Multiple Regression

Environmental Statistics 1

C.F. Dormann

Table of contents

1 Exercise 1

Set up a Quarto document in RStudio, with your name and a meaningful title and render it as PDF. Save as template for all exercises this term. To learn more about Quarto see https://quarto.org.

You can set global options like so:

```
library(knitr)
opts_chunk$set(fig.path='figure/', fig.align='center', fig.width=4, fig.height=4,
               fig.show='hold', cache=TRUE, tidy=F, tidy.opts=list(width.cutoff=60),
               size="small")
```

!!! toc and number_sections need to be specified differently (or pdf?)

Useful shortcuts:

- `<ctrl/cmd> + shift + k`: render document
- `<ctrl/cmd> + <alt gr/option> + i`: new R-code chunk

Prevent code from running over the margin by manual line breaks in the R-code (see above chunk for example).

2 Exercise 2

The data set `insulgas` (in package **faraway**, from now on written as `package::data`) contains measurements of gas consumption of a house for several weeks before and after insulation.

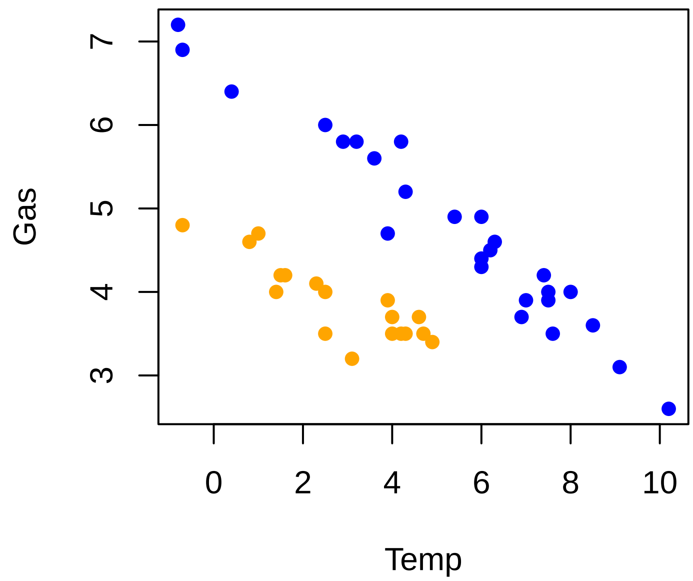
2.1 Plot

Plot the relationship between outside temperature `Temp` and gas consumption `Gas` as a **square** scatter plot. Use different colours or symbols to represent the phase before and after insulation (`Insulate`).

```
library(faraway)
data(insulgas)
summary(insulgas)
```

```
   Insulate        Temp              Gas       
 After :18   Min.   :-0.800   Min.   :2.600  
 Before:26   1st Qu.: 2.500   1st Qu.:3.675  
             Median : 4.200   Median :4.150  
             Mean   : 4.311   Mean   :4.398  
             3rd Qu.: 6.225   3rd Qu.:4.825  
             Max.   :10.200   Max.   :7.200  
```

```
plot(Gas ~ Temp, data=insulgas, col=ifelse(insulgas$Insulate=="Before", "blue", "orange"),
```



2.2 Univariate Regressions

Analyse, in two separate regressions, the effect of `Temp` and `Insulate` on gas consumption. Present the p-value of the effects, e.g. using `summary`. Summarise the result of the two regressions in a sentence, particularly the effect's direction, its strength and its significance. Do **not** refer to the variables by their variable names, but by their correct terms (e.g. "temperature", not "Temp")!

```
fTemp <- lm(Gas ~ Temp, data=insulgas)
summary(fTemp)
```

```
Call:
lm(formula = Gas ~ Temp, data = insulgas)

Residuals:
     Min       1Q   Median       3Q      Max
-1.45941 -0.76960 -0.02772  0.65826  1.69810

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.32908    0.24269  21.959  < 2e-16 ***
Temp        -0.21602    0.04773  -4.526 4.89e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8533 on 42 degrees of freedom
Multiple R-squared:  0.3278,	Adjusted R-squared:  0.3118 
F-statistic: 20.48 on 1 and 42 DF,  p-value: 4.886e-05
```

```r
fInsu <- lm(Gas ~ Insulate, data=insulgas)
summary(fInsu)
```

```
Call:
lm(formula = Gas ~ Insulate, data = insulgas)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1500 -0.5042 -0.1694  0.5153  2.4500 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      3.8889     0.2230  17.437  < 2e-16 ***
InsulateBefore   0.8611     0.2901   2.968  0.00493 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9462 on 42 degrees of freedom
Multiple R-squared:  0.1734,	Adjusted R-squared:  0.1537 
F-statistic: 8.809 on 1 and 42 DF,  p-value: 0.004932
```

Both effects are significant, with the negative temperature effect explaining twice as much variation in gas consumption as the positive effect of insulation.

2.3 Multivariate Regression

Now run a multiple regression, in this case also called ANCOVA (analysis of co-variance) with both predictors in the same model. Compare the output to the two previous outputs: state what has changed, and in which direction.

```r
fm <- lm(Gas ~ Temp + Insulate, data=insulgas)
summary(fm)
```

```
Call:
lm(formula = Gas ~ Temp + Insulate, data = insulgas)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.58314 -0.18098  0.04439  0.20251  0.62716 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     4.92249    0.08829   55.76   <2e-16 ***
Temp           -0.36769    0.01889  -19.46   <2e-16 ***
InsulateBefore  1.79462    0.10354   17.33   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2993 on 41 degrees of freedom
Multiple R-squared:  0.9193,    Adjusted R-squared:  0.9153 
F-statistic: 233.5 on 2 and 41 DF,  p-value: < 2.2e-16
```

Together both variables explain much more than the univariate regressions combined (90% rather than 30%+15%); their effects have become stronger (in absolute terms), the errors much smaller and the significances even stronger.

2.4 Multivariate Regression with Interaction

Finally, run the multiple regression but allow for a statistical interaction between the two predictors. Again: describe what changed, possibly in several sentences and with reference to the initial plot.

```r
fint <- lm(Gas ~ Temp * Insulate, data=insulgas)
summary(fint)
```

```
Call:
lm(formula = Gas ~ Temp * Insulate, data = insulgas)

Residuals:
     Min       1Q   Median       3Q      Max 
```

```
-0.62020 -0.18011  0.03405  0.16379  0.59778

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.59062    0.13016  35.269  < 2e-16 ***
Temp                -0.24963    0.04039  -6.180 2.64e-07 ***
InsulateBefore       2.26321    0.17278  13.099 4.71e-16 ***
Temp:InsulateBefore -0.14361    0.04455  -3.224  0.00252 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2699 on 40 degrees of freedom
Multiple R-squared:  0.9359,    Adjusted R-squared:  0.9311
F-statistic: 194.8 on 3 and 40 DF,  p-value: < 2.2e-16
```

This model adds another estimate for the interaction, which is significant. The model fit has improved again, but not by much. The effect of insulation has increased again, suggesting that it is much stronger at cold than at warm temperatures. This is visible in the initial plot, where the distance between gas consumption before and after insulation is largest at the lowest temperature.

2.5 Plot Fitted Model into Data

Make a "production quality" plot, i.e. including proper labels and units. Add regression lines and 95%-confidence intervals to the initial plot. (Remember that the 95%-CI is 2 times the standard error of the prediction in either direction of the expected value. The standard error is returned, by some R-functions, using the argument `se.fit=TRUE`.)

```r
plot(Gas ~ Temp, data=insulgas, col=ifelse(insulgas$Insulate=="Before", "blue", "orange"),
newTempBefore <- seq(min(insulgas$Temp[insulgas$Insulate=="Before"]), max(insulgas$Temp[in
newTempAfter <- seq(min(insulgas$Temp[insulgas$Insulate=="After"]), max(insulgas$Temp[insu
newTempBefore.df <- data.frame("Temp"=newTempBefore, "Insulate"="Before")
newTempAfter.df <- data.frame("Temp"=newTempAfter, "Insulate"="After")
preds.Before <- predict(fint, newdata=newTempBefore.df, se.fit=T)
preds.After <- predict(fint, newdata=newTempAfter.df, se.fit=T)

lines(newTempBefore, preds.Before$fit, col="blue")
lines(newTempBefore, preds.Before$fit + 2*preds.Before$se.fit, col="blue", lty=2)
lines(newTempBefore, preds.Before$fit - 2*preds.Before$se.fit, col="blue", lty=2)

lines(newTempAfter, preds.After$fit, col="orange")
```

```r
lines(newTempAfter, preds.After$fit + 2*preds.After$se.fit, col="orange", lty=2)
lines(newTempAfter, preds.After$fit - 2*preds.After$se.fit, col="orange", lty=2)
```