

Lecture 9

Simple Linear Regression: Part 1

Scatter plots - relations between numerical variables

Dependent and independent variables

The least squares regression line

Assumptions of the linear model

Testing the slope of the least squares regression line

In the Last Lecture....

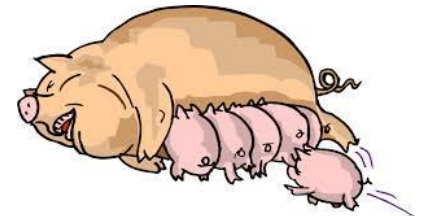
- We compared population means:
 - We used a **paired t-test** (ie. a one sample t-test on the difference between pairs) to compare the average scores for two sets of matched measurements recorded on a target population. The test assumption is that the differences between the pairs are drawn from a normal population.
 - We used a **two sample t-test** to compare the average scores in two independent populations. The assumptions are that both samples are drawn from normal populations **and** that the two populations have the same standard deviation.



Review Quiz 1

Identify an appropriate hypothesis test, and write down a null and an alternative hypothesis for the following:

Two pigs from each of 15 litters are selected. One is given diet A, the other diet B. The pig farmer wants to determine whether the average weight gain will differ for pigs fed on the two different diets.





Review Quiz 2

Identify an appropriate hypothesis test, and write down a null and an alternative hypothesis for the following:

An anthropologist has taken samples of skeletons at sites of two different prehistoric North American tribes. She wants to determine whether there is a difference between the average skeletal heights of females in the two tribes, because this will give valuable indirect information about dietary habits of the two cultures.



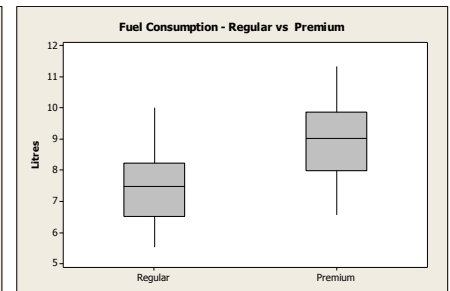
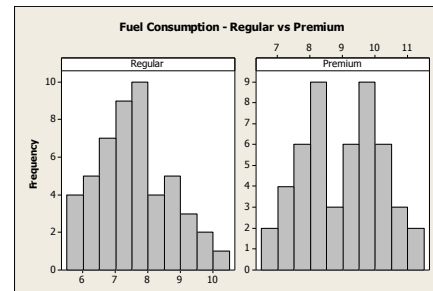
Review Quiz 3

Research Question: Is there a difference between the average fuel consumption for cars using regular unleaded petrol and cars using premium unleaded petrol?

The following output was obtained from a study comparing fuel economy for cars using regular unleaded petrol and cars using premium unleaded petrol. 100 fleet cars were randomly allocated into two groups. 50 cars had their tanks filled with regular unleaded petrol and the other 50 were filled with premium unleaded petrol. Each car was driven 100km and the amount of fuel (in litres) used by each car was recorded. Use the output below, obtained from this study, to write up a hypothesis test to answer the research question.

Two-sample T for Regular vs Premium

	N	Mean	StDev	SE Mean
Regular	50	7.54	1.15	0.16
Premium	50	8.94	1.20	0.17



Difference = μ (Regular) - μ (Premium)

Estimate for difference: -1.396

95% CI for difference: (-1.864, -0.928)

T-Test of difference = 0 (vs not =): T-Value = -5.92 P-Value = 0.000 DF = 98

Both use Pooled StDev = 1.1789

Source: adapted from Black et al, *Australasian Business Statistics*, (2013), Wiley



Solution to Review Quiz 3



Review Quiz 4

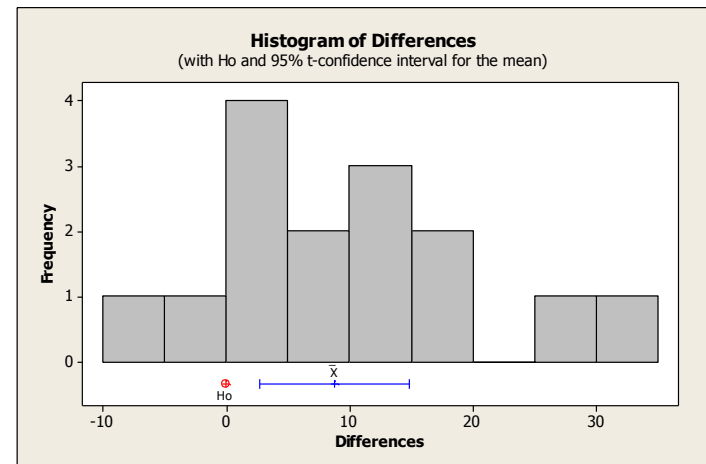
Research Question: Does the oral contraceptive pill incidentally affect the blood pressure of the user?

A study to address this research question involved recording the initial blood pressure of 15 women. After taking the pill regularly for six months blood pressures were again recorded. The following output was obtained from the study. Use this output to write up an appropriate hypothesis test.

Paired T-Test and CI: Before, After

Paired T for Before - After

	N	Mean	StDev	SE Mean
Before	15	75.87	6.86	1.77
After	15	67.07	6.67	1.72
Difference	15	8.80	10.98	2.83



95% CI for mean difference: (2.72, 14.88)

T-Test of mean difference = 0 (vs not = 0): T-Value = 3.11 P-Value = 0.008

Source: Johnson et al, *Statistics Principles & Methods*, (2010), Wiley



Solution to Review Quiz 4

Scatter plots – Relations between numerical variables

Relations between Variables

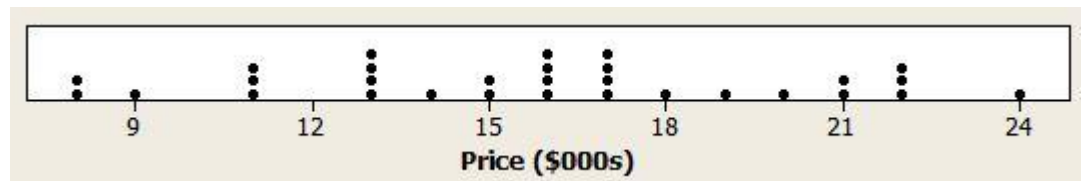
- Say we have two variables X and Y , and both of them are measured on a numerical scale.
- What sort of test would we use to determine the relation between the two numerical variables X and Y ?
- We'll be looking at the relation between two numerical variables in Lectures 9 and 10.

Used Toyota Corollas

Price of 30 Used Toyota Corollas

\$17,000	\$15,000	\$22,000
\$16,000	\$14,000	\$17,000
\$16,000	\$17,000	\$13,000
\$9,000	\$11,000	\$11,000
\$24,000	\$17,000	\$20,000
\$16,000	\$21,000	\$8,000
\$19,000	\$15,000	\$22,000
\$21,000	\$13,000	\$11,000
\$18,000	\$13,000	\$16,000
\$13,000	\$8,000	\$22,000

We would like to estimate the price of a used Toyota Corolla. We have taken a random sample of 30 used Toyota Corollas advertised on carsales.com.au. The best prediction we can make is the sample mean which is \$15,833. We could improve our estimate if we consider any factors that may affect price.



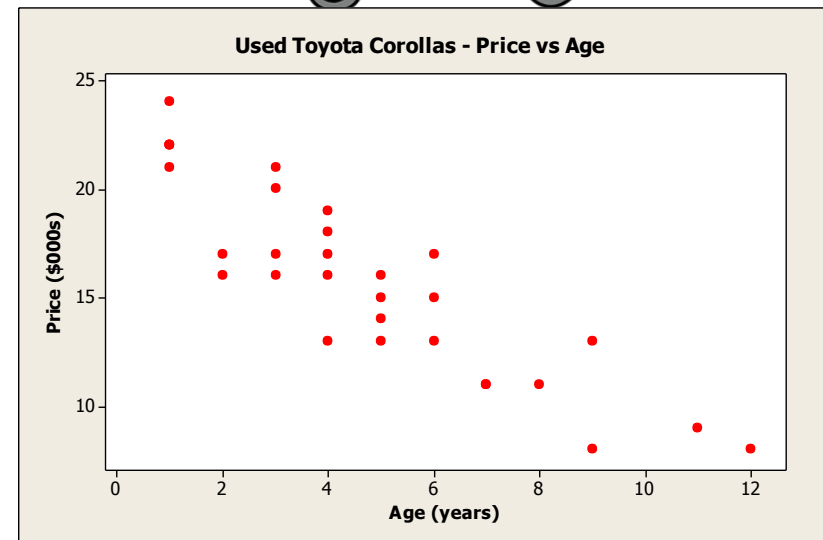
Source: carsales.com.au, 30th January, 2014(selected randomly)

Used Toyota Corollas:

Relation between Price and Age

Here is the sample again, recorded in thousands of dollars. Next to the price of each car, we have recorded its age since age is possibly a determinant of price. Since price and age are numerical variables and we are interested in the relation between these two variables we have constructed a scatterplot with price on the vertical (Y) axis and age on the horizontal (X) axis.

Price (\$000s)	Age (years)	Price (\$000s)	Age (years)	Price (\$000s)	Age (years)
17	2	15	6	22	1
16	3	14	5	17	3
16	4	17	6	13	5
9	11	11	7	11	8
24	1	17	4	20	3
16	2	21	1	8	9
19	4	15	5	22	1
21	3	13	9	11	7
18	4	13	4	16	5
13	6	8	12	22	1



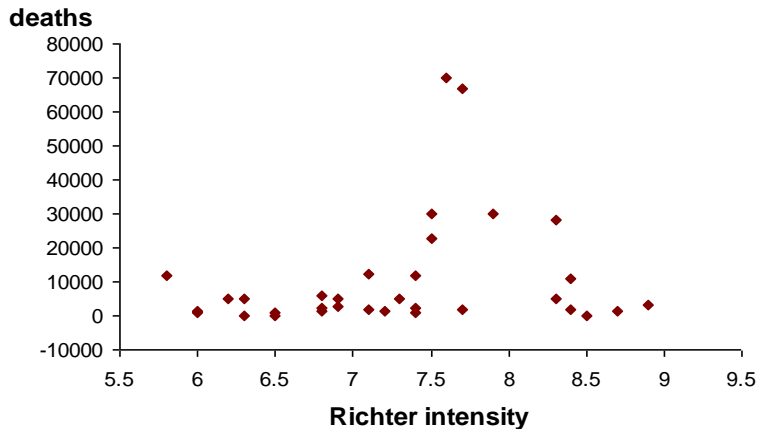


Quiz 5

Comment on any relation you see between the price and the age of used Toyota Corollas:

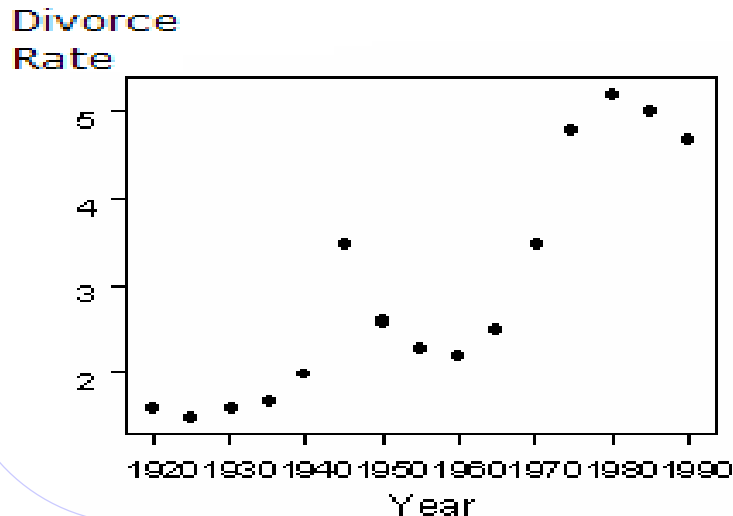
Predict the price of a used ten year old Toyota Corolla:

Scatter Plots



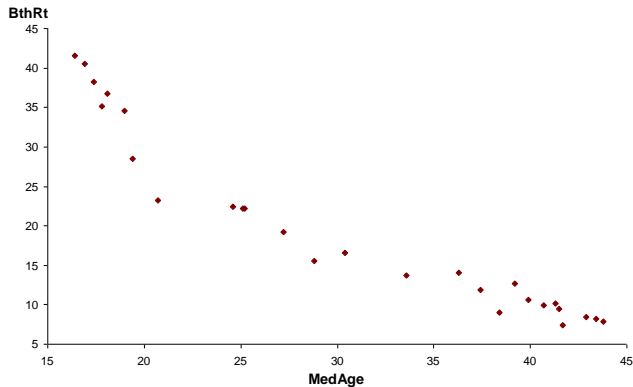
33 major earthquakes between 1932 and 1976

Does a relation exist between the number of deaths and the Richter intensity?

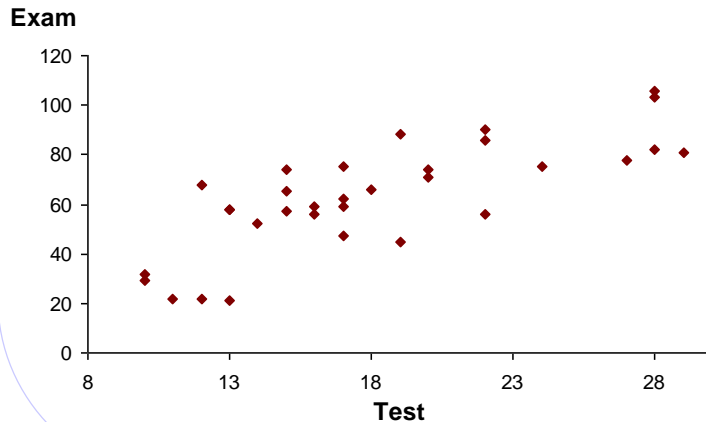


Divorce rates from 1920 to 1990.
Does any relation exist between divorce rate and the year?

Some More Scatter Plots



Birth rate tends to decrease as median age increases ie. there appears to be a negative trend/relation. The relation appears to be curved – still not linear.



Here, exam marks tend to increase with test marks – a positive trend. Students with low test marks tend to gain low exam marks, but not always! Here the relation does appear to be roughly linear.

Dependent and independent variables

Dependent and Independent Variables

In all of the examples we have looked at, we are concerned with identifying the relation between the **outcome** or **response** variable (**Y**) and a **predictor** or **determinant** variable (**X**).

We call the **outcome (Y)** the **dependent** variable and the **predictor (X)** the ***independent*** variable.

On the right we identify the **dependent (Y)** and **independent (X)** variables in each of these examples:

Cars:

Y = price of the car

X = age of the car

Earthquakes:

Y = no. of deaths

X = magnitude

Divorce Rates:

Y = divorce rate

X = year

Countries:

Y = birth rate

X = median age

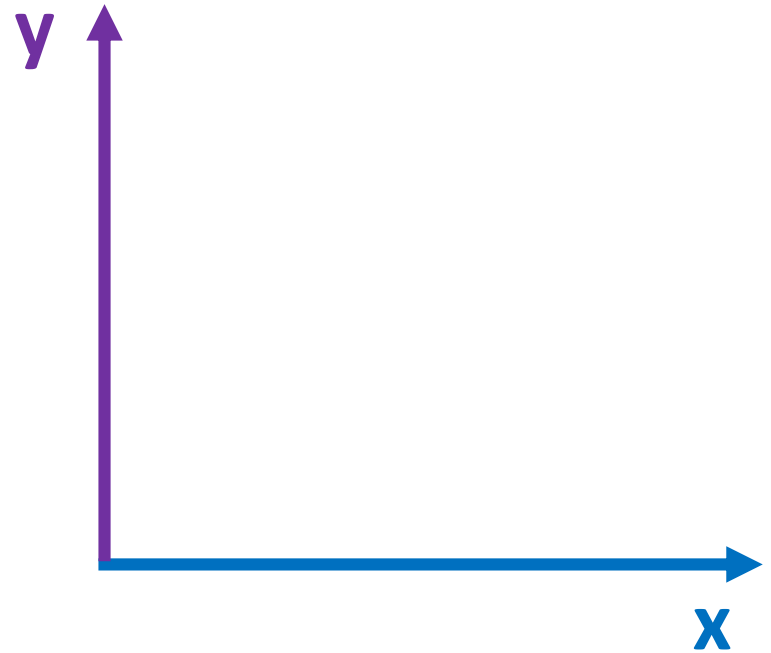
Students' marks:

Y = exam mark

X = test mark

Dependent and Independent Variables

x	y
independent	dependent
predictor	response
determinant	outcome





Quiz 6

Identify the dependent and independent variables in the following situations:

- a. Is the percentage of **pollen** that a queen bumblebee removes from a flower dependent on the **time** (seconds) it spends on the flower?
- b. Is it possible to predict the number of **manatees** killed in Miami harbour from the number of **power boats** registered?
- c. Is a student's **ATAR score** useful in determining his/her Stat170 **exam mark**?

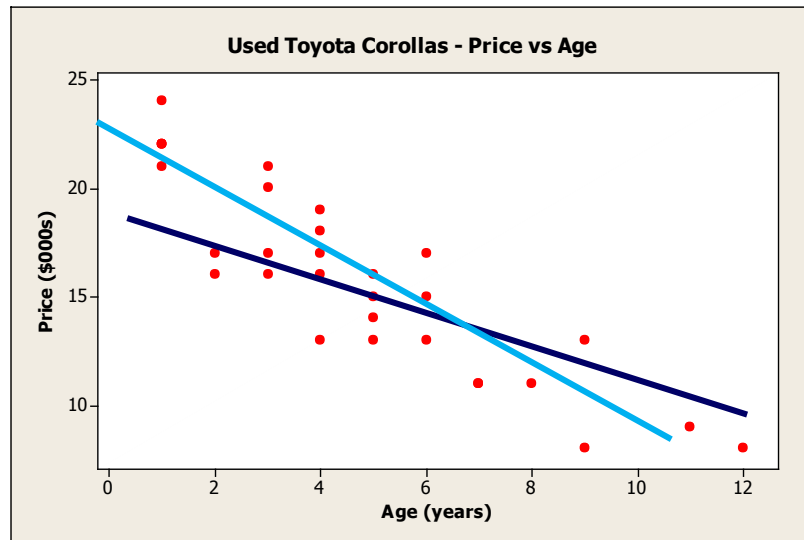
Warning about Causation

Although we call Y the dependent or outcome variable, we do not mean that it is necessarily caused by the independent variable X.

There may be a relation between values of X and Y, but they may be measuring two things that are caused by a third factor, or several other factors. We must always keep this in mind when we interpret statistical results.

The Least Squares Regression Line

Summarising a Linear Relation



Here are two lines drawn freehand that seem to roughly 'fit' the set of points.

There are an infinite number of straight lines that can be drawn through a set of points on a scatter plot.

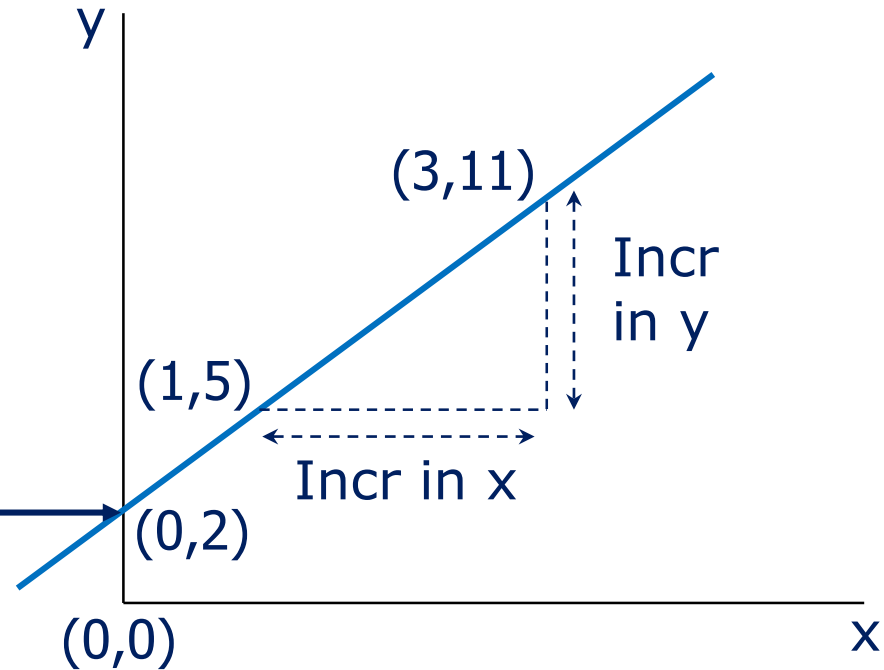
However, there is a '**BEST**' line and **Linear Regression** is the process of fitting *the best* straight line to summarise the relation between two numerical variables.

The Equation of any Straight Line:

$$y = a + bx$$

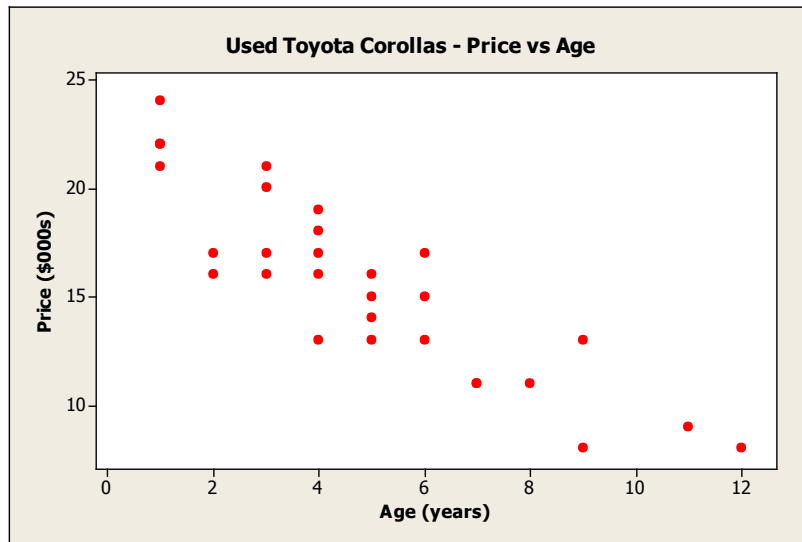
$$\begin{aligned} b &= \text{slope} \\ &= \frac{\text{incr. in } y}{\text{incr. in } x} \\ &= \frac{11-5}{3-1} \\ &= 6/2 = 3 \end{aligned}$$

$$\begin{aligned} a &= \text{y-intercept} \\ &= 2 \end{aligned}$$



$$\begin{aligned} \text{Equation of line: } y &= a + bx \\ y &= 2 + 3x \end{aligned}$$

Line of Best Fit: the Least Squares Regression Line: $\hat{y} = a + bx$



We do not know that any of the points in the scatter plot will necessarily sit on the best fitting line so we use linear regression to fit the line of best fit ie. the '**least squares regression line**'.

The '**least squares regression line**' is the line that makes the **total squared residual distances** as small as possible.

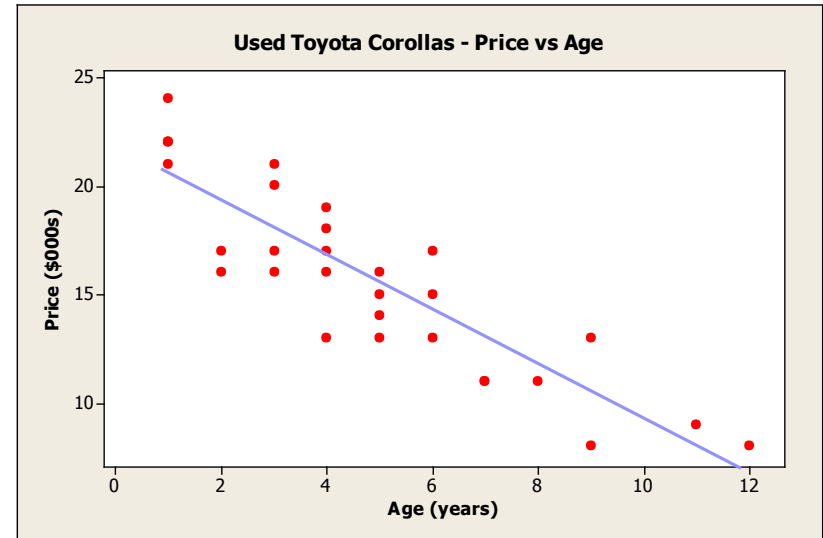
Residuals

The *residuals* in a scatter plot are the distances between the points and the line. They are obtained by subtracting the predicted values from the responses, that is,

$$\text{residual} = y_i - \hat{y}_i$$

where y_i = observed value
 \hat{y}_i = predicted value
(from the line)

These residuals can be positive or negative.



So the residuals measure the *discrepancy* between the data and the fitted line.

The Best Fitting Line

The 'best' line will be the line which has the:

- sum of its squared residuals as a minimum and
- sum of its residuals equal to zero,

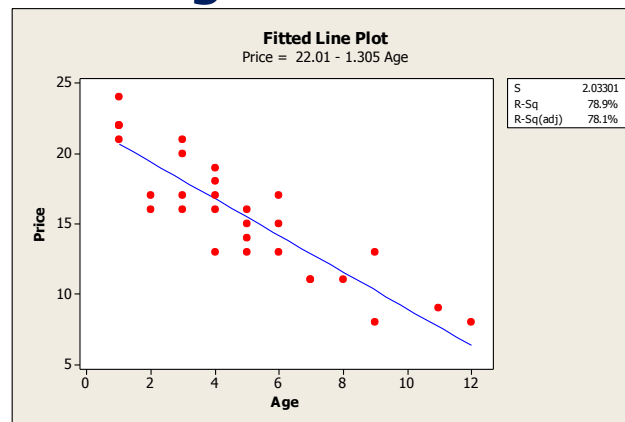
$$\sum_i (y_i - \hat{y}_i)^2 \text{ is a minimum and } \sum_i (y_i - \hat{y}_i) = 0$$

The ***least squares regression line*** (LSR) gives the *minimum possible* sum of squared residuals. It has the equation:

$$\hat{y} = a + bx$$

Using Minitab to Fit the Least Squares Regression Line

In STAT170 we will use Regression Analysis in Minitab to find the least squares regression line:



The least squares regression line fitted by Minitab is:

$$\hat{y} = 22.01 - 1.31x$$

This is the line which fits the line to the plot such that:

- the sum of the squared residuals is a minimum and
- the sum of the residuals is equal to zero.

Interpreting the Least Squares Regression Line

The least squares regression line has the equation: $\hat{y} = a + bx$
To interpret the relation between X and Y we need to look at the slope of the line, b:

For every one unit increase in X, we expect Y, the dependent variable, to change by b units, on average.

Y will increase if the slope, b is positive.

Y will decrease if the slope, b is negative.

For the car data we have the LSR line:

$$\hat{y} = 22.01 - 1.31x$$

Interpretation: For every extra one year in the age of a used Toyota Corolla, we expect the price to go down by an average of \$1310 on average.

Interpreting the Least Squares Regression Line for the Car Data

$$\hat{y} = 22.01 - 1.31x$$

For every extra one year in the age of a used Toyota Corolla, we expect the price to go down by an average of \$1310 on average.

Let's compare the price of a 5 year old car and 6 year old car:

$$\text{When } x = 5, \hat{y} = 22.01 - 1.31 \times 5 = \$15460$$

$$\text{When } x = 6, \hat{y} = 22.01 - 1.31 \times 6 = \underline{\$14150}$$

$$\text{Difference} = \$1310$$



Quiz 7

Minitab can plot the line on the scatter plot. If you needed to draw the line onto the scatter plot, you would need to plot two points on the line and then draw the line through these points. To find two points on the line, take any two easy values of x , one near each end of the range and calculate the value of \hat{y} . You don't have to use actual x -values from the data to draw the line.

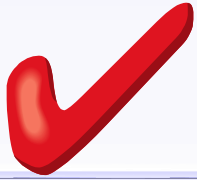
eg. to plot the LSR line for the age and price of the cars:

$$\text{When } x = 2, \hat{y} = 22.01 - 1.31 \times 2 = 19.4$$

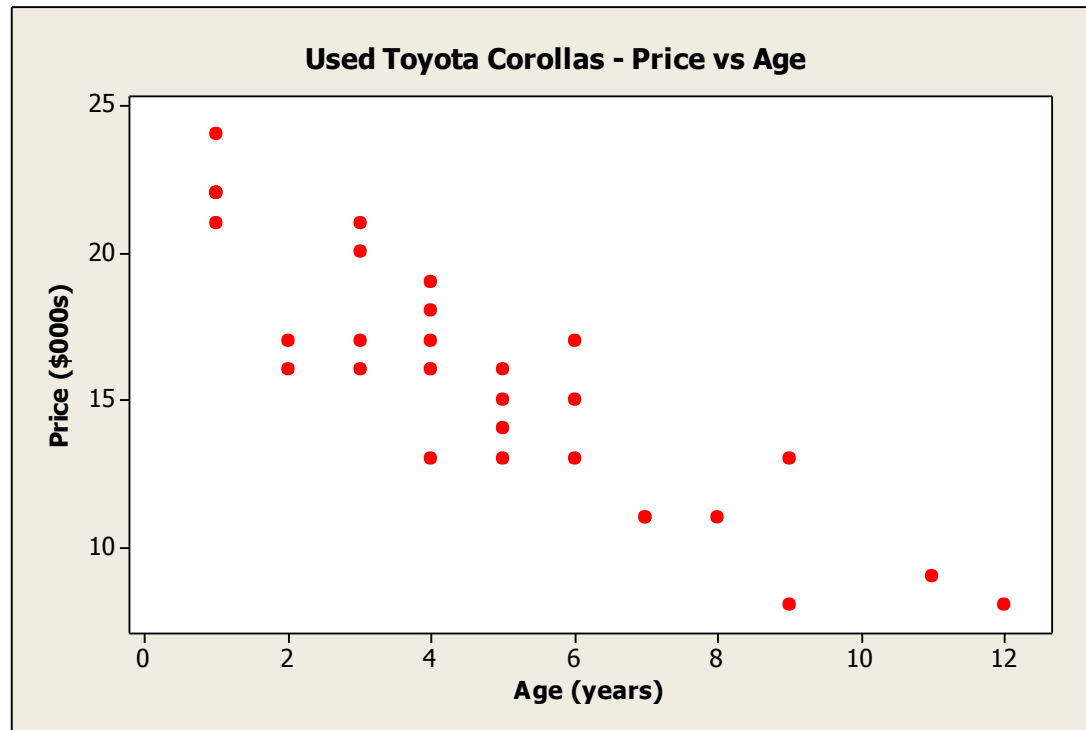
$$x = 10, \hat{y} = 22.01 - 1.31 \times 10 = 8.9$$

Plot the points $(2, 19.4)$ and $(10, 8.9)$ on the scatter plot and draw the line through them.

Check that the point $(\bar{x}, \bar{y}) = (4.7, 15.8)$ lies on the line.



Solution to Quiz 7



Assumptions of the linear model (ie. the least squared regression line)

Assumptions for a Linear Model

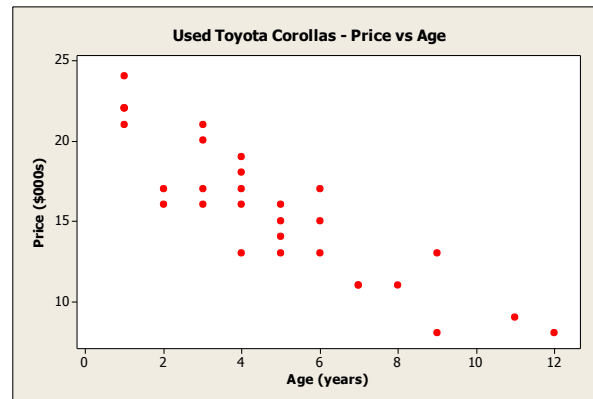
In fitting a linear model (ie. the least squared regression line) we must be able to assume that:

1. The relation in the population is *linear*. To check this, look at a scatter plot to see whether a linear model is a sensible model for the data.
2. The residuals (errors) are drawn from a *normal* distribution. To check this, look at a histogram of the residuals which we can obtain from Minitab.
3. The residuals (errors) in y (ie. the deviations from the line in the y direction) have a *constant standard deviation*. To check this, look at a plot of the residuals showing the spread of the residuals around zero. Minitab calls this plot 'Residuals vs fits' and plots the residuals against the fitted values (ie. the predicted values). If the residuals have a constant standard deviation, the spread of the residuals around zero (the horizontal line in the plot) will be constant across the range of the predicted values.

Checking the **First** Assumption for a Linear Model

We have three assumptions to check so we need to look at three different plots – one to check each assumption. We really should look at these **before** using the model!

Linearity: We have already looked at a **scatter plot** of the car data and noted that the relation between price and age in the sample appears linear. *This suggests that it should be reasonable to assume that the relation between price and age in the population may be linear.*

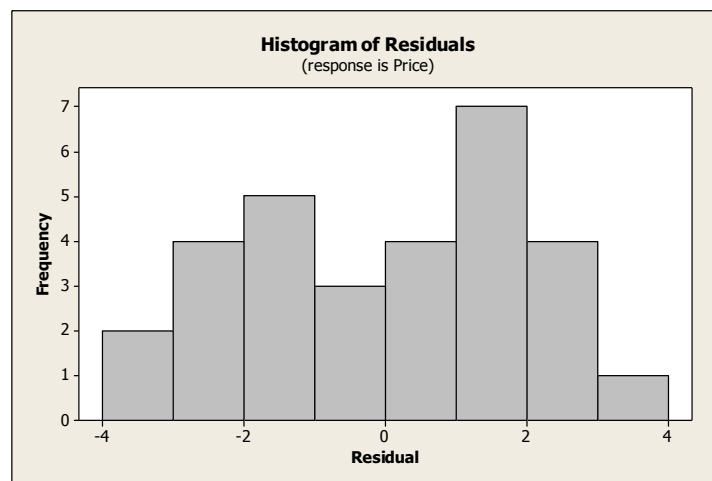


Checking the **Second** Assumption for a Linear Model

We also need to check that the **residuals are from a normal distribution**.

We can obtain a **histogram of the residuals** from Minitab, as shown below.

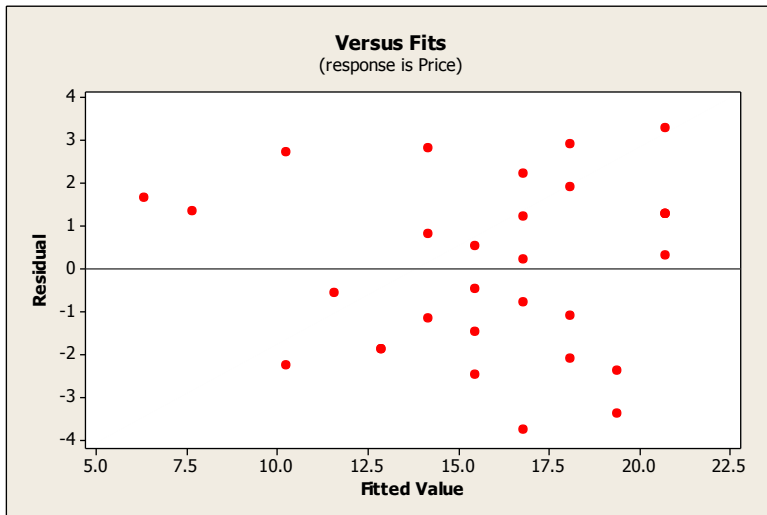
The histogram of the residuals below indicates that, for the car data, it is reasonable to assume that the residuals are drawn from a normal distribution.



Checking the **Third** Assumption for a Linear Model

Lastly, we need to check that the **residuals have a constant standard deviation**.

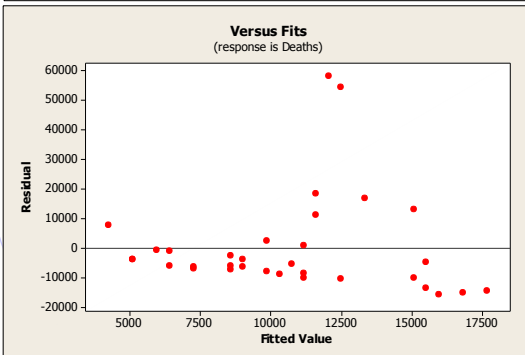
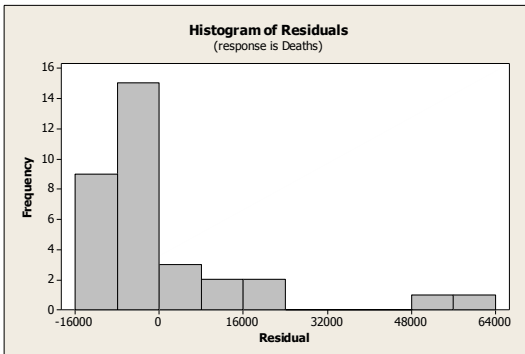
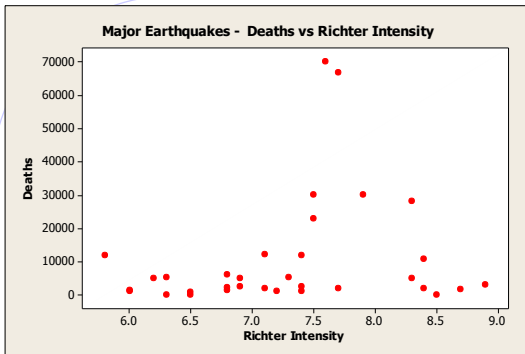
Minitab will provide the 'Residuals vs fits' plot which plots the residuals against the fitted values (ie. the predicted values). If the residuals have a constant standard deviation, the spread of residuals around zero (the horizontal line in the plot) will be constant across the range of the predicted values.



The plot on the left shows the residuals vs the fitted values for the car data. The residuals are fairly evenly scattered around zero across the range of the fitted values. There may be a little more scatter at the high end than the low end but it is probably still reasonable to assume that the residuals will be fairly evenly spread across the range of the data.

Earthquake Data

So a linear model is not appropriate for modelling the earthquake data



We have used Minitab to fit a linear model for the earthquake data. The regression equation is:

$$\text{Deaths} = -20914 + 4336 \text{ Richter Intensity}$$

The scatter plot and residual plots are shown on the left.

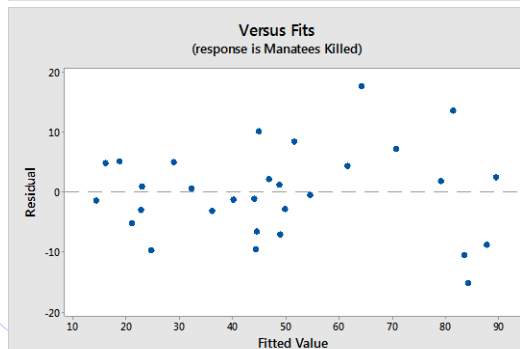
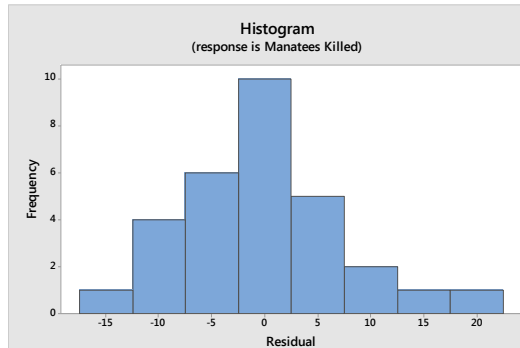
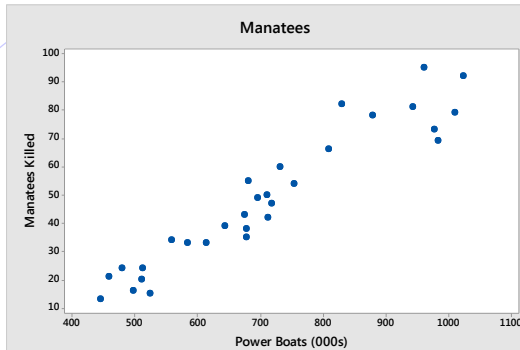
The **scatter plot** clearly does **not** indicate a **linear** relationship.

The **histogram** of the **residuals** is very right skewed, indicating that the residuals are **not** drawn from a **normal** distribution.

The plot of the **residuals versus the fitted values** does **not** show an even scatter of points around zero across the range of the fitted values, indicating that the residuals do not seem to have a **constant standard deviation**.



Quiz 8



The manatees data were recorded in Miami harbour over a 30 year period ($n = 30$). A linear model has been fitted to determine whether the number of manatees killed in a year can be predicted by the number of boats registered in that year. The model is:

$$\text{Manatees Killed} = -43.7 + 0.1301 \text{ Boats (000s)}$$

Use these plots to comment on the assumptions for a linear model and, if a linear model appears to be appropriate, interpret the least squares regression line:

Testing the slope of the line










Relation Between X and Y in a Sample: Estimating Relation in a Population

- The relation in the population is *fixed* and takes the form $Y = \alpha + \beta x$, where α and β are fixed population parameters.
- The relation in the sample is $\hat{y} = a + bx$, and *varies* from sample to sample.
- The sample estimates of α and β are a and b which may vary from sample to sample.
- We focus our interest on the slope, b , since that is what tells us whether there is a relation between X and Y.
- **b** is the estimated change in Y for every one unit change in X.

Sample Statistics

estimate

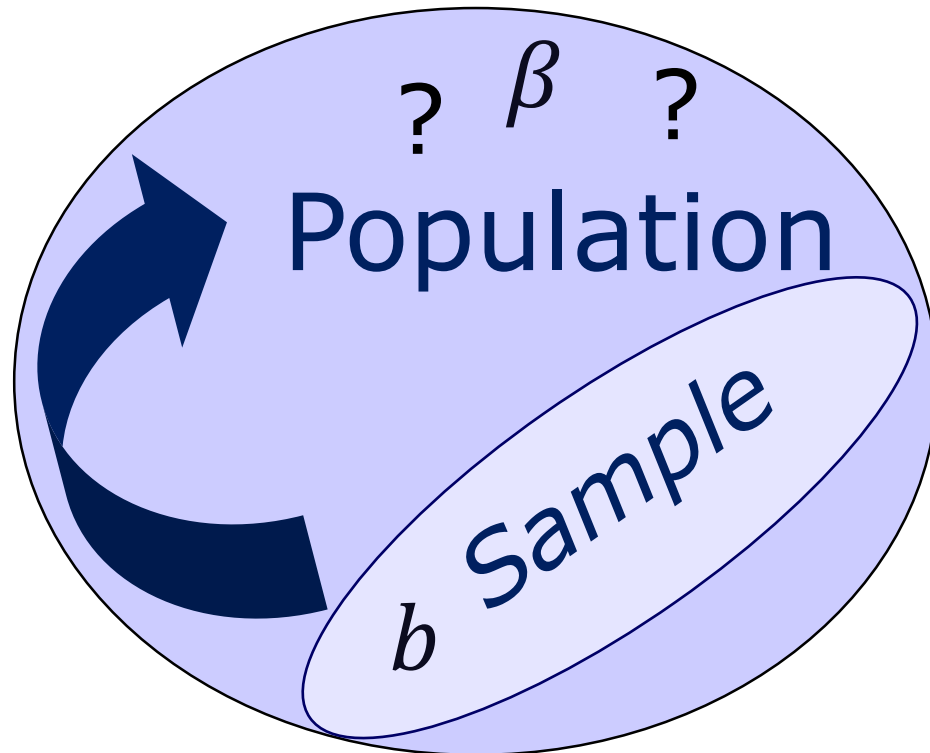
Population Parameters

Mean	\bar{y}		μ
Median	\tilde{y}		$\tilde{\mu}$
Std.dev	s		σ
Variance	s^2		σ^2
se (mean)	s/\sqrt{n}		σ/\sqrt{n}
Proportion	p		π
se (p)	$\sqrt{\frac{p(1-p)}{n}}$		$\sqrt{\frac{\pi(1-\pi)}{n}}$
equation	$\hat{y} = a + bx$		$Y = \alpha + \beta x$
Slope	b		β

Answering Research Questions

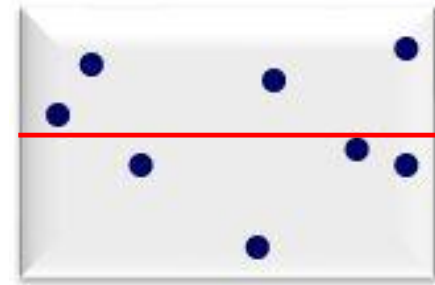
Research Questions

We use a
SAMPLE to
answer
a question
about a target
POPULATION

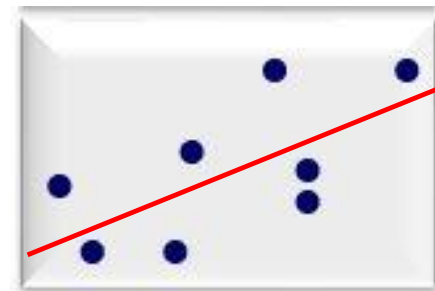
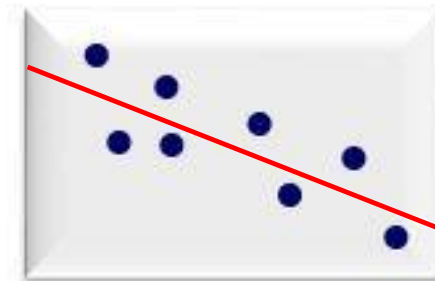


Hypothesis Test: Testing for a Linear Relation

If there *is no* linear relation between X and Y , then β must be equal to 0, and it would follow that $Y = \alpha = \mu_y$, giving the same expected value of Y for every value of X .



However, if there *is* a linear relation between X and Y , then β is not equal to 0 and it would follow that $Y = \alpha + \beta x$, in which case Y changes as X changes.



Hypothesis Test: Testing for a Linear Relation

To test for a linear relation between X and Y we need to test the slope of the line. The slope of the line in the sample, b , is an estimate of the population slope, β . We use the sample slope, b , to test whether the population slope, β , is significant.

The null and alternative hypotheses are:

$H_0: \beta = 0$ (ie. population slope = 0 ie. **no** significant linear relation between X and Y ie. X is **not** a useful predictor of Y)

$H_1: \beta \neq 0$ (ie. population slope $\neq 0$ ie. there **is** a significant linear relation between X and Y ie. X **is** a useful predictor of Y)

This test follows a **t -distribution with $(n-2)$ degrees of freedom.**

The test statistic is $t = \frac{b - \beta_0}{se_b}$ and since our hypotheses have the null value, $\beta_0 = 0$, the test statistic becomes $\frac{b}{se_b}$.

If the null hypothesis is rejected, there is a statistically significant relation between the two variables, X and Y and we can say that there is evidence indicating that X is a useful predictor of Y .

Regression Output from for the Car Data

Regression Analysis: Price versus Age

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.03301	78.89%	78.13%	75.72%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	22.011	0.709	31.05	0.000	
Age	-1.305	0.128	-10.23	0.000	1.00

Regression Equation

Price = 22.011 - 1.305 Age

Example: Used Toyota Corollas

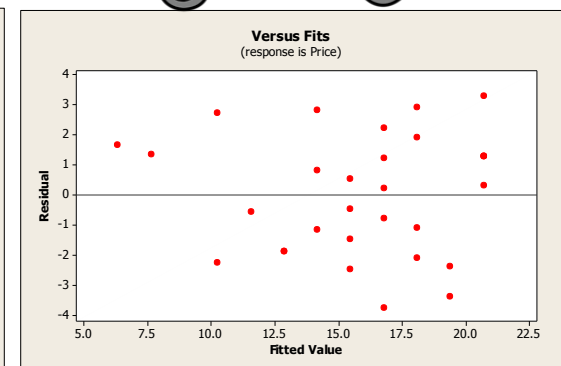
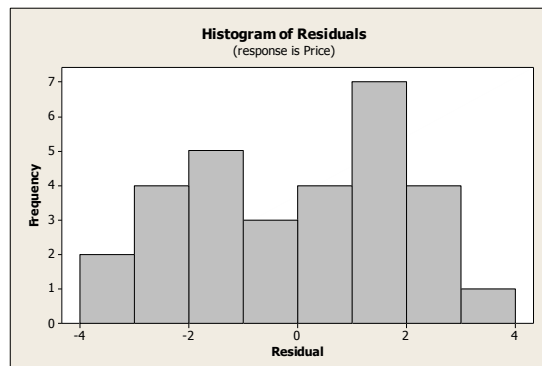
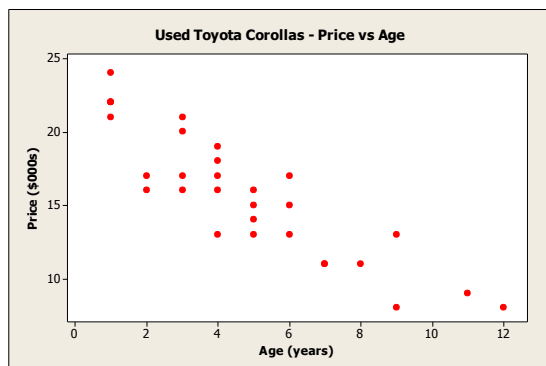
Some Sample Information:

Variables: Y = price X = age

n=30 (not given on the Minitab output, but we knew that from the description of the data set earlier)

Minitab Output:

Predictor/Term	Coef	SE Coef	T-Value	P-Value
Constant	22.0110	0.7089	31.05	0.000
Age	-1.3051	0.1276	-10.23	0.000



Example: Used Toyota Corollas

Hypothesis Test

H

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

A

The scatter plot indicates the relation may be linear in the population. The histogram indicates the residuals could be from a normal distribution. The residuals vs fits plot indicates the residuals could have a constant standard deviation.

T

$$t = \frac{b}{se_b} = \frac{-1.31}{0.1276} = -10.23 \text{ with } n - 2 = 28df$$

p-value ≈ 0

P

Since p-value < 0.05 , reject H_0

D

There is evidence of a negative linear relation between the price and the age of used Toyota Corollas. For each extra year in age, prices decrease by \$1310, on average

C

95% Confidence Interval for β

We can calculate a 95% confidence interval for β , the population slope. Minitab does not give this confidence interval with the regression output but it is useful because it estimates the average change in the Y for each one unit change in X.

95% CI for $\beta = b \pm t_{crit} \times se_b$ where t_{crit} is the critical value which cuts off 5% in the two tails of the t-distribution with $n - 2$ df and se_b is the estimated standard error for the slope which we will get from Minitab.

To estimate the slope of the regression line in the target population:

$$95\% \text{ CI for } \beta = -1.31 \pm 2.048 \times 0.1276 = (-1.57, -1.05)$$

We can be 95% confident that for each extra year in age, the average price of used Toyota Corollas decreases between \$1050 and \$1570.

(Note the CI does not contain 0, confirming the decision to reject the null hypothesis that the population slope was 0)



Quiz 9

Research Question: Is it possible to predict the number of manatees killed in Miami harbour from the number of power boats registered?

The scatter plot and residual plots from the Manatees study are reproduced on the right. The Minitab output for the regression analysis is also provided below. Use any of this output to carry out an appropriate hypothesis test to answer the research question. (remember $n = 30$)

Regression Analysis: Manatees Killed versus Power Boats (000s)

Model Summary

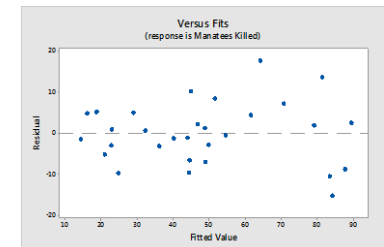
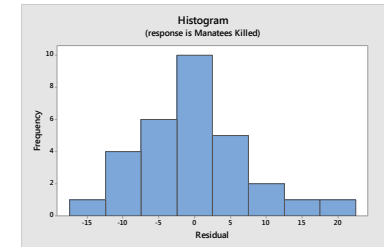
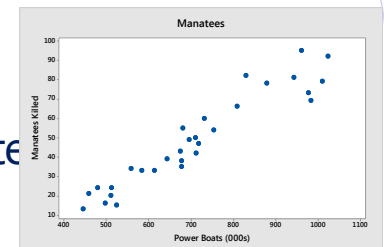
S	R-sq	R-sq(adj)	R-sq(pred)
7.53233	90.61%	90.28%	88.99%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-43.70	5.78	-7.55	0.000	
Power Boats (000s)	0.13010	0.00791	***	*****	1.00

Regression Equation

Manatees Killed = $-43.70 + 0.13010$ Power Boats (000s)





Solution to Quiz 9



Quiz 10

For the manatees data, calculate and interpret a 95% confidence interval to estimate the slope of the line in the target population.
(remember $n = 30$)

(Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-43.70	5.78	-7.55	0.000	
Power Boats (000s)	0.13010	0.00791	16.44	0.000	1.00



Quiz 11

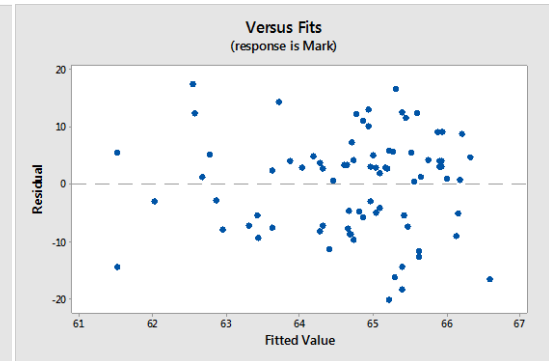
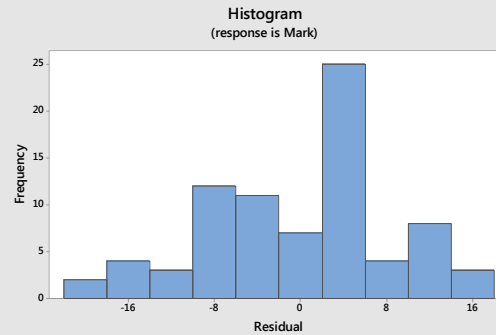
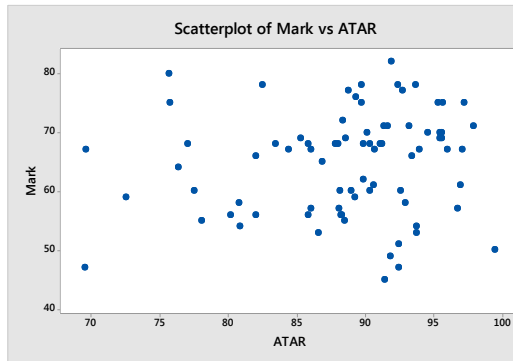
Research Question: Is there a linear relation between a student's ATAR score and their mark in an Introductory Statistics exam?

Use the output on the following slide, obtained from a sample of 79 Introductory Statistics students, to answer the research question.





Quiz 11 Solution



Regression Analysis: Mark versus ATAR

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
8.71799	1.66%	0.38%	0.00%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	49.7	13.2	3.76	0.000
ATAR	0.170	0.149	****	*****



Quiz 11 Solution

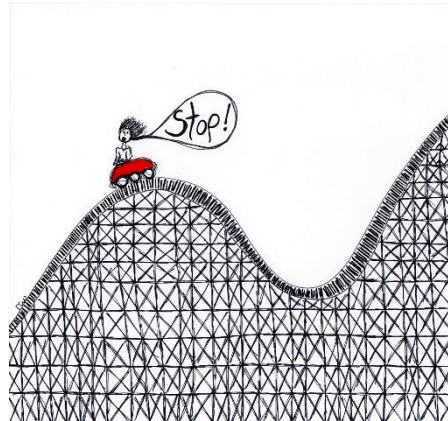
Homework Questions



Homework Question 1

Research Question: Is there a linear relation between the maximum height and the top speed of roller coasters?

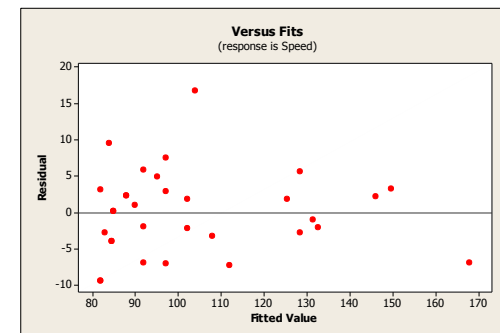
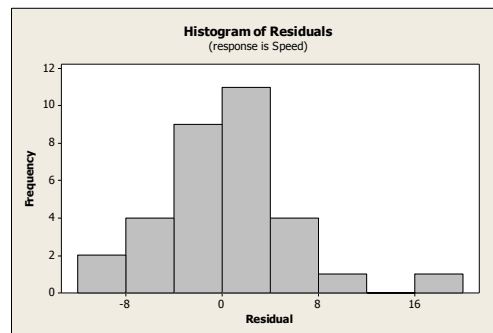
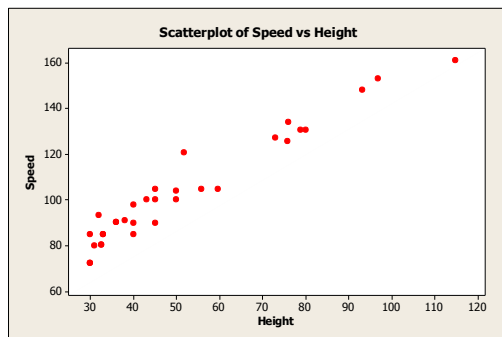
Recall the roller coaster data from practical classes. The maximum height (in metres) and the maximum speed (in kilometres per hour) were recorded for a sample of 32 roller coasters. Use the Minitab output obtained from this sample which is provided on the following page to carry out an appropriate hypothesis test to address this research question.





Homework Question 1

Research Question: Is there a linear relation between the maximum height and the top speed of roller coasters?



Regression Analysis: Speed versus Height

The regression equation is
 $\text{Speed} = 51.4 + 1.01 \text{ Height}$

Predictor	Coef	SE Coef	T	P
Constant	51.429	2.590	19.85	0.000
Height	1.01283	0.04612	*****	*****

S = 5.83160 R-Sq = 94.1% R-Sq(adj) = 93.9%



Homework Question 1 Solution



Homework Question 2

The EESEE story “Blood Alcohol Content” describes a study in which 16 student volunteers at The Ohio State University drank a randomly assigned number of cans of beer. Thirty minutes later a police officer measured their blood alcohol content (BAC) in grams of alcohol per decilitre of blood.

The students were equally divided between men and women and differed in weight and usual drinking habits.

The data presented here are adapted from the nine male students who participated in the study.

student	cans drunk (X)	BAC (Y)
Barry	1	0.01
Brian	2	0.04
Jun	3	0.02
Andrzej	4	0.05
Maurizio	5	0.05
Graham	6	0.14
Peter	7	0.09
David	8	0.13
Stephen	9	0.19

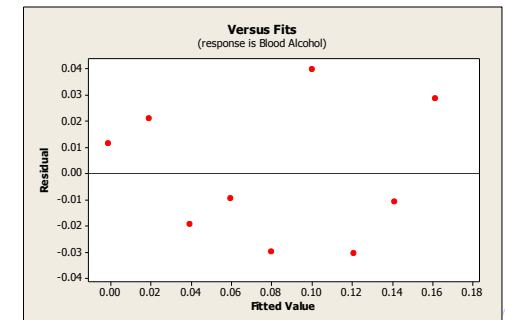
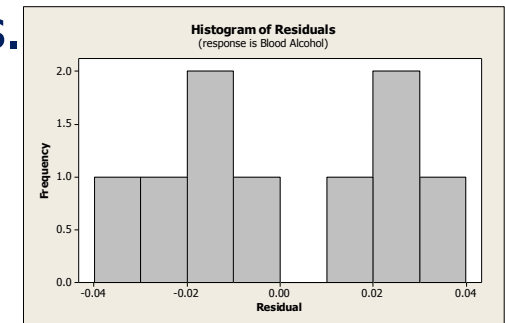
Source: Baldi, B. & Moore, D.S. (2009), *The Practice of Statistics in the Life Sciences*: W.H. Freeman & Company, New York. (adapted)



Homework Question 2 continued

Some Minitab output from the study is provided.

- Draw the least squares regression line on the scatter plot provided over the page.
- Use the scatter plot (over) and the residual plots on the right to check model assumptions.
- Interpret the least squares regression line
- Calculate a 95% confidence interval for the population slope and interpret this interval
- Use any relevant information provided to find the value of the residual for Maurizio.



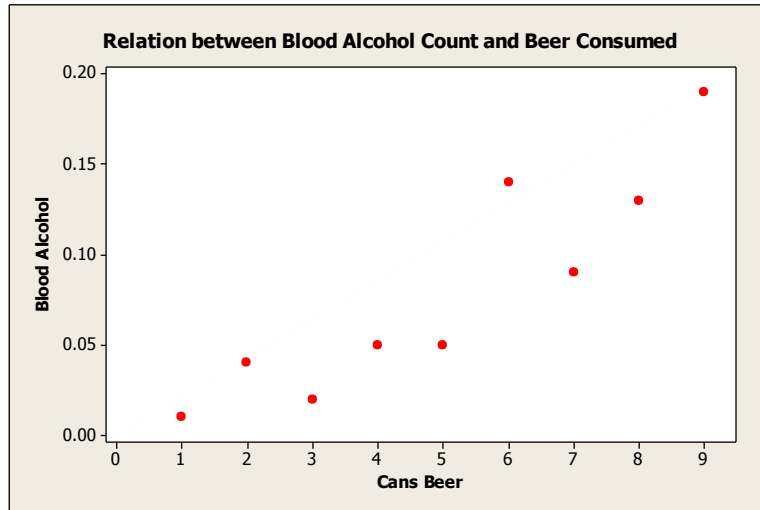
The regression equation is

$$\text{Blood Alcohol} = -0.0217 + 0.0203 \text{ Cans Beer}$$

Predictor	Coef	SE Coef	T	P
Constant	-0.02167	0.02017	-1.07	0.318
Cans Beer	0.020333	0.003583	5.67	0.001



Homework Question 2 Solution





Homework Question 3

Research Question: Can the number of salespeople on duty at a car dealership be used to predict the number of cars sold in a week?

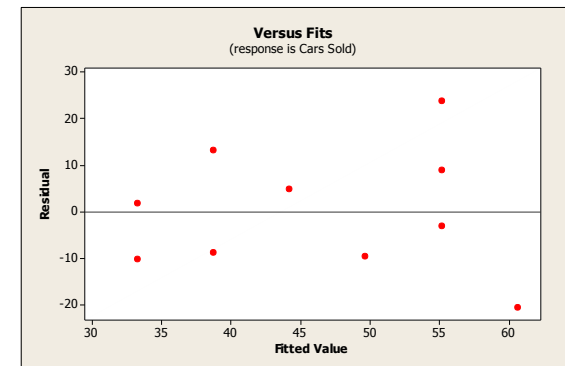
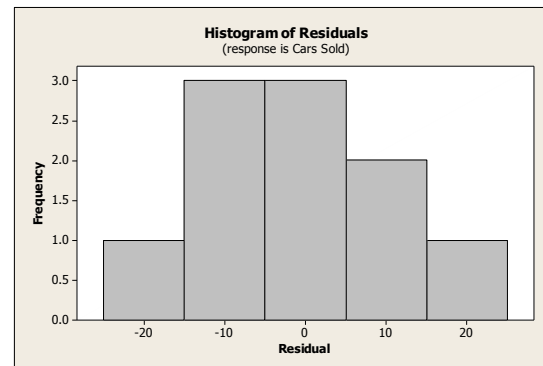
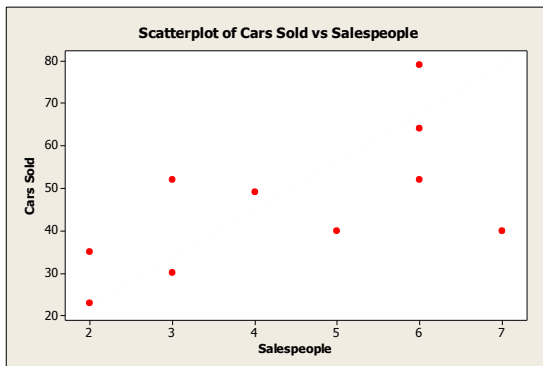
The manager of a car dealership believes there is a relationship between the number of salespeople on duty and the number of cars sold during the week. The manager records the number of cars sold and the number of salespeople employed during 10 randomly selected weeks and fits a simple linear regression model using Minitab. The output is given on the following page. Use this output to answer the research question.





Homework Question 3

Research Question: Can the number of salespeople on duty at a car dealership be used to predict the number of cars sold in a week?



Regression Analysis: Cars Sold versus Salespeople

The regression equation is
$$\text{Cars Sold} = 22.3 + 5.47 \text{ Salespeople}$$

Predictor	Coef	SE Coef	T	P
Constant	22.32	11.99	1.86	0.100
Salespeople	5.474	2.532	***	*****



Source: Black et al (2013), *Australasian Business Statistics*, (2013) Wiley (adapted)

Lecture 9 Summary

- Identify the X (independent or determinant) and Y (dependent or outcome) variables.
- Display the data graphically (scatter plot), and comment on linearity, direction of relation, extent of scatter and any outliers.
- Obtain the least squares regression line.
- Plot the line on the graph.
- Check the assumptions of the linear model.
- Test the slope of the line.
- Calculate a 95% confidence interval for the population slope
- Interpret the regression line (for one unit increase in x we expect a b unit increase/decrease in y , on average)

Textbook References

Further information on the topics discussed in this lecture can be found in:

Modern Statistics: An Introduction
by Don McNeil and Jenny Middledorp
(ISBN 9781486007011).

- Chapter 9: Pages 190 – 217