Predictive Statistics

Probability and Statistical Inference

Bojan Božić

TU Dublin

Winter 2020

Predictive Models

- Used to explore the relationship between one outcome variable and a set of independent variables (predictors).
- You should have a sound theoretical or conceptual reason for exploring the relationship and the order of the variables entering the model.
- For variables that have not been previously shown to contribute to the variation in the outcome variable try to establish statistical evidence for their inclusion in advance.

Linear Regression

Predictive model - what does it allow you do?

- Prediction:
  - Really what we are looking at is the variance in the outcome variable and how much of the variance could be considered to be explained by the predictor variables.
- Questions it allows you to answer:
  - How well a set of variables is able to predict an outcome variable?
  - Which variable in a set is the best predictor?
  - Whether a variable is still able to predict an outcome when controlling for other variables?

What is Linear Regression?

- It is a hypothetical model of the relationship between two variables.
    - The model used is a linear one.
- Theoretical assumption:
    - For every one unit of change in the independent variable (predictor) there will be a consistent and uniform change in the dependent variable (predicted/outcome).
    - Therefore, we describe the relationship using the equation of a straight line.
    - This can be seen as a way of predicting the value of one variable from another.

Before Regression

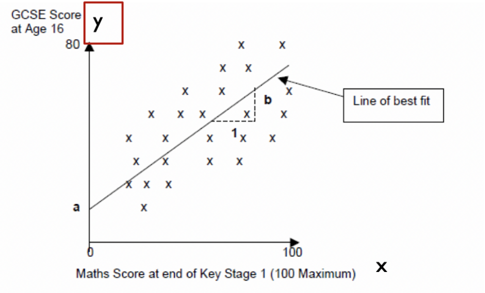
- We need to establish evidence to support going ahead with building a predictive model.
- If we are asserting a relationship:
  - We need to investigate if there is any evidence of a relationship using correlation and make a decision based on the results (strength, direction etc.).
- If we are asserting a differential effect for different groups:
  - We need to investigate if there is any difference using either the appropriate test and make a decision based on the result.

Simple Example - Linear Regression

- Suppose we want to look at what variables predict a child's maths score aged 16 in the UK (GCSE).
- We have a theory that their achievement on a standard maths test at aged 7 is a good predictor.

Simple Example

$$y = a + bx + e$$



Relationship between Maths score age 7 with GCSE result at age 16 (for 25 students)

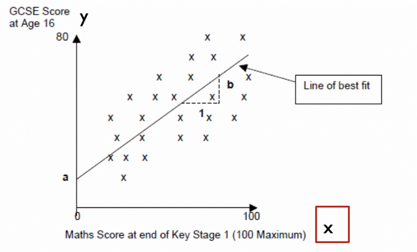
For a simple one predictor model.

'y' is the response variable to be predicted (in this case GCSE Score) - the dependent variable.

This is a simple correlation.

Simple Example

$$y = a + bx + e$$

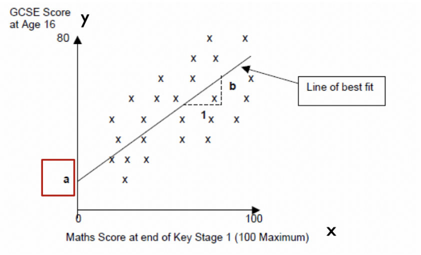


Relationship between Maths score age 7 with GCSE result at age 16

'x' is the predictor variable (in this case Maths Score aged 7) – independent variable.

Simple Example

$$y = a + bx + e$$

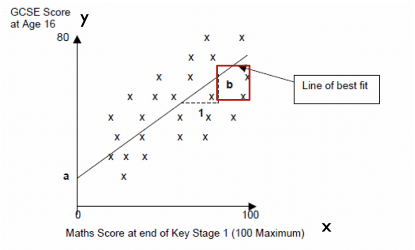


Relationship between Maths score age 7 with GCSE result at age 16

'a' is the intercept or the point where the line crosses the y-axis;

Simple Example

$$y = a + bx + e$$



Relationship between Maths score age 7 with GCSE result at age 16

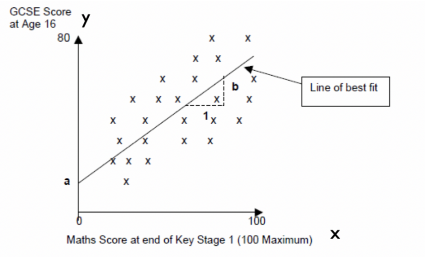
'b' is the gradient of the line.

Represents amount of response variable change for one unit in predictor variable.

(e.g. for every one percentage point increase in a child's Maths Test score, the line suggests that the child's predicted GCSE Score will increase by 'b' points);

Simple Example

$$y = a + bx + e$$



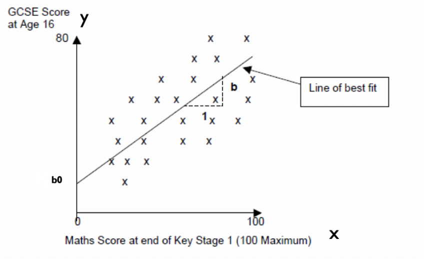
Relationship between Maths score age 7 with GCSE result at age 16

'e' is an error term reflecting the fact that the line of best fit does not perfectly model the data.

Each child will have her/his own value for 'e' - called the residual - which is simply calculated by subtracting the child's actual GCSE Score from the score predicted for them from the model.

Simple Example

$$y = a + bx + e$$



Relationship between Maths score age 7 with GCSE result at age 16

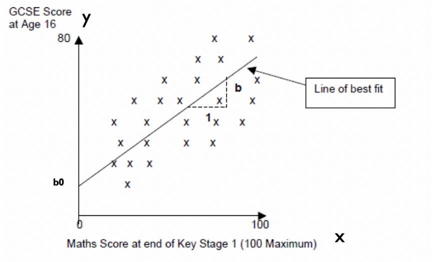
If the line did completely model the data then all of the points would rest exactly on the line.

Because there is a difference between what the line predicts (expected values) each student will achieve in their GCSE Score and what they actually achieved (observed values).

This difference is basically the vertical distance between each point and the line itself and this distance is 'e'.

Simple Example

$$y = a + bx + e$$



Relationship between Maths score age 7 with GCSE result at age 16

'e' obviously varies for each young person.

Without any further information (i.e. additional predictor variables) we cannot model this additional variation and so just treat it as random error.

Residuals

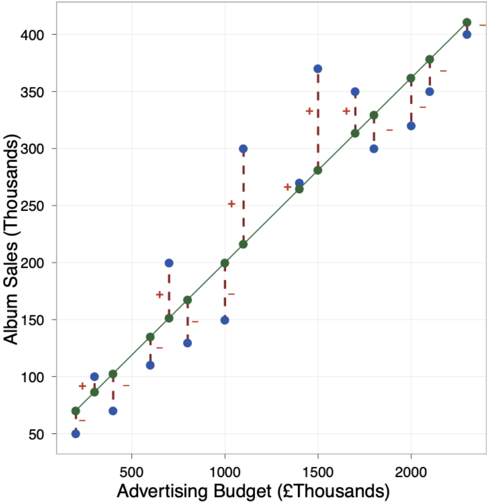


FIGURE 8.4
A scatterplot of some data with a line representing the general trend. The vertical lines (dotted) represent the differences (or residuals) between the line and the actual data

Using our model

- We can re-write the overall model as an equation:

$$GCSE\ Score = a + (b * Maths\ Score) + error$$

- To use the model to predict a child's future GCSE Score we would tend to drop the 'e' term and thus use the formula:

$$Predicted\ GCSE\ Score = a + (b * Maths\ Score)$$

- We are now referring to the response variable as the 'Predicted GCSE Score' rather than the actual GCSE Score.

Using our model

- Suppose we have plugged our numbers into our statistical tool and it has given us the following values:
    - $a = 15; b = 0.7$
- Our model becomes:
    - $Predicted\ GCSE\ Score = 15 + 0.7 * Maths\ Score$
- Based on our data this represents our best estimate of what a child's future GCSE Score is going to be, on average, given their Maths Score at age 7.
    - It is not going to be 100% correct as we know there is an error term.

Using our model

- So suppose a child got just 10% in their maths test aged 7.
  - Predicted GCSE Score = 15 + 0.7 x 10 (maths test aged 7) = 15 + 7 = 22.0
- A child who gets 60% in their maths test aged 7:
  - Predicted GCSE Score = 15 + 0.7 x 60 = 15 + 42 = 57
- Looking at a student who gets 61% we can see what the gradient b is doing in practice:
  - Predicted GCSE Score = 15 + 0.7 x 61 = 15 + 42.7= 57.7
- From our model we can see that on average one unit increase in Maths score aged 7 is predicted to lead to a 0.7 increase in GCSE Maths score aged 16.

Using our model

The gradient therefore represents the average increase in the response variable (GCSE Score) with one unit increase in the predictor variable (Maths Score).

Sample Dataset (Regression.sav)

- The dataset comprises a sample of 4,059 young people (aged 16) selected from 65 difference secondary schools from six inner London Education Authorities.
- This is a sub-sample from a much larger study undertaken by Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., et al. (1993) A multilevel analysis of school examination results, *Oxford Review of Education*, 19, pp. 425-433.
- The dataset has been specifically prepared to accompany Rasbash, J. et al. (2005) *A User's Guide to MLwiN 2.0* (Bristol, Centre for Multilevel Modelling).

Simple Linear Regression

Variable	Description
School	A unique numeric identifier for each school.
Student	A unique numeric identifier for each student.
Normexam	Student's exam score at age 16, normalised to have approx. a standard normal distribution and a mean of 0 and standard deviation of 1. (Note that the normalisation was carried out on a larger sample, so the mean in this subsample is not exactly equal to 0 and the variance is not exactly equal to 1.
Cons	A column of 1s. This is used in the multilevel modelling package MLwiN to represent the intercept in a statistical model.
Standlrt	Student's score at age 11 on the London Reading Test (LRT). standardised using z-scores
Girl	1 = girl, 0 = boy.
Schgend	school's gender (1 = mixed, 2 = boys, 3 = girls).
Avsrlt	Average LRT score in school.
Schav	Average LRT score in school (1 = bottom 25%, 2 = middle 50%, 3 = top 25%).
Vrband	Student's score in test of verbal reasoning at age 11 (1 = top 25%, 2 = middle 50%, 3 = bottom 25%).

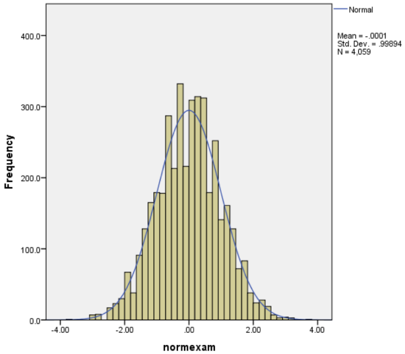
Simple Linear Regression

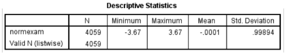
- The theory we are exploring is whether the main predictor of 'normexam' is a student's prior performance in a reading test at age 11 (the variable 'standlrt').
- The response (outcome or dependent) variable in this case is the student's exam score at age 16 ('normexam') that has been converted into a standardised score (with mean = 0 and standard deviation = 1).
- The predictor (independent) variable in this case is the students prior performance in a reading test at age 11 (standlrt) which is also a standardised score).

Simple Linear Regression

- It is generally good practice to standardise your response and predictor variables before including them in the model.
- It helps in the interpretation of the model.
  - For example, if all the predictor variables have been standardised then we know that if they all have the value '0' this would represent the average student.
  - As such, the constant in the model comes to represent the mean score on the response variable for the average student in the sample.
  - If the response variable has been standardised we can immediately gain a sense of where the 'average student' fits in the distribution of the response variable:
  - (i.e. if the constant is negative then we know that they have a mean score below the mean for the sample as a whole, and so on).

Simple Linear Regression





Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
normexam	4059	-3.67	3.67	-.0001	.99894
Valid N (listwise)	4059				

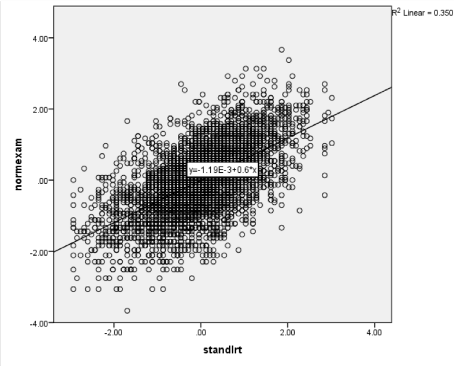
Start by looking at the variable:

Create a histogram.

Create some descriptive statistics .

The response variable (normexam) is approximately normally distributed and approximately standardised (with mean = -0.0001 and standard deviation = 0.9989).

Simple Linear Regression



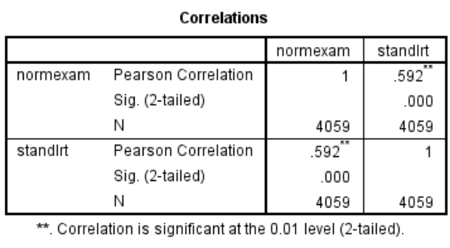
Before we create a linear regression model it is worth exploring the nature of the relationship between the response variable and this predictor variable.

Look at the scatterplot.

Calculate the correlation for these two variables.

There appears to be a strong linear relationship. $R^2 = 0.35, r = 0.592$.

Simple Linear Regression



Correlations

		normexam	standlrt
normexam	Pearson Correlation	1	.592**
	Sig. (2-tailed)		.000
	N	4059	4059
standlrt	Pearson Correlation	.592**	1
	Sig. (2-tailed)	.000	
	N	4059	4059

**. Correlation is significant at the 0.01 level (2-tailed).

Increase scores in standlrt appear to associated with increase scores in normexam.

There appears to be a moderately strong correlation between the two ($r = 0.592, p < 0.001$).

Therefore there is good justification for using regression to look at prediction.

Note: we say moderately strong as Cohen says 0.5 is strong so we are qualifying out result.

Simple Linear Regression

- To formally model this relationship now we can create a simple linear regression model (assuming we have read our data file into a dataframe called regression):
- $model1 < -lm(regression\$normexam \sim regression\$standlrt)$
- We now have access to all the elements of the regression output in the variable model which we can unpack and examine.

How Good is the Model?

- The regression line is only a model based on the data.
- This model might not reflect reality.
  - We need some way of testing how well the model fits the observed data.
- How has the model helped improve our understanding in the absence of any model?

How Good is the Model?

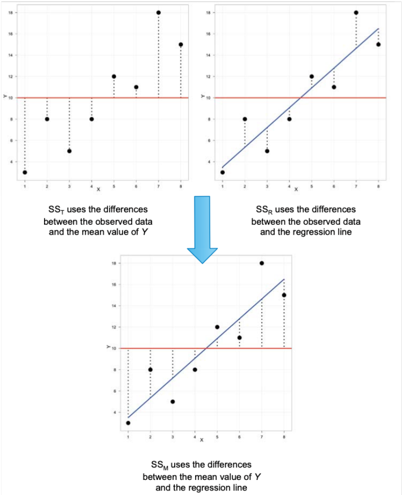
What would we use in the absence of the model?

- The common approach would be to use the mean of the sample:
    - This would be wrong in the vast majority of the cases – depending on the variation in the sample.
    - i.e. how much each case varies from the mean.
    - But we can use it as a baseline to see if our model has improved our understanding.

How good is the model?

- So how do we go about it?
- We square the differences from each data point to our mean for our outcome variable.
  - Squaring these gives us the total **sum of squares** = all variation that exists.
- But we will still experience some error:
  - This is the variation between our observed data and our regression line.
  - Squaring these errors gives use the **residual sum of squares** (unexplained sum of squares) = variation which is left over after the model is fitted to the data.
- Difference between the two is the amount of **variation accounted for by the model**.

Sums of Squares



SS_T uses the differences between the observed data and the mean value of *Y*

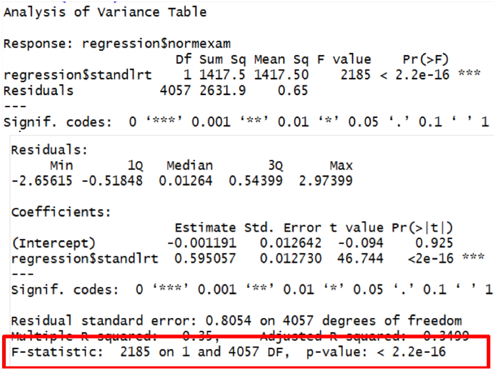
SS_R uses the differences between the observed data and the regression line

SS_M uses the differences between the mean value of *Y* and the regression line

FIGURE 8.5 Diagram showing from where the regression sums of squares derive

- SS_T: Total variability (variability between scores and the mean).
- SS_R: Residual/Error variability (variability between the regression model and the actual data).
- SS_M: Model variability (difference in variability between the model and the mean).

Simple Linear Regression



```
Analysis of Variance Table

Response: regression$normexam
                     Df Sum Sq Mean Sq F value    Pr(>F)
regression$standlrt   1 1417.5 1417.50    2185 < 2.2e-16 ***
Residuals          4057 2631.9    0.65
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.65615 -0.51848  0.01264  0.54399  2.97399

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.001191   0.012642  -0.094    0.925
regression$standlrt  0.595057   0.012730  46.744   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8054 on 4057 degrees of freedom
Multiple R-squared:   0.35,     Adjusted R-squared:  0.3499
F-statistic:  2185 on 1 and 4057 DF,  p-value: < 2.2e-16
```

To get this information in R:

$anova(model1)$

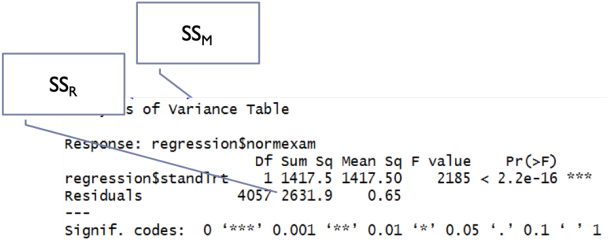
$summary(model1)$

The F statistic looks at whether the model as whole is statistically significant.

This allows us to answer the question:

Do the independent variables, taken together, predict the dependent variable better than just predicting the mean for everything?

Simple Linear Regression



```
s of Variance Table

Response: regression$normexam
                     Df Sum Sq Mean Sq F value    Pr(>F)
regression$standlrt    1 1417.5 1417.50    2185 < 2.2e-16 ***
Residuals           4057 2631.9    0.65
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- SS_R: Residual/Error variability (variability between the regression model and the actual data).
- SS_M: Model variability (difference in variability between the model and the mean).
- SS_T: Total variability (variability between scores and the mean).

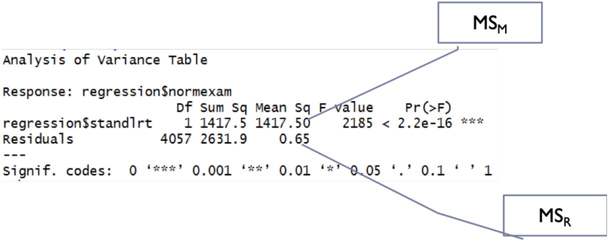
Testing the Model: ANOVA

Mean Squared Error:

- Sums of Squares are total values.
- They can be expressed as averages.
- These are called Mean Squares, MS.

$$\boxed{F = \frac{MS_M}{MS_R}}$$

Simple Linear Regression



```
Analysis of Variance Table

Response: regression$normexam
                      Df Sum Sq Mean Sq F value    Pr(>F)    
regression$standlrt    1 1417.5 1417.50    2185 < 2.2e-16 ***
Residuals           4057 2631.9    0.65                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

MS_M

MS_R

- MS_M: Mean Square Error of the Model.
- MS_R: Mean Square Error of the Residuals.

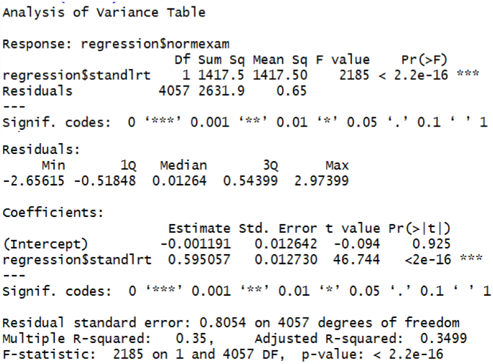
Testing the Model: R^2

- The proportion of variance accounted for by the regression model.
- The Pearson Correlation Coefficient Squared.

$$\boxed{R^2 = \frac{SS_M}{SS_T}}$$

Simple Linear Regression

anova(*model*)



```
Analysis of Variance Table

Response: regression$normexam
                    Df Sum Sq Mean Sq F value    Pr(>F)    
regression$standlrt    1 1417.5 1417.50    2185 < 2.2e-16 ***
Residuals         4057 2631.9    0.65                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residuals:
     Min       1Q   Median       3Q      Max 
-2.65615 -0.51848  0.01264  0.54399  2.97399 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -0.001191   0.012642  -0.094    0.925    
regression$standlrt  0.595057   0.012730  46.744   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8054 on 4057 degrees of freedom
Multiple R-squared:   0.35,	Adjusted R-squared:  0.3499 
F-statistic:  2185 on 1 and 4057 DF,  p-value: < 2.2e-16
```

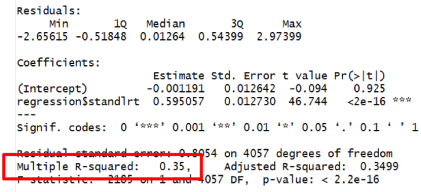
In this case there is only one independent variable, standrlrt.

The F-test is testing if this one variable predicts the exam scores at age 16 (normexam) better than if we used the average score of normexam to predict values for all students.

It seems it does as it is statistically significant $p < 0.001$.

But how useful is it?

Simple Linear Regression



```
Residuals:
     Min       1Q   Median       3Q      Max
-2.65615 -0.51848  0.01264  0.54399  2.97399

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.001191   0.012642  -0.094    0.925
regression$standlrt  0.595057   0.012730  46.744   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8054 on 4057 degrees of freedom
Multiple R-squared:   0.35,     Adjusted R-squared:  0.3499
F-statistic:  2185 on 1 and 4057 DF,  p-value: < 2.2e-16
```

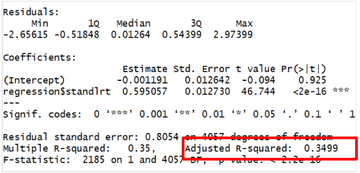
The R^2 is the proportion of variance in the outcome (dependent) variable **normexam** which can be explained by the predictor (independent) variables. This is the usefulness of the model.

An R^2 of 1 means the independent variable explains 100% of the variance in the dependent variable. Conversely 0 means it explains none.

In this case we have a value of 0.350 which means that standlrt explains 35% of the variance in normexam.

The remaining variance can be explained by variables not currently included in the model.

Simple Linear Regression

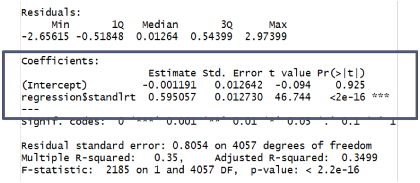


```
Residuals:
     Min       1Q   Median       3Q      Max
-2.65615 -0.51848  0.01264  0.54399  2.97399

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -0.001191   0.012642  -0.094    0.925
regression$standlrt  0.595057   0.012730  46.744   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8054 on 4057 degrees of freedom
Multiple R-squared:   0.35,     Adjusted R-squared:  0.3499
F-statistic:  2185 on 1 and 4057 DF,  p-value: < 2.2e-16
```

We have only one predictor at the moment so R^2 and adjusted R^2 are the same.

When we include more predictors it is the adjusted R^2 we will use.

Simple Linear Regression



```
Residuals:
     Min       1Q   Median       3Q      Max 
-2.65615 -0.51848  0.01264  0.54399  2.97399 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -0.001191   0.012642  -0.094    0.925    
regression$standlrt  0.595057   0.012730  46.744   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8054 on 4057 degrees of freedom
Multiple R-squared:   0.35,     Adjusted R-squared:  0.3499 
F-statistic:  2185 on 1 and 4057 DF,  p-value: < 2.2e-16
```

Note: The intercept is the constant.

The final thing to look at is 'Coefficients' gives us all the information we need to construct the actual model using unstandardized co-efficients.

The significance levels for each of the main terms (in this case the coefficients associated with the constant and 'standlrt') also tell us if there is evidence that each of these terms are adding something to the model (they are statistically significant).

Using the Coefficients from this table we get the following model:

Predicted 'normexam' = -0.001 + 0.595 x 'standlrt'

Simple Linear Regression

```
Call:
lm(formula =
regression$normexam ~
regression$standlrt)

Standardized Coefficients:
(Intercept) regression$standlrt
0.0000000      0.5916496

Function: lm.beta(model1)
```

Standardised Coefficients:

Not hugely useful for our model since the variables are all standardized.

What is Multiple Regression?

- Simple linear Regression is a model to predict the value of one variable from another.
- Multiple Regression is a natural extension of this model:
    - We use it to predict values of an outcome from several predictors.
    - It is a hypothetical model of the relationship between several variables.

Multiple Regression as an Equation

- With multiple regression the relationship is described using a variation of the equation of a straight line.

$$\boxed{y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n + \epsilon_i}$$

- b_0 is the intercept.
  - The intercept is the value of the Y variable when all Xs = 0.
  - This is the point at which the regression plane crosses the Y-axis (vertical).
- b_1 to b_n are the regression coefficient for variable 1 to n.

Dummy variables

- Many variables we are interested in prediction are categorical.
  - E.g. we are interested in the effect gender or religion or income category has.
- Because they have no scale it makes no sense to think of the effect of a unit of increase in these as it would for a continuous variable.
- But we can think in terms of the differential effect for groupings within the category.
- We can transform the categorical variable into a series of *dummy variables* which indicate whether a particular case has that particular characteristic.
- Dummy variables may also be referred to as indicator variables.

Our regression dataset (Regression.sav)

- Previously looked at standlrt as a predictor of normexam.
- We might also be interested in gender and whether it has an influence.
  - There is significant research that gender has an influence in educational achievement.
- We might also be interested in exploring the type of school a student attends (either mixed-sex or single-sex boys or girls schools) as this might also have an effect on a student's examination performance ('normexam').
- We are interested in exploring if there is a differential effect for students of different genders and for students of different genders attending different types of school.

Dummy Variables

- If we just added them into the model as they are what would happen?
- The values of these two variables would be treated as *real numerical values* rather than just arbitrary numbers representing specific categories.
- So we need to transform these into *dummy variables* before we can add them into the regression model.

Dummy Variables

- Recode to 0 (reference category) and 1 (category of interest).
- Aim is to explore if there is a differential effect for the category of interest when compared to the reference category.
- Indicator variable: Switch effect ON (1) or OFF (0).
- Before including in the regression model need to first establish if this makes sense to include as a predictor.
- Investigate using an independent t-test.

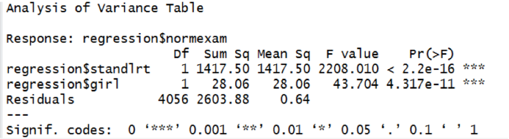
Recoding Variables

$library(car)$
$regression\$gender = recode(regression\$GIRL,' 0 = 1; 1 = 2')$
(this creates a variable gender which recodes GIRL, if 0 gender is set to 1 and if 1 gender is set to 2)

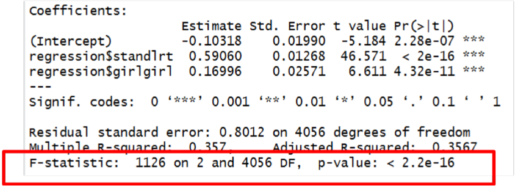
- This assumes that you have called your dataset regression and there is a variable gender (there isn't in our data, it is already coded for us in a variable Girl).
- **This is already done in the dataset you have been given**

Adding Girl to the model

```
model2<-lm(normexam~standlrt+girl ,data=regression)
anova(model2)
summary(model2)
```



```
Analysis of Variance Table

Response: regression$normexam
                    Df  Sum Sq Mean Sq  F value    Pr(>F)    
regression$standlrt    1 1417.50 1417.50 2208.010 < 2.2e-16 ***
regression$girl        1   28.06   28.06   43.704 4.317e-11 ***
Residuals           4056 2603.88    0.64                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

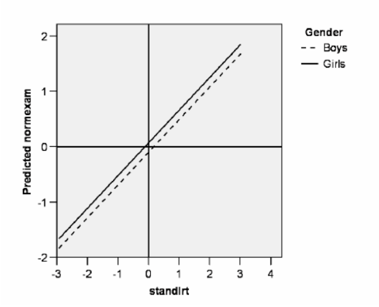


```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -0.10318    0.01990  -5.184 2.28e-07 ***
regression$standlrt  0.59060    0.01268  46.571  < 2e-16 ***
regression$girlgirl  0.16996    0.02571   6.611 4.32e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8012 on 4056 degrees of freedom
Multiple R-squared:  0.357,	Adjusted R-squared:  0.3567 
F-statistic:  1126 on 2 and 4056 DF,  p-value: < 2.2e-16
```

Our model is still significant.

Example

In effect the consequence of adding the dummy variable 'girl' to the model is to create two lines of best fit that have the same gradient (0.590) but different intercepts (constants) (i.e. -0.103 for boys (the constant term) and 0.066 for girls (the constant + the co-efficient for girl).



In other words, and as illustrated here, this model can be represented as two parallel lines with the vertical distance between both lines being 0.170 for the average student (value of 0 for standlrt) when you calculate the equation for boys and girls.

Boys normexam = -.103 +.590 = 0.487

Girls normexam = -.103 + .590 + .167 = 0.654

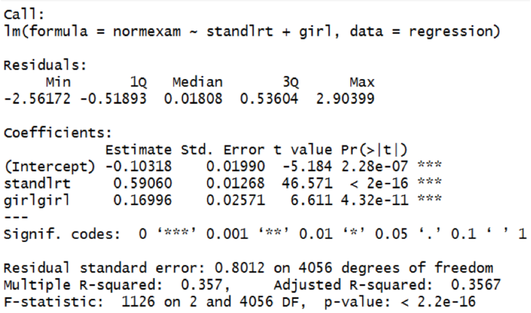
Multiple Linear Regression

Our regression equation:

$$Normexam = -0.10 + 0.59 * standlrt + 0.17 * girl$$

Multiple Linear Regression

To investigate the impact gender is having we include our recoded variable into the model:



```
Call:
lm(formula = normexam ~ standlrt + girl, data = regression)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.56172 -0.51893  0.01808  0.53604  2.90399 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.10318    0.01990  -5.184 2.28e-07 ***
standlrt     0.59060    0.01268  46.571  < 2e-16 ***
girlgirl     0.16996    0.02571   6.611 4.32e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8012 on 4056 degrees of freedom
Multiple R-squared:  0.357,	Adjusted R-squared:  0.3567 
F-statistic:  1126 on 2 and 4056 DF,  p-value: < 2.2e-16
```

Notice the slight changes in the coefficient for standlrt from the first model.

How good a fit is this second model?

- In this case our Adjusted R2 (from the summary) Adjusted R-squared: 0.3567
- So this model explains 35.67% of the variance in comparison to 35% for model1.

Multiple Linear Regression

- Our regression equation:
- $\textit{Normexam} = -0.10 + 0.59 * \textit{standlrt} + 0.17 * \textit{girl}$
- Lets explore what this means:
  - For a boy their score for normexam on average increases by 0.49 from their standlrt score:
    - $\textit{Normexam} = -0.10 + 0.59 * \textit{standlrt} + 0$ #since girl is 0 for boys
  - For a girl their score increases by 0.66 from their standlrt score:
    - $\textit{Normexam} = -0.10 + 0.59 * \textit{standlrt} + 0.17$
    - So being a girl is on average adding 0.17 to your score at age 16.
    - So there is a positive differential effect.

Linear Regression

Four important statistics:

- F statistic:
  - Whether the model as a whole predicts the dependent variable.
  - Its statistical significance is the significance of the model.
- Regression coefficients (Beta values):
  - Measure the strength and direction of relationships between independent variables and the dependent variance.
- Significance scores for the regression coefficients:
  - Tell us whether the contribution of each variable is statistically significant.
- R^2 statistic or Adjusted R^2 Statistic:
  - Measures the model's overall predictive power and the extent to which the variables explain the variation found in the dependent variable.

How to Interpret Beta Values

- Beta values:
  - the change in the outcome associated with a unit change in the predictor.
- Standardised beta values:
  - tell us the same but expressed as standard deviations.
- If we have standardised our variables in advance this will be virtually the same.

R^2 vs Adjusted R^2

- Every time you add a predictor to a model, the R-squared increases, even if due to chance alone. It never decreases.
  - Therefore, a model with more terms may appear to have a better fit simply because it has more terms.
- If a model has too many predictors it begins to model the random noise in the data.
  - This condition is known as overfitting the model and it produces misleadingly high R-squared values and a lessened ability to make predictions.
- Adjusted R^2:
  - Adjusts value of R^2 based on number of variables in the model.
  - Report Adjusted R2 for multiple linear regression.

15mins break

Lab Exercise

- Load Julie Pallant's survey data (data is in Brightspace, documentation is here: `http://spss.allenandunwin.com.s3-website-ap-southeast-2.amazonaws.com/data-files.html`)
- Build the following models including summary and plot:
  - Model 1: Baseline model (for tpcoiss) with optimism (toptim) and social desirability (tmarlow) as predictors.
  - Model 2: Add stress (tpstress).
  - Model 3: Add gender (sex).
- Discuss the models and decide which is the best one.