Ejercicio 3 - Práctico 2

Valentina Caldiroli

Modelos lineales

Para comenzar con el analisis de los datos, lo primero será cargarlos:

Parte a:

Por letra, podemos plantear el modelo de regresión lineal como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

De donde se sabe:

y = Cantidad de vapor

x_1 = Temperatura del tanque

x_2 = Temperatura de la gasolina

x_3 = Presión del vapor del tanque

x_4 = Presión del vapor de la gasolina

Entonces para poder hallar los estimadores β_i por el método de mínimos cuadrados usamos la función $lm()$, que nos modeliza los datos, y luego con la función $coef()$ del modelo obtenemos los coeficientes β_i que buscábamos, que es igual a la matriz b, entonces podemos llamar b al resultado que nos da $coef()$.

```
modelo <- lm(y~x1+x2+x3+x4, datos)
b <- coef(modelo)
b
```

```
## (Intercept)          x1          x2          x3          x4
##  1.01501756 -0.02860886  0.21581693 -4.32005167  8.97488928
```

Entonces nos queda un modelo con la siguiente forma:

$$y = 1.015 - 0.028x_1 + 0.215x_2 - 4.32x_3 + 8.974x_4$$

Debemos calcular también σ^2, que tiene la forma:

$$\sigma^2 = \frac{SCR}{n-k-1}$$

SCR está definido como la suma de los cuadrados residuales, que podemos calcularlo como:

```r
residuos <- residuals(modelo)
SCR <- sum(residuos^2)
SCR
```

```
## [1] 201.2276
```

Sabiendo que $n = 32$ *número de observaciones*, $k = 4$, *número de variables explicativas*, podemos entonces plantear

$$\sigma^2 = \frac{201.2361}{32 - 4 - 1}$$

```r
sigma2 <- SCR/df.residual(modelo)
sigma2
```

```
## [1] 7.452874
```

Entonces:

$$\sigma^2 = 7.453$$

Parte b:

Como sabemos b = a la matriz de los β_i. Que ya la tenemos calculada, podemos entonces elevar al cuadrado los errores estándar del resumen del modelo que lo obtenemos con la función *summary*().

```r
summary(modelo)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = datos)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.586 -1.221 -0.118  1.320  5.106
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.01502    1.86131   0.545  0.59001
## x1          -0.02861    0.09060  -0.316  0.75461
## x2           0.21582    0.06772   3.187  0.00362 **
## x3          -4.32005    2.85097  -1.515  0.14132
## x4           8.97489    2.77263   3.237  0.00319 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.73 on 27 degrees of freedom
## Multiple R-squared:  0.9261, Adjusted R-squared:  0.9151
## F-statistic: 84.54 on 4 and 27 DF,  p-value: 7.249e-15
```

```r
2.73^2
```

```
## [1] 7.4529
```

Obtenemos que $Var(b) = 2.73^2 = 7.4529$.

Parte c:

Para obetner el R^2 y el R^2_{aj} podemos aplicar la función $summary()$ al modelo que estimamos con $lm()$, resumen que ya tenemos del punto anterior y nos brinda los siguientes resultados.

Entonces obtuvimos como resultado que $R^2 = 0.9261$ y $R_{aj}2 = 0.9151$.

Parte d:

Para realizar el test de significación del modelo, lo que debemos hacer es el siguiente planteo para el contraste de hipótesis:

$$H_0)\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 H_1)\beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq 0$$

Para obtener la significación se utilizarán datos brindados en el $summary(modelo)$, observandp e $p - valor$ y los estadísticos F.

Este summary nos devuelve un $p - valor = 7.249e - 15$, con un estadístico F del modelo igual a 84.54. Analizando lo pequeño que es el p-valor, podemos concluir que tanto a un 5%, como a un 10%, el modelo es significativo, es decir que en su conjunto las variables logran explicar a y.

Parte e:

Para realizar ese test, también usarmos la función $lm()$

```r
modelo1 <- lm(y ~x1+x3, datos)
summary(modelo1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x3, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5722  -2.7084   0.7061   3.2237   8.3333
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.33771    2.75010   2.305  0.02855 *
## x1          -0.02771    0.14888  -0.186  0.85364
## x3           5.96808    2.00018   2.984  0.00573 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4.775 on 29 degrees of freedom
## Multiple R-squared:  0.757,  Adjusted R-squared:  0.7402
## F-statistic: 45.17 on 2 and 29 DF,  p-value: 1.235e-09
```

En este caso el modelo sigue siendo significativo a un 5%, dado que el $p-valor = 1.235e-09$, menor a $\alpha = 0.05$

Parte f:

Para analizar una hipótesis tal como $H_0) \beta_j = 0 \ \forall \ j = 1,2,3,4$, podemos utilizar el mismo $summary(modelo)$ que en la parte d. Se analiza cada regló comparando los p-valores. Observamos entonces que solo $x2$ y $x4$ son significativas al 5%.

Parte g:

La construccipon de los intervalos de confianza al $(1-\alpha)\%$, puede hacerse con la función $confint(modelo)$, de donde podemos añadir el nivel de confianza al que se quiere trabajar con el argumento $level$, por defecto trabaja con $\alpha = 0.05$.

```
confint(modelo)
```

```
##                   2.5 %     97.5 %
## (Intercept)  -2.80407162  4.8341067
## x1           -0.21450786  0.1572901
## x2            0.07687106  0.3547628
## x3          -10.16975341  1.5296501
## x4            3.28591939 14.6638592
```

Nos devuelve entonces el intervalo de confianza al 95% para cada β_i, esto indica que:

β_0 se encuentra entre $[-2.8040; \ 4.8341]$ con un 95% de confianza.

β_1 se encuentra entre $[-0.2145; \ 0.1572]$ con un 95% de confianza.

β_2 se encuentra entre $[0.0768; \ 0.3547]$ con un 95% de confianza.

β_3 se encuentra entre $[-10.1697; \ 1.5296]$ con un 95% de confianza.

β_4 se encuentra entre $[3.2859; \ 14.6638]$ con un 95% de confianza.

Parte h:

Para trabajar en esta parte, utilizaremos una función de la librería *car*, llamada $linearHypothesis()$, que nos permitirá testear otras hipótesis nulas, y se observan a continuación:

- $H_0) \ \ \beta_1 = \beta_2 = \frac{\beta_3}{12} = \frac{\beta_4}{12}$

```
A <- matrix(c(0,1,-1,1/12,-1/12), ncol=5)
linearHypothesis(modelo, A, test = 'F')
```

```
## Linear hypothesis test
##
## Hypothesis:
## x1 - x2  + 0.0833333333333333 x3 - 0.0833333333333333 x4 = 0
##
## Model 1: restricted model
## Model 2: y ~ x1 + x2 + x3 + x4
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     28 293.22
## 2     27 201.23  1    91.996 12.344 0.001578 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obtuvimos entonces un $p-valor = 0.0015$, lo cual nos indica que bajo esta hipotesis el modelo es significativo al 5%.

- $H_0) \quad \beta_1 = \beta_2$

```
A <- matrix(c(0,1,-1,0,0), ncol=5)
linearHypothesis(modelo, A, test = 'F')
```

```
## Linear hypothesis test
##
## Hypothesis:
## x1 - x2 = 0
##
## Model 1: restricted model
## Model 2: y ~ x1 + x2 + x3 + x4
##
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1     28 227.98
## 2     27 201.23  1    26.749 3.589 0.06892 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En este caso, obtuvimos un $p - valor = 0.0689$, hablamos entonces de un contraste no significativo al 5%, puede serlo al 10%.

- $H_0) \quad \beta_2 = \frac{\beta_3}{12}$

```
A <- matrix(c(0,0,1,12*(-1/12),0), ncol =5)
linearHypothesis(modelo, A, test = 'F')
```

```
## Linear hypothesis test
##
## Hypothesis:
## x2 - x3 = 0
##
## Model 1: restricted model
## Model 2: y ~ x1 + x2 + x3 + x4
```

```
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 220.58
## 2     27 201.23  1    19.349 2.5962 0.1188
```

Acá el $p - valor = 0.0118$, por lo que este contraste de hipótesis no es significativo al 5%, tampoco lo es al 10%.

- $H_0) \quad \beta_3 = \beta_4$

```
A <- matrix(c(0,0,0,1,-1), ncol=5)
linearHypothesis(modelo, A, test = 'F')
```

```
## Linear hypothesis test
##
## Hypothesis:
## x3 - x4 = 0
##
## Model 1: restricted model
## Model 2: y ~ x1 + x2 + x3 + x4
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     28 244.81
## 2     27 201.23  1    43.585 5.8481 0.02261 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En este caso, el $p - valor = 0.02261$ también el modelo es significaivo al 5% para este contraste.

- $H_0) \quad \beta_1 = \beta_2$ y $\beta_3 = \beta_4$

```
A <- matrix(c(0,1,-1,0,0,0,0,0,1,-1), nrow = 2)
linearHypothesis(modelo, A, test = 'F')
```

```
## Linear hypothesis test
##
## Hypothesis:
## - x1  + x4 = 0
## (Intercept) - x4 = 0
##
## Model 1: restricted model
## Model 2: y ~ x1 + x2 + x3 + x4
##
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1     29 281.41
## 2     27 201.23  2    80.179 5.379 0.01081 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Habalmos de un $p - valor = 0.01081$, por lo que el modelo es significativo al 5%, donde en realidad lo que testeamos es que todos los $beta_i$ son iguales entre sí.