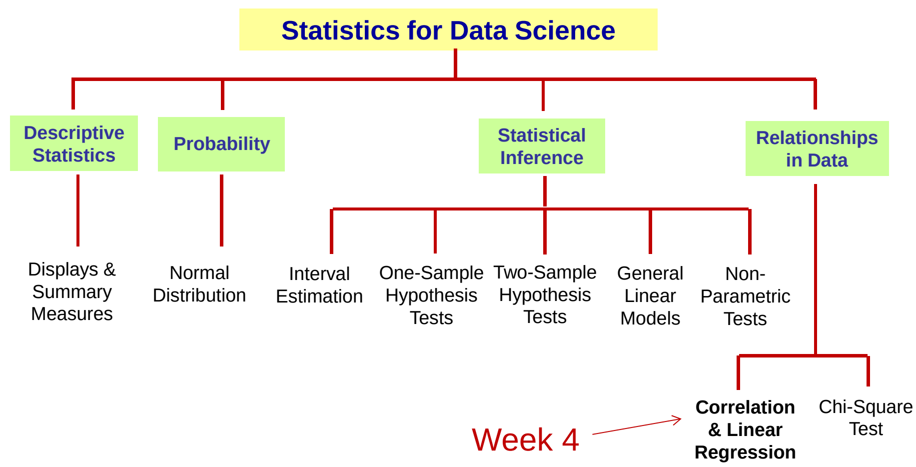
MATH 4044
Statistics for Data Science

Correlation & Regression

Course outline



Field, A & Miles, J, *Discovering Statistics Using SAS*,
Chapter 6 (sections 6.1-6.6) & Chapter 7 (sections 7.1-7.4)

Topics to be covered

- Measuring relationships
  - Pearson's Correlation Coefficient
  - Nonparametric measures: Spearman's Rho and Kendall's Tau
  - Partial correlations
- Simple *linear* regression
  - Modelling data for prediction using one variable.
- Assumption checking for linear regression models.

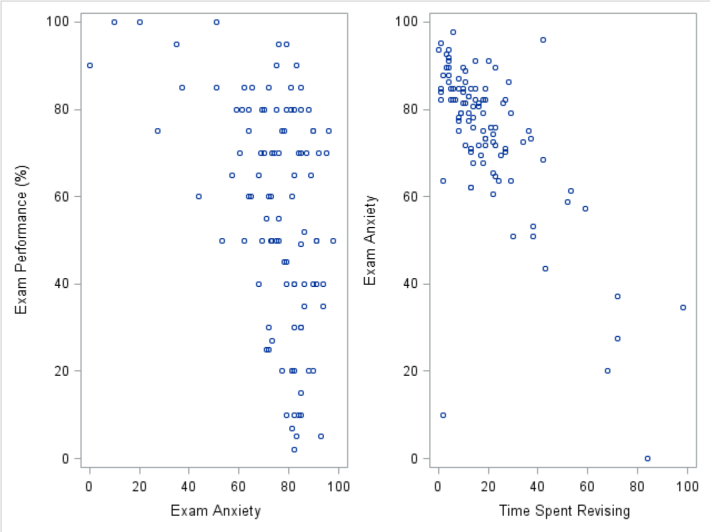
Regression modelling process

Nonlinear or no association	Is there an association between the variables?	*Correlation*
Issues with explanatory variables	Fit a regression model	*Simple or Multiple Linear Regression*
Too many variables? Missing Variable? Transformation needed?	Test the regression model	*Hypothesis Tests and Assumption Checking*
Is the model sensible?	Use the model for prediction	*Interpolation or Extrapolation*

Example: Anxiety and exam performance

- What are the effects of exam stress and revision on exam performance?
- Study participants are 103 students.
- Variables measured:
  - Gender;
  - Time spent revising (hours);
  - Exam performance (percentage score);
  - Exam Anxiety (score out of 100):
    - Based on a purposely developed and validated exam anxiety questionnaire, anxiety measured before an exam.

Example: Scatterplots

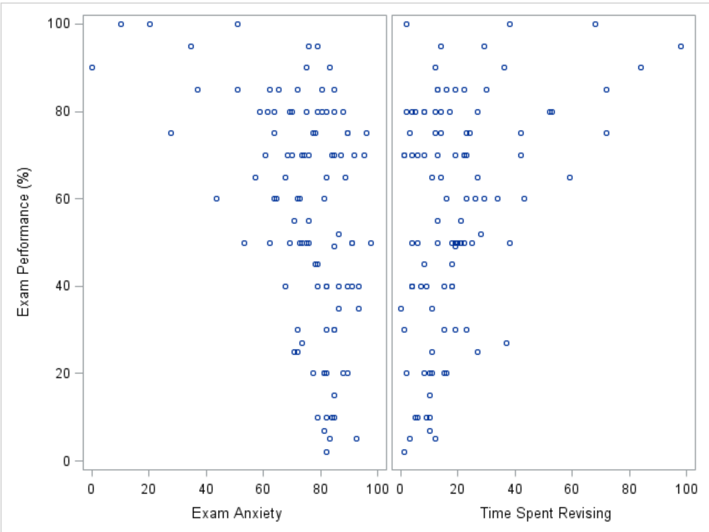


Example: Scatterplots

Using PROC SGSCATTER to display multiple plots on the same page:

```
proc sgscatter data=work.examanxiety;
     plot exam * anxiety anxiety * revise;
run;
```


Example: Scatterplots

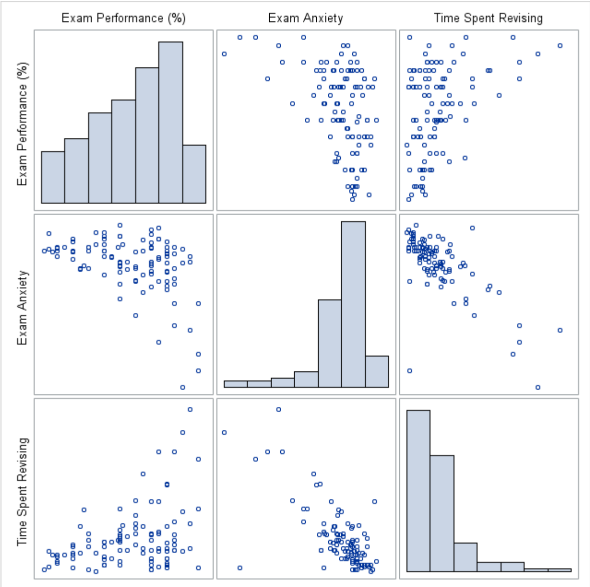


Example: Scatterplots

Using COMPARE statement with PROC SGSCATTER, to plot exam scores against anxiety scores and time spent revising:

```
proc sgscatter data=work.examanxiety;
     compare y=exam x=(anxiety revise);
run;
```


Example: Scatterplot matrix



Example: Scatterplot matrix

Using PROC SGSCATTER to plot every variable against every other variable:

```
proc sgscatter data=work.examanxiety;
     matrix exam anxiety revise /
                    diagonal=(histogram);
run;
```


Correlation analysis

- The consideration of whether there is a relationship or association between two numerical variables is called *correlation analysis*.
- A *correlation coefficient* is an index which defines the strength and direction of the relationship between two numerical variables.
- Visual impression can be formed using a *scatterplot.*
- We will see two types of correlation: Pearson and Spearman.
- The Pearson product moment correlation coefficient (linear correlation coefficient) measures the strength of the linear association between two quantitative variables.
- We use Spearman's Rho for non-parametric statistics.

Pearson correlation coefficient

- The covariance is the average cross-product deviations:

$$\mathrm{Cov}(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

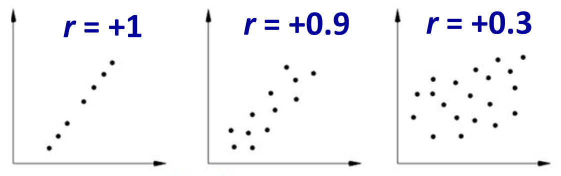
- The correlation coefficient is the standardized version of covariance:

$$r = \frac{Cov(x,y)}{s_x s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

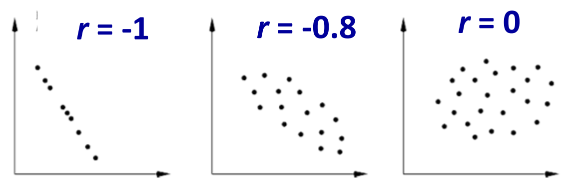
- Correlation coefficient *r* has *no units* and it is always a number *between -1 and 1*.

Positive and negative correlation

- If two variables *x* and *y* are *positively* correlated:
  - large (small) values of *x* are associated with large (small) values of *y*.

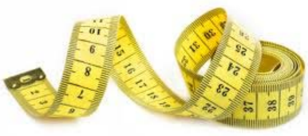


- If two variables *x* and *y* are *negatively* correlated:
  - large (small) values of *x* are associated with small (large) values of *y*.



Interpreting correlation

- Correlation coefficient is an effect size:
  - ±0.1 = small effect;
  - ±0.3 = medium effect;
  - ±0.5 = large effect.



- The square of the Pearson's correlation coefficient is the coefficient of determination R^2:
  - It measures the amount of variance in one variable that is shared by another variable.

Assumptions behind Pearson's *r*

- If we want to establish whether a correlation coefficient is statistically significant, we need the following:
  - We are working with interval variables:
    - Equal intervals on the continuous scale being measured represent equal differences in the property being measured.
  - The sampling distribution is Normal:
    - This can be assumed when both variables are Normally distributed or we have a large sample.

Example: Anxiety and exam performance

The CORR Procedure

3 Variables:	REVISE EXAM ANXIETY

Pearson Correlation Coefficients, N = 103
Prob > |r| under H0: Rho=0

	REVISE	EXAM	ANXIETY
REVISE Time Spent Revising	1.00000	0.39672 <.0001	-0.70925 <.0001
EXAM Exam Performance (%)	0.39672 <.0001	1.00000	-0.44099 <.0001
ANXIETY Exam Anxiety	-0.70925 <.0001	-0.44099 <.0001	1.00000

Significant correlation with the intended response variable means the variable should be included in a regression model.

P-values

$H_0: \rho = 0$
$H_1: \rho \neq 0$
$\alpha = 0.01$

Exam performance is significantly correlated with exam anxiety, r = -0.44, and time spent revising, r = 0.40 (both *P*-values < 0.0001). The time spent revising was also correlated with exam anxiety, r = -0.71 (P-value < 0.0001).

Example: Correlation

Use PROC CORR to obtain all pairwise correlations:

```
PROC CORR data=work.examanxiety nosimple;
      VAR revise exam anxiety;
RUN;
```

Option to suppress simple descriptive statistics output

Example: Anxiety and exam performance

The CORR Procedure

1 With Variables:	EXAM
2 Variables:	ANXIETY REVISE

Pearson Correlation Coefficients, N = 103
Prob > |r| under H0: Rho=0

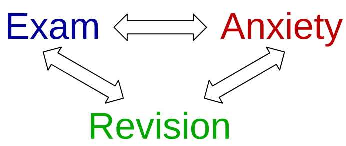
	ANXIETY	REVISE
EXAM		
Exam Performance (%)	-0.44099	0.39672
	<.0001	<.0001

Correlations between every variable in the `var` list with every variable in the `with` list.

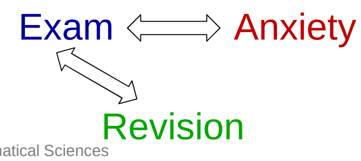
```
proc corr data=work.examanxiety nosimple rank;
     var anxiety revise;
     with exam;
run;
```


Partial and semi-partial correlations

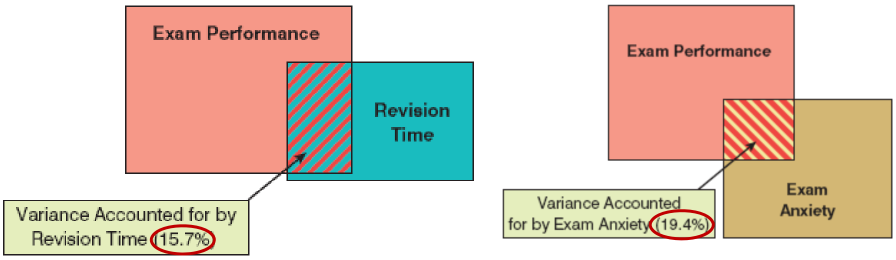
- Partial correlation:
  - Measures the relationship between two variables, controlling for the effect that a third variable has on them both.



- Semi-partial correlation:
  - Measures the relationship between two variables controlling for the effect that a third variable has on only one of the others.

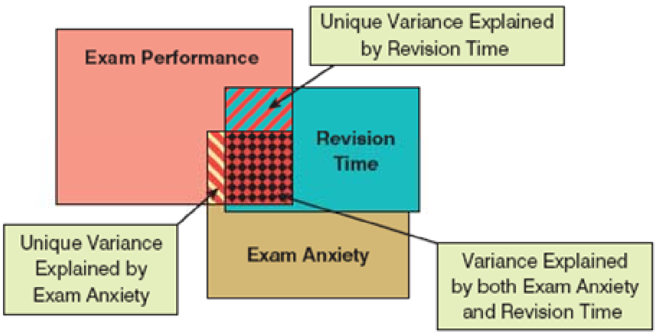


Partial correlations



- R^2 is the coefficient of determination which measures the amount of variance in one variable that is shared by another variable.
- It is the square of the Pearson's correlation coefficient.

Partial correlations – complete picture



Exam anxiety alone does explain some of the variation in exam scores, but there is a complex relationship between anxiety, revision and exam performance that might have otherwise been ignored.

Example: Exam performance

The CORR Procedure

1 Partial Variables:	REVISE
2 Variables:	EXAM ANXIETY

Pearson Partial Correlation Coefficients, N = 103
Prob > |r| under H0: Partial Rho=0

	EXAM	ANXIETY
EXAM	1.00000	-0.24667
Exam Performance (%)		0.0124
ANXIETY	-0.24667	1.00000
Exam Anxiety	0.0124	

The partial correlation between exam performance and exam anxiety is -0.247, which is considerably less when the effect of revision time is not controlled (r = -0.44). This correlation is still statistically significant, but the relationship is diminished.

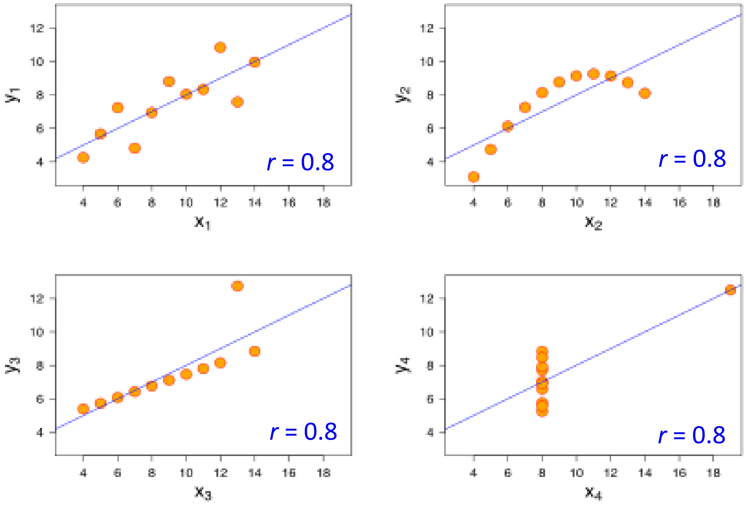
Example: Partial correlation

Use PROC CORR:

```
PROC CORR data=chapter6.examanxiety;
      VAR exam anxiety;
      PARTIAL revise;
      RUN;
```

Anscombe's Quartet

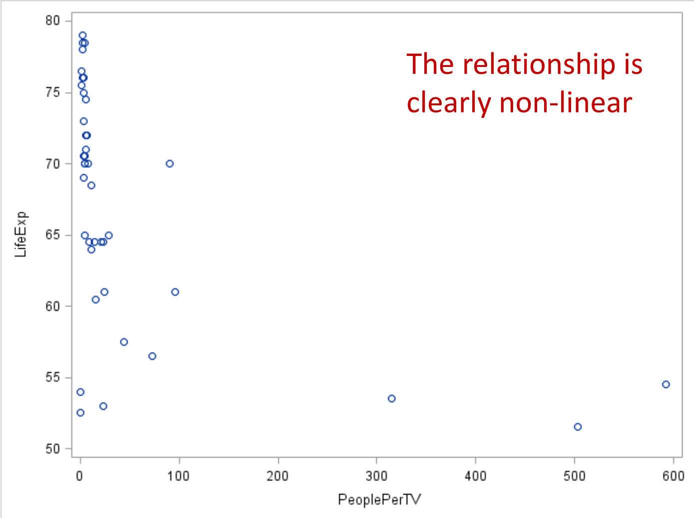
The importance of looking at data



Spearman's correlation coefficient

- The Spearman's rank correlation coefficient measures the strength of curved relationships between two quantitative variables that are strictly increasing or decreasing.
  - Also used when outliers are present.
- It is denoted by r_s or ρ (rho) and calculated by first ranking the data for each quantitative variable and then applying the linear correlation coefficient formula.
- A non-parametric alternative to Pearson's correlation coefficient (also Kendall's Tau for small samples).

Example: TVs and life expectancy



Example: TVs and life expectancy

Spearman Correlation Coefficients, N = 40
Prob > |r| under H0: Rho=0

	LifeExp	PeoplePerTV
LifeExp LifeExp	1.00000	-0.62609 <.0001
PeoplePerTV PeoplePerTV	-0.62609 <.0001	1.00000

Kendall Tau b Correlation Coefficients, N = 40
Prob > |tau| under H0: Tau=0

	LifeExp	PeoplePerTV
LifeExp LifeExp	1.00000	-0.53787 <.0001
PeoplePerTV PeoplePerTV	-0.53787 <.0001	1.00000

H_0: $\rho_S = 0$
H_1: $\rho_S \neq 0$
$\alpha = 0.01$

There is a statistically significant negative non-linear association between life expectancy and the number of people per TV set.

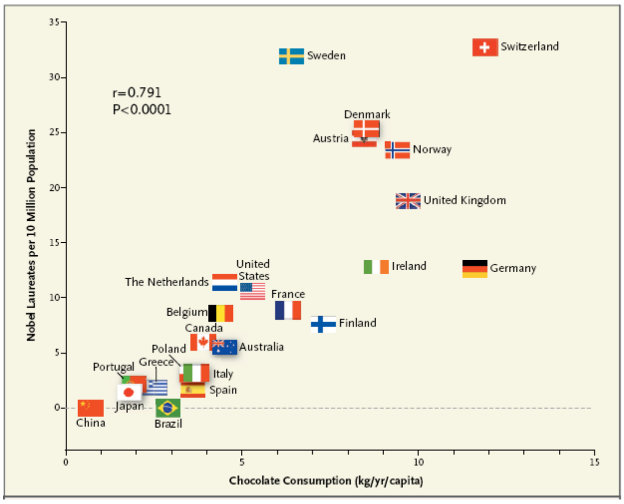
```
proc corr
data=work.life_expectancy_tvs
      spearman kendall;
      var LifeExp PeoplePerTV;
run;
```

Correlation and causality

- If two variables are significantly correlated, this *does not imply* that one must be the cause of the other.

- Does *x* 'cause' *y* ?
  - Temperature and weight of clothing worn?
  - Ice cream sales and number of drownings?
  - Shoe size and spelling ability?
  - Height and salary?

Chocolate consumption, cognitive function and Nobel Laureates



Does chocolate consumption improve the overall cognitive function of a country?

Messerli, FH 2012, 'Chocolate Consumption, Cognitive Function and Nobel Laureates', *The New England Journal of Medicine*, DOI: 10.1056/NEJMon1211064.

Simple linear regression

- Suppose that a scatter diagram shows a reasonably strong, *linear* association between *x* and *y* variables.
- It is then natural to represent linear association by a straight line. A regression model is of the form:

$$\boxed{\text{outcome} = \text{(model)} + \text{error}}$$

- For simple linear regression we have one explanatory variable (x):

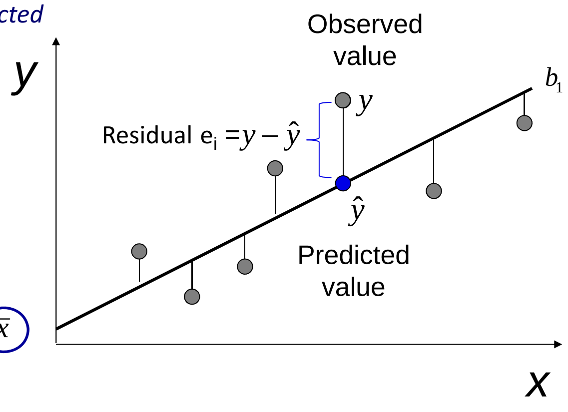
$$\underbrace{\hat{y}}_{\text{outcome/dependent/response}} = \underbrace{b_0 + b_1 \overbrace{x}^{\text{independent/explanatory}}}_{\text{model}} + \underbrace{e_i}_{\text{error}}$$

- We can use this simple linear regression model to make predictions.

Least squares regression

- Minimise the sum of squares of residuals, which are the vertical distances from line to points.

variable to be predicted



slope

$$b_1 = \frac{\sum[(x_i - \bar{x})(y_i - \bar{y})]}{\sum(x_i - \bar{x}^2)}$$

$$\hat{y} = b_0 + b_1 x$$

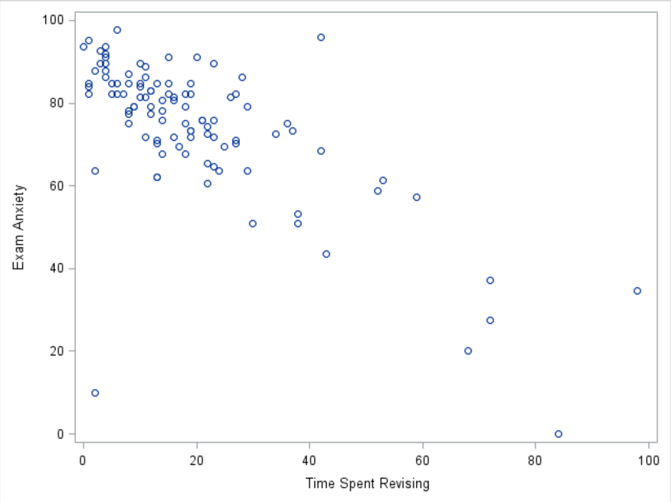
intercept

$$b_0 = \bar{y} - b_1\bar{x}$$

variable used to predict

Is exam anxiety affected by revising?

The relationship appears to be linear.



Is exam anxiety affected by revising?

The REG Procedure
Model: MODEL1
Dependent Variable: ANXIETY Exam Anxiety

Number of Observations Read	103
Number of Observations Used	103

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	15147	15147	102.23	<.0001
Error	101	14965	148.16489		
Corrected Total	102	30112			

Root MSE	12.17230	R-Square	0.5030
Dependent Mean	74.34367	Adj R-Sq	0.4981
Coeff Var	16.37301		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	87.66755	1.78184	49.20	<.0001
REVISE	Time Spent Revising	1	-0.67108	0.06637	-10.11	<.0001

Results for a simple linear regression model

Tasks > Regression > Linear Regression…

Or PROC REG

Dependent variable is *Anxiety* and explanatory variable is *Revise*.

Is exam anxiety affected by revising?

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	87.66755	1.78184	49.20	<.0001
REVISE	Time Spent Revising	1	-0.67108	0.06637	-10.11	<.0001

$$\hat{y} = b_0 + b_1 x$$

$$Anxiety = b_0 + b_1\ Revise$$

$$Anxiety = 87.67 - 0.67\ Revise$$

Interpretation of b_0 and b_1

- The intercept b_0 identifies the value of y when x is zero but it can be meaningless.
- The slope b_1 is the 'rate of change' of y with respect to x.
  - The slope b_1 determines how much the variable y will change when x increases by one unit.
- For the Exam Anxiety vs Revision Time regression model:
- Slope b_1 = -0.67
  - For every unit increase in x (revision time) there is a 0.67 decrease in y (decrease in exam anxiety score).
  - On average, exam anxiety decreases by 0.67 for each 1 hour increase in revision time.
- Intercept b_0 = 87.67
  - When x = 0 (no revision), y = 87.67 (exam anxiety score).
  - On average, exam anxiety score is 87.67 when revision time is 0.

How good is the regression model?

R^2 value (%)	Strength of linear association	Quality of simple linear regression model
>90	Very strong	Excellent
75-90	Strong	Very good
50-75	Reasonable	Good
25-50	Weak	Weak
<25	Very little	Poor

- R^2 is the coefficient of determination which measures the proportion of variance among the original *y* observations, which is 'explained' by the linear regression model that uses *x*.

Coefficient of determination

Root MSE	12.17230	R-Square	0.5030
Dependent Mean	74.34367	Adj R-Sq	0.4981
Coeff Var	16.37301		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	87.66755	1.78184	49.20	<.0001
REVISE	Time Spent Revising	1	-0.67108	0.06637	-10.11	<.0001

Variance explained by the model

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

Total amount of variance

The coefficient of determination R^2 is 50.3%. The line appears to be a *good* fit to the data. Revision time explains 50.3% of variability in exam anxiety scores.

$\hat{y} = b_0 + b_1 x$ *Anxiety* = *87.67 – 0.67* *Revise*

Checking the Linear Regression Model

- The simple linear regression model

$$\hat{y} = b_0 + b_1 x + e_i$$

is a sample-based estimate of the population regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Before using the sample-based model for prediction, test the model against the population regression model to determine if it is valid.
- We need to test the slope β_1 using b_1 and check assumptions.
- If the model passes these tests, we can use it for prediction. Otherwise we may need to revise the model structure.

Inference about the slope β_1: *t*-test

- Null and alternative hypotheses

H_0: $\beta_1 = 0$ (no linear relationship)

H_1: $\beta_1 \neq 0$ (linear relationship does exist)

- Test statistic

$$t = \frac{b_1}{SE_{b_1}} \quad \text{where} \quad SE_{b_1} = \frac{S}{\sqrt{\sum (x - \bar{x})^2}}$$

Standard error

- Confidence interval

$$b_1 \pm t^*_{\alpha/2, n-2} \times SE_{b_1}$$

Critical value

Degrees of freedom

Inference for the slope β_1

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	87.66755	1.78184	49.20	<.0001	84.13285	91.20225
REVISE	Time Spent Revising	1	-0.67108	0.06637	-10.11	<.0001	-0.80274	-0.53942

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

with the test statistic $t_{103-2} = t_{101} = -10.11$ (for n=103)

Since the p-value < 0.0001, we reject H_0. At 5% significance level, we conclude there is a relationship between exam anxiety and time spent revising. The slope is significantly different from zero.

We are 95% confident that the population value of the slope is between -0.803 and -0.539.

We can similarly test the intercept β_0, i.e. H_0: $\beta_0 = 0$, H_1: $\beta_0 \neq 0$

Inference for overall model fit

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	15147	15147	102.23	<.0001
Error	101	14965	148.16489		
Corrected Total	102	30112			

F-ratio $$F = \frac{MS_M}{MS_R}$$

$$\frac{\text{Improvement due to the model}}{\text{Difference between the model and the observed data}}$$

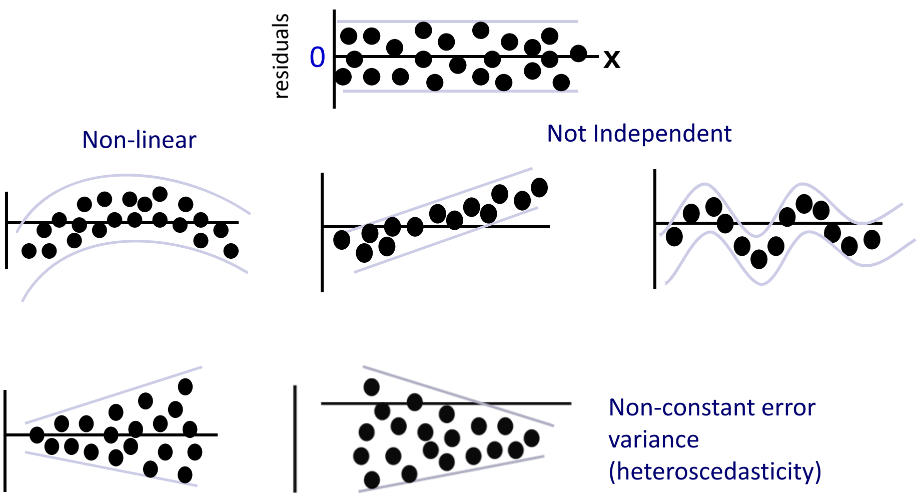
A good model has a large F-ratio and a small P-value.

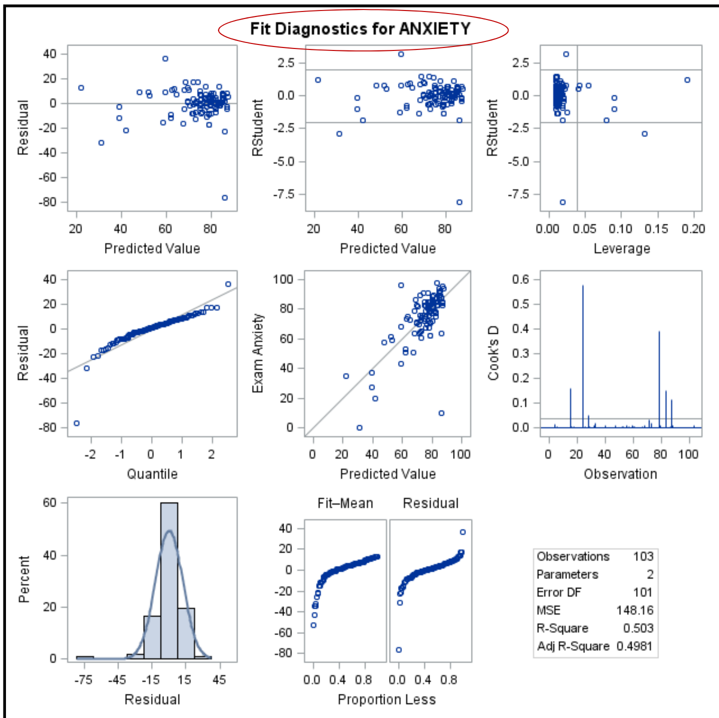
Linear regression assumptions

- Best remembered using the acronym **LINE**:
  - **L**inearity: The relationship between y and x is linear.
  - **I**ndependent errors: the residuals are independent.
    - In particular, repeated observations on the same individual are not allowed.
  - **N**ormality: the residuals are Normally distributed for any given value of x – use a P-P or Q-Q Plot.
  - **E**qual Variance (homoscedasticity): the residuals have constant variance around the 0 line.
- Check these assumptions using plots.

Residual Analysis: L, I and E.

- The plots should resemble random scatter, with no apparent pattern.

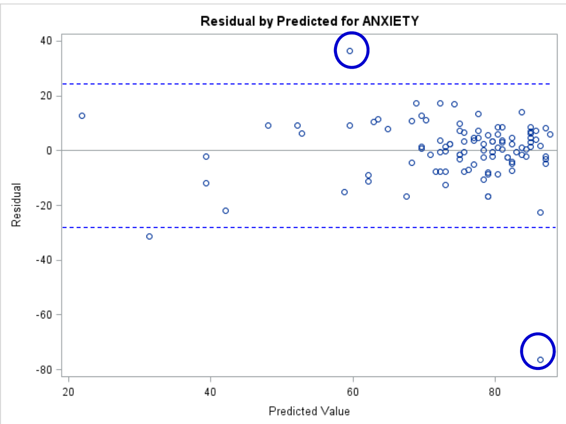




More about these plots later

For now, we focus on the plots in the first column

Assumption checking

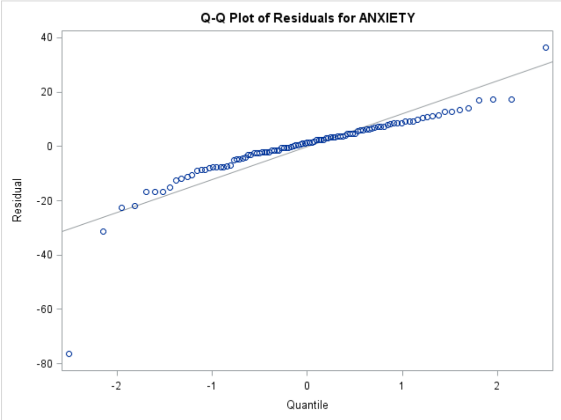


Linearity, Independence, Error Variance:

The residual-versus-fitted values plot shows no apparent pattern so linearity and independence should be OK.

However there is a somewhat un-equal vertical spread from left to right, so constant error variance may not be OK.

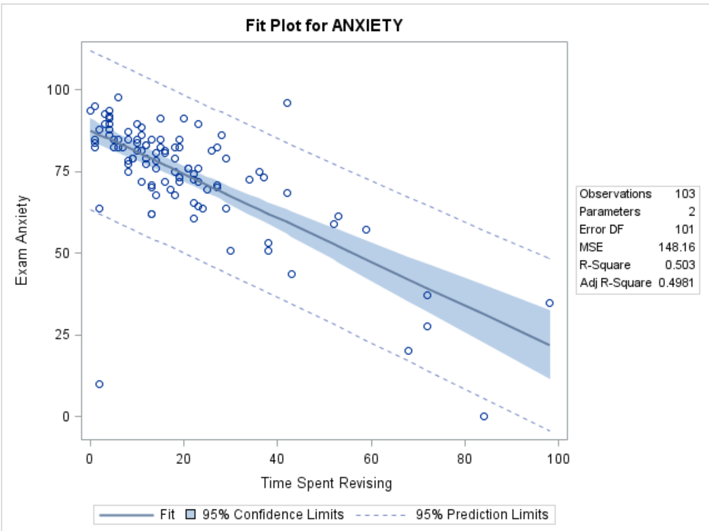
Assumption checking



Normality:

Normal probability plot of residuals shows no curved pattern.

Confidence and prediction limits



Confidence and prediction intervals for regression response

- Confidence interval for the mean response μ_Y when x takes the value x^*:

$$\hat{y} \pm t^* \times SE_{\hat{\mu}}$$

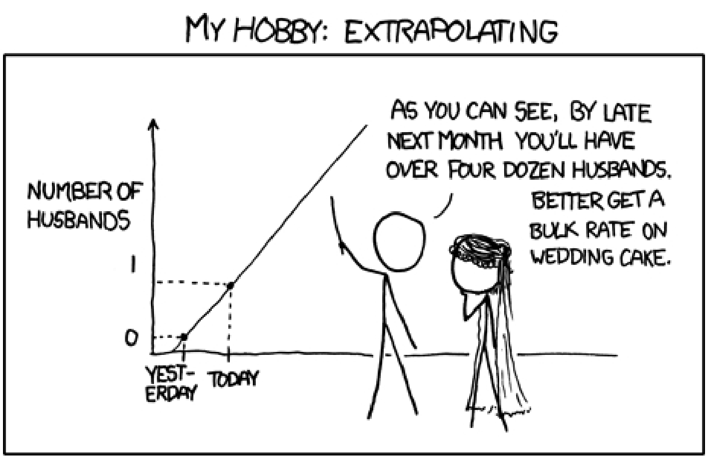
- Prediction interval for a single observation y when x takes the value x^*:

$$\hat{y} \pm t^* \times SE_{\hat{y}}$$

Different standard errors

- The prediction interval is always wider than the confidence interval.
  - Individuals are always more variable than averages.

Making predictions



Making predictions

- We can use the regression model to predict the value of *y* for a specific value of each *x*.
- If the regression line is a poor fit to the data the prediction will be of little use.
- Even if the regression line is a good fit to the data, a prediction can still be 'suspect'.
  - Preferred prediction is based on *interpolation*.
  - It is always dangerous to make a prediction based on *extrapolation.*
  - *Extrapolation* involves at least one *x*-value outside the limits of the *x*-values used in producing the regression model.

Example: Interpolation

- In the original data set:
  - Time spent revising ranged between 1 and 98 hours.
- To predict the anxiety score of a student who spent 20 hours revising, we use our linear regression model:

Anxiety = 87.67 – 0.67 Revise

Anxiety = 87.67 – 0.67 x 20

Anxiety = 74.25

- The predicted anxiety for this individual is 74.25.
- Since R^2 is moderate (50.3%) and this is an interpolation, the prediction is likely to be reasonably trustworthy.

Example: Extrapolation

- What is the predicted anxiety score for a student who spends 120 hours revising?
- Using our linear regression model we now have:

$$Anxiety = 87.67 - 0.67\ Revise$$
$$Anxiety = 87.67 - 0.67 \times 120$$
$$Anxiety = 7.14$$

- The predicted anxiety for this individual is 7.14.
- Since R^2 is moderate (50.3%) and this is an extrapolation, the prediction may not be reliable.