Research Methods for Quantitative Data
Linear Regression – Computing Practical for R
Peter Martin, April 2018

Getting started:

- **Log in to moodle** http://www.ucl.ac.uk/moodle using your UCL username and password
- Find the course: **Research Methods for Quantitative Data**
- Go to **Linear Regression** and download the dataset **ncds-r.zip** to your workspace.

The example is inspired by Elliott et al (2014), who used data from the National Child Development Study (NCDS) to investigate predictors of mental wellbeing. NCDS is a birth cohort of UK residents born in 1958. The data used in this example are from a 2008 survey, conducted when the cohort members were about 50 years old.

The data used in this practical are a subset of the NCDS data set. Data were cleaned to remove records with missing values, and a **subsample of 500 participants** was taken to make data visualization easier for the purpose of a learning exercise. For this reason, **the results you obtain in this practical won't exactly match the results from the lecture**, since the lecture slides are based on a larger extract from the NCDS data (n = 7603). Although you will be using parts of the same data as Elliott et al (2014), this practical does not follow their methods exactly.

Variables in the dataset "ncds":

Variable name	What it measures	Description
MWB	Mental Wellbeing	A summary score from a psychometric scale based on eight survey questions (Warwick-Edinburgh Mental Wellbeing Scale)
NHC	Neighbourhood Cohesion Scale	A summary score measuring perceived cohesiveness of neighbourhood
SUPPORT	Social Support Scale	A summary score representing perceived social support
SOCPART	Social Participation Index	A summary score representing the degree to which participants participate in social groups (leisure activity clubs, political parties, …)
NDHNVQ	Highest Qualification: numeric codes	0 = No qualification 1 = NVQ level 1 2 = NVQ level 2 3 = NVQ level 3 4 = NVQ level 4 5 = NVQ level 5
hqual	Highest Qualification: six categories	The same variable as NDHNVQ, but as a factor (string)
hq3	Highest Qualification: three categories	hqual reduced to three categories (as used in the lecture slides)
limill	Presence of a limiting illness: categorical / string	
limill2	Presence of a limiting illness: numeric	0 = No 1 = Yes

Part 1: Read in data and load packages

#Read in the data
```
ncds = readRDS("ncds.Rds")
```

#Load libraries
```
library(MASS)
library(ggplot2)
library(car)
```

#You may need to install libraries before loading, via:
```
install.packages("package_name")
```

Part 2: Data exploration

#Get an overview of the data set and look at correlations between numeric variables
```
head(ncds)
summary(ncds)
cor(ncds[,c("MWB", "NHC", "SUPPORT", "SOCPART")])
```

#Make a scatterplot of Mental Wellbeing by Neighbourhood Cohesion
```
plot(MWB ~ NHC, ncds)
```

Question 1: Consider the scatterplot of MWB by NHC, and also consider the correlation between MWB and NHC, obtained above. What can you say about the relationship between the two variables in this data set?

#Tip: a nicer-looking scatterplot with added regression line and prediction interval can be obtained using the `ggplot2` package:
```
ggplot(ncds, aes(NHC, MWB)) + geom_point() +
  geom_smooth(method = "lm") + theme_classic() +
  scale_y_continuous(limits = c(0, 56))
```

Part 3: Simple linear regression (Model 1)

Let's predict Mental Wellbeing (MWB) by a single predictor, Neighbourhood Cohesion (NHC), and let's call this simple linear regression **Model 1**.

#You can fit a simple linear regression of Mental Wellbeing on Neighbourhood Cohesion as follows:
```
mod_1 = lm(MWB ~ NHC, ncds)
summary(mod_1)
```

Question 2: Write down the estimated regression equation. Do you have evidence that Neighbourhood Cohesion is a predictor of Mental Wellbeing in the population?

Question 3: What does the R-squared statistic tell you? What is the relationship between R-squared and the Pearson correlation coefficient?

Part 4: Residual diagnostics

Let's check whether basic model assumptions appear to hold for the simple linear regression above.

#The standardised residuals can be calculated using the stdres command from the MASS package:

```
sr_1 = stdres(mod_1)
```

#Predicted values are stored in the model object (which we named mod_1), and can be found like so:

```
pred_1 = fitted(mod_1)
```

#Histogram of standardised residuals

```
x = seq(-3.5, 3.5, 0.01)  # Define the x-axis for plotting
n = length(sr_1)  # Store the sample size in the object n
bin_width = 0.5 # Define the width of the histogram bars

hist(sr_1, main = "Histogram of standardised residuals", xlab
= "Standardised residuals", breaks = seq(-3.5, 3.5,bin_width))

curve(n*bin_width*dnorm(x, mean = 0, sd = 1), add = TRUE, col
= "orangered", lwd = 2)
```

#Spread-level plot of standardised residuals against predicted values

```
plot(pred_1, sr_1, main = "Spread-level plot: standardised
residuals and predicted values", cex.main = 0.9,
     xlab =  "Predicted values", ylab ="Standardised
residuals")
```

Question 4: Interpret the histogram of the standardised residuals, and the spread level plot. Which model assumptions do these plots allow you to check? Do the assumptions appear justified?

#Tip: you can also get quick diagnostic plots like so:

```
par(mfrow = c(2,2))
plot(mod_1)
```

Part 5: Multiple linear regression: adjusting for social support (Model 2)

We will now adjust the regression of Mental Wellbeing on Neighbourhood Cohesion by adding another predictor, Social Support. **Let's call this Model 2.**

```
# We can fit a multiple linear regression like so:
mod_2 = lm(MWB ~ NHC + SUPPORT, ncds)
summary(mod_2)

# You can obtain confidence intervals for the regression coefficients like so:
confint(mod_2)
```

Question 5: Write down the estimated regression equation, and interpret the estimated slope coefficients.

Question 6: Interpret the confidence intervals of the slope coefficients for Neighbourhood Cohesion and Social Support.

Question 7: Compare the estimated slope coefficients of Neighbourhood Cohesion in Model 1 and Model 2. Comment on the difference between the two numbers.

Question 8: Check the assumptions of normality, homoscedasticity, linearity, and absence of outliers in Model 2. Write the code for this yourself.

Part 6: Multiple regression with dummy variables (Model 3)

Let's add another predictor to the model, Highest Qualification. We will use the three-category version (hq3). If a predictor variable has the class 'factor' (rather than numeric), the lm command in R will automatically estimate a regression with dummy variables. The model can be estimated like so:

```
mod_3 = lm(MWB ~ NHC + SUPPORT + hq3, ncds)
summary(mod_3)
```

Question 9: Write down the estimated regression equation and interpret the coefficients of the dummy variables.

Part 7: Model comparison

Does Model 3 improve the prediction of Mental Wellbeing relative to Model 2? We can look at the Analysis of Variance table and conduct an F-test of model comparison like so:

```
anova(mod_2, mod_3)
```

Question 10: What is the null hypothesis of the F-test? What is the alternative hypothesis? Report the result of the F-test: write down the F-statistic, its associated degrees of freedom, and the p-value. Interpret the result.

Question 11: Which other statistic could you use to compare Model 2 and Model 3? What conclusion does this statistic suggest?

Part 8: Exercise

Estimate a multiple linear regression predicting Mental Wellbeing, using the following predictors:

- Neighbourhood Cohesion
- Social Support
- Social Participation
- Limiting Illness
- Highest Qualification (with three categories)

Let's call this model, which is the largest we have estimated so far, Model 4. Write down the estimated regression equation, carry out regression diagnostics, and assess individual predictors: is there evidence that all predictor variables are predictive of Mental Wellbeing in the population, in the presence of the others? Compare this model to Model 3. Which model would you choose, and why?