Q-STEP R 'HOW TO' GUIDES:

LINEAR REGRESSIONS

Creator: Dr Patrick English

This guide takes you through two examples of linear regression analysis - including one bivariate and one multiple linear regression. Data comes from the OECD: https://stats.oecd.org/.

Gross Domestic Product (GDP) and Health Spending

The following example looks at the bivariate relationship between a country's wealth and their healthcare expenditure. We will go through the process of fitting an Ordinary Least Squares (OLS) regression model to some real world data - including testing and examining the linear modelling assumptions - and how to interpret and report the findings.

Let's suppose we have a research question which is as follows: does greater average wealth in a country mean higher rates of spending on public healthcare?

As we would have pretty sensible grounds to assume that countries with more money will be able to spend more on health (making health care a priority over say poverty reduction or infrastructure investment), we hypothesise that: countries with greater wealth will spend more on healthcare. We must test this hypothesis in a regression model in order to answer our research question.

Hypothesis testing involves collecting data from the world (our 'population') and using statistical methods to examine the relationship we are hypothesising about. The 'null hypothesis' is that there is no relationship between wealth and health care expenditure. The null hypothesis could either be rejected by the data (suggesting that there is some relationship), or not rejected (implying that there is no relationship at all between our variables).

.

Q-Step A step-change in quantitative social science skills Funded by the Nuffield Foundation, ESRC and HEFCE
Nuffield Foundation
Sheffield Methods Institute.
The University Of Sheffield.

To fit a regression model and test our hypothesis, we need some data. We can call up the following CSV file named 'OECDData' which contains information on the socio-economic situation in 40 countries around the world from the year 2015. The file contains information on country GDP per capita in thousands of US dollars (GDPPC), reported unemployment rates (Unemployment), average life expectancy at birth (LifeExpectancy), and public spending on health as a proportion of GDP (HealthSpending). The data comes from the OECD's country statistical databases: https://stats.oecd.org/

```
my_data <-
  read.csv("https://dl.dropboxusercontent.com/s/i0qtus70l0n57wd/OECDdata.csv?dl=0")

head(my_data)
```

```
##     Country    GDPPC Unemployment LifeExpectancy HealthSpending
## 1 Australia 46.64403     6.057946           82.5          9.321
## 2   Austria 49.95926     5.723468           81.3         10.343
## 3   Belgium 45.56118     8.481071           81.1         10.106
## 4    Canada 44.64737     6.908333           81.9         10.377
## 5     Chile 22.42970     6.213719           79.9          8.022
## 6  Colombia 13.83318     8.960834           76.2          6.191
```

Assumptions for Linear Models

When statistically analysing data we must be confident that we are using the correct tools - horses for courses, if you will. Some statistical techniques are simply inappropriate to use on certain types of data, and linear regressions are no different. These are the three 'most important' assumptions for modeling using OLS:

1) The dependent variable is continuous.
2) The residuals (errors) from the fitted model are non-patterned and independent.
3) The residuals from the fitted model follow a normal distribution.

Briefly, a model residuals is the 'error' between the estimated value of the dependent variable that the linear regression predicts for a given (set of) value(s) of the independent variable(s), and the actual value that the dependent variable takes at this fitted value. OLS models work by fitting regression equations which a) reduce the overall size of the total errors in the model to be as small as possible, but also b) keeping the 'mean value' of these residuals at 0 (so that the errors 'balance out' above and below the regression line).

To check the residual assumptions, we plot them directly after running our models. The two most common problems regarding non-patterned residuals are clusters/patterns in the residuals - such as a scattered or heaped collections of residuals in different sections of the plot - and non-constant variance (known as heteroscedasticity) - whereby as we move along the fitted values (x axis), the range of residuals significantly changes and meaning that the model is explaining substantially different amounts (or ranges) of variation.

As we move through the two examples below, we will address each assumption in turn. If the assumptions are not met, then we must either make changes to our model (by adding in new variables), or use a different type of model (more on this below).

Examining the Dependent Variable

In order to examine whether or not our data is suitable for linear regression modeling, we should first investigate the summary statistics of our data and plot its distribution. This checks the first assumption: The dependent variable is continuous.

In this example, as we are testing the relationship between country wealth and health spending, we will be using HealthSpending as our dependent variable and GDPPC as our independent variable. Our OLS regression analysis will thus report to us the impact of a one-unit change in GDP per capita on health spending, according to our sample data.

To familiarise ourselves with the data, and test the first assumption, we summarise our dependent variable and then plot a histogram using the ggplot2 package:

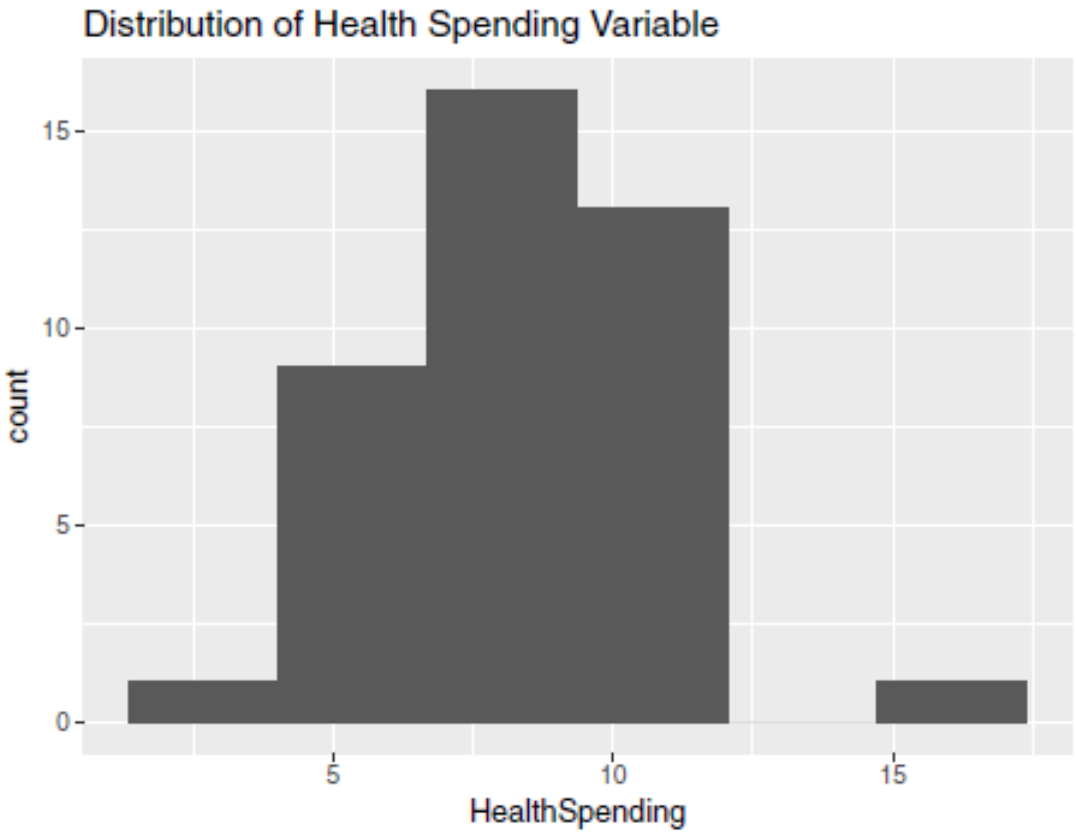
```
summary(my_data$HealthSpending)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.433   6.771   8.415   8.510  10.150  16.816
```

```
library(ggplot2)
g <- ggplot(my_data, aes(x=HealthSpending))

## Set the number of bins using the Freedman-Diaconis rule
x <- my_data$HealthSpending
bw <- diff(range(x)) / (2 * IQR(x) / length(x)^(1/3))

## Plot graph
(graph <- g + geom_histogram(bins=bw)
  + labs(title="Distribution of Health Spending Variable"))
```



From the summary statistics, we can see a nice continuous range which looks approximately normally distributed around a fairly centrally placed mean and median. From the histogram, though the variable does not follow a perfect normal distribution (i.e. we cannot draw a perfect 'bell curve' over it), it does show a central cluster of values with a decreasing count of observations either side. This is encouraging for later requirements regarding a normal and unpatterned distribution of errors. We can confirm that our dependent variable is continuous, and move on from here.

Fitting a Bivariate OLS

We now move to fit an OLS model to the data, using GDPPC as our independent variable. The R command for linear models is lm(), and inside the command we must specify the dependent (y) variable, which is predicted (~) by an independent (x) variable, and the data frame from which the variables originate (data).

```
l_model <- lm(HealthSpending ~ GDPPC, data=my_data)

summary(l_model)
```

```
##
## Call:
## lm(formula = HealthSpending ~ GDPPC, data = my_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5668 -1.2652  0.4618  1.0927  7.1095
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.93198    0.86981   6.820 4.32e-08 ***
## GDPPC        0.06690    0.02067   3.237  0.00251 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.212 on 38 degrees of freedom
## Multiple R-squared:  0.2162, Adjusted R-squared:  0.1955
## F-statistic: 10.48 on 1 and 38 DF,  p-value: 0.002505
```

The regression output contains a number of important pieces of information. The first row of text displays the command we used to generate the model. The second line reports the distribution and average value of the model residuals. Recall that a residual is the distance between the predicted value of the dependent variable (according to the line of best fit) and the actual position of observed values. The 'median' value of the residuals is 0.5, which we can be quite pleased with. The wide range (relative to the distribution of the variables) in residual sizes however (between around -6.5 and 7.0) suggests that more work could be done (and, specifically, that we might have outliers to look into).

The third set of information concerns the model coefficients and their associated reliability, robustness, and significance estimates.

The 'Estimate' column contains two coefficient estimates. The first is for the model intercept (which is the value of our dependent variable y when the independent variable x is zero. Note that this may or may not be a sensible thing to consider - do countries have GDP per capita equal to zero?!). The second is the coefficient associated with the independent variable itself: the estimated (average) impact on our dependent variable (health spending)from a one-unit increase in our independent variable (GDP per capita).

So, from the model summary we can see that - according to our sample data - there is an average 0.1% increase in health spending (as a proportion of GDP) for each one-thousand extra dollars in GDP per capita (note how important it is to know the units of measurement for the dependent and independent variables in order to be able to interpret the magnitude of the regression coefficients).

Importantly, there are standard error (2nd column) and significance estimates (or p-values) (4th column) for both coefficients. The significance estimate [Pr(>|t|)] for our relationship between GDP per capita and health spending is given as a value of 0.00251, meaning we have a statistically significant relationship at the $p < 0.01$ level. This is represented by two stars (significance codes). This result means that we have enough evidence to reject the null hypothesis (that there is no real relationship between the two variables) in favour of there being a positive relationship between these variables.

P-values represent the probability that we would observe the same or a more extreme (larger) estimated relationship (as in our regression output) under the 'null hypothesis' situation. So, this means we have a probability estimate regarding how likely it is that our estimated effect of the independent variable on the dependent variable does not represent a 'real' relationship, but rather a 'sampling event' (a random-chance event which, if repeated on a different sample, would most likely not occur again). If the probability is sufficiently small (i.e. passes the 'lower than 0.05 convention), we can reasonably reject the suggestion that the relationship would also be observed under the null hypothesis situation.

Finally, the last two lines show us statistics about the model itself regarding how well it describes the data. Firstly, the R-Squared statistics show us the proportion of the variance in the dependent variable which is explained by the variation in the independent variable in our model. In this case, it is around 20%.

The second line gives a p-value for the model overall. This tests the hypothesis that our specified model is better at explaining the variance in the dependent variable than a null (completely empty) model. Here, a p-value of 0.0025 tells us that our specified model is significantly better than a null model (as $p < 0.01$).

Finally, a word of caution. The relationshiop between GDP per capita and health expenditure that we have identified in the data can only be interpreted as a postive association. We cannot ascribe any causal relationship between the two variables ('correlation does not imply causation'). While we may think that the relatioship goes from richer countries being able to spend a greater share of their income on health, and that is the relationship that we have implicitly estimated, it is also plausible that countries which spend more on health are likely to grow richer (for example because they have a healthier workforce able to work more productively and longer into old age. That is, the relationship may run rom health expenditure to income as well as from income to health care. In practice, the relationship may go both ways.

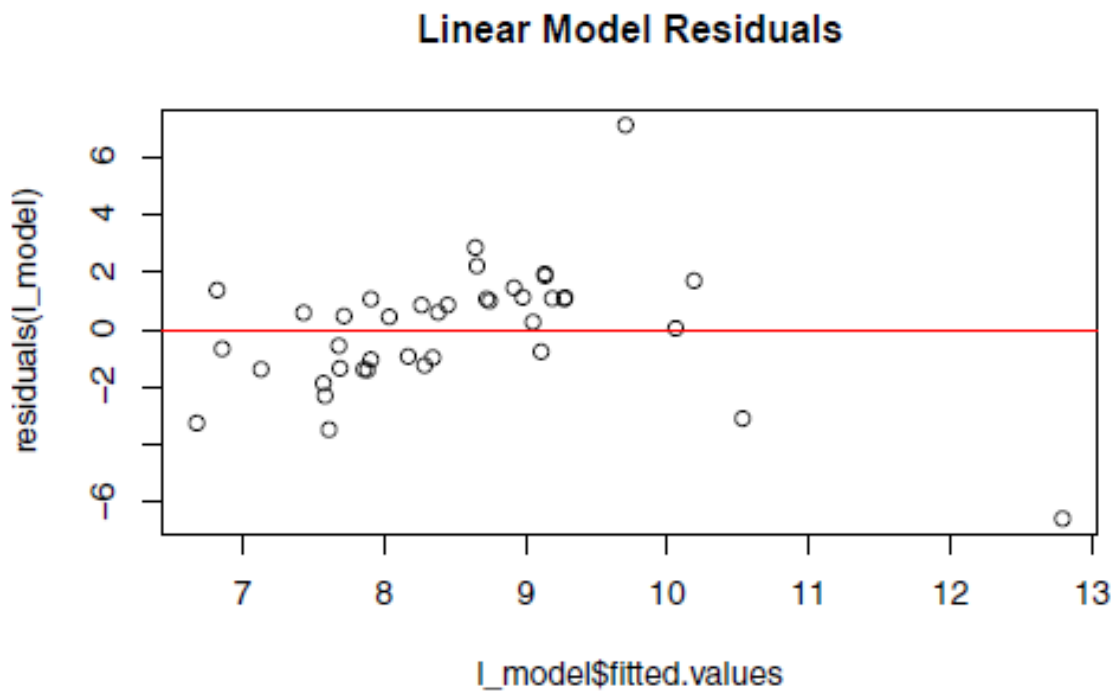
Methods for identifying causal relationships are more advanced and beyond the scope of this note.

Examining Model Residuals

Recall the second of the two most important assumptions required for OLS modelling: The residuals (errors) are non-patterned and independent.

To test this particular assumption, we need to inspect a residuals plot. We will use a base R residuals plot and add an abline (a straight line across y=0). Remember, what we are looking for here is any patterning or clustering around the plot (which would violate linearity), or any changes in the amount (or range) of variance explained across the fitted values (which would violate homoscedasticity).

```
plot(l_model$fitted.values, residuals(l_model), main="Linear Model Residuals");abline(0,0, col="red")
```

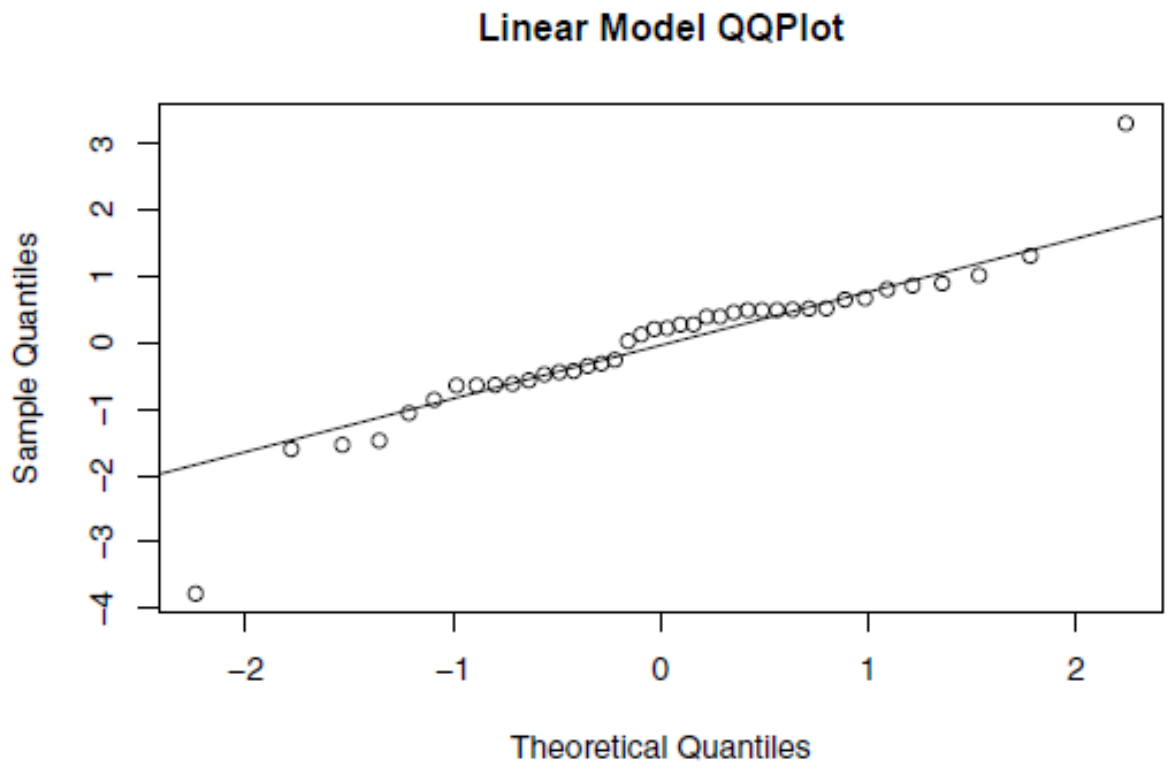


As we can see, there isn't a perfect random patterning to the residuals, but overall the distribution is not too bad. We have two significant outliers in the plot which ought to be investigated further, but the centrsl cluster of resdiuals looks fairly randomly assigned across the fitted-values.

Moving on, we recall the third 'most important' assumption from above: The residuals follow a normal distribution.

We can check this by using a QQNorm (Quantile-Quantile plot for normal distributions) plot, which tells us whether or not the error term in our model is (approximately) normally distributed. This is important for the tests of statistical significance described above (if the errors are not normally distributed, we cannot trust the p-values reported). Here we are looking for our residual values to mostly line up well along an 'ideal fit' value line. If there is some level of deviance, we want it to be mostly balanced on either side.

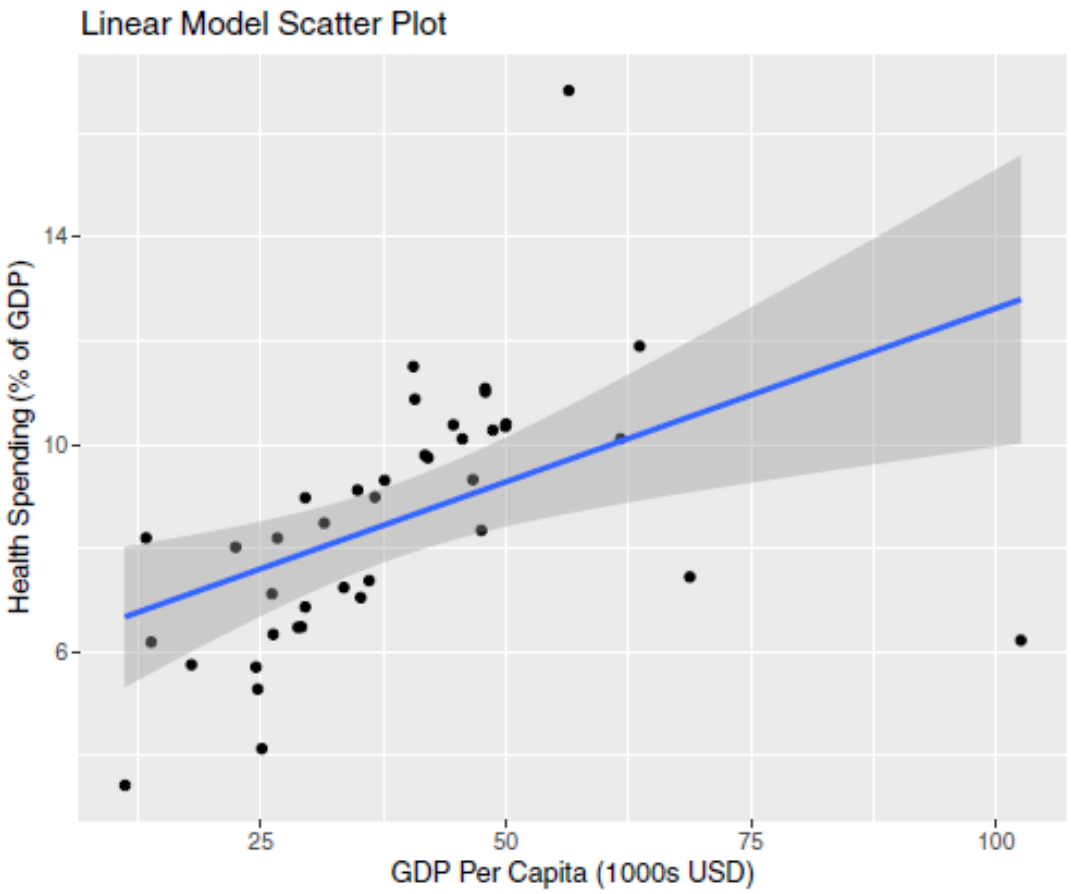
```
l_resids <- rstandard(l_model)
qqnorm(l_resids, main="Linear Model QQPlot");qqline(l_resids)
```



Again, we can see our outliers represented strongly at the top and tail of the QQNorm plot, but all other residual values lie close to the 'normality line' fitted through the middle, thus suggesting a normally distributed model error term.

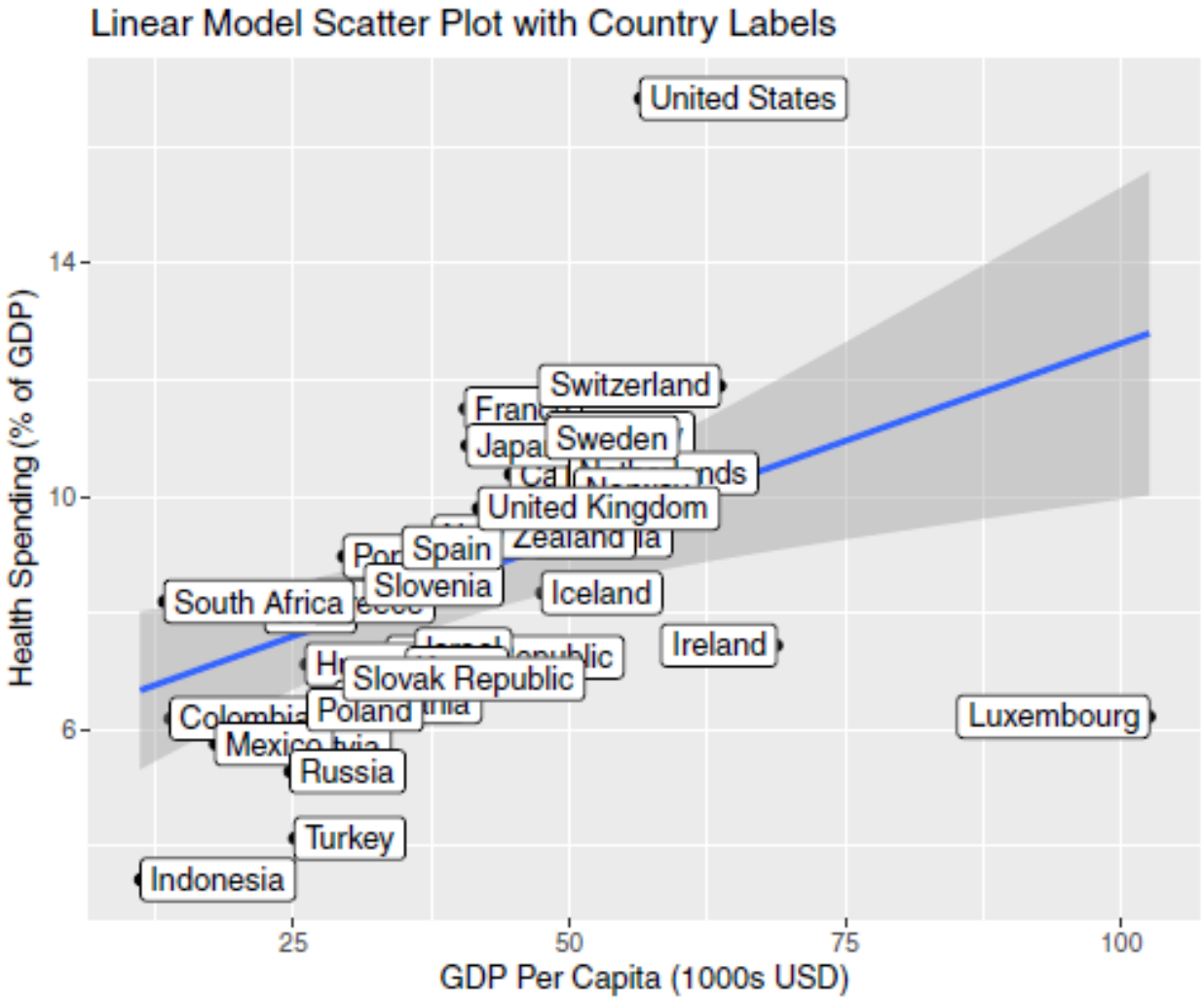
Now that we are confident our model and data passes the linear and normality assumptions, we can visualise the relationship between our dependent and independent variables by fitting a scatter plot with a linear line of best fit (geom_smooth).

```
g <- ggplot(my_data, aes(x=GDPPC, y=HealthSpending))
(graph <- g + geom_point()
  + geom_smooth(method="lm")
  + labs(x = "GDP Per Capita (1000s USD)",
         y = "Health Spending (% of GDP)",
         title ="Linear Model Scatter Plot"))
```



The scatter plot and smoothed line of best fit nicely illustrate the relationship reported in the regression estimate. By and large, there is a linear association between increasing GDP per capita and health spending as a proportion of GDP. There are two outliers in our graph which are of note. We can add country labels to the scatter plot to give further information about them.

```
g <- ggplot(my_data, aes(x=GDPPC, y=HealthSpending))
(graph <- g + ggplot2::geom_point()
  + geom_smooth(method="lm")
  + geom_label(label=my_data$Country, label.size = 0.1, hjust = "inward")
  + labs(x = "GDP Per Capita (1000s USD)",
         y = "Health Spending (% of GDP)",
         title ="Linear Model Scatter Plot with Country Labels"))
```



We can see that the United States and Luxembourg (and also to some extent, Indonesia) are very much outliers to the bivariate linear relationship. Though they are significant outliers, as they are almost polar opposites of each other the overall 'net effect' on the regression estimate will actually be fairly close to zero. Therefore, we do not have to be too concerned about them.

Plotting the descriptive statistics is an important part of the process of data analysis. Not only does it allow us to check for the performance of the line of best fit and to inspect outliers and potential issues with model residuals, it also is important for convincing the reader that our statistically estimated associations can be easily spotted in the raw data.

Multiple Linear Regression

Though our bivariate linear model reported a statistically significant relationship which passed the assumptions required for linear modeling, we cannot reasonably conclude that we have uncovered a real world relationship through our data without first also testing for (or read: controlling for) other factors which might affect health care spending. In other words, we must also examine other factors which might cause increases in health spending and see if our GDP effect 'survives'.

Adding additional variables into our model turns a bivariate OLS into a multiple linear regression. Given the power and flexibility of regression analysis, this process is very simple to do but it is important to understand and be able to interpret the impact on the regression outputs that additional variables will have.

Consider that we might expect higher life expectancy to also correlate with higher health spending - if people are living longer, we will need to spend more on health care (particularly as older people often require more care than their younger counterparts). Also, let's imagine that we think unemployment rates might also have something to do with healthcare spending - perhaps when unemployment is higher, so a population's quality of life and ability to remain healthy might be negatively impacted, and so the state ends up spending more on healthcare.

N.B. The above model specification is not supposed to represent an exhaustive list of factors which relate to healthcare spending, but constitute a demonstration of the thought process for the purposes of this exercise. Additional variables for linear regression models should also be added with the aid of previously published work and theoretical argument wherever possible. Keep in mind the rule of thumb on overfitting regression models - there should never be less than around 10x as many observations as there are variables (including the dependent variable) in a regression.

We can specify a further regression model to test the extent to which life expectancy and unemployment also explain variance in health spending. Here, we not only estimate the relationship between multiple variables and the dependent variable, but the coefficients for each variable become partial coefficients which are accounting for the effects of other variables in the regression on the dependent variable. This allows us to isolate variation (and significance) associated with our variable(s) of interest, while 'controlling for' the effects of other, competing explanations.

.

The command to run a multivariate linear regression is exactly the same as for the bivariate regression, the only difference is that we include additional variables from our dataframe after GDPPC:

```
ml_model <- lm(HealthSpending ~ GDPPC + LifeExpectancy + Unemployment, data = my_data)

summary(ml_model)
```

```
##
## Call:
## lm(formula = HealthSpending ~ GDPPC + LifeExpectancy + Unemployment,
##     data = my_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.702 -1.073 -0.043  1.081  7.691
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -5.32061    6.89443  -0.772    0.445
## GDPPC           0.04750    0.02525   1.881    0.068 .
## LifeExpectancy  0.14595    0.09046   1.613    0.115
## Unemployment    0.05269    0.07333   0.718    0.477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.191 on 36 degrees of freedom
## Multiple R-squared:  0.2714, Adjusted R-squared:  0.2107
## F-statistic: 4.471 on 3 and 36 DF,  p-value: 0.009081
```

There are some substantial changes in the results when we include life expectancy and unemployment rate in the model. Most importantly, both the coefficient and statistical significance for GDP per capita have changed. The model R-Square value has increased as it must do - as more variables are included in the model, more of the variance in the dependent variable will be 'explained'. Better to look at the Adjusted R-squared, since this measure accounts (compensates) for the number of added variables - it has increased slightly as compared to the bivariate regression above.

The reason why the coefficients for the model intercept and GDP have changed is because they are now being calculated with regard to levels of the other additional variables (in our case, life expectancy and unemployment). This creates a multi-dimensional space of data points through which a line of best fit is placed.

.

In other words, this means that our reported coefficient for each variable (and associated reliability and significance estimates) constitutes the direct relationship between said variable and the dependent variable while also controlling for levels of all other independent variables in the regression. This allows us to test our hypothesis between GDP and heathcare whilst controlling for (taking into account) competing determinants of the level of health care expenditure.

So, what has happened in terms of our model results?

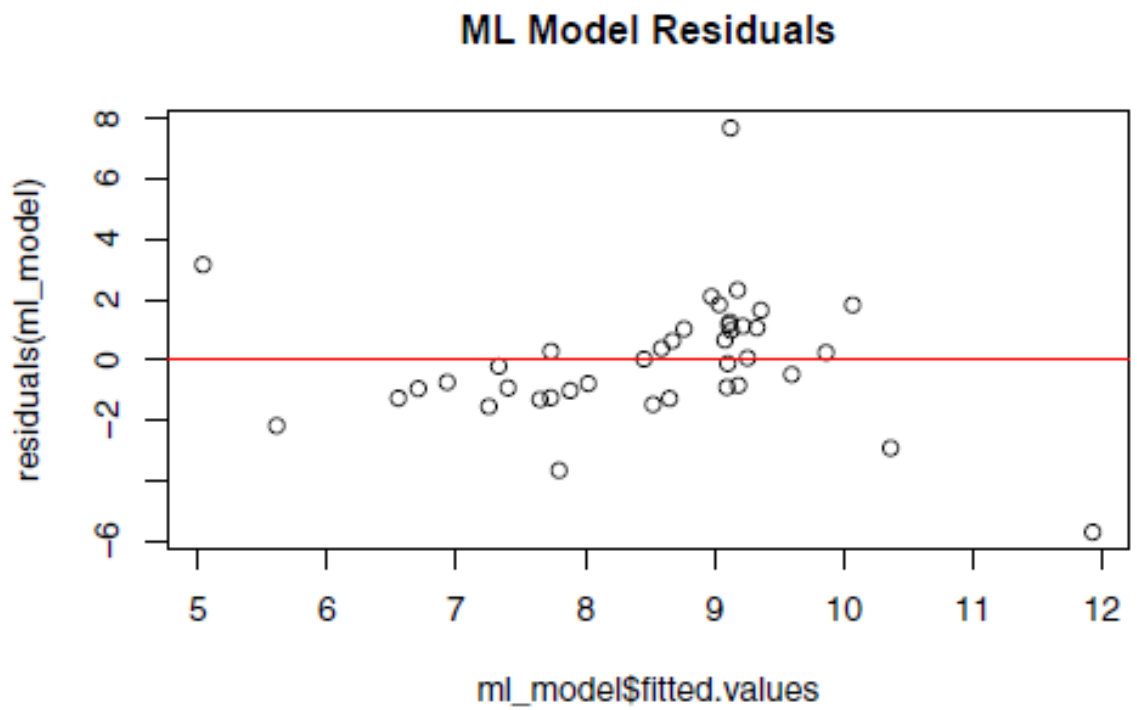
Firstly, it appears that neither life expectancy nor unemployment levels significantly relate to variance in health spending. Though we have positive coefficient estimates, they are not statistically significantly different from zero. Therefore, the null hypothesis is not rejected and we cannot suggest that the positive coefficients observed in our data are caused by other than chance.

Secondly, our GDPPC coefficient has been reduced by around 30%, from 0.07 to 0.05. This has had an important consequence for the significance estimate. The p-value for our GDPPC coefficient has increased to 0.068, meaning our relationship is now only statistically significant at the $p < 0.1$ (10%) level. This is represented by the . significance code in the table. Though normally failure to reject the null hypothesis is defined when $p > 0.05$ (5% sifnificance level), given we are working with only a small sample of data (40 observations), we would probably still report this as a finding.

In all, the addition of further variables has improved the variance explained by our regression model but has had a detrimental impact in terms of statistical significance - reducing certainty of null hypothesis rejection from 99% ($p < 0.01$) to 90% ($p < 0.1$). The p-value of the model itself is still statistically significant at the $p < 0.01$ level.

Again, we need to inspect the residuals plot to check that the model is still an appropriate fit for our data:

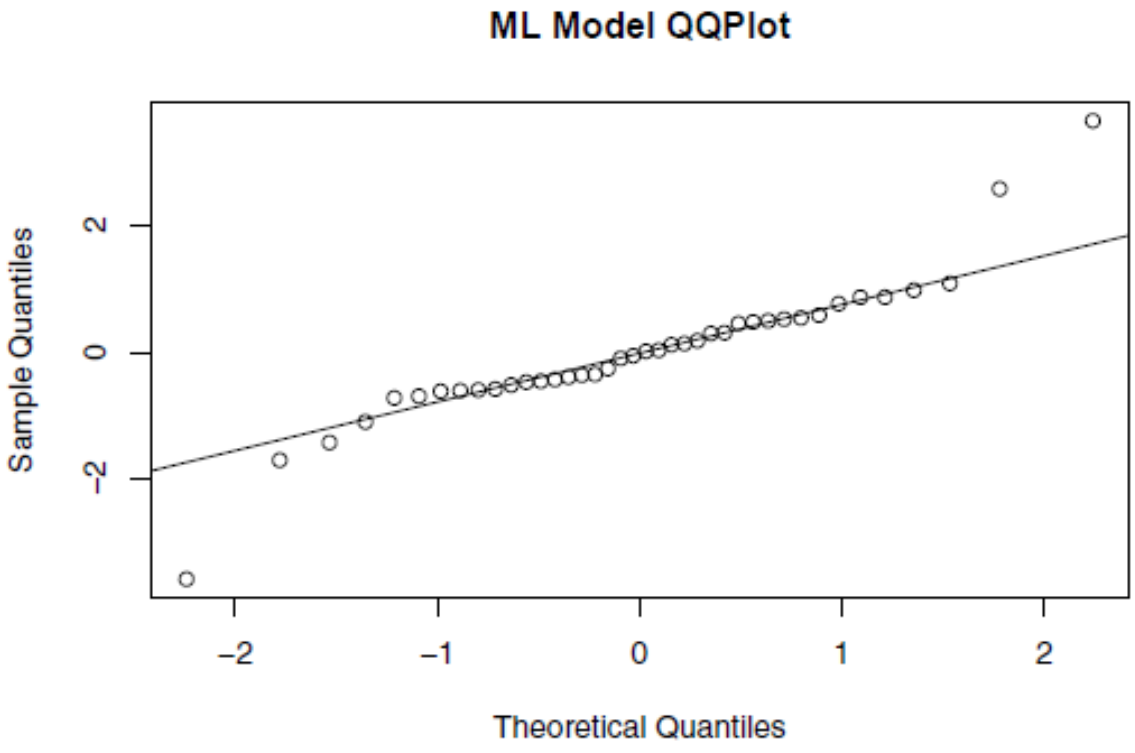
```
plot(ml_model$fitted.values, residuals(ml_model), main="ML Model Residuals");abline(0,0, col="red")
```



The residual plot this time looks a little more problematic, as we can see a fairly evidence different in the range of variance explained between the fitted values around 6.5 and 9 as we can 9 and 10.5. Substantively speaking, the former cluster is mostly 'undershot' by the fitted values (consistently lower than the actual values, represented by the 0,0 line), while the former cluster is mostly overshooting. What this means is that our model estimates are potentially unreliable, and the finding regarding the effect of GDP being significant at the $p < 0.1$ level should be treated with scepticism.

The QQNorm plot for the model looks as follows:

```
ml_resids <- rstandard(ml_model)
qqnorm(ml_resids, main="ML Model QQPlot");qqline(ml_resids)
```

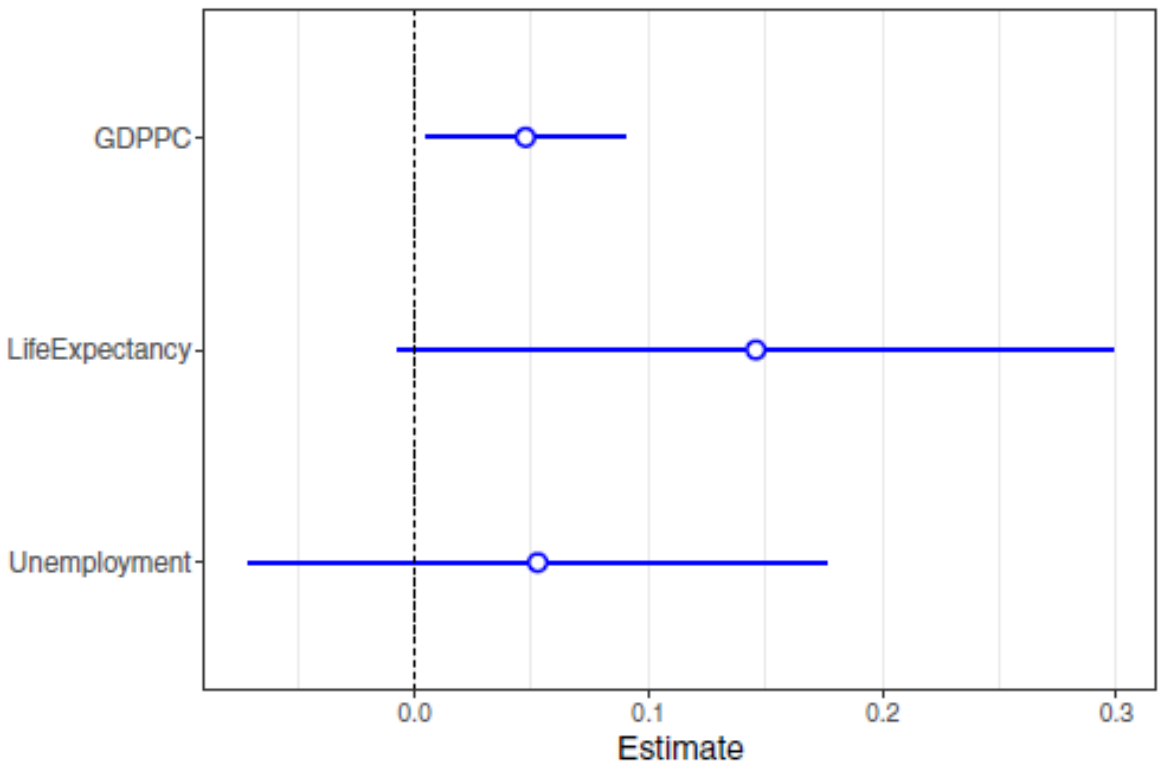


However, though ideally we would prefer not to see large outlier values, as they are mostly cancelling each other out (and there are so few of them) there is nothing substantially wrong with the QQNorm plot. We can thus be confident in our standard errors. That said, given that our estimates may be biased (as above), it would be unwise to make too-strong conclusions from this model. Instead, the onus would be on us as researchers to test new combinations of variables to try and eliminate this problem and come to more reliable conclusions.

When fitting multiple linear regressions, ao aid our interpretation of the results we can call upon another plot. As the 'line of best fit' estimated by the OLS is operating through a multi-dimensional space, we can't make use of a scatter plot. Instead, we can visualise the regression results through a summary coefficient plot. This requires the "jtools" package:

```
library(jtools)

plot_summs(ml_model, ci_level=0.9, color.class = "blue")
```



The plot demonstrates nicely the findings. We can see the statistically significant positive impact of GDP per capita, with a relatively tightly bounded confidence interval which does not include zero (using the ci_level of 90%). The reported effect of the other two variables cross zero (and so do not reject the nul hypothesis of no effect) and have very wide intervals.

When Assumptions Fail

Of course, most often in social science we will not end up working with data which conform nicely the necessary assumptions for linear modelling - as indeed arguably happened in this example. Though OLS models can handle a fair amount of non-normality (i.e. things don't have to be perfect for OLS estimates to be reasonably reliable), there are certain situations which require us to use different approaches.

For example, oftentimes we work with time series data. This violates the assumption of independent errors, as the values (and model residuals) of a time series are heavily correlated upon previous values. This requires us to use specialised time series analysis techniques in order for us to avoid spurious results.

As another example, sometimes our dependent variable will not be continuous, and instead come in the form of say dummies or integer counts or as a discrete Likert scale. An example of a dummy dependent variable would be whether a person voted at a given election or not - this can only take two values, yes (1) or no (0). An example of a count variable would be the number of goals scored in football matches by a given Football Club - the distribution of this variable can never pass below zero, is unlikely to take significantly higher values than around 5, and will be heavily clustered around an off-centred mean (toward 1). Measures of well-being and satisfaction are often recorded on a Likert-type scale ranging from 1 (completely unsatisfied) to 5 (completely satisfied), which we could not consider to be continuous (even if normally distributed). Working with these types of variables require fitting generalised linear models of whcih there are many variants.

Finally, we might find that observations in our data are not at all independent from one another, and that there is some form of 'grouping' or 'clustering' going on around our observations which will impact the extent to which their characteristics and thus model residuals can actually be independent of one another. For instance, if we are examining student performance in tests across multiple schools, we must consider that a given students' performance may depend on the class and then school in which they are taught. This means that students in the same class will have a lot more similarity (and dependency) on one another in terms of their test outcomes than students from different classes - and even more so from different schools. This 'hierarchical' structure of data requires us to make use of multilevel models.

There are many other types of data that we might come across, and it is important to always examine and re-examine dependent variables and residual plots in order to make sure we are fitting the right kind of models to test our hypotheses.