Computer exercise 1: solutions

Anna Lindgren

11 mars 2019

Exercise 1: Air pollution

Read in the data

```
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)
  )
```

(a) Response variable

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) Linear relationship

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

(c) Estimated line

```
model1 <- lm(pollution ~ vehicles, data = emission)
```

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) Confidence interval for line

```
x0 <- data.frame(vehicles = seq(-25, 65, 5))
mu0 <- predict(model1, x0, interval = "confidence")
x0 <- cbind(x0, mu0)
```

When the change in vehicle flow is X_0, a confidence interval for the expected (average) change in pollution, $E(Y_0) = \beta_0 + \beta_1 X_0$, is given by

$$I_{E(Y_0)} = (\hat{Y}_0 \pm t_{0.05/2,12-2} \cdot SE(\hat{Y}_0))$$

where

$$SE(\hat{Y}_0) = s\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}},$$

see Figure 1(c). The interval seems reasonable.

(e) Prediction interval for observations

```
pred0 <- predict(model1, x0, interval = "prediction")
x0 <- cbind(x0, pred = pred0)
```

When the change in vehicle flow is X_0, a prediction interval for the observed change in pollution, $Y_{\text{pred},0} = \beta_0 + \beta_1 X_0 + \epsilon_0$, is given by

$$I_{Y_{\text{pred},0}} = (\hat{Y}_0 \pm t_{0.05/2,12-2} \cdot SE(\hat{Y}_{\text{pred},0}))$$

where

$$SE(\hat{Y}_{\text{pred},0}) = s\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}},$$

see Figure 1(d). The interval looks reasonable. It should be expected to contain 95% of the observations (11.4), so zero or, at most, one observation should fall outside.

(f) Residual analysis

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.

(g) Parameter estimates

```
summodel1 <- summary(model1)
summodel1
#> 
#> Call:
#> lm(formula = pollution ~ vehicles, data = emission)
#> 
#> Residuals:
#>     Min      1Q  Median      3Q     Max 
#> -3.2002 -1.2097 -0.3325  1.0236  3.3210 
#> 
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)    
#> (Intercept) -0.83623    0.71842  -1.164    0.271    
#> vehicles     0.83435    0.02474  33.728 1.24e-11 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
#> Residual standard error: 2.043 on 10 degrees of freedom
#> Multiple R-squared:  0.9913, Adjusted R-squared:  0.9904 
#> F-statistic:  1138 on 1 and 10 DF,  p-value: 1.241e-11
```

```
beta0 <- model1$coefficients["(Intercept)"]
beta1 <- model1$coefficients["vehicles"]
s2 <- summodel1$sigma^2
```

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

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

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

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = 2.043^2 = 4.174.$$

(h) Test of intercept

```
pvalue0 <- summodel1$coefficients["(Intercept)", "Pr(>|t|)"]
```

We want to test $H_0 : \beta_0 = 0$ against $H_1 : \beta_0 \neq 0$. Since the P-value $0.271 \nless \alpha = 0.05$ we can not reject H_0. It is possible that the change in pollution is zero when there is no change in the vehicle flow.

(i) Test of slope

```
pvalue1 <- summodel1$coefficients["vehicles", "Pr(>|t|)"]
```

We want to test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. Since the P-value $1.24 \times 10^{-11} < \alpha = 0.05$ we can reject H_0. The change in pollution is not the same regardless of the change in vehicle flow.

(j) Confidence interval for slope

```
sebeta1 <- summodel1$coefficients["vehicles", "Std. Error"]
tquantile <- qt(1 - 0.05 / 2, nrow(emission) - 2)
ci <- confint(model1)["vehicles", ]
```

A 95% confidence interval for the slope is given by

$$I_{\beta_1} = (\hat{\beta}_1 \pm t_{0.05/2,12-2} \cdot SE(\hat{\beta}_1)) = (0.834 \pm 2.23 \cdot 0.025) = (0.779, 0.889)$$

(k) Test of slope = 1

Since $H_0 : \beta_1 = 1$ is not included in the 95% confidence interval for β_1, we can reject H_0 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.

```
t1 <- abs((beta1 - 1) / sebeta1)
```

Alternatively, we can reject H_0 since

$$\frac{|\hat{\beta}_1 - 1|}{SE(\hat{\beta}_1)} = 6.696 > t_{\alpha/2,n-2} = 2.23.$$

```
t1pvalue <- 2 * pt(t1, nrow(emission) - 2, lower.tail = FALSE)
```

Or reject H_0 since the P-value $5.39 \times 10^{-5} < \alpha = 0.05$.

Plots

```r
# 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"
fig1cap = "Confidence (blue dashed) and prediction (red dotted) intervals"
fig2cap = "Residual analysis"
```

Data and intervals

```r
par(mfrow = c(2, 2))
with(emission, plot(pollution ~ vehicles,
                    xlab = xtext, ylab = ytext,
                    main = "(a) Pollution vs traffic")
     )
with(emission, plot(pollution ~ vehicles,
                    xlab = xtext, ylab = ytext,
                    main = "(b) Pollution vs traffic with fitted line"))
abline(model1)
with(emission, plot(pollution ~ vehicles,
                    xlab = xtext, ylab = ytext,
                    main = "(c) Confidence interval"))
with(x0, {
  lines(fit ~ vehicles)
  lines(lwr ~ vehicles, lty = 2, col = "blue")
  lines(upr ~ vehicles, lty = 2, col = "blue")
})
with(emission, plot(pollution ~ vehicles,
                    xlab = xtext, ylab = ytext,
                    main = "(d) Confidence and prediction intervals"))
with(x0, {
  lines(fit ~ vehicles)
  lines(lwr ~ vehicles, lty = 2, col = "blue")
  lines(upr ~ vehicles, lty = 2, col = "blue")
  lines(pred.lwr ~ vehicles, lty = 3, col = "red")
  lines(pred.upr ~ vehicles, lty = 3, col = "red")
})
```

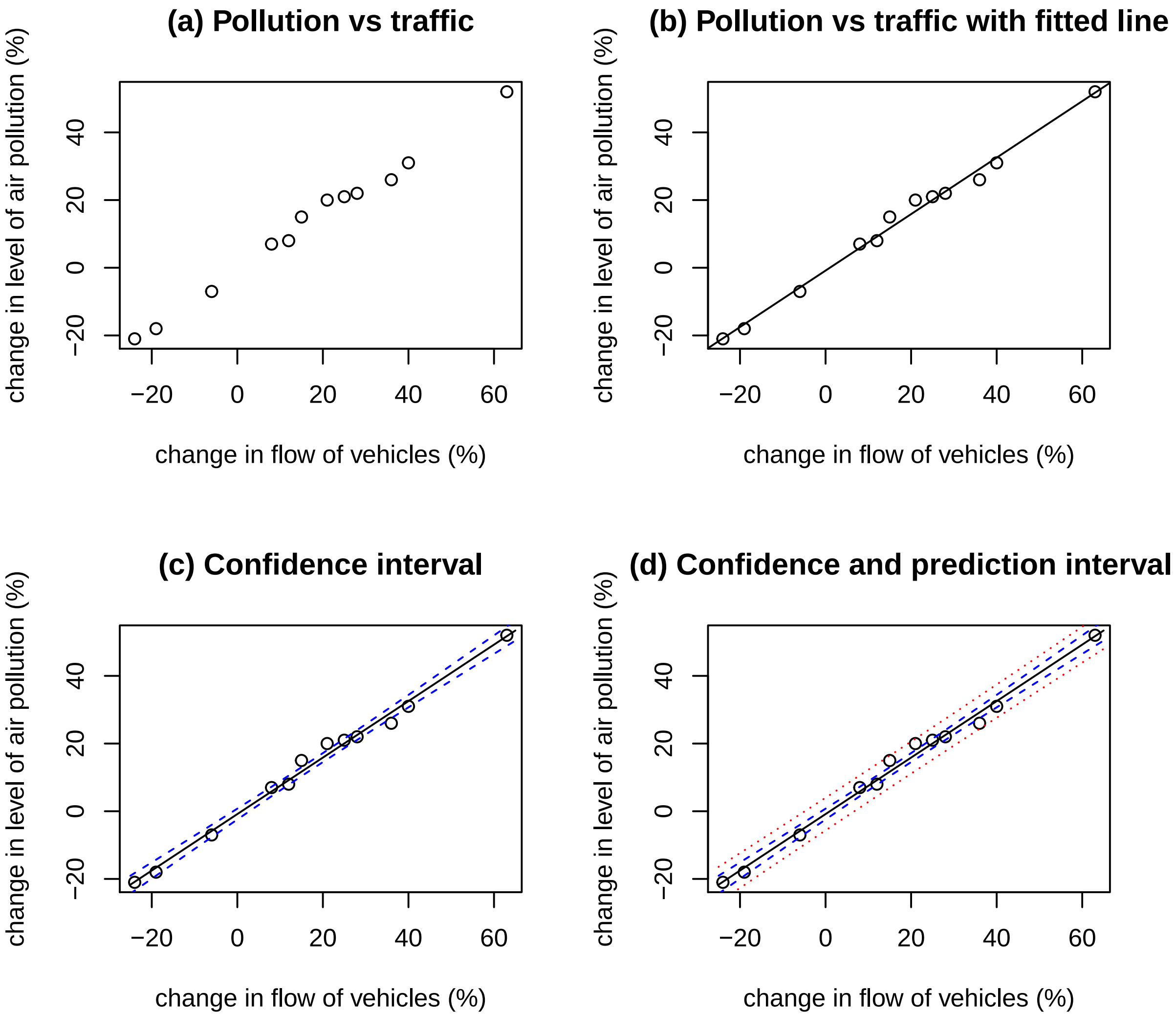

Figure 1: Confidence (blue dashed) and prediction (red dotted) intervals

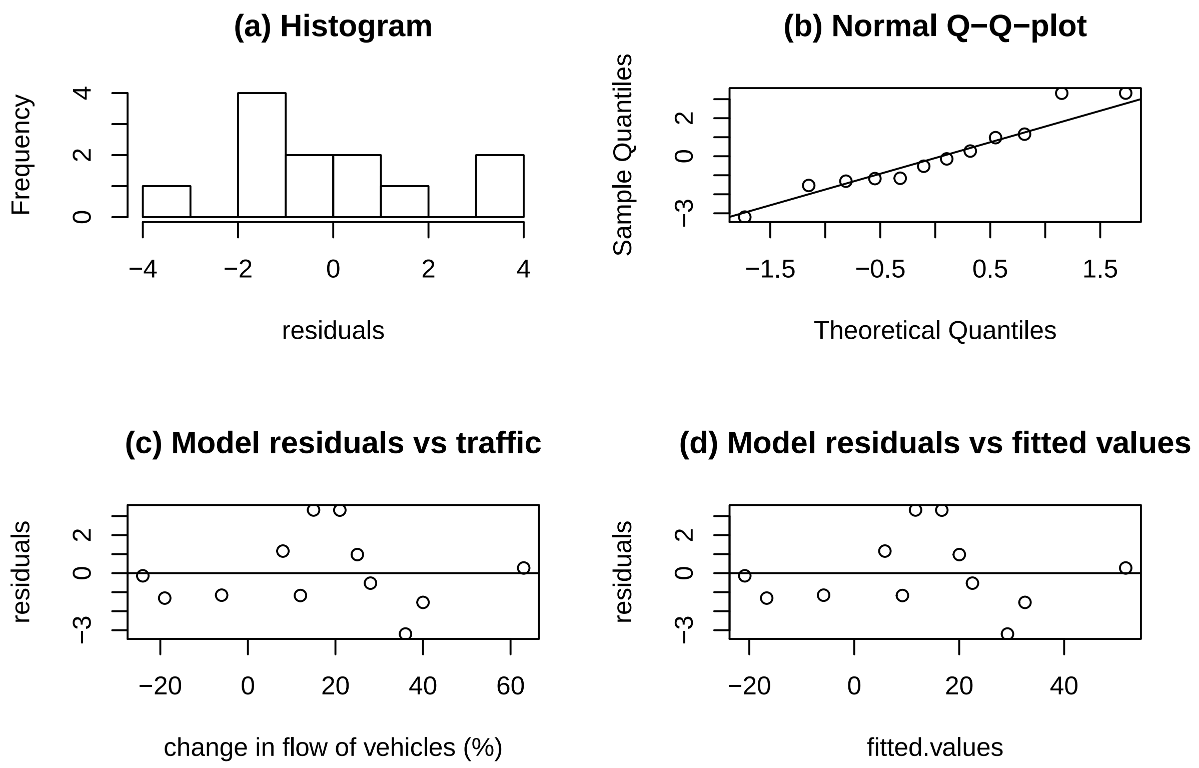

Figure 2: Residual analysis

Residual analysis

```
par(mfrow = c(2, 2))
hist(model1$residuals, xlab = etext,
     main = "(a) Histogram")
with(model1, {
  qqnorm(residuals, main = "(b) Normal Q-Q-plot")
  qqline(residuals)
  })
plot(model1$residuals ~ emission$vehicles,
     xlab = xtext, ylab = etext,
     main = "(c) Model residuals vs traffic")
abline(h = 0)
with(model1, plot(residuals ~ fitted.values,
                  main = "(d) Model residuals vs fitted values"))
abline(h = 0)
```

Note: this document is generated with Rmarkdown which allows you to mix text with R-code and its output in the same document. If you want to use it to, e.g., produce your project reports, you must install the `rmarkdown` package. In RStudio you can use `Tools -- Install Packages...`. Then create a new document with `File --  New File -- R Markdown...` If you want to use the pdf format you need a LaTeX installation on your computer. Otherwise, you can use the word format and then produce the pdf from the word file, in whatever way you usually do that. In order to compile the document you should "knit" it by pressing the Knit button. It looks like a ball of yarn with a knitting needle stuck through it.

You should also go through "Get Started" at https://rmarkdown.rstudio.com/

For an example, see `lab1_vt19_solutions.Rmd` which was used to generate this pdf.