Practical Stats CW3 - 10171119

Robin Curnow - 10171119

November 2021

A Figure 1 suggests that Ozone concentration are positively correlated with radiation and temperature. This means that on a day when radiation or temperature are high, Ozone concentration is likely to be high also. The strongest trend is between ozone and temperature. There is also a weak, negative correlation between wind speed and Ozone. The coursework question asked for the *pairs* function. The plot also shows that radiation and temperature are positively correlated to a similar degree to ozone and radiation. I prefer the ggpairs function used to make Figure 1, however I have also put the plot made using the pairs function as an appendix at the end of this document.

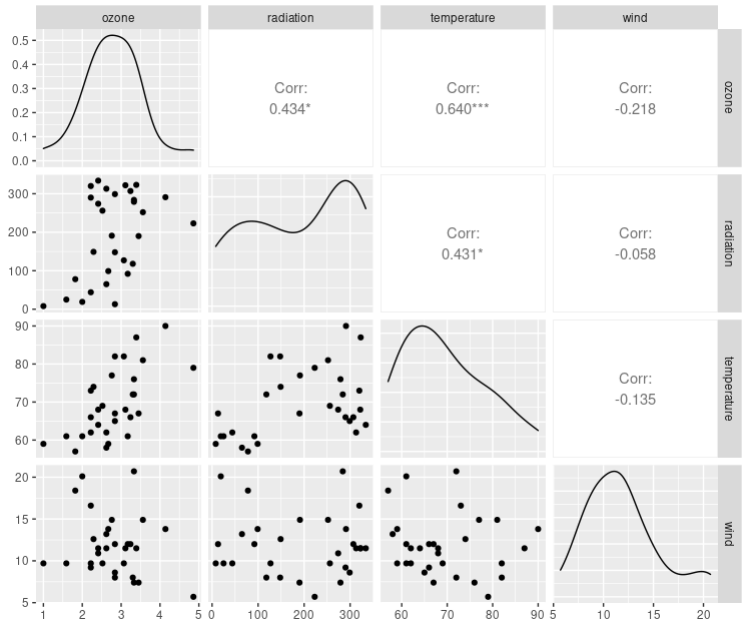

Figure 1: Scatter plots of Ozone concentration against Solar Radiation, Temperature and Wind speed

R code used for part A:

```
#Checks that all required packaged have been installed
packages <- c("GGally")

installed_packages <- packages %in%
rownames(installed.packages())
if (any(installed_packages == FALSE)) {
    install.packages(packages[!installed_packages])
}


envData <- read.table(file = "Environmental_data.txt",
                          header = TRUE)
Y <- envData$ozone
X1 <- envData$radiation
X2 <- envData$temperature
X3 <- envData$wind

n <- length(Y)

#(A)
pairs(envData)

library(GGally)
ggpairs(envData)
```

B (i) The model to be fitted is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

Where Y is the surface concentration of ozone in New York city, X_1 is the solar radiation, X_2 is the observed temperature, in degrees Fahrenheit and X_3 is the wind speed in miles per hour. $\beta_0 - \beta_3$ are unknown parameters to be estimated. The number of parameters, p, is equal to 4. The assumptions made about the model are:

- Ozone concentration is a random variable
- samples of ozone concentration are independent and identically distributed (iid)
- additivity & homogeneity of the variance of errors in the model
- errors, ϵ, are iid with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. For some unknown σ.

(ii)

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 2.86 & 6.17 \times 10^{-4} & -0.0354 & -0.0409 \\ 6.17 \times 10^{-4} & 3.34 \times 10^{-6} & -1.81 \times 10^{-5} & 1.41 \times 10^{-8} \\ -0.0354 & -1.81 \times 10^{-5} & 5.34 \times 10^{-4} & 1.44 \times 10^{-4} \\ -0.0409 & 1.41 \times 10^{-8} & 1.44 \times 10^{-4} & 2.61 \times 10^{-3} \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 83.8 \\ 17100 \\ 5950 \\ 974 \end{pmatrix}$$

$$\mathbf{y}^T\mathbf{y} = 251^*$$

Where $\mathbf{y}$ is a vector containing the observed values of Ozone concentration.

R Code used for part (ii)

```
 #(B), (ii)
model <- lm(Y ~ X1 + X2 + X3)
X <- cbind(rep(1,n),X1, X2, X3)
XTXInv <- solve(t(X) %*% X)
XTY <- t(X) %*% Y
YTY <- t(Y) %*% Y
```

(iii) $\hat{\boldsymbol{\beta}}$ below contains the vector of LS estimates of the parameters $\beta_0 - \beta_3$. Positive signs relate to upwards slopes in figure 1 and negative signs to downwards slopes

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{pmatrix} -0.295 \\ 0.00131 \\ 0.0456 \\ -0.0278 \end{pmatrix}$$

R Code used for part (iii)

```
#(iii)
betaHat <- XTXInv %*% XTY
q = length(betaHat)
```

(iv)

$$SSE = \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$SSR = (\mathbf{X}\hat{\boldsymbol{\beta}})^{-1}\mathbf{y}$$

$$SST_C = \mathbf{y}^T\mathbf{y} - n\bar{y}$$

where $\bar{y}$ is the mean of the observed values and n is the number of observed values.

*I have put all answers to 3 significant figures in my report.

Source	Sum of Squares	Degrees of Freedom	Mean Squares
SSE	9.07	3.00	3.02
SSR	242	26.0	9.31
SST_C	16.7	29.0	

Table 1: ANOVA summary for air pollution data

R Code used for part (iv)

```
#(iv)
SSR <- t(betaHat) %*% t(X) %*% Y
SST <- t(Y) %*% Y
SSE <- SST - SSR
SSTC <- SST-(sum(Y)^2)/n

SS <- c(SSE, SSR, SSTC)
DF <- c(q-1, n-q, n-1)
varSum <- data.frame(Source = c("SSE", "SSR", "SSTC"),
                             SS = SS, DF = DF,
                             meanSq= c(SS[1]/DF[1], SS[2]/DF[2], NA))
library(xtable)
anova_table <- xtable(varSum,caption = "Anova sumarry for
air pollution data", align=c(rep("|c",length(names(varSum))),"|c|"))
```

(v) The vector **p** below shows the p-values for each regression coefficient. It suggests that only temperature is likely to be significant in predicting the level of Ozone concentration. The p-values of 0.237 and 0.365, associated with radiation and wind speed respectively, suggest that there is less than 75% chance that the true mean of $\hat{\beta}_1$ & $\hat{\beta}_3$ are different from zero.

$$\mathbf{p} = \begin{pmatrix} 0.237 \\ 0.00253 \\ 0.365 \end{pmatrix}$$

The p value, p_i associated with the ${\beta_i}^{th}$ parameter is calculated in the following way:

$$\hat{\sigma}^2 = \frac{SSE}{(n-p)}$$

$$t_i = \frac{\hat{\beta}_i}{\sigma\sqrt{(X^TX)^{-1}_{i+1,i+1}/(n-p)}}$$

$$p_i = 2(1 - F_{n-p}(|t_i|))$$

Where t_i is the t statistic for $\hat{\beta}_i$ with a null hypothesis that $\hat{\beta}_i = 0$, and F(x) is the cdf of the student t distribution on n-p degrees of

freedom.

The p-value for the entire regression model is 1.03×10^{-3} which suggests at least one of the parameters is likely to be different from zero, and therefore have an effect in predicting the level of ozone concentration. This is calculated in the following way

$$F = \frac{(SST_C - SSE)/(p-1)}{SSE/(n-p)}$$

$$p = 1 - F(F)$$

Where F(F) is the cdf of the F distribution on p-1 and n-p degrees of freedom.

R Code used for part (v)

```
#(v)
sigmaSq <- SSE/(n-q)
SE <- rep(0,q-1)
for (i in seq(2:q)){
  SE[i] <- sqrt(sigmaSq*XTXInv[i+1,i+1])
  }

tStat <- 0
for(i in seq(1:length(SE))){
  tStat[i] <- betaHat[i+1]/SE[i]
}

pVals <- (1 - pt(abs(tStat), n-q))*2

FStat <- (varSum[3,2]-varSum[1,2])*varSum[2,3]/(varSum[1,2]*varSum[1,3])
pVal <- 1- pf(FStat, q-1, n-q)
```

(vi) The coefficient of determination for the model is 0.458 which suggests that less than half of the variation in the data is explained by the regression model. The formula for R^2 is given by:

$$R^2 = \frac{SST_C - SSE}{SST_C}$$

R Code used for part (vi)

```
#(vi)
Rsq <- (SSTC-SSE)/SSTC
```

(vii) The predictions for ozone concentration, $\hat{y_0}$ $\hat{y_1}$, for the two days are shown below. Since each one incurs an error, assuming the errors are additive, the calculation for the prediction interval is:

$$\mathbf{x_0} = (100, 70, 10)$$

$$\mathbf{x_1} = (50, 80, 10)$$

$$y_0 = \hat{\beta}\mathbf{x_0} + \epsilon_0$$

$$y_1 = \hat{\beta}\mathbf{x_1} + \epsilon_1$$

Therefore, There is a 95% probability that

$$\epsilon_0 \leq t_{30-4,0.025}\hat{\sigma}\sqrt{1 + \mathbf{x_0}(\mathbf{X^T X})^{-1}\mathbf{x_0}}$$

$$\epsilon_1 \leq t_{30-4,0.025}\hat{\sigma}\sqrt{1 + \mathbf{x_1}(\mathbf{X^T X})^{-1}\mathbf{x_1}}$$

so the 95% CI for x_0 is $(y_0 - \epsilon_0, y_0 + \epsilon_0) = (1.49, 4.01)$ and for similarly for x_1, $(y_1 - \epsilon_1, y_1 + \epsilon_1) = (1.80, 4.09)$. So the minimum the difference could be with a 95% certainty is zero and the maximum is 2.60.

R Code used for part (vii)

```
#(vii)
#calculates the predicted values of ozone concentration fo the two days
x0 <- c(1, 100, 70, 10)
x1 <- c(1, 50, 80, 10)
y0Hat <- t(x0) %*% betaHat
y1Hat <- t(x1) %*% betaHat
yDiffHat <- abs(y0Hat-y1Hat)
conf <- 0.975

epsilon_0 <- qt(conf, n-q)*sqrt((sigmaSq)*(1+t(x0) %*% XTXInv %*% x0))
epsilon_1 <- qt(conf, n-q)*sqrt((sigmaSq)*(1+t(x1) %*% XTXInv %*% x1))

CI_0 <- c(y0Hat - epsilon_0,y0Hat + epsilon_0)
CI_1 <- c(y1Hat - epsilon_1, y0Hat + epsilon_1)
```

(viii) Figure 2 shows that the residuals have a negative correlation with ozone concentration. This may suggest that the assumption that the variance of the errors in our model equal for different samples is incorrect. However, there is no obvious pattern between the residuals and any of the dependant variables suggesting that all of the necessary parameters are present in the model.

R Code used for part (viii)

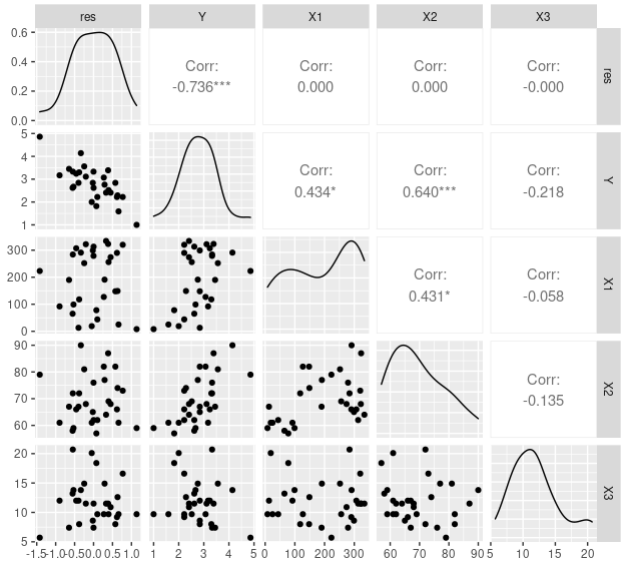

Figure 2: Scatter plots of the residuals against the variables in the model

```
#(viii)
residuals(model)
ggpairs(data.frame(res, Y, X1, X2, X3))
```

(C) (i) The null hypothesis, H_0 is that parameter $\beta_2 = 0$. The alternative hypothesis, H_1 is that $\beta_0 \neq 0$

(ii) A 95% CI for the temperature parameter is obtained using

$$P(\hat{\beta}_i - t_{n-p,0.025}\hat{\sigma}\sqrt{g_{ii}} \leq \beta_i \leq \hat{\beta}_i + t_{n-p,0.025}\hat{\sigma}\sqrt{g_{ii}}) = 0.95$$

Where $t_{n-p-1,0.025}$ is the 0.025^{th} quantile of the student t distribution on n-p degrees of freedom, and g_{ii} is the i, i^{th} element of $G^{-1} = (X^T X)^{-1}$. Using R the CI is $(0.0175, 0.0737)$.

R Code used for part (ii)

```
#(C)
#(ii)
c(confint(model, level = 0.95)[3],confint(model, level = 0.95)[7])
```

(iii) Since $0 \notin (0.0175, 0.0737)$, the inclusion of temperature almost definitely does improve the fitting and should not be removed from the model.

(D) (i)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

where $\mathbf{X} = \begin{pmatrix} \mathbf{1}, & \mathbf{x_1}, & \mathbf{x_2}, & \mathbf{x_3}, & \mathbf{x_1 \cdot x_2}, & \mathbf{x_1 \cdot x_3}, & \mathbf{x_2 \cdot x_3} \end{pmatrix}$

and $\mathbf{x_i}$ corresponds to the vector of observed values for the i^{th} variable.

```
#D
#(i)
model_int <- lm(Y~X1+ X2+ X3 + X1*X2 + X1*X3 + X2*X3)
model_int[["coefficients"]]
```

(ii) R^2 for this model, calculated using the same formula shown in (B)(iv),(vi), is 0.482. This is only slightly higher than in the model in part (B) which suggests that roughly the same amount of variation in this model is explained by the new parameters.

R Code used for part (ii)

```
#(ii)
summary(model_int)
summary(model_int)$r.squared
```

(iii) Table 2 sows the 95% CI for the parameters in the new model, calculated in the same way as in part (C)(ii) except for the number of parameters, p, equal to 7 and $\mathbf{X}$ is the the matrix described in part (D)(i)

	2.5 %	97.5 %
(Intercept)	-12.58	5.33
X1	-0.02	0.02
X2	-0.05	0.23
X3	-0.44	0.99
X1:X2	-0.00	0.00
X1:X3	-0.00	0.00
X2:X3	-0.02	0.01

Table 2: Confidence intervals for parameters $\beta_0 - \beta_6$ in the updated regression model

R Code used for part (iii)

```
#(iii)
confint(model_int, level = 0.95)
CI_table <- xtable(confint(model_int, level = 0.95), caption =
                   "Confidence intervals for parameters beta_0-beta_6 in
                   the updated regression model", align = c("|c", "|c", "|c|"))
```

iv) None of the parameters in the new model are non-zero with 95% probability or more which suggests that it is possible that none of the prediction variables are useful for predicting the level of Ozone concentration. However, the interaction parameters between temperature and radiation, and between radiation and wind speed are zero to three significant figures and so it may be useful to exclude these terms in the regression model, as they are in the model in (B). The model from part (B) almost certainly has a positive parameter associated with temperature and so may be more useful in predicting the level of ozone concentration.

(v) To compare models from (B) and (D), an F statistic will be using the null hypothesis that the parameters associated with the interaction terms are 0:

$$H_0 : \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \boldsymbol{\beta} = \mathbf{0}$$

$$\text{i.e. } H_0 : \boldsymbol{\beta}_2 = \begin{pmatrix} \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix} = \mathbf{0} \text{ vs } H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$$

$$F_{stat} = \frac{SSE_B - SSE_D/3}{SSE_D/(n-p)}$$

$$p - value = 1 - F_{3,n-p}(F_{stat})$$

The p-value for this F test is 0.752 and so there is strong evidence in favour of the null hypothesis, suggesting that the interaction terms do not likely have a significant effect on predicting the level of ozone concentration..

R Code used for part (v)

```
#(v)
SSE_int <- sum(residuals(model_int)^2)
F <- (SSE- SSE_int)*(n-q)/(SSE_int*3)
pVal_comp <- 1-pf(F, 3,n-q)
```

Appendix

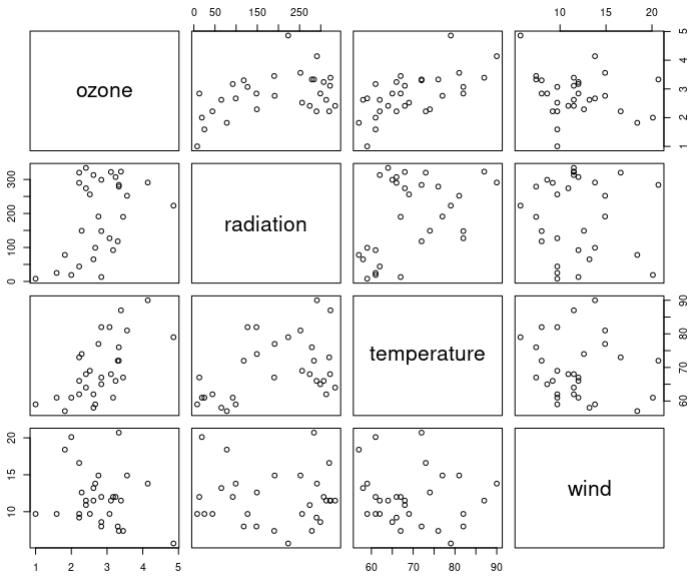

Figure 3: Scatter plots of Ozone concentration against Solar Radiation, Temperature and Wind speed using the Pairs function in R