- **Module 9 - Simple Linear Regression and Correlation**

Overview

- 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.

Linear Regression

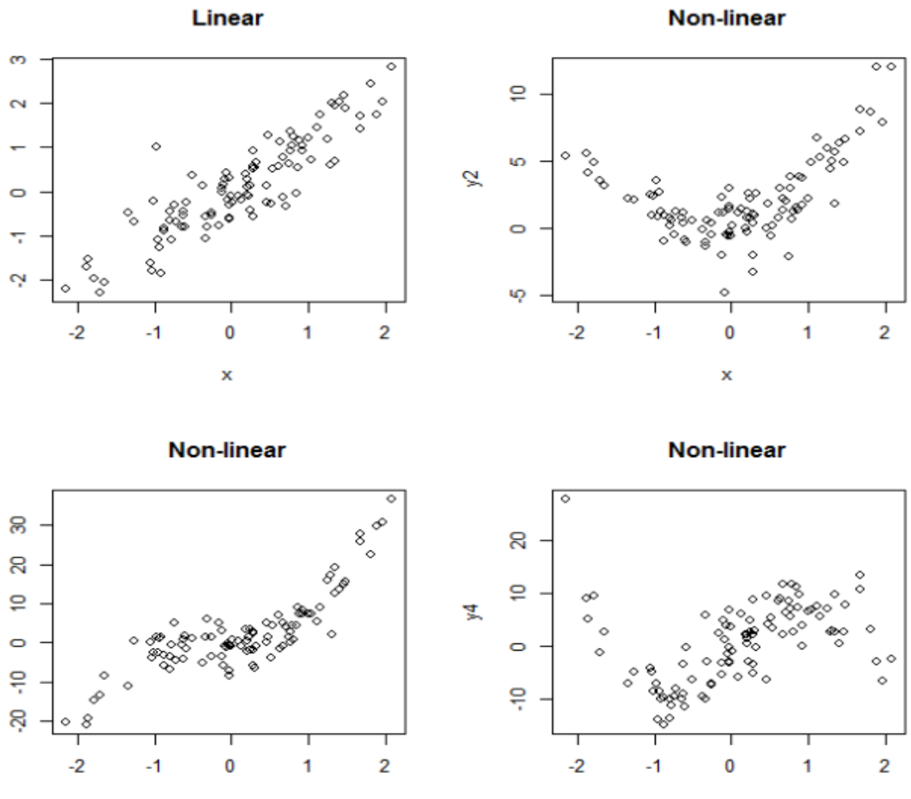
- Correlation and simple linear regression are used to examine the relationship between two quantitative (discrete or continuous) variables.
- **Predictor variable**, x, provides information about some **dependent variable**, y.

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

Linear Regression

- 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.
- The error ε is assumed to be normally distributed with μ = 0 and σ. Linear regression also assumes that the variance of ε is constant and unchanging across the range of the predictor variable, x.
- The slope represents change in Y when X is increased by one unit. If slope is positive then we will have increase in Y and if it negative we would observe decrease in Y.

Should we fit Linear Regression? Check the scatter plot between Y and X Scatter plot must be linear

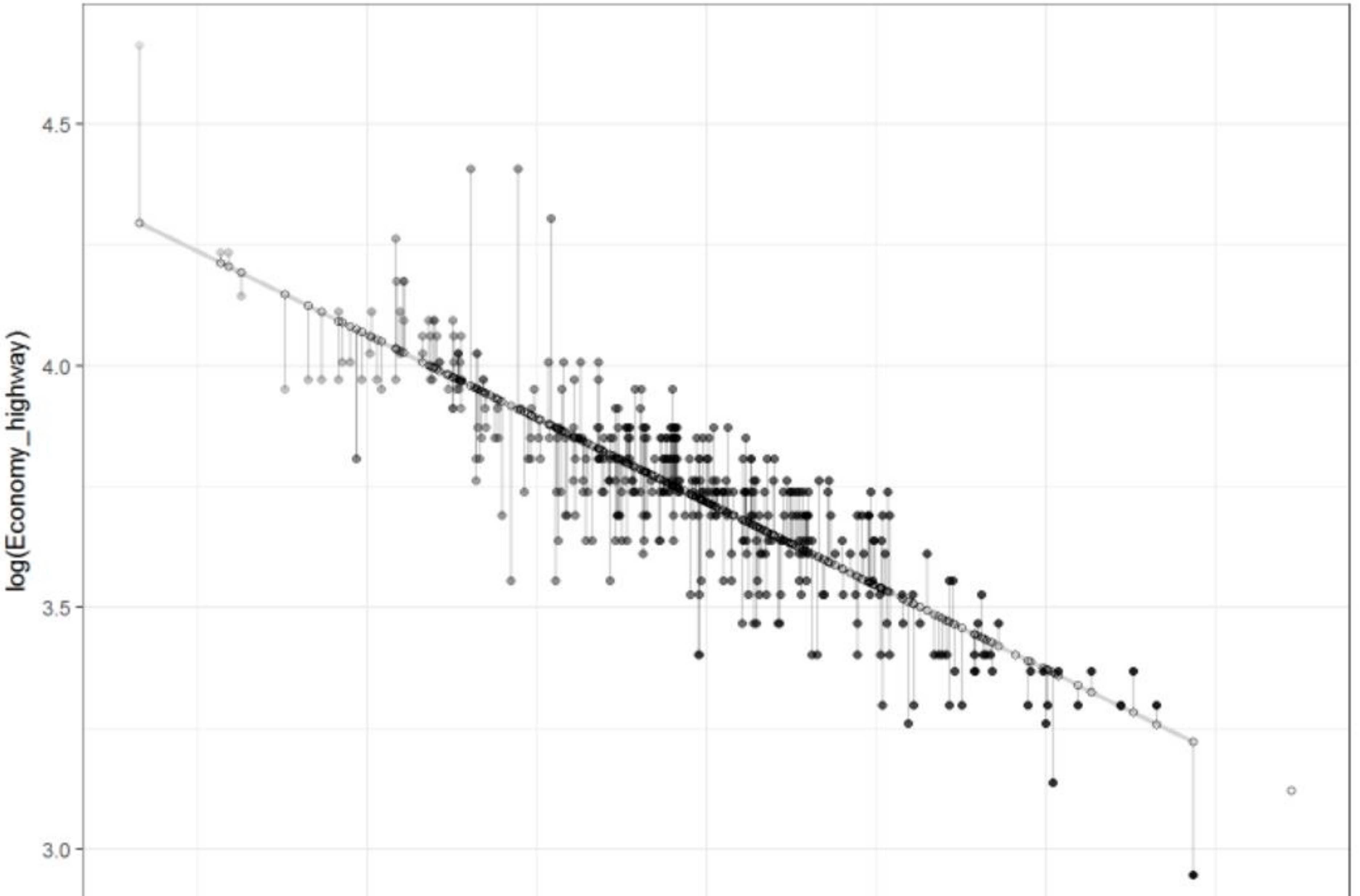


Fitting Linear Regression model to sample data using R

- 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:

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

Fitting Linear Regression model to



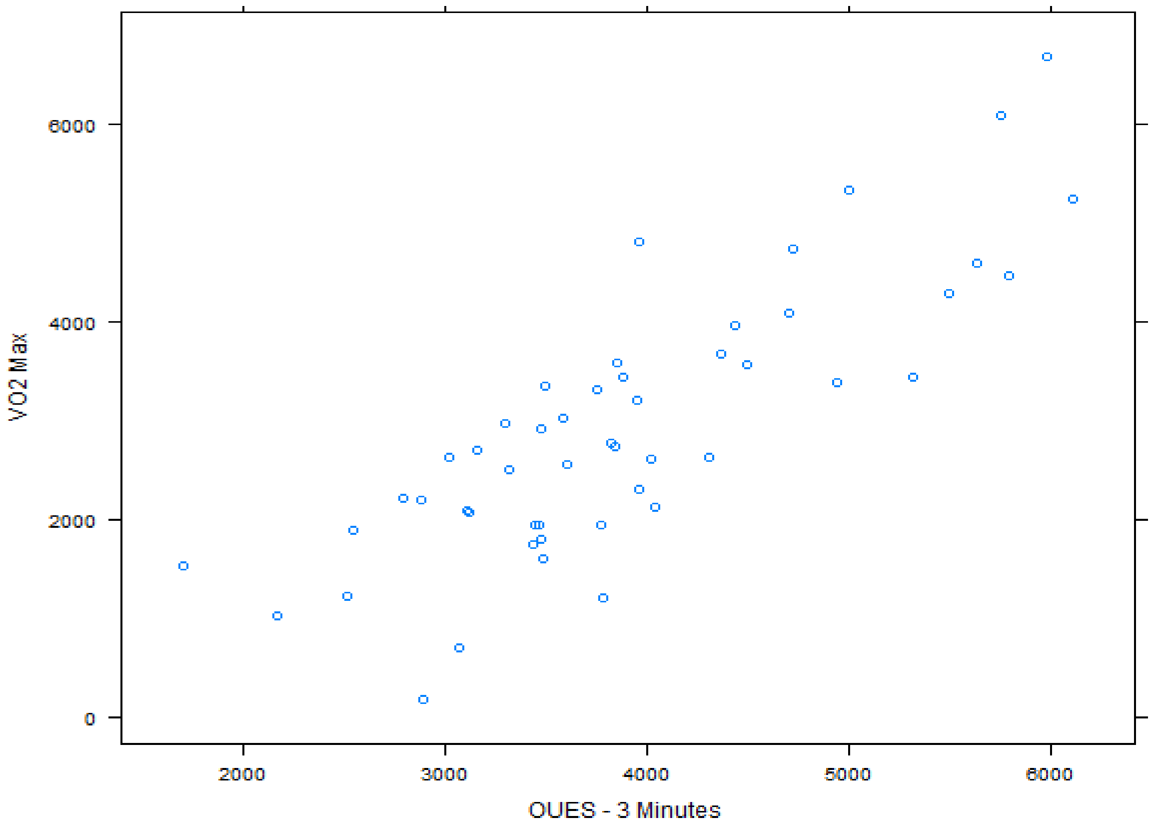
Example - Oxygen Uptake Efficiency

- 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.
- 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.

Example - Oxygen Uptake Efficiency

D	y = VO2 Max	x = OUES 3 Minutes	y2	x2	xy	
	1	4310.56	2626.44	18580927.51	6898187.074	11321427.21
	2	3157.98	2701.46	9972837.68	7297886.132	8531156.651
	3	3751.89	3308.54	14076678.57	10946436.93	12413278.14
	4	2791.33	2206.55	7791523.169	4868862.903	6159209.212
	5	4022.37	2608.16	16179460.42	6802498.586	10490984.54
	6	3107.35	2090.39	9655624.023	4369730.352	6495573.367
	7	2892.19	172.82	8364762.996	29866.7524	499828.2758
	8	4037.79	2122.35	16303748.08	4504369.523	8569603.607
	9	3854.39	3584.09	14856322.27	12845701.13	13814480.66
	10	3072.82	705.59	9442222.752	497857.2481	2168151.064
	11	5498.97	4279.19	30238671.06	18311467.06	23531137.43
	12	3023.52	2621.46	9141673.19	6872052.532	7926036.739
	13	4718.86	4726.43	22267639.7	22339140.54	22303361.47
	14	3782.19	1194.22	14304961.2	1426161.408	4516766.942
	15	3958.14	4794.13	15666872.26	22983682.46	18975837.72
	16	3319.66	2500.85	11020142.52	6254250.723	8301971.711
	17	5980.62	6678.83	35767815.58	44606770.17	39943544.27
	18	3475.77	1800.94	12080977.09	3243384.884	6259653.224
	19	2509.71	1214.04	6298644.284	1473893.122	3046888.328
	20	2881.11	2190.51	8300794.832	4798334.06	6311100.266
	21	3578.75	3015.03	12807451.56	9090405.901	10790038.61
	22	4432.84	3959.55	19650070.47	15678036.2	17552051.62
	23	3601.55	2546.3	12971162.4	6483643.69	9170626.765
	24	3946.94	3192.09	15578335.36	10189438.57	12598987.7
	25	3448.56	1947.26	11892566.07	3791821.508	6715242.946
	26	4369.16	3677.25	19089559.11	13522167.56	16066493.61
	27	6112.58	5230.23	37363634.26	27355305.85	31970199.29
	28	5749.87	6077.92	33061005.02	36941111.53	34947249.87
	29	3957.07	2307.5	15658402.98	5324556.25	9130939.025
	30	4705.48	4074.82	22141542.03	16604158.03	19173984.01
	31	5315.19	3437.67	28251244.74	11817575.03	18271869.21
	32	5632.32	4584.79	31723028.58	21020299.34	25823004.41
	33	2167.45	1030.9	4697839.503	1062754.81	2234424.205
	34	1698.4	1524.37	2884562.56	2323703.897	2588990.008
	35	3817.64	2777.68	14574375.17	7715506.182	10604182.28
	36	3114.81	2073.59	9702041.336	4299775.488	6458838.868
	37	3484.14	1605.88	12139231.54	2578850.574	5595110.743
	38	3881.28	3426.48	15064334.44	11740765.19	13299128.29
	39	2542.49	1888.01	6464255.4	3564581.76	4800246.545
	40	3477.67	2921.48	12094188.63	8535045.39	10159943.35
	41	3294.2	2974.7	10851753.64	8848840.09	9799256.74
	42	5791.01	4460.16	33535796.82	19893027.23	25828831.16
	43	3495.11	3352.15	12215793.91	11236909.62	11716132.99
	44	4493	3557.7	20187049	12657229.29	15984746.1
	45	5001.9	5333.79	25019003.61	28449315.76	26679084.2
	46	3464.99	1936.67	12006155.7	3750690.689	6710542.183
	47	4938.21	3389.12	24385918	11486134.37	16736186.28
	48	3844.41	2727.18	14779488.25	7437510.752	10484398.06
	49	3433.36	1746.94	11787960.89	3051799.364	5997873.918
	50	3767.64	1949.32	14195111.17	3799848.462	7344336.005
	∑	194705.24	146853.52	807085161.3	521621342	626812929

Example - Oxygen Uptake Efficiency



Hypothesis and Assumptions for linear regression model

- H0: The data does not fit to linear regression model.
- HA: The data fits to linear regression model.
- We test the overall model using F-test.
- Assumptions:
- Independence (check research design)
- Linearity(Check scatter plot)
- Normality of residuals(check after model is fitted)
- Homoscedasticity (check after model is fitted)
- Decision Rules:
- Reject H0 if P-value for F statistics < α.

Example - Oxygen Uptake Regression output (R output)

Linear Regression is fitted using lm() function

```
> ouesvo2maxmodel <- lm(VO2_Max ~ OUES_3, data = OUES)
> msummary(ouesvo2maxmodel)
              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 ***
```

P- value to test α =0.0

P-value to test β =0.0

The value of slope β

The value of intercept α

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 regression model is

Y=2107+0.6085X

So if we increase X by one unit, we have increase of o.6085 in Y

Example - Oxygen Uptake Efficiency (R output)

- 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.

Example - Oxygen Uptake Efficiency (R output)

- The R^2 is a measure of goodness of fit for linear regression. Higher R^2 indicates stronger relationship between Y and X.
- 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.

Example - Oxygen Uptake Efficiency (R output)

- **The F statistic is testing**
- **H_0: The data do not fit the linear regression model**
- **H_A: The data fit the linear regression model**
- the F statistic reported in the summary as $F = 103.9$, will have a F distribution with $df_1 = 1$ and $df_2 = n - 2 = 50 - 2 = 48$.

Example - Oxygen Uptake Efficiency (R output) hypothesis testing for **α and** β

- To test if constant or slope is zero or not . The out put provide the p-value for the following tests:
- **H_0: α = 0**
- **H_A: α ≠ 0**
- And
- **H_0: β = 0**
- **H_A: β ≠ 0**

Example - Oxygen Uptake Efficiency (R output) hypothesis testing for α and β

In R, we use the confint() function:

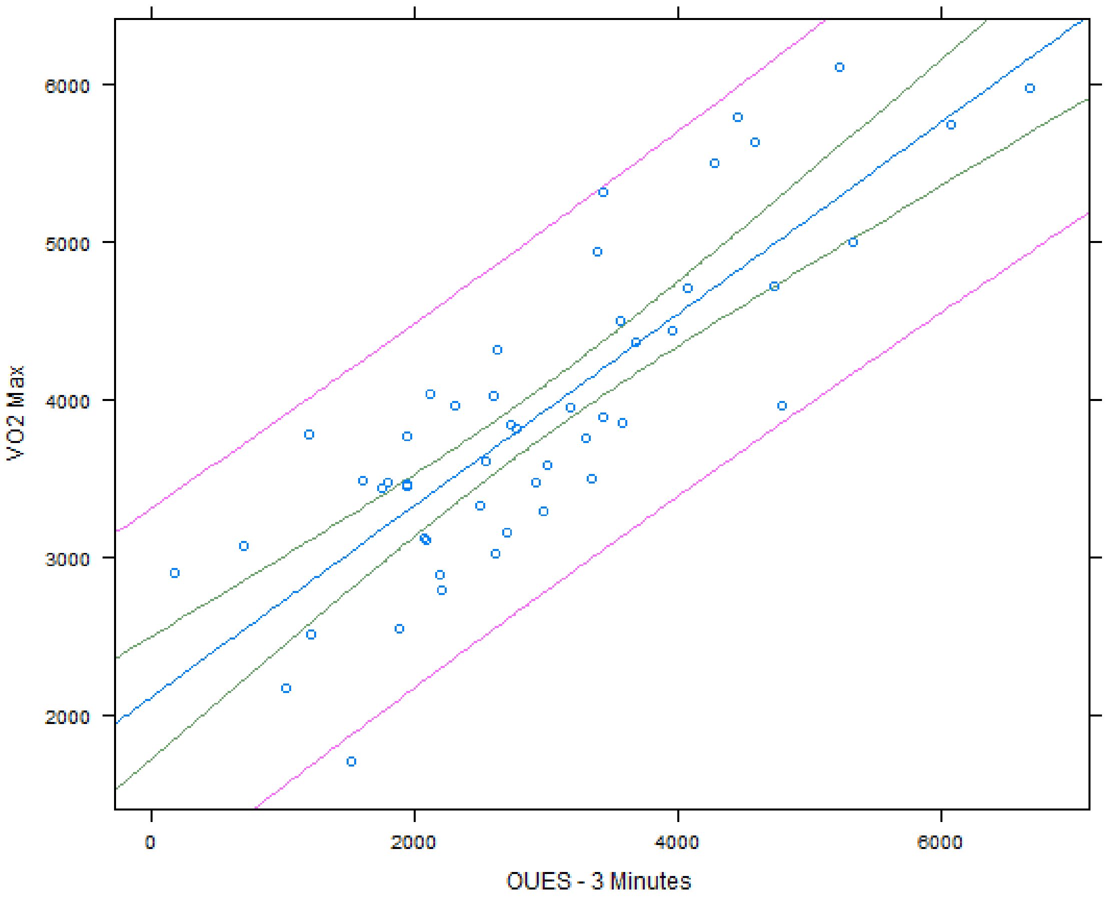
```
> confint(ouesvo2maxmodel)
                    2.5 %       97.5 %
(Intercept)  1719.2061023  2494.521512
OUES_3          0.4884909     0.728532
```

R reports the 95% *CI* for **β** to be [.488, .729].

Example - Oxygen Uptake Efficiency (R output) 95% CI for the fitted line

- In the following graph we have:
- The blue line is the line of best fit for the linear regression. The green bands represent the 95% CI of mean VO_2 max readings for the regression line.
- The pink outer lines are the prediction intervals.
- The prediction intervals are where 95% of the data will fall assuming the residuals are normally distributed.

Example - Oxygen Uptake Efficiency (R output) 95% CI for the fitted line



Checking the regression assumptions

Assumptions

- Before we report the final regression model, we must validate all the following assumptions for linear regression.
- Independence (check research design)
- Linearity(Check scatter plot)
- Normality of residuals(check after model is fitted)
- Homoscedasticity (check after model is fitted)

Example - Oxygen Uptake Efficiency Residuals

The residuals are calculated as:

- IF y_i is the observed score in the sample and $\hat{y}_i$ is the predicted score based on the fitted regression model then the residual for the ith sample point is defined by

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

- For example, the predicted score for OUES = 4000 is

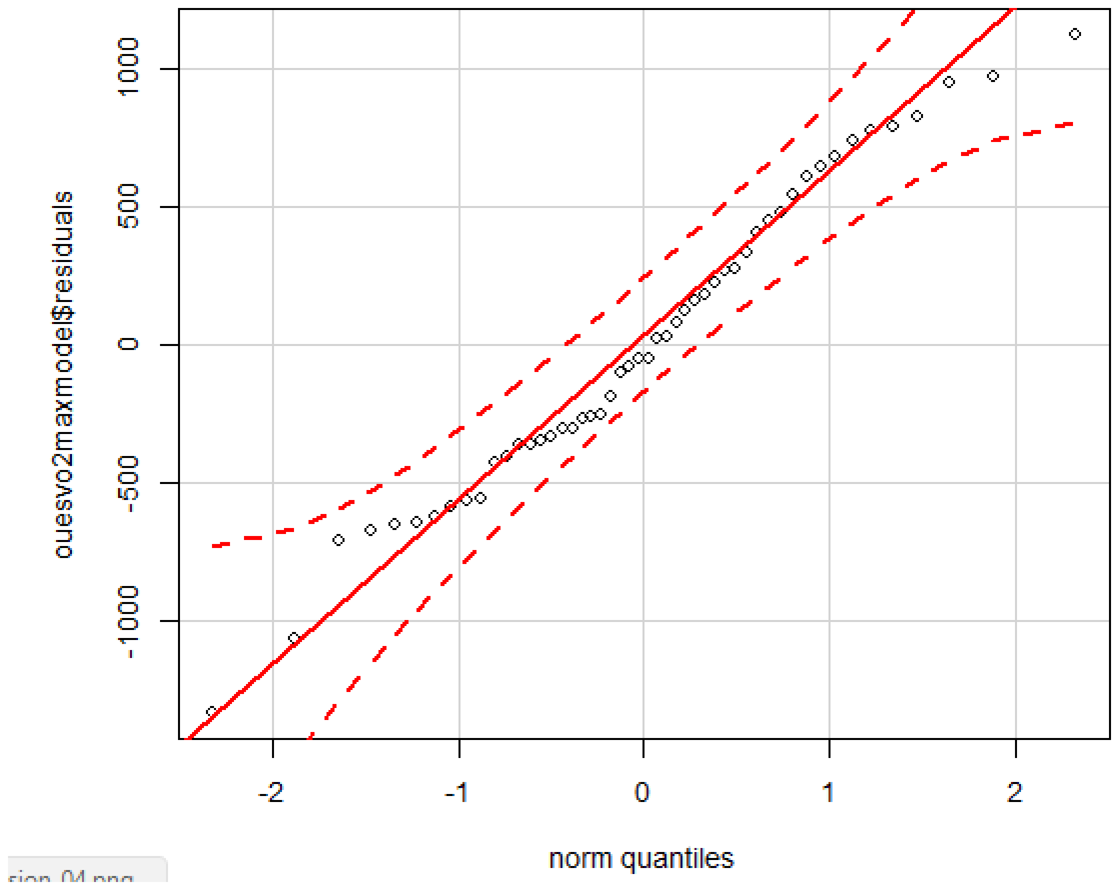
$$\hat{y} = a + bx_i = 2106.864 + 0.609(4000) = 4542.864$$

Example - Oxygen Uptake Efficiency (R output) Checking normality of residuals

Using

- qqPlot(ouesvo2maxmodel$residuals, dist="norm")
- We get

Normal plot of residuals

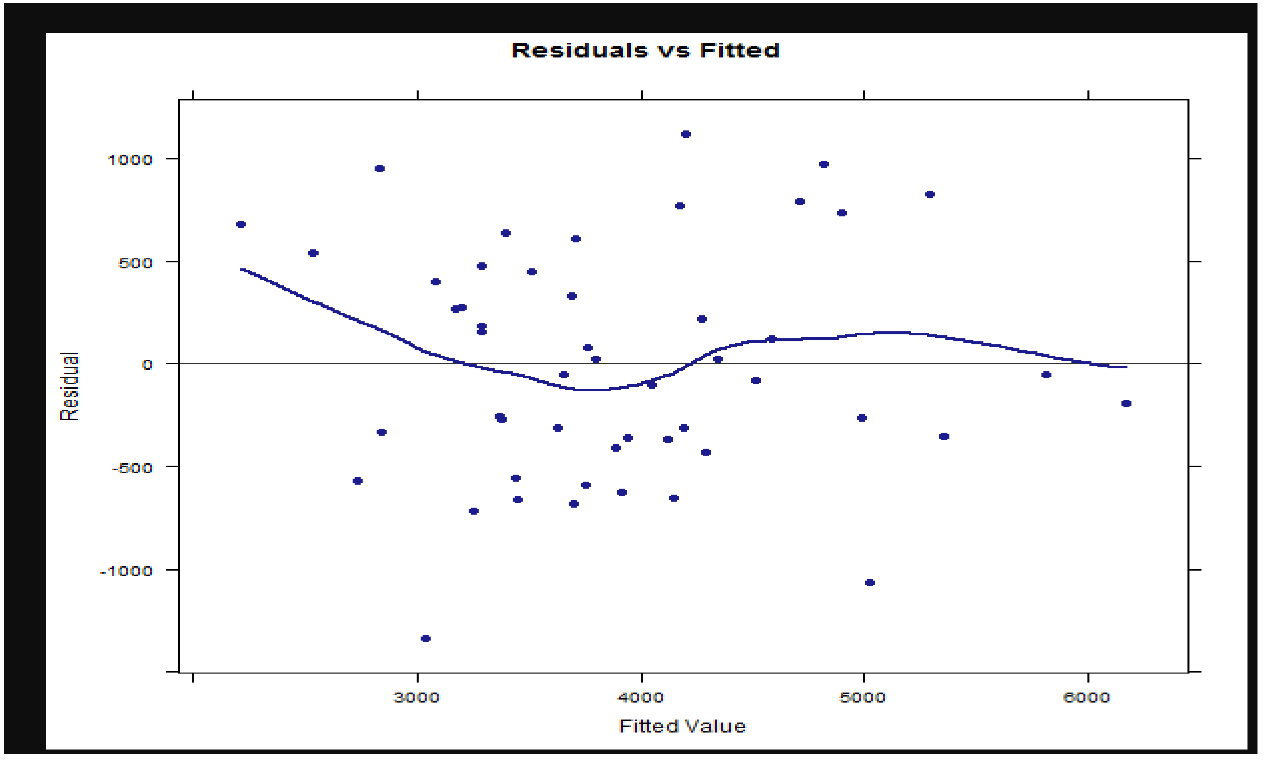


Assumption of **homoscedasticity**, or constant variance.

- **Homoscedasticity** is related to the assumption of homogeneity of variance for the two-sample t-test.
- Check this assumption by looking at a scatter plot of the predicted values on the x axis and the residuals on the y.
- As we move across predicted values, the variance in the residuals should remain constant.

Assumption of **homoscedasticity**, or constant variance.

- > mplot(ouesvo2maxmodel, 1)



Assumption of **homoscedasticity**, or constant variance

- The blue line in the plot is a non-parametric locally weighted scatterplot 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.

Assumption of **homoscedasticity**, or constant variance

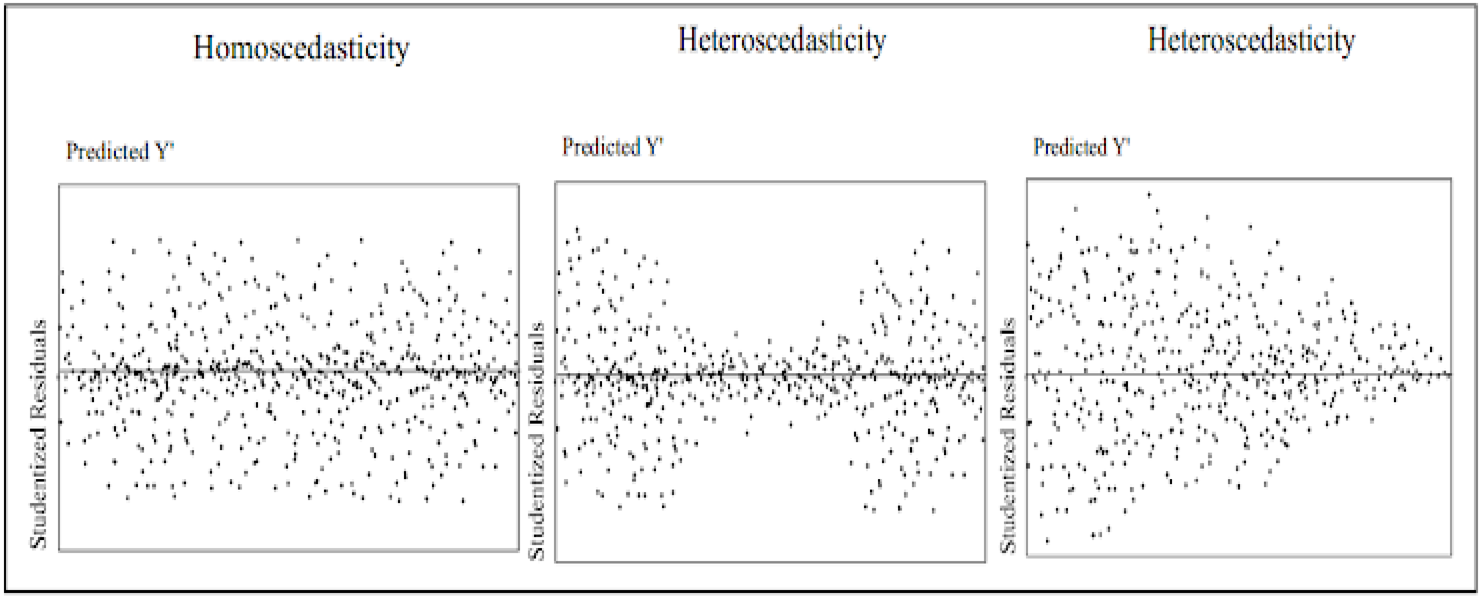


Figure 3. Examples of homoscedasticity and heteroscedasticity

Example Write-up

- Prior to fitting the regression, a scatterplot assessing the bivariate relationship between VO_2 max and OUES 3 minutes was inspected.
- The scatterplot demonstrated evidence of a positive linear relationship.
- The overall regression model was statistically significant, $F(1, 48) = 103.92$, $p < .001$
- The results show that OUES 3 minutes explains 68.4% of the variability in $V0_2$ max, $R^2 = .684$.

Example Write-up

- 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.

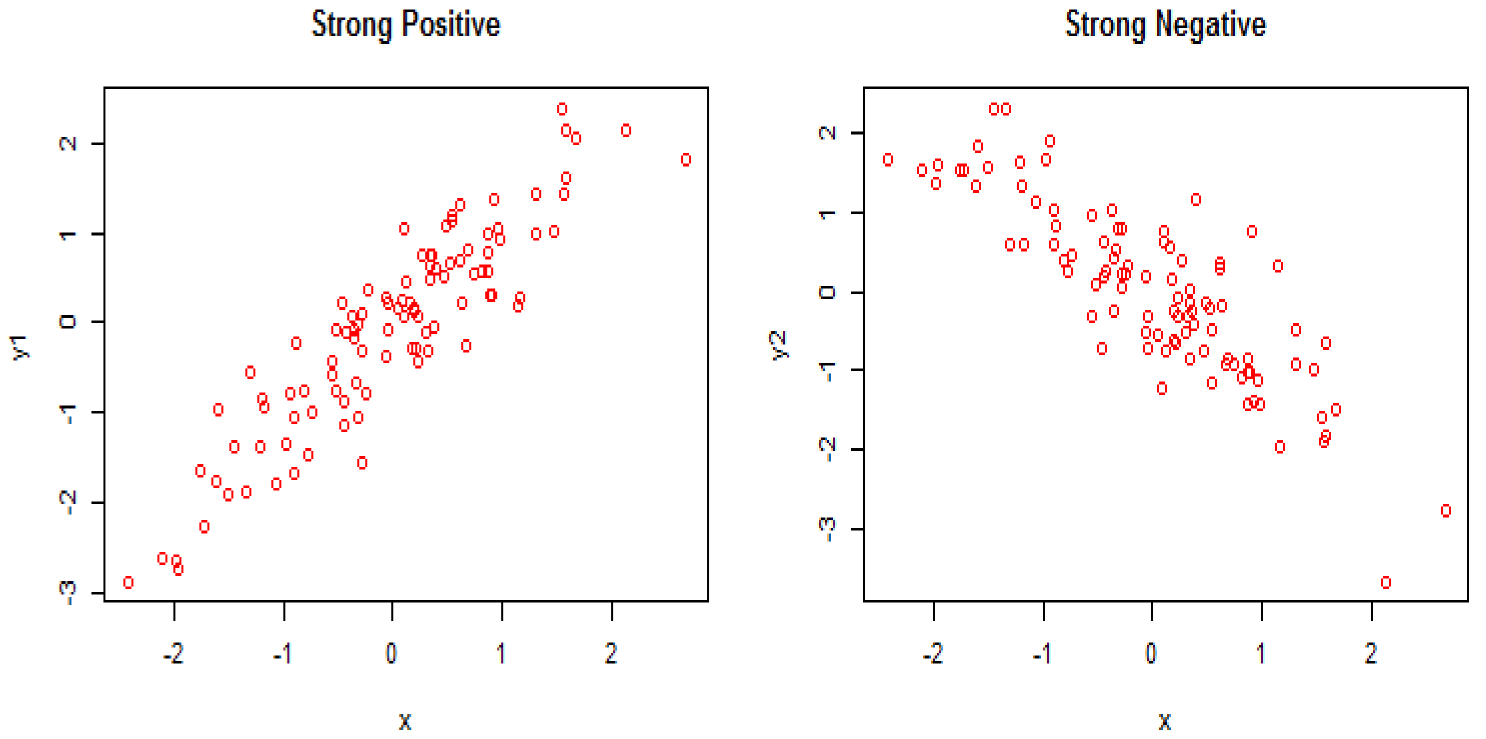
Correlation

- The **Pearson correlation coefficient**, *r*, is a standardised measure of the strength of the linear relationship between two variables.
- Its value r is sqrt (R^2 = .684)= 0.827
- 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.
- We can calculate a quick correlation in R using the cor() function:

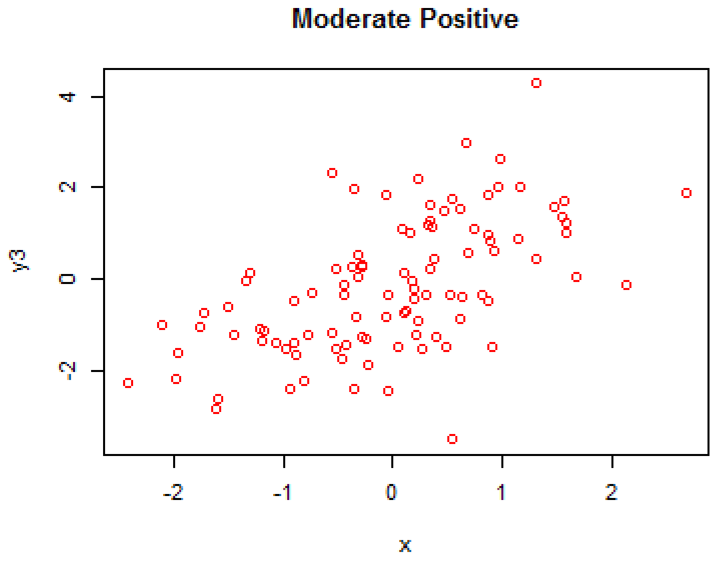
Correlation

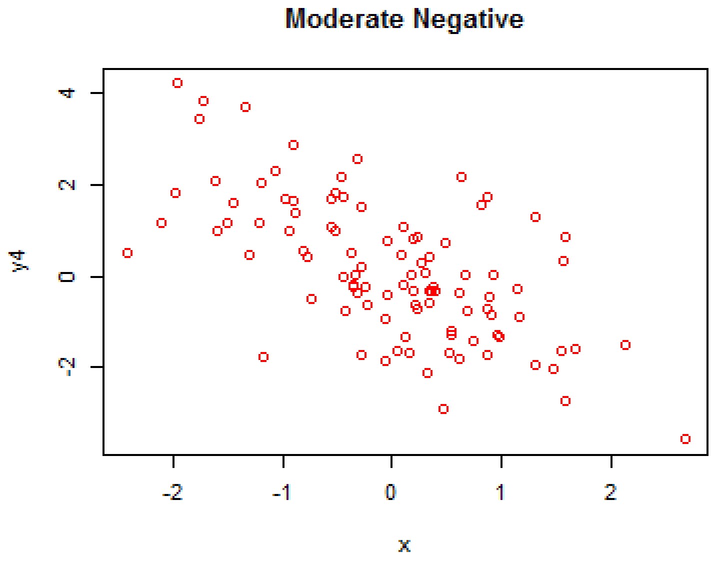
- > cor(VO2_Max,OUES_3,data = OUES)
- [1] 0.8270676

Correlation



Correlation





Correlation