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

Getting started:

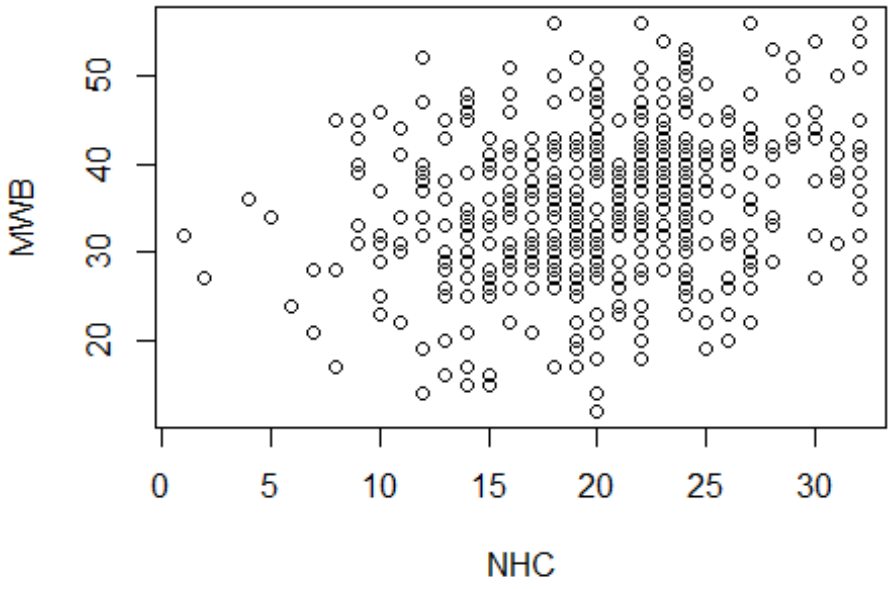
- **Log in to moodle** http://www.ucl.ac.uk/moodle using your UCL username and password
- Find the course: **Research Methods for Quantitative Data**
- Go to **Linear Regression** and download the dataset **ncds-r.zip** to your workspace.

Part 2: Data exploration

```
cor(ncds[,c("MWB", "NHC", "SUPPORT", "SOCPART")])
```

```
##                MWB       NHC   SUPPORT   SOCPART
## MWB      1.0000000 0.2431826 0.2434071 0.1657761
## NHC      0.2431826 1.0000000 0.2208204 0.1308750
## SUPPORT 0.2434071 0.2208204 1.0000000 0.1719907
## SOCPART 0.1657761 0.1308750 0.1719907 1.0000000
```

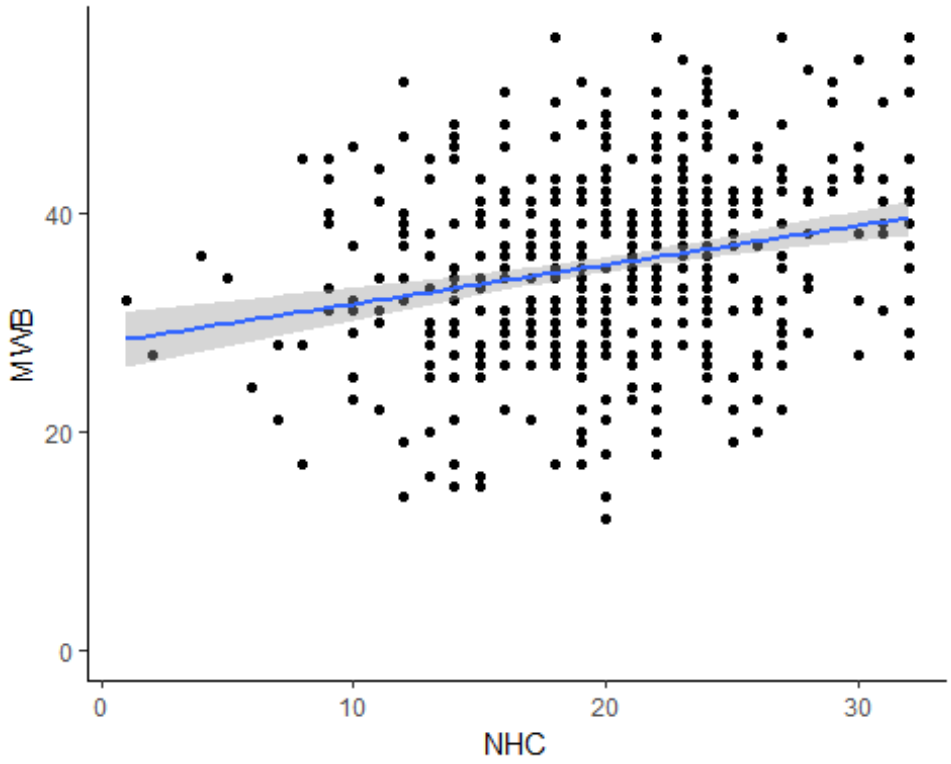
```
plot(MWB ~ NHC, ncds)
```



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?

There is a weak positive linear relationship between Neighbourhood Cohesion and Mental Wellbeing in this data set, with Pearson's r = 0.243. There is no indication of non-linearity.

```
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)

```
mod_base = lm(MWB ~ NHC, ncds)
summary(mod_base)
```

```
Call:
## lm(formula = MWB ~ NHC, data = ncds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.2070  -5.6367   0.1504   5.1133  21.5078
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.05877    1.34902  20.799  < 2e-16 ***
## NHC          0.35741    0.06388   5.595 3.65e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.942 on 498 degrees of freedom
## Multiple R-squared:  0.05914,    Adjusted R-squared:  0.05725
## F-statistic:  31.3 on 1 and 498 DF,  p-value: 3.65e-08
```

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

$$\widehat{MWB} = 28.06 + 0.36 \times NHC$$

The t-test of the coefficient of NHC assesses the null hypothesis that this coefficient is zero in the population. With t = 5.595, on 498 degrees of freedom, we have a p-value of 0.0000000365. Thus there is evidence against the null hypothesis. So we are justified in believing that NHC is a predictor of MWB in the population.

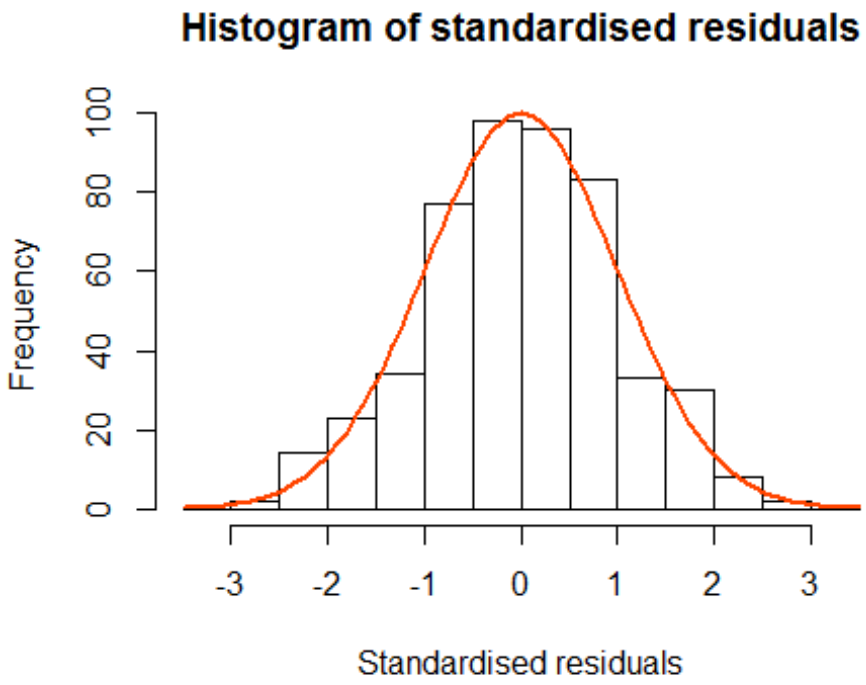
Question 3: What does the R-squared statistic tell you? What is the relationship between R-squared and the Pearson correlation coefficient?

R^2 = 0.059. This suggests that Neighbourhood Cohesion 'accounts for' 5.9 % of the variation in Mental Wellbeing in this data set. (More formally, this means that the prediction error is reduced by 5.9%, relative to a 'null model' that tries to predict Mental Wellbeing without any predictor variable.)
R^2 is equal to the square of Pearson's correlation coefficient: $r^2 = 0.243^2 = 0.059 = R^2$.

Part 4: Residual diagnostics

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?

```
sr_1 = stdres(mod_1)
pred_1 = fitted(mod_1)
summary(sr_1)

##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -2.9250000 -0.7108000  0.0189500  0.0000636  0.6448000  2.7110000

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 = "Standardi
sed residuals", breaks = seq(-3.5, 3.5,bin_width)) # draw histogram
curve(n*bin_width*dnorm(x, mean = 0, sd = 1), add = TRUE, col = "orangered
", lwd = 2) # draw a standard normal curve
```

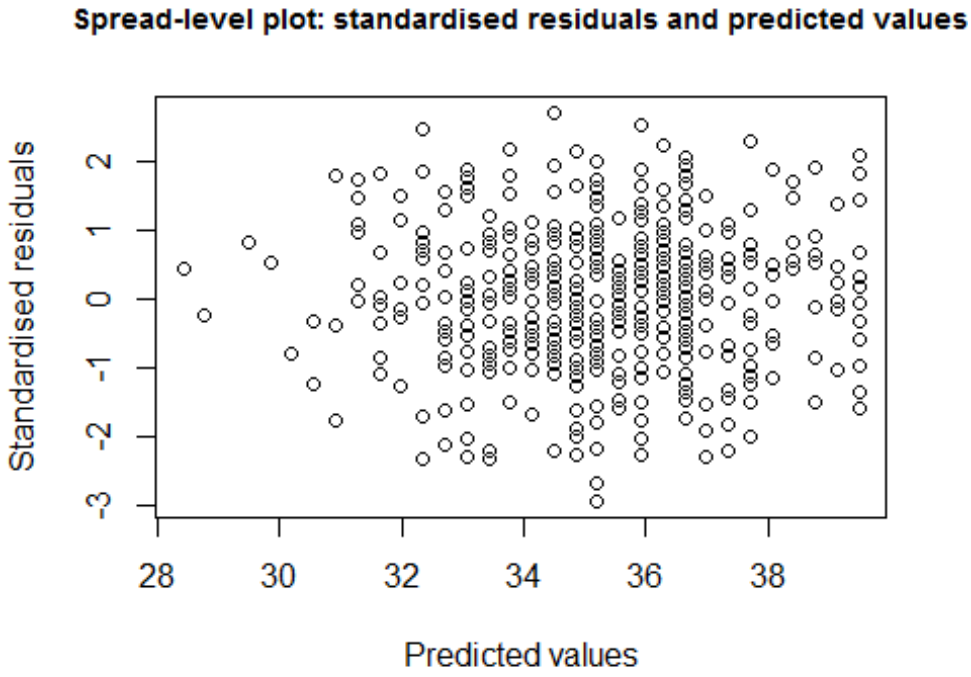


Using the histogram, we can check the following assumptions:

- Normality of errors
- Absence of extreme outliers

The histogram shows that the standardised residuals approximately follow a normal distribution, and that all values are between -3 and +3, thus suggesting that there are no extreme outliers in these data.

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

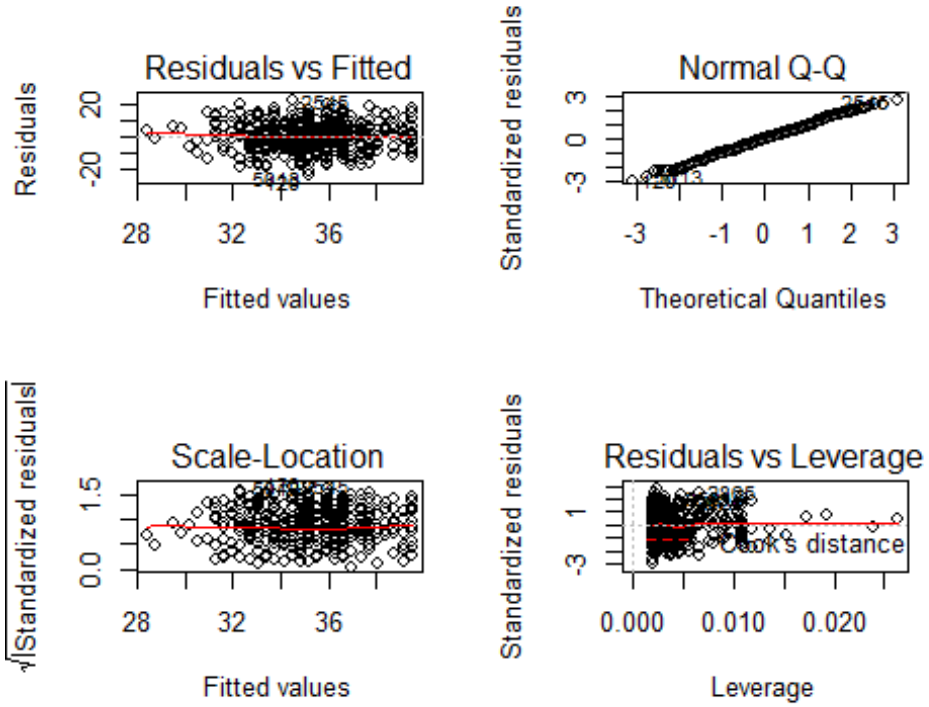


Using the spread-level plot, we can check the assumptions:

- Homoscedasticity of errors
- Linearity of the relationship

The spread-level plot does not suggest any obvious non-linear pattern. The standardised residuals appear to be homoscedastic, by and large, although the residual variance is maybe smallest for the lowest predicted values.

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



The normal q-q plot confirms that the residuals are approximately normally distributed.

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

```
mod_2 = lm(MWB ~ NHC + SUPPORT, ncds)
summary(mod_2)

##
## Call:
## lm(formula = MWB ~ NHC + SUPPORT, data = ncds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8882  -5.0766  -0.1166   4.8495  21.8122
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.46982    1.66178  14.123  < 2e-16 ***
## NHC          0.29269    0.06423   4.557 6.55e-06 ***
## SUPPORT      0.64115    0.14050   4.563 6.36e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.788 on 497 degrees of freedom
## Multiple R-squared:  0.09697,    Adjusted R-squared:  0.09334
## F-statistic: 26.69 on 2 and 497 DF,  p-value: 9.813e-12

confint(mod_2)

##                  2.5 %     97.5 %
## (Intercept) 20.2048379 26.7348116
## NHC          0.1664845  0.4188927
## SUPPORT      0.3650904  0.9172025
```

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

$$\widehat{\text{MWB}} = 23.47 + 0.29\text{NHC} + 0.64\text{SUPPORT}$$

We estimate that a one-point difference in the Neighbourhood Cohesion Scale is associated with a 0.29-point difference in Mental Wellbeing, holding Social Support constant.
We estimate that a one-point difference in Social Support is associated with a 0.64-point difference in Mental Wellbeing, holding Neighbourhood Cohesion constant.

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

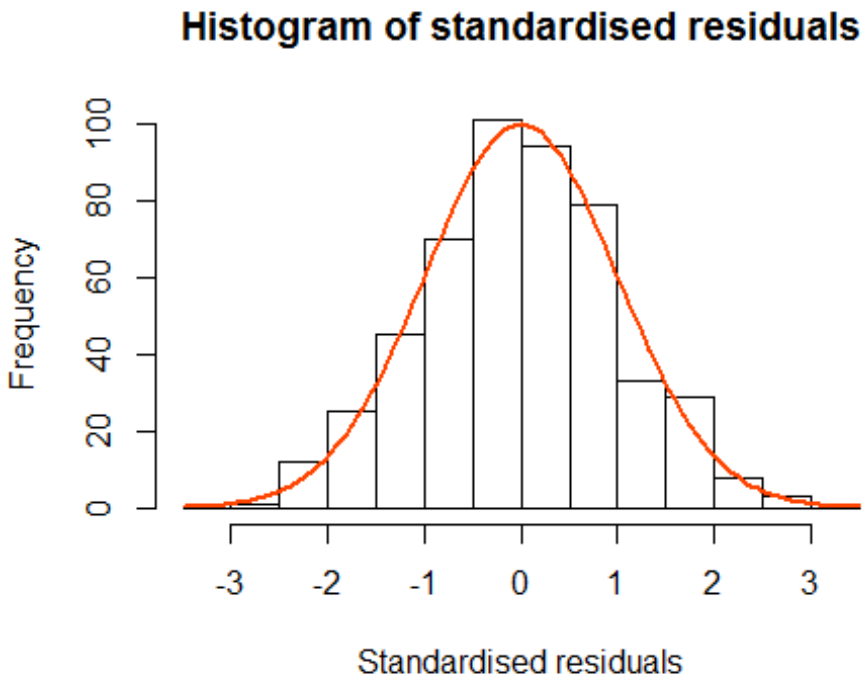
The estimated slope of NHC is 0.29, with a confidence interval from 0.17 to 0.42. So we estimate that the interval from 0.17 and 0.42 contains the true coefficient, with 95 % confidence. Similarly, we estimate that the interval from 0.37 and 0.92 contains the true coefficient for SUPPORT.
Neither confidence interval contains the value 0, suggesting that both variables are predictors of Mental Wellbeing in the population, in the presence of each other. (Equivalently, we might have looked at the p-values of the t-statistics to come to this conclusion.)

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

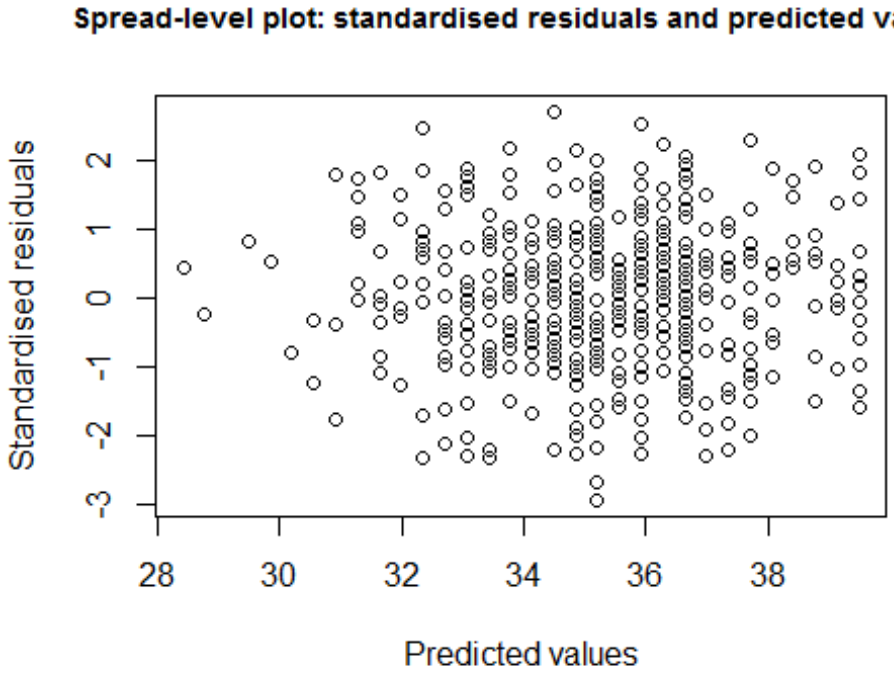
In the simple linear regression (mod_1), the slope of NHC is estimated as 0.36. When adjusting for Social Support, the estimated slope is 0.29. If we take the slope as an estimate of the 'effect' of Neighbourhood Cohesion on Mental Wellbeing, the downward adjustment suggests that some of the correlation between NHC and MWB is not uniquely due to the effect of Neighbourhood Cohesion itself, but may be due to the effect of Social Support, which is correlated with both NHC and MWB.

Question 8: Check the assumptions of normality, homoscedasticity, linearity, and absence of outliers in mod_2. Write the code for this yourself.

```
sr_2 = stdres(mod_2)

pred_2 = fitted(mod_2)

par(mfrow = c(1,1))

hist(sr_2, main = "Histogram of standardised residuals", xlab = "Standardi
sed residuals")
curve(n*bin_width*dnorm(x, mean = 0, sd = 1), add = TRUE, col = "orangered
", lwd = 2) # draw a standard normal curve
```




```
plot(pred_2, sr_2, main = "Spread-level plot: standardised residuals and p
redicted values", cex.main = 0.9,
     xlab =  "Predicted values", ylab ="Standardised residuals")
```



The histogram and spread-level plot lend support to the assumptions of normality, homoscedasticity, linearity, and absence of outliers.

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

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

##
## Call:
## lm(formula = MWB ~ NHC + SUPPORT + hq3, data = ncds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.4574  -5.1675   0.1218   4.7464  22.0477
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              22.00842    1.64783  13.356  < 2e-16 ***
## NHC                       0.27756    0.06281   4.419 1.22e-05 ***
## SUPPORT                   0.63570    0.13739   4.627 4.75e-06 ***
## hq3A-level                2.91058    0.94725   3.073  0.00224 **
## hq3Degree or equivalent   3.71937    0.75973   4.896 1.33e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.605 on 495 degrees of freedom
## Multiple R-squared:  0.1424, Adjusted R-squared:  0.1354
## F-statistic: 20.54 on 4 and 495 DF,  p-value: 1.13e-15
```

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

$$\widehat{MWB} = 22.01 + 0.28NHC + 0.64SUPPORT + 2.91Alevel + 3.72Degree$$

People with an A-level qualification have an estimated MWB score that is 2.91 points higher than people with a lower qualification, controlling for Neighbourhood Cohesion and Social Support.

People with a degree have an estimated MWB score that is 3.72 points higher than people with a qualification below A-level, controlling for Neighbourhood Cohesion and Social Support.

Part 7: Model comparison

```
anova(mod_2, mod_3)

## Analysis of Variance Table
##
## Model 1: MWB ~ NHC + SUPPORT
## Model 2: MWB ~ NHC + SUPPORT + hq3
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    497 30147
## 2    495 28632  2      1515 13.096 2.869e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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, it's associated degrees of freedom, and the p-value. Interpret the result.

Null hypothesis of the F-test:
This can be phrased in various ways (all are correct – they are equivalent):

- Model 2 and Model 3 give equally good prediction of Mental Wellbing.
- Both additional parameters contained in Model 3, but not in Model 2, are zero.
- Highest qualification (as represented by the two dummy variables, Alevel and degree) does not add to the prediction of Mental Wellbeing, when controlling for Neighbourhood Cohesion and Social Support.

Alternative hypothesis:
The three phrases below are all equivalent ways to state the alternative hypothesis:

- Model 3 is superior to Model 2 in the prediction of Mental Wellbeing.
- At least one of the additional predictors contained in Model 3 adds to the prediction of Mental Wellbeing.
- At least one of the dummy variables representing highest qualification adds to the prediction of Mental Wellbeing, when controlling for Neighbourhood Cohesion and Social Support.

Reporting the F-test: **$F_{2,495}$ = 13.096, p = 0.000** to three decimal points. This suggests that the null hypothesis should be rejected: Model 3 is superior to Model 2.

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

We can compare the adjusted R^2 values of Model 2 and Model 3.

- Model 2: adj R^2 = 0.09
- Model 3: adj R^2 = 0.14

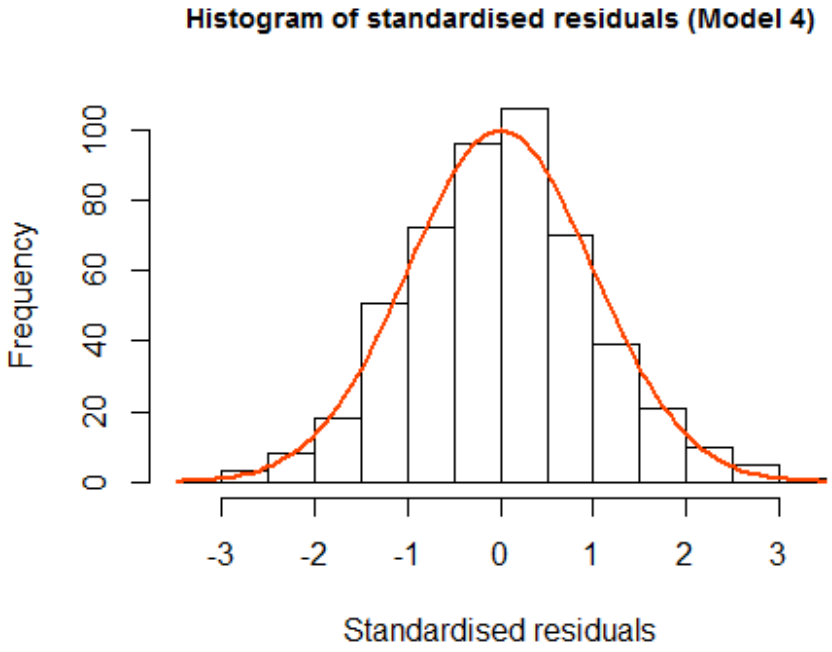
Model 3 has the higher adjusted R2, which suggests that Model 3 is superior to Model 2.

Part 8: Exercise (Model 4)

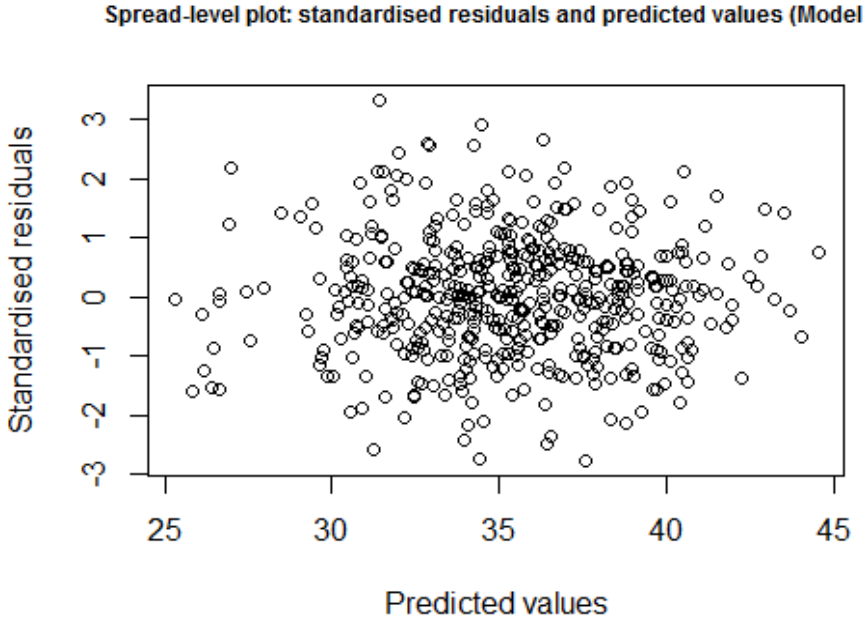
R-code and results for Exercise 8:

```
#Estimate the model and store standardised residuals and predicted values
mod_4 = lm(MWB ~ NHC + SUPPORT + SOCPART + limill + hq3, ncds)
sr_4 = stdres(mod_4)
pred_4 = fitted(mod_4)

# Model summary and confidence intervals
summary(mod_4)

##
## Call:
## lm(formula = MWB ~ NHC + SUPPORT + SOCPART + limill + hq3, data = ncds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.6121  -4.9756   0.0299   4.4398  24.6002
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              22.50206    1.62175  13.875  < 2e-16 ***
## NHC                       0.27398    0.06184   4.430 1.16e-05 ***
## SUPPORT                   0.65398    0.13681   4.780 2.32e-06 ***
## SOCPART                   0.20713    0.16160   1.282  0.20053
## limillYes                -3.88159    0.88075  -4.407 1.29e-05 ***
## hq3A-level                2.70871    0.93261   2.904  0.00384 **
## hq3Degree or equivalent   3.15752    0.76818   4.110 4.63e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.458 on 493 degrees of freedom
## Multiple R-squared:  0.1786, Adjusted R-squared:  0.1686
## F-statistic: 17.86 on 6 and 493 DF,  p-value: < 2.2e-16

confint(mod_4)

##                               2.5 %     97.5 %
## (Intercept)              19.3156607 25.6884613
## NHC                       0.1524755  0.3954846
## SUPPORT                   0.3851670  0.9227867
## SOCPART                  -0.1103798  0.5246470
## limillYes                -5.6120730 -2.1511038
## hq3A-level                0.8763229  4.5410997
## hq3Degree or equivalent   1.6482212  4.6668246
```

```
#Regression diagnostics

hist(sr_4, main = "Histogram of standardised residuals (Model 4)", xlab =
"Standardised residuals", cex.main = 0.9, breaks = seq(-3.5, 3.5,bin_width
))
curve(n*bin_width*dnorm(x, mean = 0, sd = 1), add = TRUE, col = "orangered
", lwd = 2) # draw a standard normal curve
```




```
plot(pred_2, sr_2, main = "Spread-level plot: standardised residuals and p
redicted values (Model 4)", cex.main = 0.8,
     xlab =  "Predicted values", ylab ="Standardised residuals")
```



```
#Compare Model 4 to Model 3
anova(mod_3, mod_4)

## Analysis of Variance Table
##
## Model 1: MWB ~ NHC + SUPPORT + hq3
## Model 2: MWB ~ NHC + SUPPORT + SOCPART + limill + hq3
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    495 28632
## 2    493 27423  2    1209.3 10.871 2.399e-05 ***
## ---
##

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answers to Exercise 8:

Regression equation:

$$\widehat{MWB} = 22.50 + 0.27NHC + 0.65SUPP + 0.21SocPart - 3.88\,LimIll + 2.71Alevel + 3.16Degree$$

The t-tests of slope coefficients suggest that all variables are predictors of Mental Wellbeing in the population, with the exception of Social Participation, whose p-value (0.20) suggests that it does not add to the prediction of Mental Wellbeing, when the other predictors are controlled for.

Regression diagnostics suggests an adequate model, although we now have one large positive standardised residual above 3, which may suggest an outlier.

The ANOVA table suggests that Model 4 is superior to Model 3, with F2, 493 = 10.87, p = 0.000. Also, Model 4 has a higher adjusted R^2 (0.17) compared to Model 3 (0.14).