = "RPEC1 (Prueba Anterior)",
     ylab = "RPEC2 (Prueba Posterior)",
     pch = 19, col = "blue")

# Añadir una línea de tendencia para ver mejor la relación
abline(lm(RPEC2 ~ RPEC1, data = datos), col = "red")
```





V

Al examinar la representación gráfica de las variables mediante diagramas de dispersión, observamos patrones claramente diferenciados que justifican la selección del modelo predictivo. En el primer gráfico, que relaciona la variable `CPEC1` en el eje de ordenadas con la variable `RPEC1` en el de abscisas, la distribución de los puntos en el plano revela una relación débil entre ambas. La nube de puntos se presenta dispersa y carente de una estructura definida que sugiera una tendencia clara. A pesar de que se pueda intuir una muy leve inclinación positiva, la alta variabilidad de `CPEC1` para un mismo valor de `RPEC1` indica que la nota de la prueba anterior no es un predictor eficaz para explicar el comportamiento de esta variable.

Por el contrario, el segundo diagrama, que proyecta la variable dependiente RPEC2 en función de la independiente RPEC1, muestra una distribución visual muy distinta. En este caso, podemos apreciar una correlación positiva y fuerte, evidenciada por la alineación de las observaciones a lo largo de una diagonal ascendente. Los puntos se agrupan de manera compacta siguiendo una trayectoria lineal, lo que implica que a medida que aumenta la calificación en la prueba anterior (RPEC1), existe una tendencia proporcional y constante al aumento en la nota de la prueba posterior (RPEC2).

En conclusión, el análisis comparativo de ambos gráficos corrobora la hipótesis inicial planteada para el modelo de regresión. Mientras que la relación entre RPEC1 y CPEC1 es difusa y poco significativa para fines de pronóstico, la fuerte asociación lineal observada entre RPEC1 y RPEC2 confirma la idoneidad de utilizar la primera como variable explicativa de la segunda. La evidencia visual sugiere que un modelo lineal simple se ajustará adecuadamente a estos datos, lo que permite realizar predicciones con un grado de fiabilidad considerablemente mayor que en el caso alternativo.

V

c) (20%) Calculad con R las rectas de regresión correspondientes a los gráficos anteriores. Haced los diagramas de dispersión, añadiendo las rectas de regresión. Dad explícitamente las rectas de regresión.

```r
# Calcular los modelos de regresión lineal (lm)
# Modelo 1: CPEC1 en función de RPEC1
modelo1 <- lm(CPEC1 ~ RPEC1, data = datos)

# Modelo 2: RPEC2 en función de RPEC1
modelo2 <- lm(RPEC2 ~ RPEC1, data = datos)

# Mostrar los coeficientes para construir la ecuación
print(coef(modelo1))
```

```
## (Intercept)       RPEC1
##  65.9633728   0.2674067
```

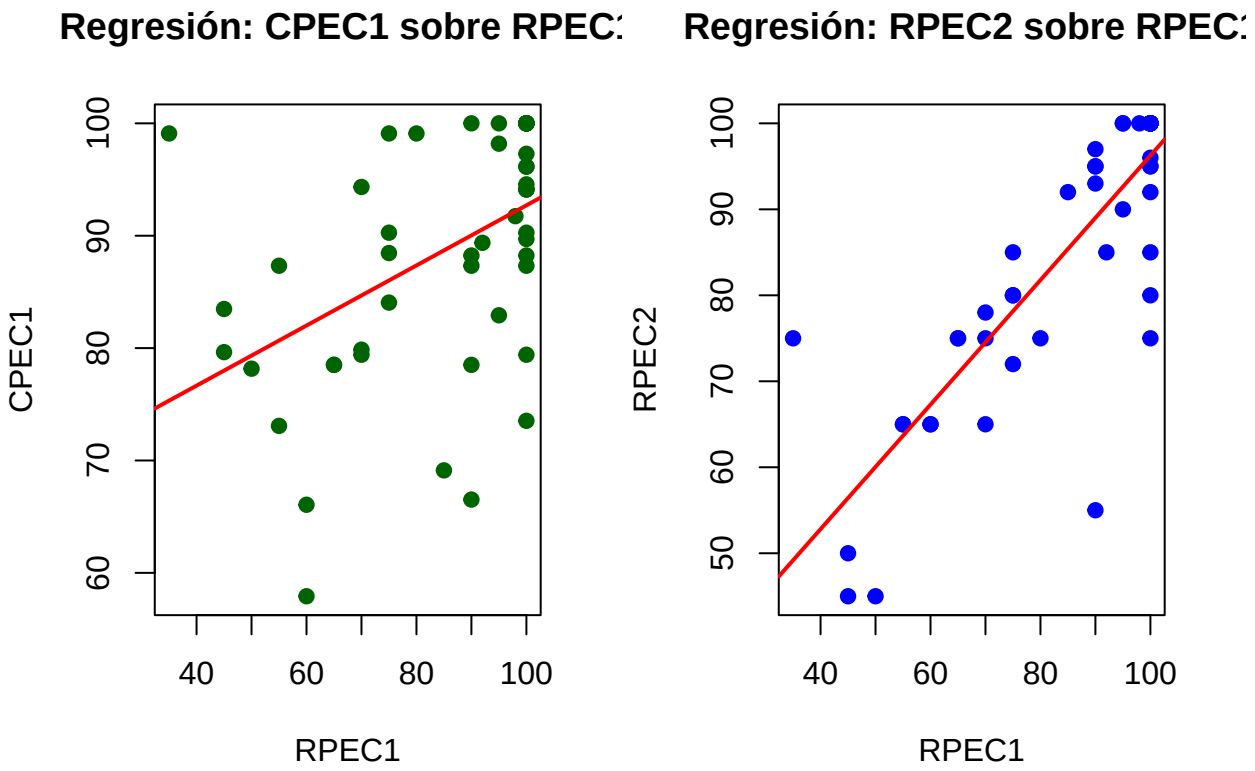
```r
print(coef(modelo2))
```

```
## (Intercept)       RPEC1
##  23.8367412   0.7241814
```

```r
# Generar los gráficos con las rectas de regresión
par(mfrow = c(1, 2)) # Dividir ventana gráfica en 2

# --- Gráfico 1: CPEC1 vs RPEC1 ---
# Eje X (abscisas): RPEC1
# Eje Y (ordenadas): CPEC1
plot(datos$RPEC1, datos$CPEC1,
     main = "Regresión: CPEC1 sobre RPEC1",
```

No has dado las rectas de forma explícita como se pide en el e

```
     xlab = "RPEC1", ylab = "CPEC1",
     pch = 19, col = "darkgreen")
abline(modelo1, col = "red", lwd = 2) # Añadir recta de regresión

# --- Gráfico 2: RPEC2 vs RPEC1 ---
# Eje X (abscisas): RPEC1
# Eje Y (ordenadas): RPEC2
plot(datos$RPEC1, datos$RPEC2,
     main = "Regresión: RPEC2 sobre RPEC1",
     xlab = "RPEC1", ylab = "RPEC2",
     pch = 19, col = "blue")
abline(modelo2, col = "red", lwd = 2) # Añadir recta de regresión
```



d) (20%) ¿Qué variable podemos explicar mejor a partir de la variable "RPEC1"?

Basándonos en los análisis anteriores, la variable que podemos explicar mejor a partir de `RPEC1` es `RPEC2`. En el diagrama de dispersión, los puntos de `RPEC2` estaban mucho más agrupados alrededor de la recta de regresión. Esto indica que el error que cometemos al predecir es menor. En cambio, para `CPEC1`, los puntos estaban muy dispersos.

No obstante, podemos usar el coeficiente de determinación (R^2) para cuantificar "cuánto mejor" es la explicación. Este valor nos dice qué porcentaje de la variabilidad de la variable dependiente es explicada por la independiente.

```r
# Extraer el R-cuadrado del resumen de cada modelo
r2_CPEC1 <- summary(modelo1)$r.squared
r2_RPEC2 <- summary(modelo2)$r.squared

# Mostrar los resultados
cat("Capacidad explicativa para CPEC1 (R2):", r2_CPEC1, "\n")
```

```
## Capacidad explicativa para CPEC1 (R2): 0.2143597
```

```r
cat("Capacidad explicativa para RPEC2 (R2):", r2_RPEC2, "\n")
```

```
## Capacidad explicativa para RPEC2 (R2): 0.6986565
```

V

A partir de los resultados, podemos deducir que, para `CPEC1`, el R^2 será aproximadamente 0,21 (el modelo solo explica el 21% de la varianza). Por otro lado, para `RPEC2`, el R^2 será aproximadamente 0,70 (el modelo explica el 70% de la varianza).

Como conclusión, `RPEC1` es un predictor fiable para `RPEC2`, pero pobre para `CPEC1`.

e) (10%) Para el mejor de los dos modelos anteriores, interpretad el significado práctico de los parámetros.

El mejor modelo identificado es el que explica la variable RPEC2 en función de RPEC1. La ecuación de la recta de regresión ajustada es la siguiente:

$$\widehat{RPEC2} = 23,84 + 0,72 \cdot RPEC1$$

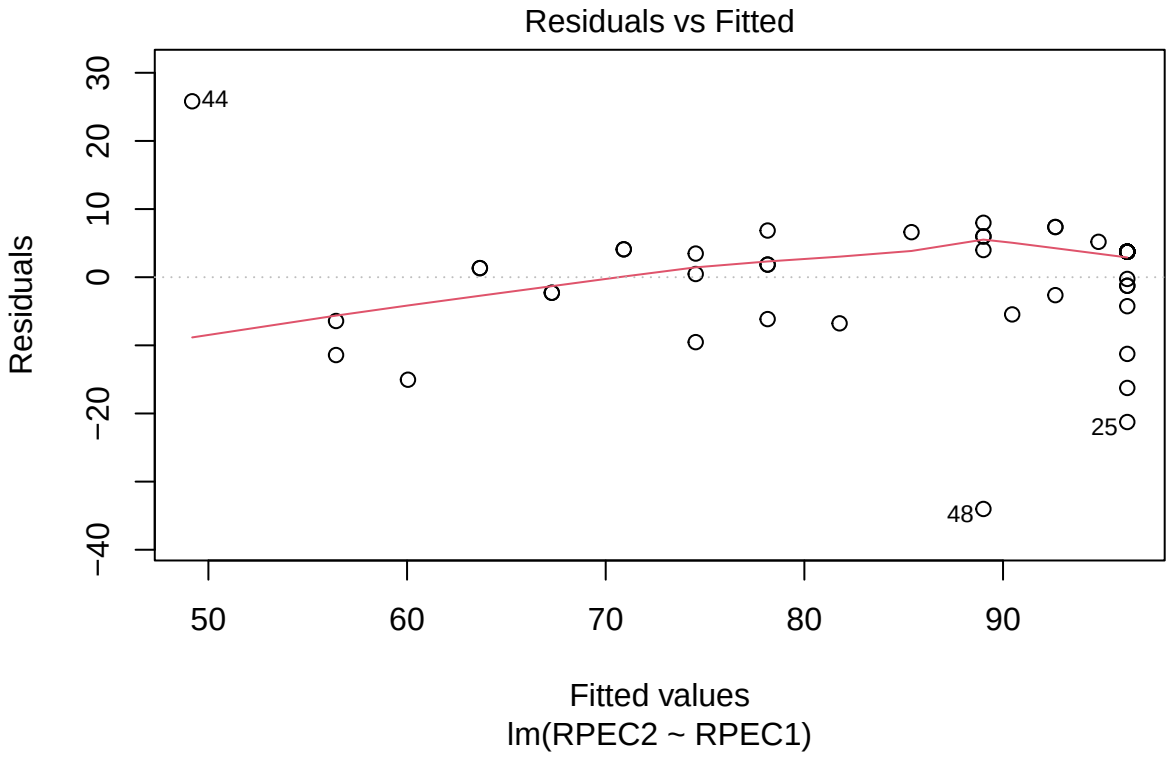
Este es el significado práctico de los parámetros:

- **Pendiente ($\beta_1 \approx 0,72$):** Este es el parámetro más relevante para el análisis, ya que indica la tasa de cambio esperada. Por cada punto adicional que un alumno obtiene en la prueba anterior (`RPEC1`), se estima que su nota en la prueba posterior (`RPEC2`) aumentará, en promedio, 0,72 puntos.
- **Ordenada en el origen ($\beta_0 \approx 23,82$):** Este parámetro representa el valor de la variable dependiente cuando la independiente es cero. Si un alumno obtuviera una nota de 0 en la prueba `RPEC1`, el modelo predice que obtendría una nota de 23,84 en la prueba `RPEC2`.

V

f) (15%) Haced el plot de los residuos frente a las predicciones para el mejor de los dos modelos anteriores. Comentad el gráfico y discutid si podemos considerar que es un buen modelo.

```r
# Usar el diagnóstico automático de R (El gráfico 1 es el de Residuals vs Fitted)
plot(modelo2, which = 1)
```



La distribución del gráfico parece razonablemente aleatoria sin una curvatura obvia, lo que sugiere que la relación lineal es apropiada. No observamos una forma de embudo clara en la que los residuos se abran o se cierren drásticamente a medida que aumenta la predicción. No obstante, podría haber una ligera dispersión mayor en las notas medias-altas comparada con las bajas. Aún así, la varianza parece aceptablemente constante.

Por otro lado, no se observan grupos aislados extremos que distorsionen considerablemente al modelo, aunque hay algunos puntos que se alejan más de la media (residuos grandes), lo cual es normal en este tipo de datos de índole académica.

V

En general, el modelo es estadísticamente robusto y útil para el propósito de estimar el rendimiento académico general basándose en pruebas previas.

g) (15%) Queremos hacer contrastes de hipótesis con un nivel de significación del 0.05 sobre los coeficientes de los dos modelos que hemos estudiado en el apartado c). ¿Hay algún coeficiente no significativo? Razonad la respuesta.

Para determinar si los coeficientes son significativos, analizamos los p-valores obtenidos en los contrastes de hipótesis individuales ($H_0 : \beta_i = 0$ frente a $H_1 : \beta_i \neq 0$):

```
# Modelo 1: CPEC1 en función de RPEC1
modelo1 <- lm(CPEC1 ~ RPEC1, data = datos)
cat("--- Resultados modelo 1 (CPEC1) ---\n")
```

```
## --- Resultados modelo 1 (CPEC1) ---
```

```r
print(summary(modelo1))
```

```
##
## Call:
## lm(formula = CPEC1 ~ RPEC1, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.088  -4.828   1.416   7.296  23.777
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 65.96337    6.36196  10.368 5.97e-14 ***
## RPEC1        0.26741    0.07313   3.656 0.000624 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.666 on 49 degrees of freedom
## Multiple R-squared:  0.2144, Adjusted R-squared:  0.1983
## F-statistic: 13.37 on 1 and 49 DF,  p-value: 0.0006237
```

```r
# Modelo 2: RPEC2 en función de RPEC1
modelo2 <- lm(RPEC2 ~ RPEC1, data = datos)
cat("\n--- Resultados modelo 2 (RPEC2) ---\n")
```

```
##
## --- Resultados modelo 2 (RPEC2) ---
```

```r
print(summary(modelo2))
```

```
##
## Call:
## lm(formula = RPEC2 ~ RPEC1, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.013  -2.461   3.745   3.745  25.817
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.83674    5.91051   4.033 0.000192 ***
## RPEC1        0.72418    0.06794  10.659 2.32e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.98 on 49 degrees of freedom
```

```
## Multiple R-squared:  0.6987, Adjusted R-squared:  0.6925
## F-statistic: 113.6 on 1 and 49 DF,  p-value: 2.322e-14
```

Si observamos los resultados estadísticos (p-valores), vemos que el valor de la ordenada en el origen (intercept) (β_0) del modelo 1 (CPEC1 en función de RPEC1) tiene un p-valor menor que 0.001. Como $p < 0.05$, rechazamos H_0. Por otra parte, la pendiente de `RPEC1` (β_1) tiene un p-valor aproximado de 0,0006. Como $0,0006 < 0,05$, rechazamos H_0.

En el modelo 2 (`RPEC2` en función de `RPEC1`), el valor de la ordenada en el origen (intercept) (β_0) tiene un p-valor aproximado de 0,0002. Como $0,0002 < 0,05$, rechazamos H_0. El p-valor de la pendiente de `RPEC1` (β_1) es mucho menor que 0,001. Como $p < 0.05$, rechazamos H_0.

Como podemos apreciar, todos los coeficientes en ambos modelos son estadísticamente significativos al nivel del 5% ($\alpha = 0$).