PAUTA PREGUNTA 2 CONTROL 3, 21.6.2004

a) El estimador MCO para β se obtiene multiplicando la matriz $(X^tX)^{-1}$ con el vector X^tY, cuyo resultado es:

$$\hat{\beta} = \begin{pmatrix} 764{,}842 \\ 0{,}0806 \\ -30{,}023 \\ -7{,}222 \end{pmatrix} \qquad \text{(Valores aproximados)}$$

NOTA: Recordar que (X^tX) es simétrica.

Los residuos se obtienen como $Y - X\hat{\beta}$, donde:

$$X = \begin{pmatrix} 1 & 286 & 8{,}6 & 78 \\ 1 & 238 & 3{,}4 & 81 \\ 1 & 223 & 11{,}5 & 68 \\ 1 & 148 & 8 & 82 \end{pmatrix} \qquad Y = \begin{pmatrix} 44 \\ 168 \\ 20 \\ 23 \end{pmatrix}$$

En este punto, dadas las aproximaciones a decimales, los resultados pueden diferir bastante. Sin embargo, teóricamente los residuos debieran dar 0 (como veremos a continuación).

Si en un modelo general $Y = X\beta + \varepsilon$ se cumple que el número de observaciones es igual al número de coeficientes, es decir si $n = k$, la matriz X es cuadrada, se tiene que:

$$(X^tX)^{-1} = X^{-1}(X^t)^{-1} \Rightarrow \hat{\varepsilon} = Y - X\hat{\beta} = Y - XX^{-1}(X^t)^{-1}X^tY = Y - I_n \cdot Y = 0_n$$

Es decir todos los residuos son 0. Luego, la suma residual $SSR = \hat{\varepsilon}^t\hat{\varepsilon}$ es 0, y por ende el valor de R^2 es igual 1, es decir, ajuste perfecto. Sin embargo, este resultado no es muy bueno, ya que al tener muy pocas observaciones el modelo no es muy robusto, es más, si cambiamos arbitrariamente cualquiera de las observaciones por otra con valores muy diferentes, el ajuste será nuevamente perfecto, lo que hace perder confiabilidad al modelo. Además, numéricamente, el tener esta cantidad de observaciones no deja ningún grado de libertad para estimar algún otro parámetro, ya que $n-k = 0$. En términos geométricos este resultado equivale a encontrar la ecuación de la recta exacta que pasa por k puntos (sistema de k incógnitas y k puntos).

b) La ecuación que relaciona los distintos valores de la tabla 4 es:

$$\frac{\hat{\beta}_i}{\tilde{\sigma}_{ii}} = t_{\hat{\beta}_i} \quad \text{es decir} \quad \frac{\text{Estimación}}{\text{Desviación típica}} = \text{t-Student}$$

Usando esta ecuación se encuentra que:

- t-Student de Constante = 1.295
- Estimación de Radsolar = 0.10812
- Desviación típica de Temp = 1.653

En el caso de los cuadros ANOVA se debe cumplir que:

- gl(Regresión)+gl(Residuos)=gl(Total)
- Suma cuadrados(Regresión)+Suma cuadrados(Residuos)=Suma cuadrados(Total)
- Cuadrados medios = Suma cuadrados/gl
- F = Cuadrados medios(Regresión)/Cuadrados medios(Residuos)

De donde:

- gl(Residuos) = 11
- Suma cuadrados(Total) = 31169
- Cuadrados medios(Residuos)= 834.73

c) Al analizar el primer modelo podemos observar que si bien el coeficiente de determinación R^2 no es bajo, y el modelo es globalmente significativo, ya que F=8,78 con P-valor=0,00295$<5\,\%$, la mayoría de las variables son no significativas (P-valor mayor al $5\,\%$). En el caso del segundo modelo, el R^2 es un poco más bajo, y el modelo es significativo (F=13,7 con P-valor=0,000787$<5\,\%$), pero salvo Radsolar, el resto de las variables son significativas. Si examinamos los residuos del primer y segundo modelo: 9182 y 9467 respectivamente (o los coeficientes de determinación), veríamos que no existe una diferencia apreciable, lo que hace sospechar que la variable Temperatura no es importante en el modelo.

NOTA: Este último hecho se puede probar formalmente haciendo un test de Fisher.

$$F = \frac{\frac{SSR_r - SSR_c}{k_c - k_r}}{\frac{SSR_c}{n-k_c}} = \frac{\frac{9467-9182}{4-3}}{\frac{9182}{15-4}} = 0{,}341$$

Donde SSR_c, k_c y SSR_r, k_r representan el modelo completo y el modelo reducido con la cantidad de sus coeficientes respectivamente. Como $0{,}341 < F_1, 11(5\,\%) = 4{,}844$, no se rechaza que el coeficiente de la variable Temp sea 0, y por ende ambos modelos hacen un ajuste similar. Además se puede observar que al remover la variable Temp, las desviaciones estándar de los coeficientes disminuyen, lo que aumenta su precisión. Finalmente notando que existe una alta correlación entre esta variable y el resto, es posible que Temp no esté haciendo un aporte significativo al modelo en términos de información. Es por esto que al parecer, desde el punto de vista de los datos es mejor quedarse con el segundo modelo.