Module 9

Simple Linear Regression and Correlation

James Baglin

Last updated: 13 July, 2020

Overview

##Summary

Statistical investigations often aim to understand the relationship between variables in order to make accurate predictions. This module will cover the use of linear regression models for modelling relationships between two quantitative variables.

Learning Objectives

The learning objectives associated with this module are:

- Define the common terms of linear regression and correlation.
- Explain the concept behind the least squares method for simple linear regression.

- Interpret simple scatter plots visualising bivariate data.
- Use technology to fit a simple linear regression and perform hypothesis tests of the various model components.
- Use technology to test the various assumptions behind linear regression analysis and identify when assumptions are in doubt.
- Interpret the output of a simple linear regression analysis.
- Use technology to compute the Pearson correlation coefficient and perform hypothesis testing.
- Interpret the Pearson correlation coefficient and determine when a correlation can be considered statistically significant.

Correlation Game

The following game is a bit of fun, but also focuses on the key concepts of this module. It might make more sense to come back to this game after you have read through the module notes. If the game does not appear your web browser might be blocking it. You can choose to run the script or access the game directly here (http://guessthecorrelation.com/).

Linear Regression

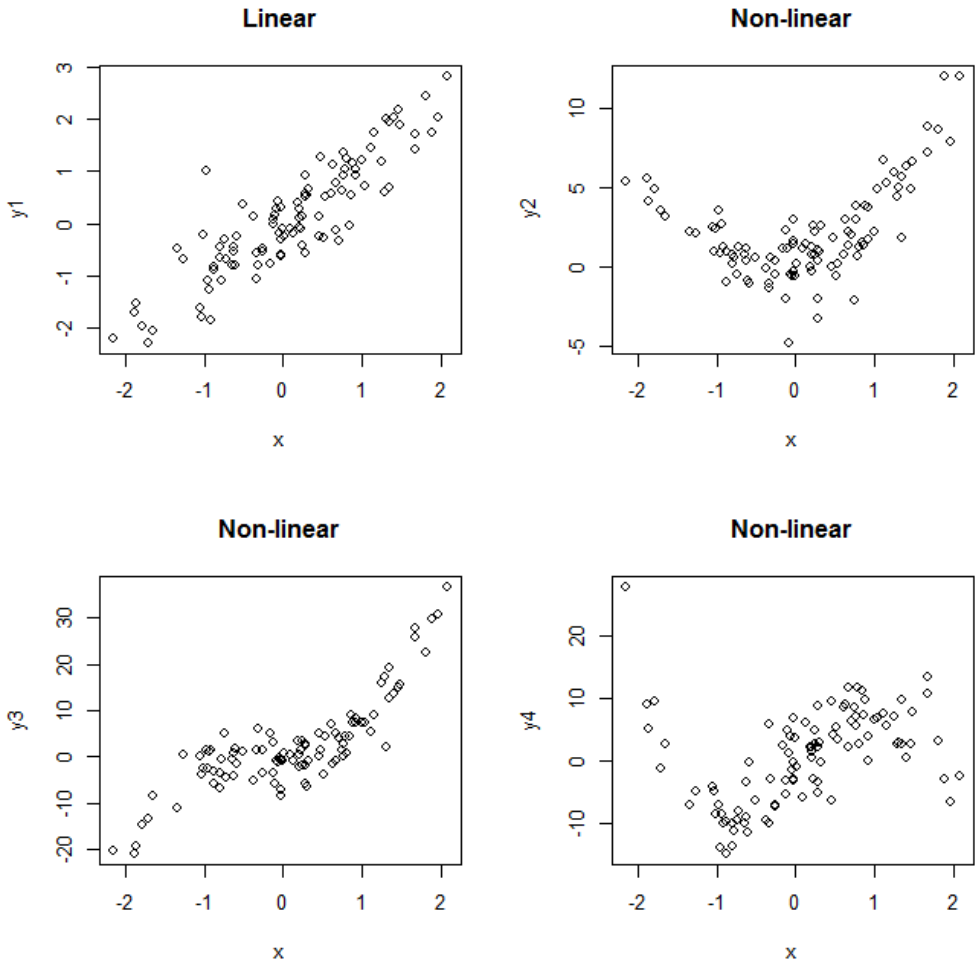
Correlation and simple linear regression are used to examine the relationship between two quantitative (discrete or continuous) variables. These methods assume that a **predictor variable**, x, provides information about some **dependent variable**, y. In practice, an investigation may have many predictors, x_i, but to introduce the basic concepts we will consider the simplest case of one predictor. We write a simple linear regression equation as:

$$y = \alpha + \beta x + \epsilon$$

where y is the dependent variable, α is the constant/intercept, β is the slope, x is the predictor and ϵ is the random error/residuals. Linear regression is a parametric method because ϵ are assumed to be normally distributed, $N(\mu, \sigma^2)$. Linear regression also

assumes that the variance of ϵ is constant and unchanging across the range of the predictor variable, x. For linear regression, we need to check these assumptions carefully, but we can only do so after the model has been fitted.

Linear regression also assumes that the relationship between the predictor and dependent variable is explained by a linear, or straight line, relationship. If the relationship is not linear, linear regression should not be used. The following scatter plots show examples of linear and non-linear bivariate relationships.



Fitting a linear regression line to sample data is done using a method known as **ordinary least squares (OLS)**. The idea behind this method is to minimise the **sum of squared distances**, S, for each (x_i, y_i) bivariate data point from a fitted regression line. The sum of squares is written as:

$$S = \sum_{i=1}^{n} d_i^2$$

where d_i refers to the (x_i, y_i) pairs' deviation from the regression line. The regression line that minimises this value is known as the **line of best fit**. Check out this excellent interactive visualisation site by Victor Powell and Lewis Lehe. The site is embedded in the following frame (if it does not appear, ensure you configure your browser to display the

page, otherwise you can access the site directly here (http://setosa.io/ev/ordinary-least-squares-regression/)). Work through the steps of the visualisations to get a better grasp of OLS.

OLS seeks to minimise the sum of squares, S, by determining a line of best fit.

Finding the line of best fit is a mathematical exercise. The formulas will be presented here for completeness, but in practice this is better left to R. To find the line of best fit, we calculate the **raw sum for x**:

$$\sum_{i=1}^{n} x_i$$

and the **raw sum for y**:

$$\sum_{i=1}^{n} y_i$$

Next, the **raw sum of squares for x**:

$$\sum_{i=1}^{n} x_i^2$$

Followed by the **raw sum of squares for y**:

$$\sum_{i=1}^{n} y_i^2$$

Then we calculate the **raw sum of the cross product**...

$$\sum_{i=1}^{n} x_i y_i$$

We correct the raw sum values by adjusting them to the **squared deviation from the mean**. First for x:

$$L_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

Then for y:

$$L_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

Finally we calculate the **corrected sum of the cross products**:

$$L_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

The **slope** of the line of best fit is computed as:

$$b = \frac{L_{xy}}{L_{xx}}$$

with an **intercept** equal to:

$$a = \bar{y} - b\bar{x}$$

As you can tell, this isn't a lot of fun by hand, especially when the sample size gets large. Fortunately, we have computers that can do all the hard work of finding the line of best fit in order to estimate the regression equation. This means we can use our precious time

to focus on understanding the concepts and testing hypotheses. We will look at the following example to help explain the meaning behind the regression concepts.

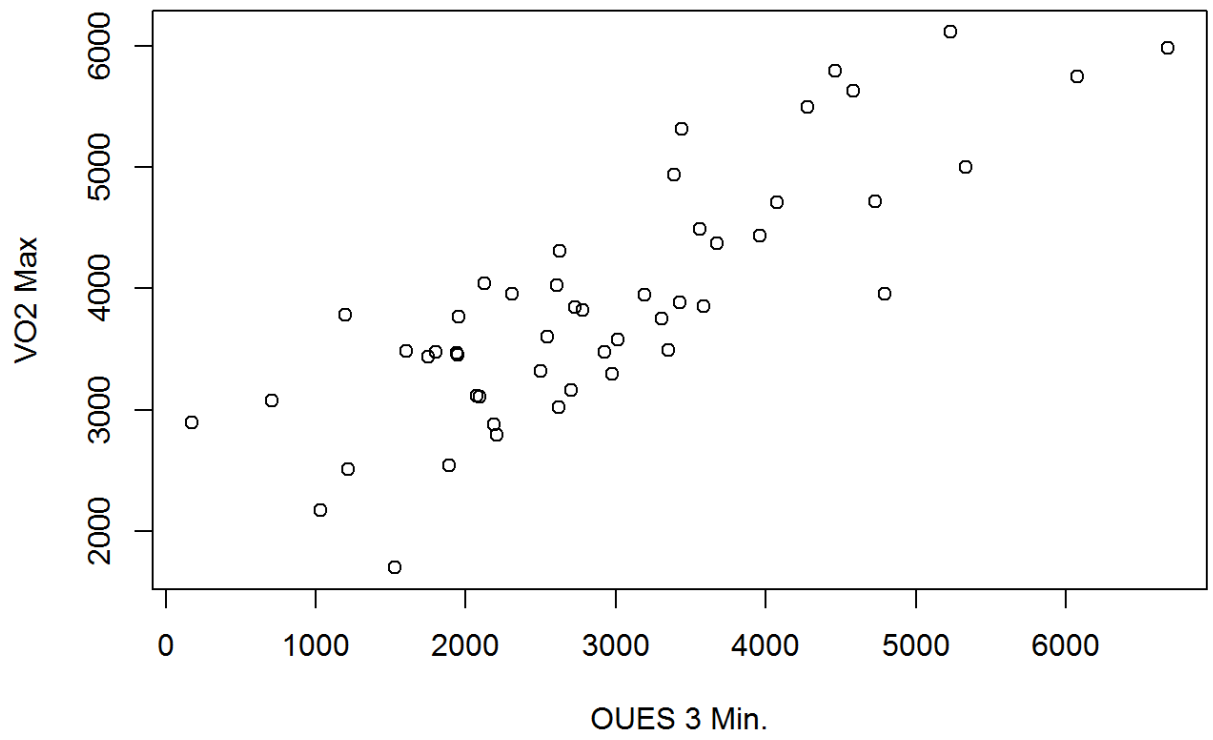
Example - Oxygen Uptake Efficiency Slope

Maximal oxygen consumption (VO_2 max) is a measure of aerobic fitness. However, measuring VO_2 max requires a subject to fully exert their aerobic system. This can be very impractical (e.g. time consuming) and dangerous in certain populations (e.g. the elderly or unwell). Researchers are interested to know if the oxygen uptake efficiency slope (OUES), an indicator of cardiopulmonary reserve, can be used as a sub-maximal predictor of VO_2 max. The advantage of OUES is that it can be measured much earlier in VO_2 max testing before the subject reaches aerobic exertion. This would make it a more practical and safe measure of aerobic fitness. The OUES.csv (data/OUES.csv) dataset contains the OUES measured at three minutes into a VO2 max test and final VO2 readings of 50 healthy adults.

```
OUES<- read.csv("data/OUES.csv")
```

What does this relationship look like? We can use a scatter plot to visualise it.

```
plot(VO2_Max ~ OUES_3, data = OUES, xlab = "OUES 3 Min.", ylab = "VO2 Ma
  x")
```



As you can see, as OUES increases, so too does VO_2 max. This is a positive relationship.

The following R code and table shows how to work through the formulas described in the previous section. However, this is better left to the built in functions in R that will be covered shortly.

First we square x_i, y_i and take the cross product $x_i y_i$.

```
OUES$y2 <- OUES$VO2_Max^2
OUES$x2 <- OUES$OUES_3^2
OUES$xy <- OUES$VO2_Max*OUES$OUES_3
```

You can see these computed values in the data frame:

Show 50 ▼ **entries** **Search:**

ID	VO2_Max	OUES_3	y2	x2	xy
1	4310.56	2626.44	18580927.5136	6898187.0736	11321427.2064
2	3157.98	2701.46	9972837.6804	7297886.1316	8531156.6508
3	3751.89	3308.54	14076678.5721	10946436.9316	12413278.1406
4	2791.33	2206.55	7791523.1689	4868862.9025	6159209.2115
5	4022.37	2608.16	16179460.4169	6802498.5856	10490984.5392
6	3107.35	2090.39	9655624.0225	4369730.3521	6495573.3665
7	2892.19	172.82	8364762.9961	29866.7524	499828.2758
8	4037.79	2122.35	16303748.0841	4504369.5225	8569603.6065
9	3854.39	3584.09	14856322.2721	12845701.1281	13814480.6551
10	3072.82	705.59	9442222.7524	497857.2481	2168151.0638
11	5498.97	4279.19	30238671.0609	18311467.0561	23531137.4343
12	3023.52	2621.46	9141673.1904	6872052.5316	7926036.7392
13	4718.86	4726.43	22267639.6996	22339140.5449	22303361.4698
14	3782.19	1194.22	14304961.1961	1426161.4084	4516766.9418
15	3958.14	4794.13	15666872.2596	22983682.4569	18975837.7182
16	3319.66	2500.85	11020142.5156	6254250.7225	8301971.711

17	5980.62	6678.83	35767815.5844	44606770.1689	39943544.2746
18	3475.77	1800.94	12080977.0929	3243384.8836	6259653.2238
19	2509.71	1214.04	6298644.2841	1473893.1216	3046888.3284
20	2881.11	2190.51	8300794.8321	4798334.0601	6311100.2661
21	3578.75	3015.03	12807451.5625	9090405.9009	10790038.6125
22	4432.84	3959.55	19650070.4656	15678036.2025	17552051.622
23	3601.55	2546.3	12971162.4025	6483643.69	9170626.765
24	3946.94	3192.09	15578335.3636	10189438.5681	12598987.7046
25	3448.56	1947.26	11892566.0736	3791821.5076	6715242.9456
26	4369.16	3677.25	19089559.1056	13522167.5625	16066493.61
27	6112.58	5230.23	37363634.2564	27355305.8529	31970199.2934
28	5749.87	6077.92	33061005.0169	36941111.5264	34947249.8704
29	3957.07	2307.5	15658402.9849	5324556.25	9130939.025
30	4705.48	4074.82	22141542.0304	16604158.0324	19173984.0136
31	5315.19	3437.67	28251244.7361	11817575.0289	18271869.2073
32	5632.32	4584.79	31723028.5824	21020299.3441	25823004.4128
33	2167.45	1030.9	4697839.5025	1062754.81	2234424.205
34	1698.4	1524.37	2884562.56	2323703.8969	2588990.008
35	3817.64	2777.68	14574375.1696	7715506.1824	10604182.2752
36	3114.81	2073.59	9702041.3361	4299775.4881	6458838.8679
37	3484.14	1605.88	12139231.5396	2578850.5744	5595110.7432
38	3881.28	3426.48	15064334.4384	11740765.1904	13299128.2944
39	2542.49	1888.01	6464255.4001	3564581.7601	4800246.5449
40	3477.67	2921.48	12094188.6289	8535045.3904	10159943.3516
41	3294.2	2974.7	10851753.64	8848840.09	9799256.74

42	5791.01	4460.16	33535796.8201	19893027.2256	25828831.1616
43	3495.11	3352.15	12215793.9121	11236909.6225	11716132.9865
44	4493	3557.7	20187049	12657229.29	15984746.1
45	5001.9	5333.79	25019003.61	28449315.7641	26679084.201
46	3464.99	1936.67	12006155.7001	3750690.6889	6710542.1833
47	4938.21	3389.12	24385918.0041	11486134.3744	16736186.2752
48	3844.41	2727.18	14779488.2481	7437510.7524	10484398.0638
49	3433.36	1746.94	11787960.8896	3051799.3636	5997873.9184
50	3767.64	1949.32	14195111.1696	3799848.4624	7344336.0048

Showing 1 to 50 of 50 entries Previous 1 Next

Now let's see how we can compute each formula in the linear regression using R:

$$\sum_{i=1}^{n} x_i$$

```
sum_x <- sum(OUES$OUES_3)
sum_x
```

```
## [1] 146853.5
```

$$\sum_{i=1}^{n} y_i$$

```
sum_y <- sum(OUES$VO2_Max)
sum_y
```

```
## [1] 194705.2
```

$$\sum_{i=1}^{n} x_i^2$$

```
sum_x_sq <- sum(OUES$OUES_3^2)
sum_x_sq
```

```
## [1] 521621342
```

$$\sum_{i=1}^{n} y_i^2$$

```
sum_y_sq <- sum(OUES$VO2_Max^2)
sum_y_sq
```

```
## [1] 807085161
```

$$\sum_{i=1}^{n} x_i y_i$$

```
sum_xy <- sum(OUES$OUES_3*OUES$VO2_Max)
sum_xy
```

```
## [1] 626812930
```

$$L_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

```
n <- length(OUES$VO2_Max) #Sample size
n
```

```
## [1] 50
```

```
Lxx <- sum_x_sq-((sum_x^2)/n)
Lxx
```

```
## [1] 90302215
```

$$L_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

```
Lyy <- sum_y_sq-((sum_y^2)/n)
Lyy
```

```
## [1] 48882552
```

$$L_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

```
Lxy = sum_xy - (((sum_x)*(sum_y))/n)
Lxy
```

```
## [1] 54949933
```

$$b = \frac{L_{xy}}{L_{xx}}$$

```
b = Lxy/Lxx
b
```

```
## [1] 0.6085115
```

$$a = \bar{y} - b\bar{x}$$

```
a = mean(OUES$VO2_Max - b*mean(OUES$OUES_3))
a
```

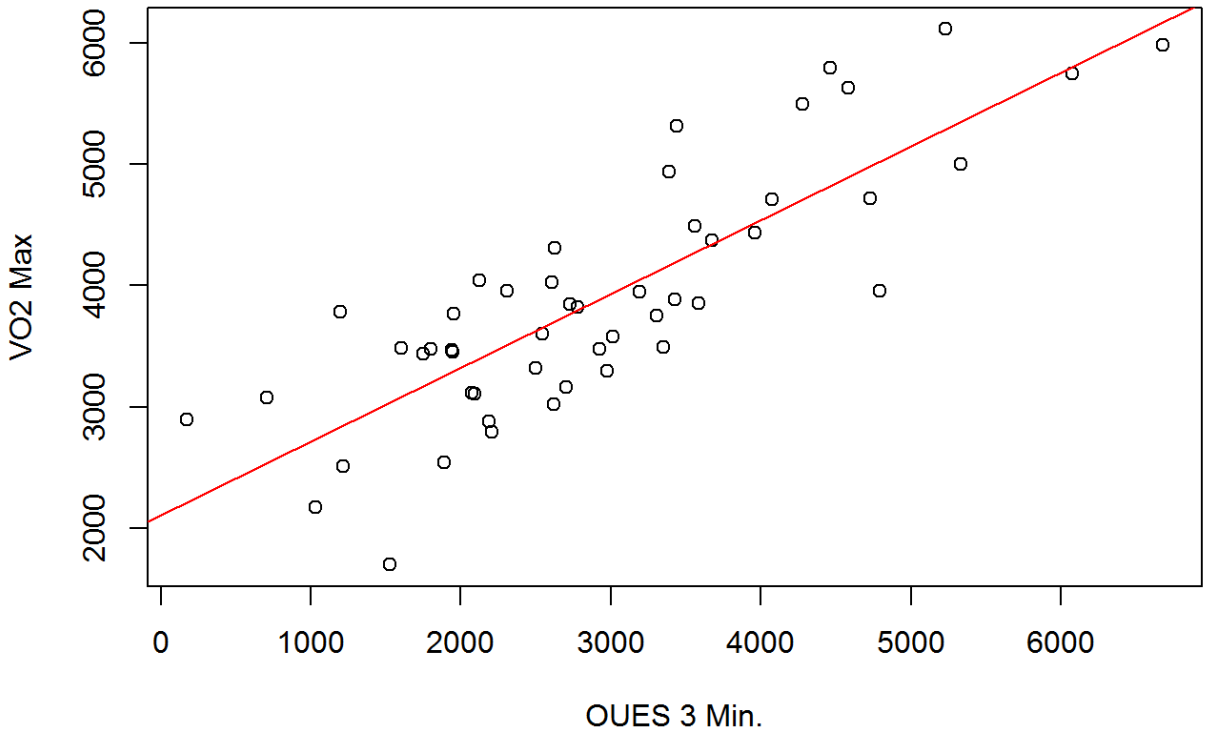
```
## [1] 2106.864
```

Now we have our linear regression model.

$$\mathrm{VO2\_Max} = 2107 + 0.609 * \mathrm{OUES\_3}$$

We can now add the line of best fit to the scatter plot.

```
plot(VO2_Max ~ OUES_3, data = OUES, xlab = "OUES 3 Min.", ylab = "VO2 Ma
  x")
abline(a = a, b = b, col= "red")
```



In the following sections, we will work through performing a simple linear regression using R functions. Along the way, we will look at checking assumptions, interpreting important output and testing the statistical hypotheses of a linear regression model. The first step for all linear regression analyses is to plot the relationship between your x and y variables, OUES and VO_2 max, to determine if linear regression is suitable. We have already done this in the previous scatter plots. The data exhibited a **positive** linear trend. A positive linear relationship occurs when as the predictor variable increases in value, so too do the values for the dependent variable. In this situation, higher OUES values are associated with higher VO_2 max. This makes scientific sense. However, if VO_2 max had decreased with increasing values for OUES, the relationship would be negative.

As the data exhibit signs of a positive linear relationship, we can proceed with fitting the linear regression model using R. We will work through the important code and output. Let's run the regression using the `lm()` function.

```
ouesvo2maxmodel <- lm(VO2_Max ~ OUES_3, data = OUES)
ouesvo2maxmodel %>% summary()
```

```
##
## Call:
## lm(formula = VO2_Max ~ OUES_3, data = OUES)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1336.06  -366.89   -55.12   434.57  1116.46
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.107e+03  1.928e+02   10.93 1.28e-14 ***
## OUES_3      6.085e-01  5.969e-02   10.19 1.35e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 567.2 on 48 degrees of freedom
## Multiple R-squared:  0.684,  Adjusted R-squared:  0.6775
## F-statistic: 103.9 on 1 and 48 DF,  p-value: 1.345e-13
```

The first part of the code assigns the linear regression model `lm()` to an object named `ouesvo2maxmodel`. This will allow you to call this object to summarise the model and later test assumptions. The `summary()` function prints out a summary of the important output required to interpret the linear regression model.

The model summary reports the R^2 statistic. To calculate R^2 look back to the linear regression worksheet and calculate the following formula.

$$R^2 = \frac{bL_{xy}}{L_{yy}}$$

Using the previous objects we computed in R for the regression formulae:

```
R2 <- (b*Lxy)/Lyy
R2
```

```
## [1] 0.6840409
```

The R^2 value can range from 0 - 1. R^2 **reflects the proportion of variability in the dependent variable that can be explained by a linear relationship with the predictor variable**. Therefore, OUES measured at 3 minutes, explained 68.4% of the variability in final VO_2 max readings. The R^2 is a measure of goodness of fit for linear regression. The better the line fits the data (i.e. the closer the data points sit on the line) the higher R^2 will be. If there is no linear relationship between the predictor and dependent variable, R^2 = 0 or close to it. You will also notice an Adjusted R^2 value. R^2

tends to overestimate the population R^2 . The adjusted R^2 takes this overestimation into account and down-scales it. Which do you report? It does not really matter, just as long as you're clear on which one you use.

The model summary also reports an F statistic which is used to test the overall regression model. The F-test for the linear regression has the following statistical hypotheses:

$$H_0 : \text{The data do not fit the linear regression model}$$

$$H_A : \text{The data fit the linear regression model}$$

This test is more useful when you deal with multiple predictors, but we will explain it here so you know what it means. Assuming the data do not fit a linear model in the population, the F statistic reported in the summary as $F = 103.9$, will have a F distribution with $df_1 = 1$ and $df_2 = n - 2 = 50 - 2 = 48$. The F distribution is positively skewed, so to calculate the p-value of the the observed F statistic, we need to find $Pr(F_{1,n-2} > F)$. This is easy using the `pf()` function in R.

```
pf(q = 103.9,1,48,lower.tail = FALSE)
```

```
## [1] 1.348976e-13
```

We confirm the p-value reported in the summary to be $p < .001$. What does this mean? As p is less than the 0.05 level of significance, we reject H_0. There was statistically significant evidence that the data fit a linear regression model. How was the actual F statistic calculated? We can compute it directly from the R2 statistic:

$$F = \frac{R^2}{1 - R^2} \times \frac{df_2}{df_1}$$

```
(R2/(1-R2)*(48/1))
```

```
## [1] 103.9184
```

We can also use the formulae objects in the previous section.

$$F = \frac{bL_{xy}/df_1}{(L_{yy} - bL_{xy})/df_2}$$

where $bL_{xy} = RegSS$, the regression sum of squares, and $L_{yy} - bL_{xy} = ResSS$, is the residual sum of squares. Therefore,

$$F = \frac{RegSS/df_1}{ResSS/df_2}$$

$RegSS/df_1 = RegMS$, or the regression mean sum of squares, and $ResSS/df_2 = ResMS$, residual mean sum of squares. As such, F can also be written as:

$$F = \frac{RegMS}{ResMS}$$

Or, we can quickly get the RegSS and ResSS values using the `anova()` function in R.

```
ouesvo2maxmodel %>% anova()
```

```
## Analysis of Variance Table
##
## Response: VO2_Max
##           Df   Sum Sq  Mean Sq F value    Pr(>F)
## OUES_3     1 33437664 33437664  103.92 1.345e-13 ***
## Residuals 48 15444888   321768
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we can solve for F:

$$F = \frac{RegSS/df_1}{ResSS/df_2} = \frac{33437664/1}{15444888/48} = \frac{33437664}{321768} = 103.9$$

The next part of the model summary is the all important one. The coefficients table reports the sample estimates for the intercept/constant, a, and slope, b. Let's begin with our interpretation of the intercept or constant.

```
ouesvo2maxmodel %>% summary() %>% coef()
```

```
##                Estimate   Std. Error  t value     Pr(>|t|)
## (Intercept) 2106.8638073 192.80364251 10.92751 1.282351e-14
## OUES_3         0.6085115   0.05969288 10.19404 1.345027e-13
```

The intercept/constant is reported as $a = 2106.864$. **The constant, or intercept, is the average value for y when $x = 0$.** In this example, this value represents the average VO_2 max score when OUES is equal to 0. Given that an OUES of 0 is impossible (assuming you're alive) the constant typically has no meaningful interpretation. To test the statistical significance of the constant, we set the following statistical hypotheses:

$$H_0 : \alpha = 0$$

$$H_A : \alpha \neq 0$$

This hypothesis is tested using a t statistic, reported as $t = 10.928, p < .001$. The constant is statistically significant at the 0.05 level. This means that there is statistically significant evidence that the constant is not 0. The t statistic for the constant was calculated as:

$$t = \frac{a}{\sqrt{s^2_{y.x}(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}})}} = \frac{2107}{\sqrt{321768(\frac{1}{50} + \frac{2937^2}{90302215})}} = \frac{2107}{193} = 10.92$$

We confirm that $p < .001$. R can also report a 95% CI for a. This is calculated as:

$$\left[a - t_{n-2,1-\alpha/2}\sqrt{s^2_{y.x}(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}})}, a + t_{n-2,1-\alpha/2}\sqrt{s^2_{y.x}(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}})}\right]$$

In R, we use the `confint()` function:

```
ouesvo2maxmodel %>% confint()
```

```
##                   2.5 %      97.5 %
## (Intercept) 1719.2061023 2494.521512
## OUES_3         0.4884909    0.728532
```

R reports the 95% CI for a to be [1719.206, 2494.522]. $H_0 : \alpha = 0$ is clearly not captured by this interval, so was rejected.

The slope of the regression line was reported as $b = 0.609$. **The slope represents the average increase in y following a one unit increase in** x. In relation to the example, a one unit increase in OUES was related to an average increase in VO_2 max of .609 units. This is a positive change. Had the slope been a negative number, VO_2 max would decrease. The hypothesis test of the slope, b, was as follows:

$$H_0 : \beta = 0$$

$$H_A : \beta \neq 0$$

The slope was also tested using a t statistic which was reported as $t = 10.194, p < .001$. t was calculated as:

$$t = \frac{b}{\sqrt{\frac{s^2_{x.y}}{L_{xx}}}} = \frac{0.609}{\sqrt{\frac{321768}{90302215}}} = \frac{.609}{0.0597} = 10.2$$

Note there is a slight rounding difference. To calculate the two-tailed p-value for the slope in this example, we compute:

```
2*pt(q = 10.20,df = 50 - 2, lower.tail=FALSE)
```

```
## [1] 1.319213e-13
```

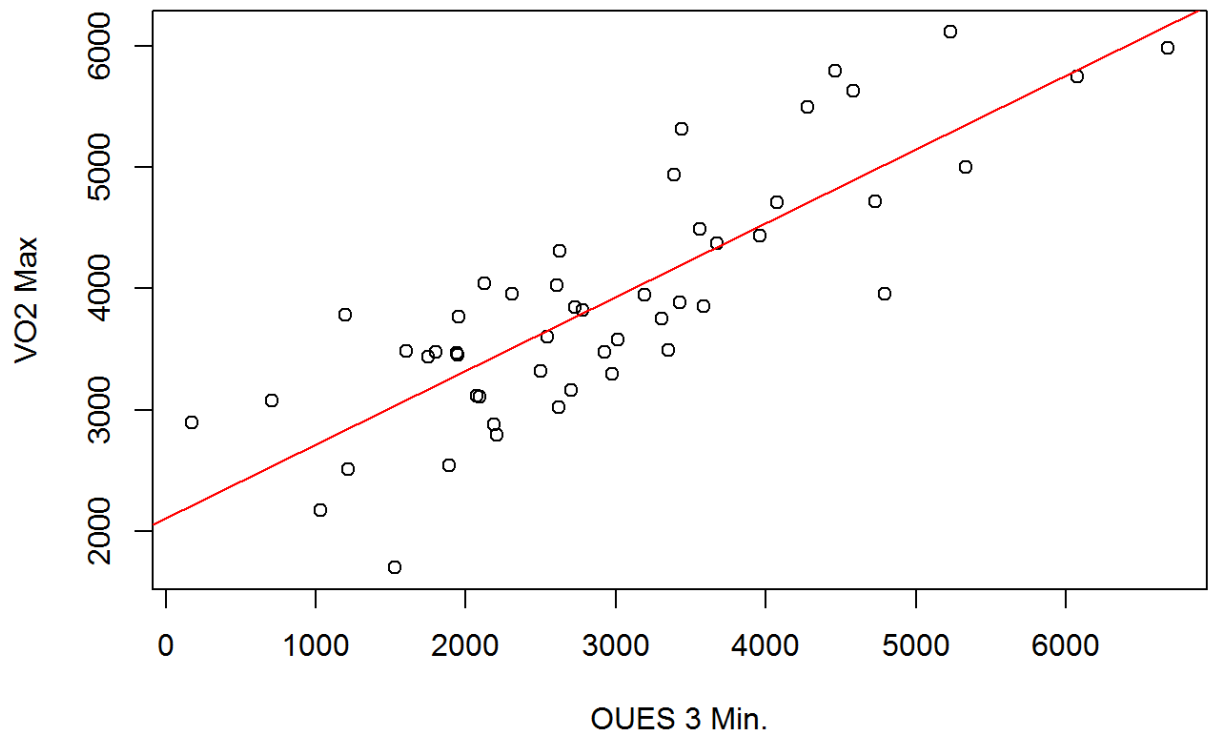
and confirm that $p < .001$. As $p < .05$, we reject H_0. There was statistically significant evidence that OUES was positively related to VO_2 max.

The 95% CI for the slope can be calculated as:

$$\left[b - t_{n-2,1-\alpha/2}\sqrt{\frac{s^2_{x.y}}{L_{xx}}}, b + t_{n-2,1-\alpha/2}\sqrt{\frac{s^2_{x.y}}{L_{xx}}}\right]$$

Looking back to the `confint()` function, R reports the 95% CI for b to be [.488, .729]. This 95% CI does not capture H_0, therefore it was rejected. There was a statistically significant positive relationship between OUES measurements taken at 3 minutes and final VO_2 max. Finally, a nice plot to summarise the linear relationship:

```
plot(VO2_Max ~ OUES_3, data = OUES, xlab = "OUES 3 Min.", ylab = "VO2 Ma
  x")
abline(ouesvo2maxmodel, col = "red")
```



Note how we can quickly plot the line of best fit by calling the `ouesvo2maxmodel` into the `abline()` function.

Assumptions

Before we report the final regression model, we must validate all the following assumptions for linear regression.

- **Independence**
- **Linearity**
- **Normality of residuals**
- **Homoscedasticity**

Independence is checked through the research design. You must ensure that all measurements between participants or observations are independent, for example, you have not included multiple measurements from the same people or knowing the measurements of one person do not share a relationship with other peoples' measurements.

We have already checked and confirmed linearity earlier in the notes. The idea is to rule out non-linear relationships.

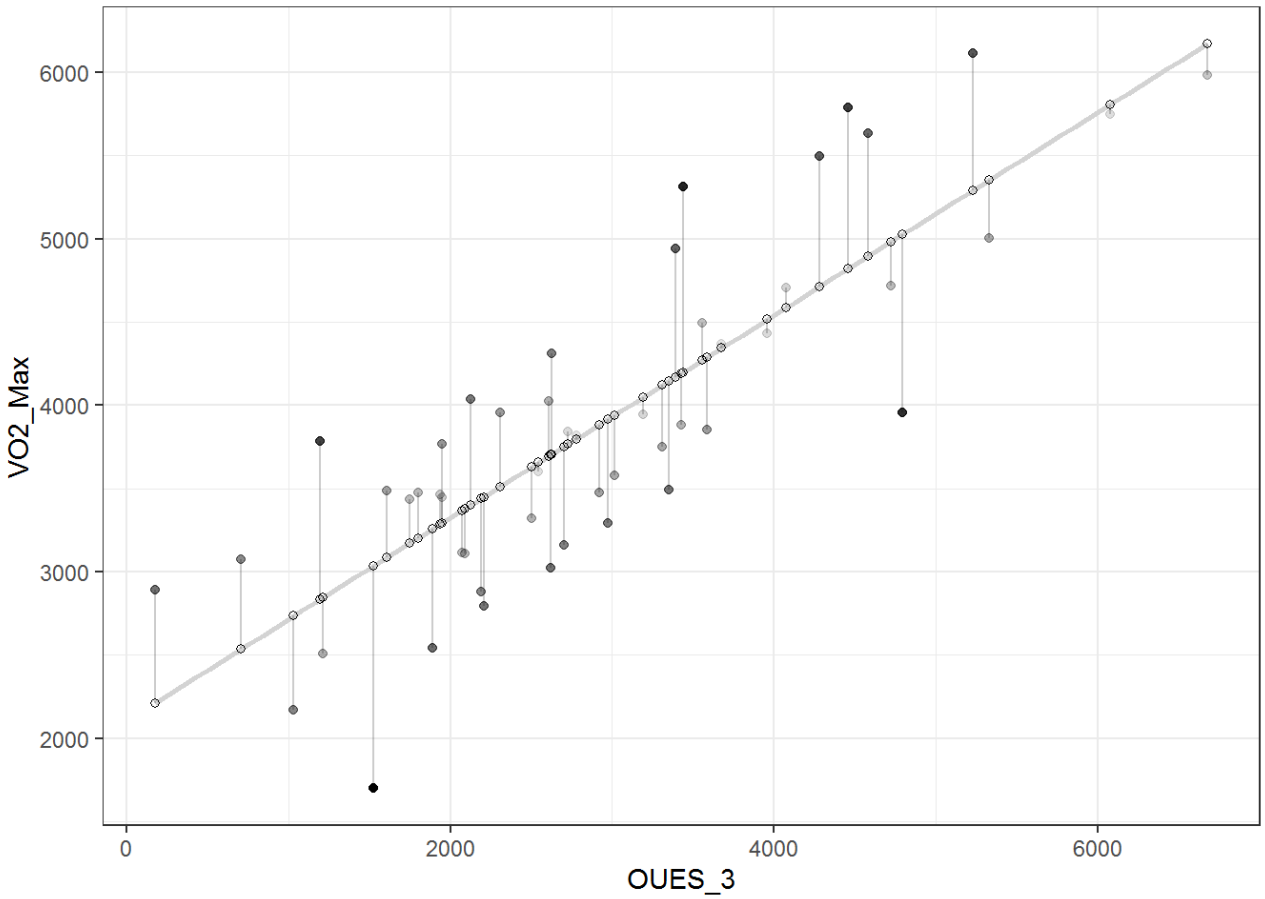
The assumption of normality for linear regression relates to the distribution of the errors or residuals for the model. The residuals are calculated as:

$$y_i - \hat{y}_i$$

where y_i was an observed score in the sample and $\hat{y}_i$ was the predicted score based on the fitted regression model. For example, the predicted score for OUES = 4000 was:

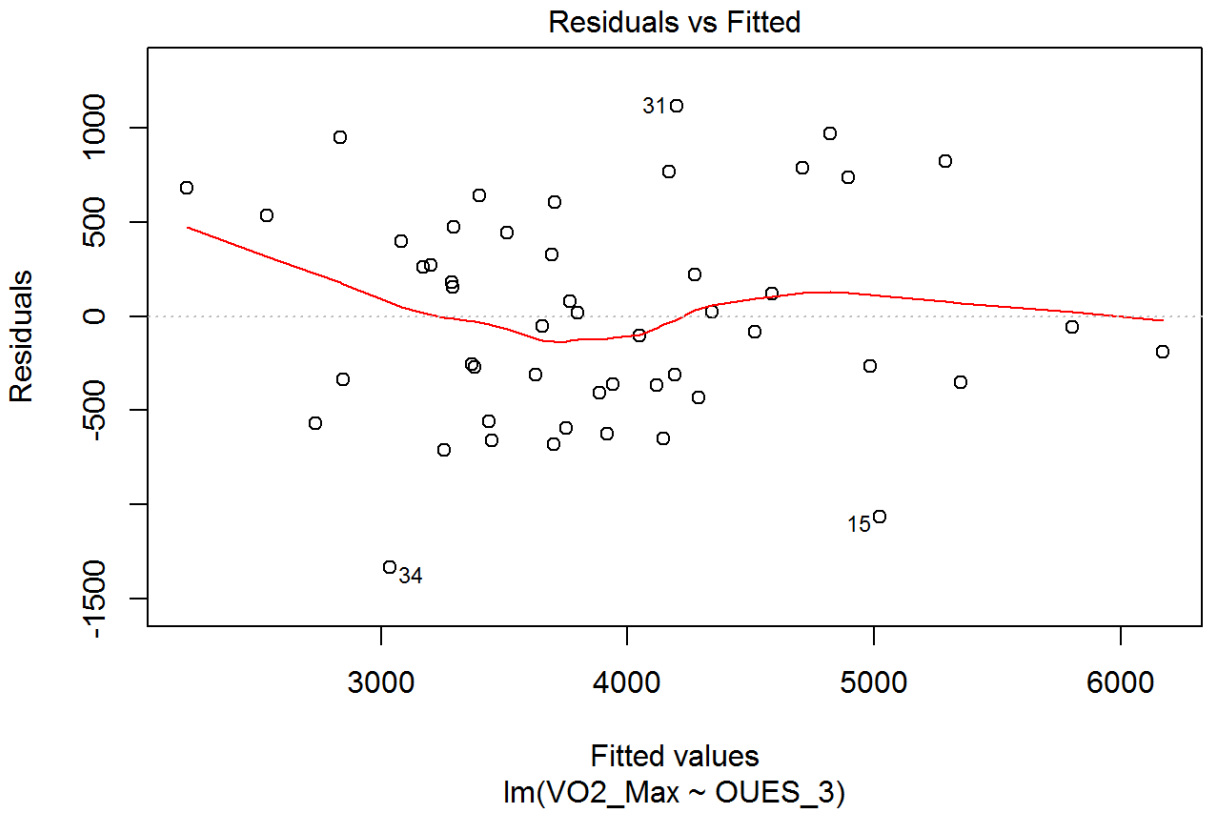
$$\hat{y} = a + bx_i = 2107 + 0.609(4000) = 4543$$

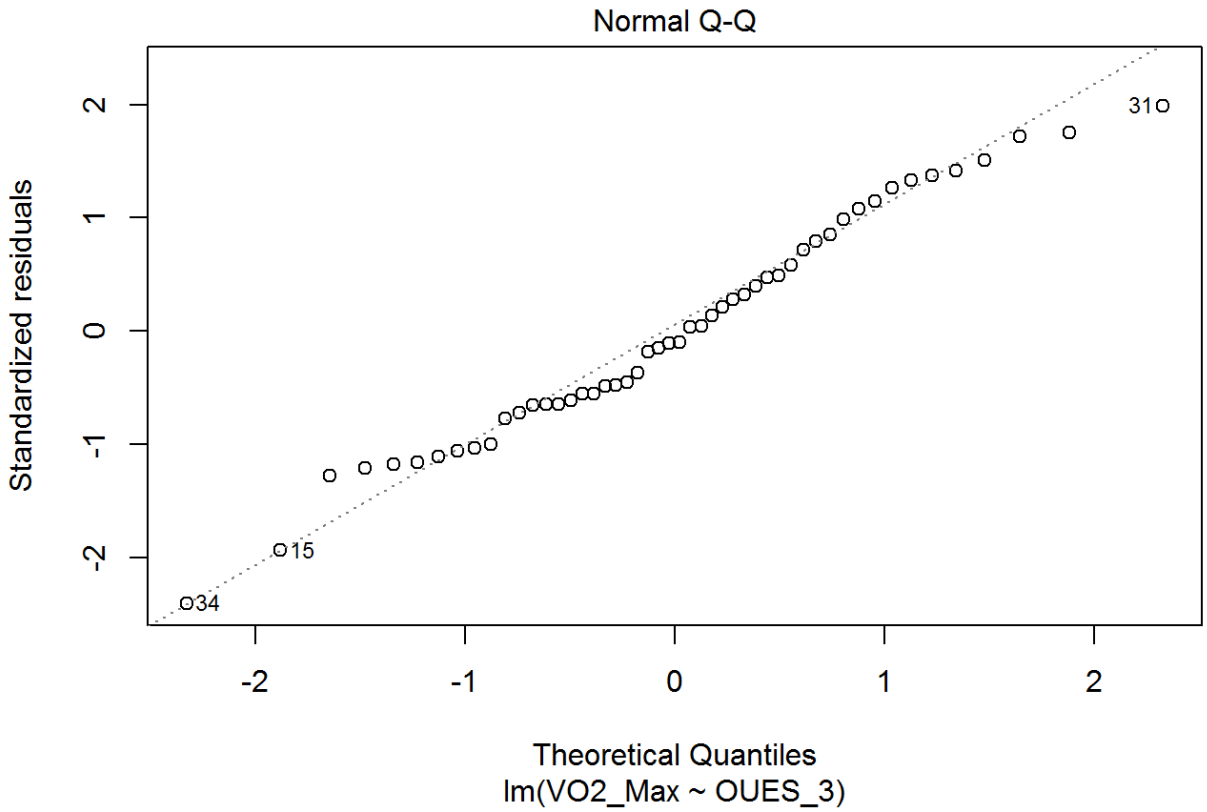
So, for each observed score in the sample, the predicted score was calculated and the residuals recorded. The residuals reflect how far an observed score y_i deviates from the $\hat{y}_i$ score predicted by the line of best fit. The following plot (credit to Simon at Blogr (https://drsimonj.svbtle.com/visualising-residuals) for this visualisation code). Each data point has a vertical line that connects to the line of best fit. The length of each point's line is a residual.

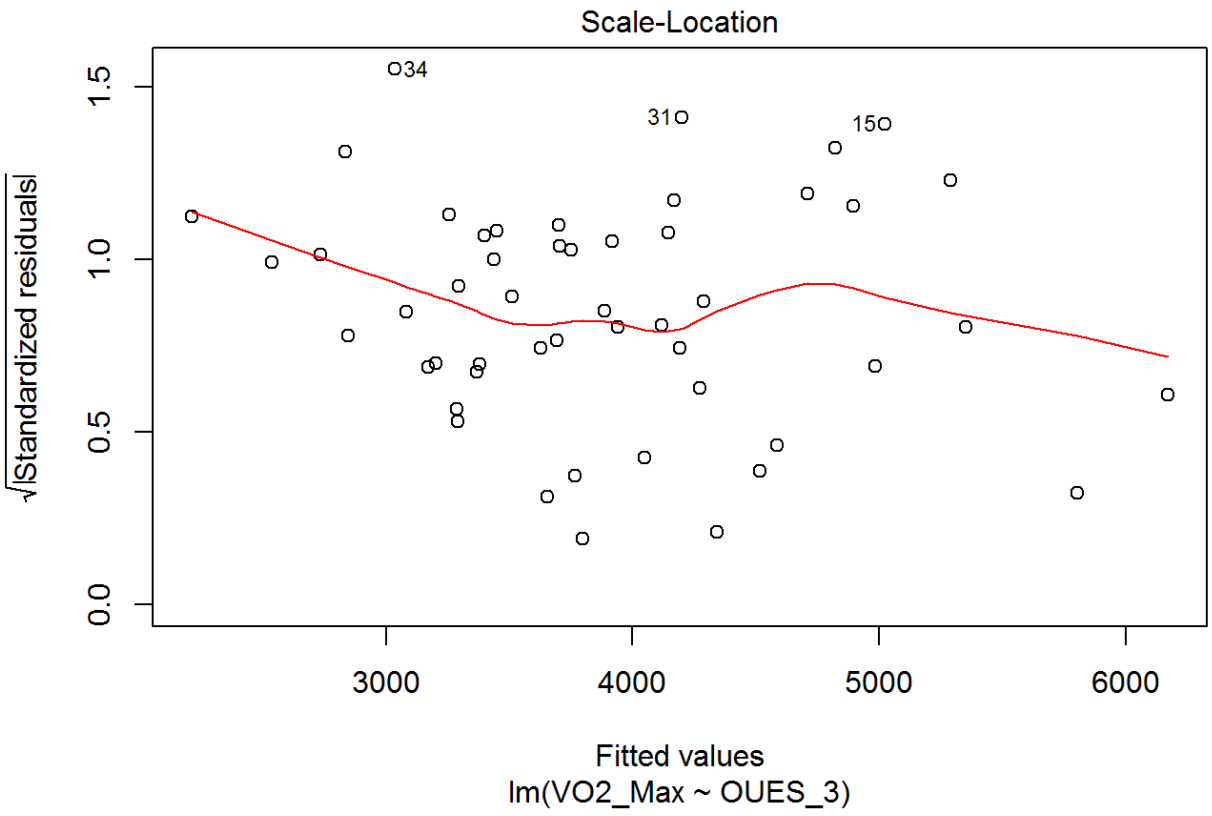


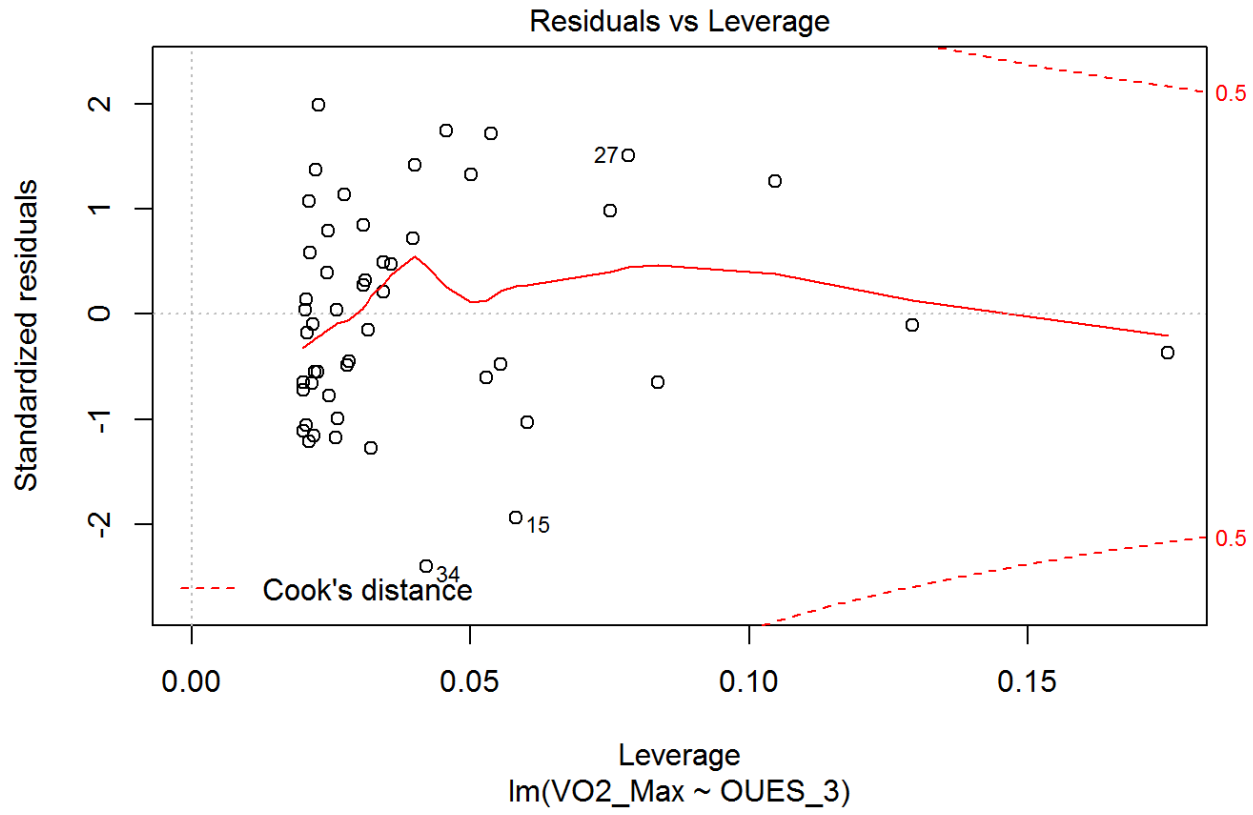
We can use the `plot()` function to quickly obtain a series of plots that can be used to check the diagnostics of a fitted regression model.

```
plot(ouesvo2maxmodel)
```


Residuals vs Fitted
Residuals
1000
500
0
-500
-1500
31
15
34
3000
4000
5000
6000
Fitted values
lm(VO2_Max ~ OUES_3)


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
31
15
34
-2
-1
0
1
2
Theoretical Quantiles
lm(VO2_Max ~ OUES_3)





Residual vs. Fitted

Check this plot for non-linear trends. If the relationship between fitted values and residuals is flat (look at the red line), this is a good indication that you are modelling a linear relationship. Of course, you can always look at the scatter plot between x and y too. This plot is better for multiple regression (more than one predictor).

We can also check the assumption of **homoscedasticity**, or constant variance. **Homoscedasticity** is related to the assumption of homogeneity of variance for the two-sample t-test. As we move across predicted or fitted values, the variance in the residuals should remain constant. In the plot above, the variance appears to remain the same. The red line in the plot is a non-parametric locally weighted scatter plot smoother (LOESS). The line fits to the data. The straighter the line, the safer the assumption of homoscedasticity.

If the variance changed across predicted values, we would call the data **heteroscedastic**. Ordinary least squares linear regression is not appropriate for heteroscedastic data. You might be wondering what heteroscedasticity looks like. The following figure taken from Osborne and Waters (2002) (http://pareonline.net/getvn.asp?n=2&v=8) provides a guide. As you can see, the OUES regression model was pretty safe.

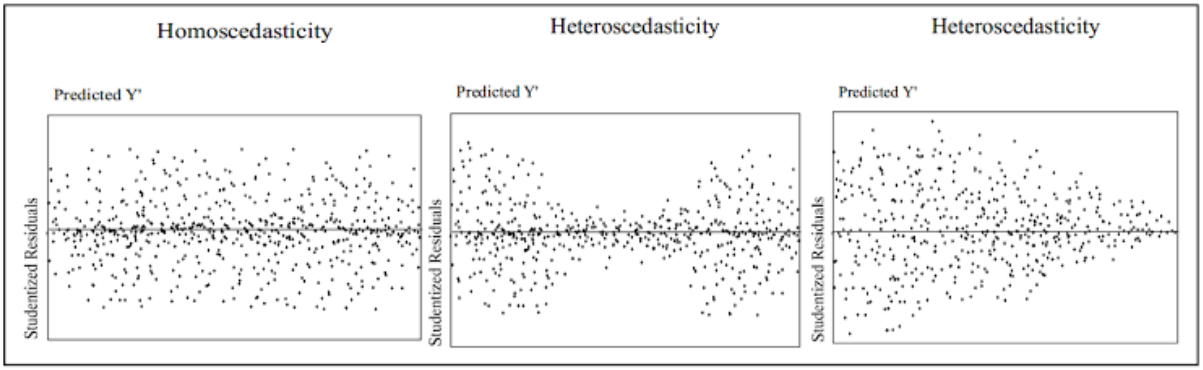


Figure 3. Examples of homoscedasticity and heteroscedasticity

Normal Q-Q

We check the normal Q-Q plot to determine if there were any gross deviations from normality (e.g obvious S shapes or non-linear trends). The plot above suggests there are no major deviations from normality. It would be safe to assume the residuals are approximately normally distributed.

Scale-Location

This is another plot used to check homoscedasticity. The red line should be close to flat and the variance in the square root of the standardised residuals should be consistent across predicted (fitted values).

Residual vs Leverage

This plot is used to identify cases that might be unduly influencing the fit of the regression model, for example, outliers. However, keep in mind that not all outliers are influential. This means that if we remove them, the regression model won't change much. However, outliers that do change the regression model substantially after removal are considered influential.

What we need to look for are values that fall in the upper and lower right hand side of the plot beyond the red bands. These bands are based on Cook's distances (https://www.ime.usp.br/~abe/lista/pdfWiH1zqnMHo.pdf). In the diagnostic plot above, there are no values that fall outside the bands, and therefore, no evidence of influential cases.

Cases that fall outside Cook's distances are labelled using their case ID. This make it easier to investigate the value in your dataset. Influential points should be considered for removal from the model.

Example Write-up

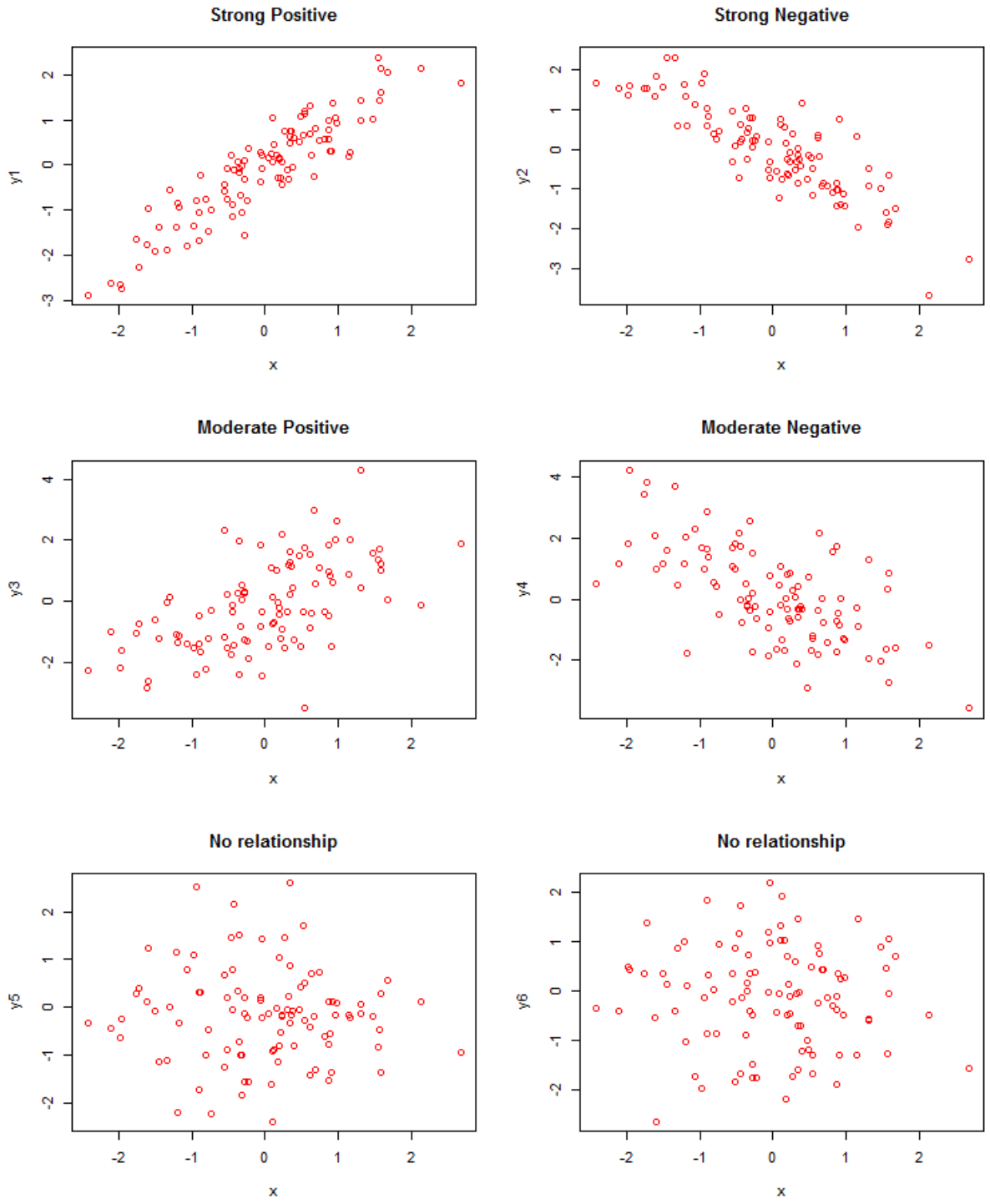
A linear regression model was fitted to predict the dependent variable, VO_2 max, using measures of OUES taken at 3 minutes as a single predictor. Prior to fitting the regression, a scatter plot assessing the bivariate relationship between VO_2 max and OUES 3 minutes was inspected. The scatter plot demonstrated evidence of a positive linear relationship. Other non-linear trends were ruled out. The overall regression model was statistically significant, $F(1, 48) = 103.92$, $p < .001$, and explained 68.4% of the variability in $V0_2$ max, $R^2 = .684$. The estimated regression equation was $VO_2 = 2106.84 + .609 * OUES$. The positive slope for OUES 3 minutes was statistically significant, $b = 0.609$, $t(48) = 10.194$, $p < .001$, 95% CI [0.488, 0.729]. Final inspection of the residuals supported normality and homoscedasticity.

Correlation

The Pearson correlation coefficient, r, is a standardised measure of the strength of the linear relationship between two variables. It can be calculated as follows:

$$r = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$$

A Pearson correlation can range from a perfect negative correlation, $r = -1$, to zero correlation, $r = 0$, and all the way through to a perfect positive correlation, $r = 1$. r and the slope, b, of a simple linear regression will have the same sign. The following plots provide examples of different strengths and types of correlations.



We can calculate a quick correlation in R using the `cor()` function:

```
cor(OUES$VO2_Max,OUES$OUES_3)
```

```
## [1] 0.8270676
```

But this does not give us a p-value or CI to test the null hypothesis.

We can perform a full correlation analysis between OUES and VO_2 in R by installing the `Hmisc` library and using the `rcorr()` function.

```
library(Hmisc)
bivariate<-as.matrix(dplyr::select(OUES, VO2_Max,OUES_3)) #Create a matri
  x of the variables to be correlated
rcorr(bivariate, type = "pearson")
```

```
##         VO2_Max OUES_3
## VO2_Max    1.00   0.83
## OUES_3     0.83   1.00
##
## n= 50
##
##
## P
##         VO2_Max OUES_3
## VO2_Max          0
## OUES_3   0
```

R reports the correlation between OUES and VO_2 max to be $r = .83$ and the p-value = 0, which we write as $p < .001$. A hypothesis test for r has the following statistical hypotheses:

$$H_0 : r = 0$$

$$H_A : r \neq 0$$

The p-value for r can be readily calculated by converting r to a t statistic:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = .827\sqrt{\frac{50-2}{1-.827^2}} = 10.191$$

Can you find this value in the output for the regression? That's correct, the t statistic was the same as the t statistic for the slope of the simple regression. Therefore, a two-tailed p-value for r can be calculated using the R formula:

```
2*pt(q = 10.191,df = 50 - 2,lower.tail=FALSE)
```

```
## [1] 1.358369e-13
```

where $df = n - 2 = 50 - 2 = 48$. We find $p < .001$. Using the 0.05 level of significance, we must reject H_0. There was a statistically significant positive correlation between OUES 3 minutes and VO2 max, $r = .827, p < .001$.

We can also test H_0 using a confidence interval approach. We can use a normal approximation to calculate a confidence interval for r. First, r is converted to a z-score using the z-transformation:

$$r = z = \frac{1}{2} ln \left(\frac{1+r}{1-r} \right) = \frac{1}{2} ln \left(\frac{1+.827}{1-.827} \right) = 1.179$$

This is easier in R...

```
0.5*(log((1+.827)/(1-.827)))
```

```
## [1] 1.178569
```

This allows a 95% CI for r to be approximated using a normal distribution, z, as $[z_1, z_2]$:

$$z_1 = z - z_{1-\alpha/2}/\sqrt{n-3} = 1.179 - 1.96/\sqrt{50-3} = 0.893$$

$$z_1 = z + z_{1-\alpha/2}/\sqrt{n-3} = 1.179 + 1.96/\sqrt{50-3} = 1.464$$

To transform z_1 and z_2 back to a 95% CI for r, $[\rho_1, \rho_2]$:

$$\rho_1 = \frac{e^{2_{z1}} - 1}{e^{2_{z1}} + 1} = \frac{e^{2*0.893} - 1}{e^{2*0.893} + 1} = 0.713$$

$$\rho_2 = \frac{e^{2_{z2}} - 1}{e^{2_{z2}} + 1} = \frac{e^{2*1.464} - 1}{e^{2*1.464} + 1} = 0.899$$

Therefore, $r = 0.827$, 95% CI [0.713, .899]. Alternatively, we can take away the hard work by using the `CIr()` function from the `psychometric()` package.

```
library(psychometric)
r=cor(OUES$VO2_Max,OUES$OUES_3)
CIr(r = r, n = 50, level = .95)
```

```
## [1] 0.7128199 0.8985565
```

This confidence interval does not capture H_0, therefore, H_0 was rejected. There was a statistically significant positive correlation between OUES 3 minutes and VO_2 max.

Example Write-up

A Pearson's correlation was calculated to measure the strength of the linear relationship between OUES 3 minutes and VO_2 max. The positive correlation was statistically significant, $r = .827$, $p < .001$, 95% CI [0.713, .899].

References