Solutions to Computer exercise 1: simple linear regression

(a) It is stated that the level of pollution varies with the flow of traffic. Since it is the traffic that, together with other sources, causes the pollution, not the pollution that causes the traffic, the response variable, $Y$, should be change in level of air pollution (pollution), while the explanatory variable, $X$ is the change in flow of vehicles (vehicles).

(b) As seen in Figure 1(a), the data lies close to a straight line, so a linear relationship seems sensible.

(c) This is also confirmed in Figure 1(b) where the estimated linear relationship $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ has been added.

(d) The residuals, $e_i = y_i - \hat{y}_i$, seems fairly close to a normal distribution, as can be seen in the histogram in Figure 2(a) and, especially, in the Q-Q-plot in Figure 2(b) where they fall close to a straight line. According to Figure 2(c)–(d), there is no obvious, unexplained non-linear trend left in either the change in the flow of vehicles, or in the predicted change in air pollution. Also, the variance seems to be constant. Since data might have been measured on consecutive days (it is not clear from the description), the residuals might be correlated. This would be the case if, e.g., yesterday’s increase in air pollution lingered on into the following day. However, Figure 2(e)–(f) show that the residuals seem uncorrelated. The conclusion is that residuals follow the assumptions.

(e) The estimates, based on the $n = 12$ observations, of the regression coefficients and the residual variance become

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = -0.836,$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{X})^2} = 0.834,$$

$$\hat{\sigma}^2 = s^2 = \frac{1}{n - 2} \sum_{i=1}^{n} e_i^2 = 4.17.$$
(f) Since the p-value for $H_0: \beta_0 = 0$ against $H_1: \beta_0 \neq 0$ is $0.27 > 0.05$ we can not reject $H_0$ at significance level $\alpha = 5\%$. The change in level of air pollution when there is no change in the flow of vehicles is not significantly different from zero.

(g) The standard error of $\hat{\beta}_1$ is given by $SE(\hat{\beta}_1) = s/\sum_{i=1}^{n} (X_i - \bar{X})^2 = 0.0247$. With $n - 2 = 10$ degrees of freedom a 95\% confidence interval for $\beta_1$ is given by

$$I_{\beta_1} = (\hat{\beta}_1 \pm t_{0.025,10} \cdot SE(\hat{\beta}_1)) = (0.834 \pm 2.23 \cdot 0.0247) = (0.78, 0.89)$$

which is the same as that produced by `confint`.

(h) Since $\beta_1 = 1$ is not included in the 95\% confidence interval for $\beta_1$, we can reject $H_0: \beta_1 = 1$ in favour of $H_1: \beta_1 \neq 1$ at significance level $\alpha = 5\%$. No, a change in the flow of vehicles does not produce an equally large change in the level of air pollution. The change is smaller than that.

(i) A 30 percent increase, $X_0 = 30$, in the flow of vehicles should give, on average, a $E(Y_0) = \beta_0 + \beta_1 X_0$ percent increase in the level of air pollution. This is estimated by $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0 = -0.836 + 0.834 \cdot 30 = 24.2$. With $SE(\hat{Y}_0) = s\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$ a 95\% confidence interval for $E(Y_0)$ is given by $I_{E(Y_0)} = (\hat{Y}_0 \pm t_{0.025,10} \cdot SE(\hat{Y}_0)) = (22.7, 25.7)$. This means that the average change in pollution, on days with a 30 percent change in the flow of vehicles, is, probably, somewhere between 22.7 and 25.7 percent.

(j) On a day with a 30 percent increase, $X_0 = 30$, in the flow of vehicles, the observed change in the level of the air pollution will be $Y_{pred} = \hat{\beta}_0 + \hat{\beta}_1 X_0 + \epsilon_0$ with a 90\% prediction interval given by

$$I_{Y_{pred}} = (\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{0.05,10} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}) = (20.3, 28.1).$$

This means that, in 90\% of the days with a 30 percent change in the flow of vehicles, the observed change in the level of air pollution will be between 20.3 and 28.1 percent. Also, on 5\% of the days it will be below 20.3, and on 5\% of the days it will be above 28.1 percent.
Figure 2: (a) Histogram of model residuals. (b) Normal Q-Q-plot of model residuals. (c) Residuals plotted against the explanatory variable. (d) Residuals plotted against the predicted values. (e) Residuals plotted against the observation number. (f) Autocorrelation function for the residuals.
### Appendix: R-code for solving Computer exercise 1, 2018 ###

# Create the data frame
emission <- data.frame(vehicles = c(28,36,15,-19,-24,8,25,40,63,12,-6,21),
                       pollution = c(22,26,15,-18,-21,7,21,31,52,8,-7,20))

# Check that it looks ok
emission
summary(emission)
View(emission)

####################
### Question (a) ###
####################

# If we save the labels for the plot axes in variables we won't have
# to type them each time we make a plot. Also makes it easier to be
# consistent.
xtext <- "change in flow of vehicles (%)"
ytext <- "change in level of air pollution (%)"
etext <- "residuals"

# Figure 1(a):
# Note the "(a)" in the main title since this will become
# Figure 1(a). Save the plot using the "Export" tab just above it.
with(emission, plot(pollution ~ vehicles, xlab=xtext, ylab=ytext,
                     main="(a) Pollution vs traffic"))

####################
### Question (b) ###
####################

# Figure 1(b):
# Add the straight line to the plot:
with(emission, plot(pollution ~ vehicles, xlab=xtext, ylab=ytext,
                     main="(b) Pollution vs traffic with fitted line"))
abline(model)

# Question (c):
# Fit the model:
model <- lm(pollution ~ vehicles, data=emission)

# Figure 1(b):
# Add the straight line to the plot:
with(emission, plot(pollution ~ vehicles, xlab=xtext, ylab=ytext,
                     main="(b) Pollution vs traffic with fitted line"))
abline(model)
### Question (d) ###

# Figure 2(a):

```
hist(model$residuals, xlab=etext, main="(a) Histogram of model residuals")
```

# Figure 2(b):

```r
# This is how you perform several commands "with" the same data frame.
with(model, {
  qqnorm(residuals, main="(b) Normal Q-Q-plot of model residuals")
  qqline(residuals)
})
```

# Figure 2(c):

```r
plot(model$residuals ~ emission$vehicles, xlab=xtext, 
     ylab=etext, main="(c) Model residuals vs traffic")
abline(h=0)
```

# Figure 2(d):

```r
with(model, plot(residuals ~ fitted.values, 
    main="(d) Model residuals vs fitted values"))
abline(h=0)
```

# Figure 2(e):

```r
plot(model$residuals, xlab="observation number", ylab=etext, 
     main="(e) Model residuals vs observation number")
abline(h=0)
```

# Figure 2(f):

```r
acf(model$residuals, main="(f) Auto-correlation function for model residuals")
```

### Question (e) ###

# Extract estimates of beta_0, beta_1, sigma^2 (and save them since we might need them later):

```r
beta0 <- model$coefficients["(Intercept)"]
beta1 <- model$coefficients["vehicles"]
s <- summary(model)$sigma
```

beta0

beta1

s^2
### Question (f) ###

## p-value for H0: beta0=0 vs H1: beta0<>0: ##
summary(model)
# or extract the p-value
summary(model)$coefficients["(Intercept)","Pr(>|t|)"]

### Question (g) ###

# Extract SE(beta1) from the summary:
se.beta1 <- summary(model)$coefficients["vehicles","Std. Error"]
se.beta1
# Extract the degrees of freedom = n-2:
df <- model$df.residual
df
# Calculate the t-quantile, alpha = 0.05:
t <- qt(1-0.05/2,df)
t
# Calculate the 95% confidence interval:
beta1+t*se.beta1*c(-1,1)
# Check that it is the same as the one R gives (vehicles line):
confint(model)["vehicles",]

### Question (h) ###

# use the interval from (g)

### Question (i) ###

# Create a new data frame for prediction:
x0 <- data.frame(vehicles=30)
# predict and get a (95%) confidence interval
predict(model,x0,interval="confidence")
# Construct a 90% prediction interval
predict(model,x0,interval="prediction",level=0.90)