PO12Q - Introduction to Quantitative Political Analysis II:
Week 7 – Solutions to Additional Exercises

Dr Florian Reiche
F.Reiche@warwick.ac.uk

1 | Load Packages and data

Before starting, we need to load libraries and install packages if not already installed. In these exercises we will be using the following packages:

1. `haven`
2. `ggplot2`
3. `modelsummary`

Set working directory and load the data. It's called simd.csv and it's the "Scottish Index of Multiple Deprivation". More details here: https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/

```
setwd("")

library(tidyverse)
library(haven)
library(modelsummary)

sim <- read.csv("simd.csv", stringsAsFactors = T)
```

2 | Inspect your data

Here you can use several basic functions. The dataset does not contain too many variables, so you can start by using `names()`, `str()`, etc.

```
names(sim)
##  [1] "data_zone"               "intermediate_zone"
##  [3] "council_area"            "health_board"
##  [5] "pct_school_attend"       "alcohol"
##  [7] "population"              "working_population"
##  [9] "pct_income_deprived"     "pct_employment_deprived"
## [11] "illness"                 "mortality"
## [13] "drugs"                   "pct_depress"
## [15] "pct_low_bw"              "hosp_emerg"
## [17] "noquals"                 "crime"
## [19] "pct_no_heating"          "pct_overcrowded"
## [21] "simd_2016_quintile"      "urban"
dim(sim)
## [1] 6976   22
```

3 | Preliminary Analysis

Now that you have a preliminary idea of the structure of the dataset, you can select two variables and test a possible relationship. The topic today is bivariate linear regression analysis, so remember that the outcome variable needs to be continuous.

Let's say our research question is whether higher levels alcohol abuse are associated with higher mortality rate (yes, not a very funny topic, but interesting nonetheless). The two variables are, respectively, `ALCOHOL` and `SMR`.

Now, formulate the working (alternative) and the null hypothesis.

H_0: Alcohol consumption has no statistically significant influence on mortality.

H_1: Higher levels of alcohol consumption significantly increase the risk of mortality.

Which is your dependent variable?

Run a frequency table on the mortality variable. What is the level of measurement?

```
table(sim$mortality)
##
##   0   6   8   9  11  12  13  15  17  18  19  20  21  22  23  24  25  26  27  28
##   4   1   1   1   2   2   1   3   7   4   2   1   5   3   2   4   2   5   5   8
##  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48
##   3   9   7   7  12   8  17  10   7  25  16  18  31  26  23  23  21  28  28  36
##  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68
##  33  43  37  39  51  54  46  49  46  57  67  59  61  73  70  71  80  67  85  57
##  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88
##  87  64  71  86  76  84  97  88  70  70  79  74  84  81  83  68  68  74  98  70
##  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
##  81  74  84  77  77  76  85  72  63  74  56  63  66  72  62  71  59  82  64  60
## 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128
##  61  63  52  62  61  47  66  53  53  42  44  47  54  46  59  52  43  39  40  45
## 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148
##  48  37  39  40  28  21  40  38  39  37  31  25  34  28  26  30  31  17  23  31
## 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168
##  22  21  21  11  20  22  19  25  18  21  21  18  25  17  13  16  17  15   9   8
## 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188
##  11  11  12  11  12  13  12   7   5   3  11   9   7  10   9   8   6   7   9   8
## 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 205 206 207 208 209
##   4   6   6   7   7   3  11   4   7   9  10   1   5   2   2   5  10   6   1   1
## 210 211 212 213 214 216 217 218 219 220 221 222 223 224 227 228 229 230 231 232
##   2   3   3   1   2   3   2   1   1   5   4   3   1   3   2   1   1   3   3   1
## 233 235 237 239 240 241 242 244 246 248 251 253 255 259 261 263 266 269 272 273
##   1   1   3   2   2   5   1   2   2   1   1   1   2   2   1   1   1   1   1   1
## 282 286 294 299 303 310 312 324 340 346 356 385 398 399 411 472 473 476 504 523
##   2   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 557 563 589 669 719 950
##   1   1   1   1   1   1
class(sim$mortality)
## [1] "integer"
```

Do the same for the other variable. And guess what is the level of measurement.

(I'm sparing you the endless table here and jump straight to `class()`)

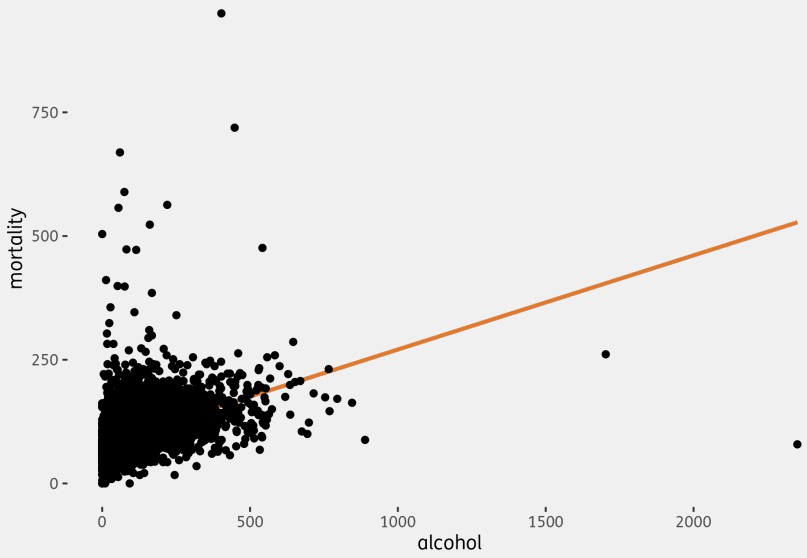
```
class(sim$alcohol)
## [1] "integer"
```

4 | Visualisation

Let's start with the visualisation of the relationship between the two variables. Use a scatterplot to visualise the relationship and add the regression line.

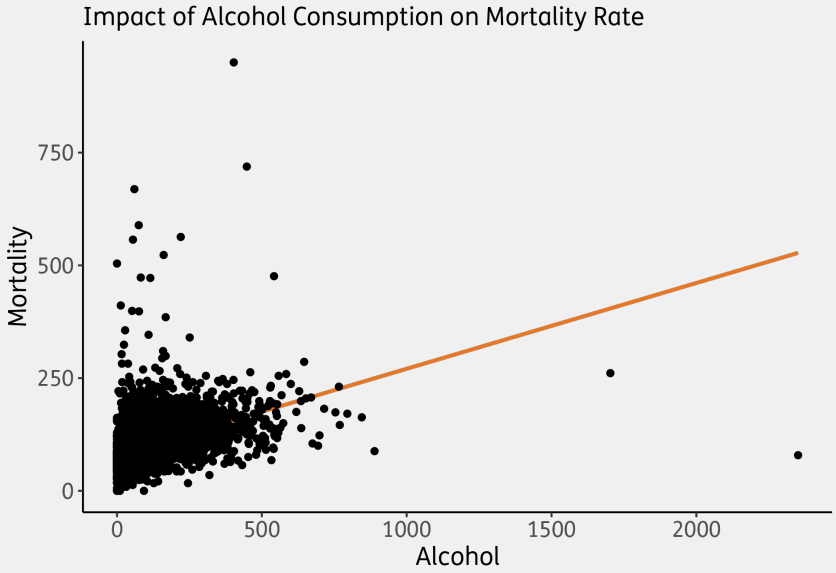
You can use `ggplot`, but also the standard `plot()` function.

```
ggplot(sim, aes(x = alcohol, y = mortality)) +
  geom_smooth(method = lm, se=FALSE) +
  geom_point()
```



Improve the graph by:

1. Adding a regression line.
2. Adding up a relevant title, also possibly a subtitle.
3. Adding axes labels and making them readable.
4. Remove the grid in the background.



```
ggplot(sim, aes(x = alcohol, y = mortality)) +
  geom_smooth(method = lm, se=FALSE, colour="orange") +
  geom_point() +
  theme_classic() +
  xlab('Alcohol') +
  ylab('Mortality') +
  ggtitle("Impact of Alcohol Consumption on Mortality Rate") +
  theme(axis.text=element_text(size=12),
      axis.title=element_text(size=14)) +
  theme(plot.title = element_text(size = 14))
```

If you have done everything correctly, you should see that the dots are rather concentrated in the bottom left corner of the scatterplot with some outliers far away from the cloud of our data. It is not a big deal, but we might want to get rid of the outliers and this way improve our visualisation.

There are several ways to do that, of course. But let's say we want to transform our variables, excluding all values over a certain point. For instance, we want to exclude the values above 750 of our alcohol variable and above 500 for our mortality variable. How would you do that?

```
# Option 1 with tidyverse
simred <- filter(sim, alcohol<750 & mortality<500)

# Option 2 with base R
simred<-subset(sim,alcohol<750 & mortality<500)
```

Hint: there are several solutions. One could be, creating a new datset subsetting the original. Another solution could be again, creating a new variable telling R to transform all the values above our treshold in NA (missing values). Try to find an apply the appropriate code. Use Google if necessary, it helps a lot.

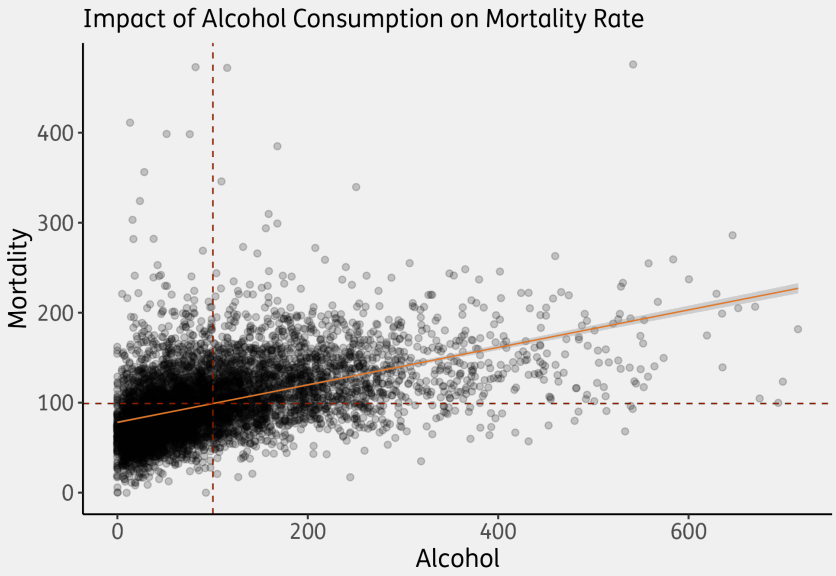
5 | Visualisation 2.0

Now, visualise the relationship using the reduced data frame. What can you see?

You can improve the scatterplot using a series of arguments (e.g., alpha) in the `geom_point()` function in ggplot. Try to improve the Aesthetics of the scatterplot playing with alpha, for example (see R Documentation).

Also, you can draw a vertical and horizontal line corresponding to the mean of your variables using `geom_hline` and `geom_vline`. You can thus check if the regression line passes through the mean of X and Y. (see R Documentation).

```
ggplot(simred, aes(alcohol, mortality)) +
  geom_point(position='jitter', alpha = 1/5) +
  xlab('Alcohol') +
  ylab('Mortality') +
  ggtitle("Impact of Alcohol Consumption on Mortality Rate") +
  theme_classic() +
  geom_smooth(method = 'lm', se=T, colour = 'orange', lwd=0.4)+
  geom_hline(yintercept = mean(simred$mortality, na.rm=TRUE), color='red', lty='dashed', lwd=0.4)+
  geom_vline(xintercept = mean(simred$alcohol,na.rm=TRUE), color='red', lty='dashed', lwd=0.4)+
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=14)) +
  theme(plot.title = element_text(size = 14))
```



6 | Saving the Scatterplot

You can also save a graph as .png, .JPG (even .pdf) that you can then import in a word document. Although there are many way to use your R output, saving a graph might be sometimes useful.

Use the function `ggsave()` to save your scatterplot. Again, there are tons of examples online, google it.

Hint: You first need to store the graph in an object.

```
# First, create an object #
scatterplot<-ggplot(simred, aes(alcohol, mortality)) +
  geom_point(position='jitter', alpha = 1/5) +
  xlab('Alcohol') +
  ylab('Mortality') +
  ggtitle("Impact of Alcohol Consumption on Mortality Rate") +
  theme_classic() +
  geom_smooth(method = 'lm', se=T, colour = 'orange', lwd=0.4)+
  geom_hline(yintercept = mean(simred$mortality, na.rm=TRUE), color='red', lty='dashed', lwd=0.4)+
  geom_vline(xintercept = mean(simred$alcohol,na.rm=TRUE), color='red', lty='dashed', lwd=0.4)+
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=14)) +
  theme(plot.title = element_text(size = 14))

# Then save the file #
save_plot("scatterplot_mortality.png", scatterplot)
```

7 | Regression Analysis (yes, finally)

Now we can finally run a linear regression with mortality as the outcome variable and alcohol as the predictor using the `lm()` function. Store the results in an object called model and visualise the regression output using `summary()`. Use the reduced data set without outliers.

```
# Store the results in an object called model #
model<-lm(mortality ~ alcohol, simred)
# Visualise the regression output using summary() #
summary(model)
```

```
##
## Call:
## lm(formula = mortality ~ alcohol, data = simred)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -122.69  -22.52   -4.71   17.27  377.77
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 78.155160   0.619648  126.13   <2e-16 ***
## alcohol      0.208270   0.004476   46.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.61 on 6956 degrees of freedom
## Multiple R-squared:  0.2374, Adjusted R-squared:  0.2373
## F-statistic:  2165 on 1 and 6956 DF,  p-value: < 2.2e-16
```

You can also extract specific blocks of the output table. One way of doing it is to use the brackets [] after the `summary()` function. For example `summary()[8]`. Try to extract the block of Coefficients from the table, like this:

```
summary(model)[4]
## $coefficients
##               Estimate  Std. Error   t value Pr(>|t|)
## (Intercept) 78.1551599 0.619648419 126.12823        0
## alcohol      0.2082697 0.004476011  46.53021        0
```

8 | Interpretation

Interpret the results, starting with model evaluation.

1. How much variation in the outcome variable does the model explain? What does this tell us about the model?
    - R^2 = 0.2451 which means that alcohol-related admissions to hospital explain 24.51% of the variation in the standardised mortality ratio. This is not bad, but suggests that there is more to explaining mortality than alcohol abuse.
2. Interpret the slope coefficient.
    - The slope is 0.217635 and is significant. This means that for a one-unit increase in alcohol-related admissions to hospital, the standard mortality ratio rises by 0.217635, on average.

3. Interpret the intercept. What does it mean in practice?
    - The intercept is 76.733653 and significant. This means that on average the standardised mortality rate would be 76.733653 if there were no alcohol-related admissions to hospital.
4. Interpret the results (in plain language) referring to the hypothesis you formulated above.
    - We fail to reject the null hypothesis, and find evidence that higher levels of alcohol consumption significantly increase the risk of mortality.
5. What is the answer to our research question?
    - Yes, we have evidence that higher levels of alcohol consumption significantly increase the risk of mortality.

9 | Exporting the Results

See companion, but do play around with `modelsummary` in the coming weeks to familiarise yourself with it.

10 | Comparing models

You can now run another regression model with a different independent variable. Can you compare your original model with the new one? How? How do you know which independent variable is doing a better job in explaining your dependent variable?