Lecture 10
Simple Linear Regression: Part 2

Using the least squares regression line to make predictions

Goodness-of-fit statistic

Correlation coefficient

In the Last Lecture….

We examined the relation between two numerical variables by:

- selecting the response variable,Y, and the determinant (or predictor) variable, X,
- constructing a scatter plot to check for a linear relation between X and Y,
- using Minitab to fit the least squares regression line:

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

- interpreting the least squares regression line,
- plotting the fitted line onto the scatter plot,
- checking the assumptions of the linear model using the scatter plot and residual plots,
- testing the **slope** of the regression line and
- calculating a 95% confidence interval to estimate the population slope

Review Quiz 1

Provide short answers to each of the following questions:

a. What does a residual of zero indicate about an observation?

b. If you had 3 positive residuals (5, 10 and 15) after fitting the least squares regression line, what would the sum of the negative residuals be?

c. In part b. how many negative residuals would there be?

Review Quiz 2

What factors may be useful in determining an adult's income? Do you think the following variables may be useful predictors of income. If you think any of these variables are possible determinants of income, indicate whether the relation is likely to be positive or negative.

a. Duration of employment in current position

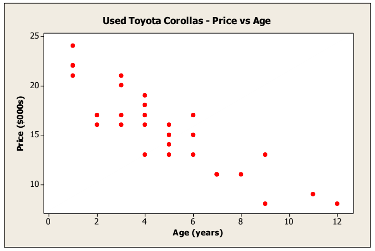
b. Shoe size

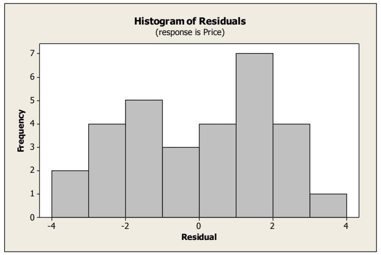
c. Amount of debt

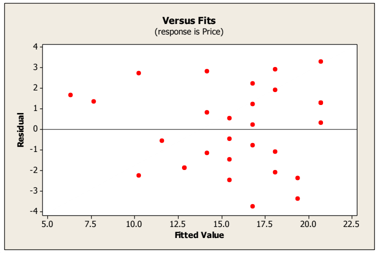
Using the least square regression line to make predictions

Predicting the Price of a Used Toyota Corolla

We used the following Minitab output in lecture 9 to determine that there was a significant negative linear relation between the age and the price of a used Toyota Corolla:







```
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
```

It follows that age can now be considered a useful predictor of price so we can start using our model to make predictions.

Making a Prediction Using the Fitted Line

In lecture 9 we attempted to predict the price of a used 10 year old Toyota Corolla using a scatter plot. That was before we fitted a least squares regression line. Since we have determined that our linear model is useful, we will be able to improve that prediction:

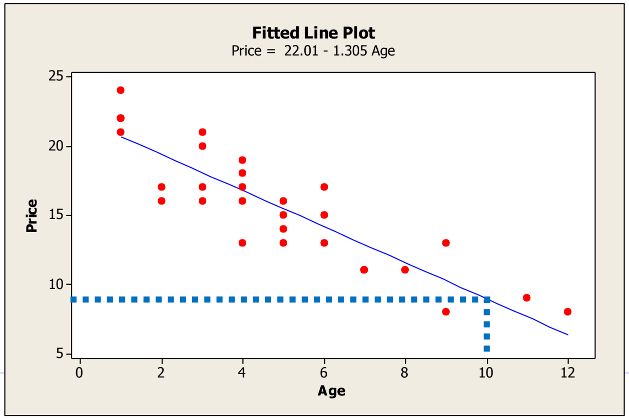
The least squares regression line is:

$$\hat{y} = 22 - 1.31x$$

So the predicted price of a 10 year old car is:

$$\hat{y} = 22 - 1.31 \times 10 = 8.9$$

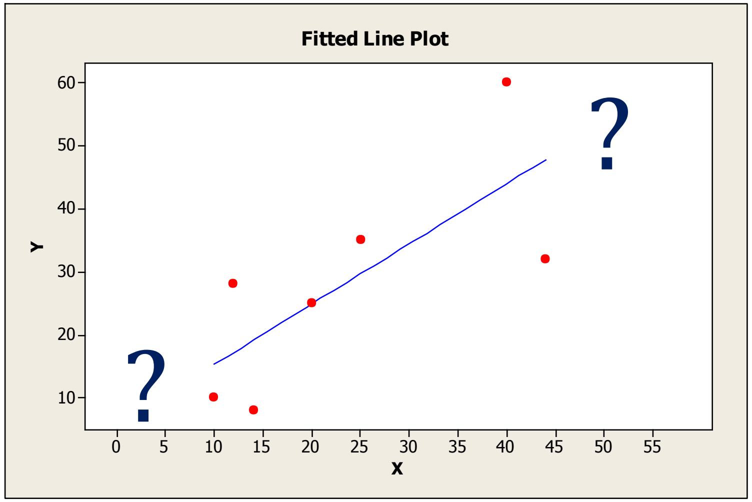
So our prediction is $8,900



Only Make Valid Predictions!

We know that we should only use the model to make predictions when a statistically significant relation exists (ie. $H_0: \beta = 0$ has been rejected).

We also need to check that any prediction we make is within the range of the x-values we used to obtain the regression line. We cannot assume that the relation between X and Y is the same outside that range as it is inside that range. For example we could not use our model to make a valid prediction for the price of a Toyota Corolla which was either less than one year old or more than 12 years old.



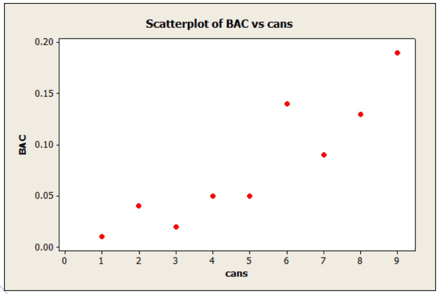
Also, our model will only give valid predictions when we predict Y (the outcome) from X (the determinant). We cannot use the model to predict X from Y.

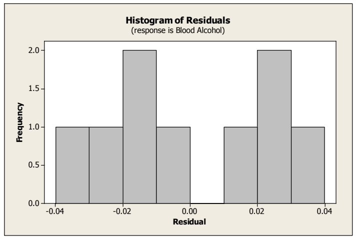
Quiz 3

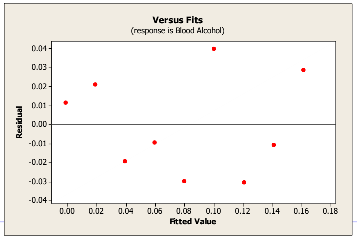
Research Question:
Is the number of alcoholic drinks consumed a useful predictor of a person’s blood alcohol content?

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

The EESEE study of Blood Alcohol Content was introduced in last week’s homework exercises. Use the computer output below, obtained from the study, to answer the research question. If appropriate, predict the BAC for Don, who believes he can consume 4 cans of beer and still drive home without being over the legal limit.


Scatterplot of BAC vs cans


Histogram of Residuals
(response is Blood Alcohol)


Versus Fits
(response is Blood Alcohol)

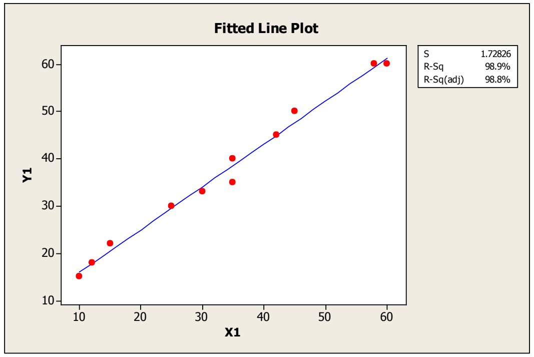
```
Model Summary

        S    R-sq  R-sq(adj)  R-sq(pred)
0.0277575  82.14%     79.59%      71.00%

Coefficients

Term          Coef  SE Coef  T-Value  P-Value   VIF
Constant   -0.0217   0.0202    -1.07    0.318
Cans       0.02033  0.00358     5.67    0.001  1.00

Regression Equation

BAC = -0.0217 + 0.02033 Cans
```

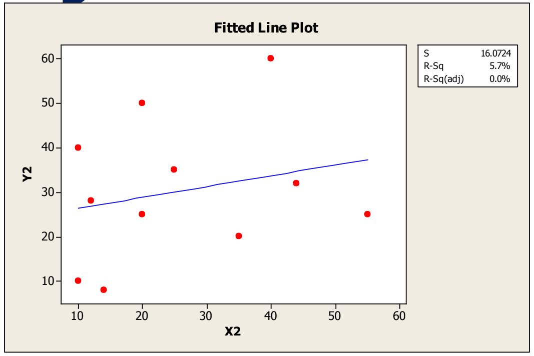
✓ Solution to Quiz 3

Goodness-of-fit statistic

Goodness-of-fit

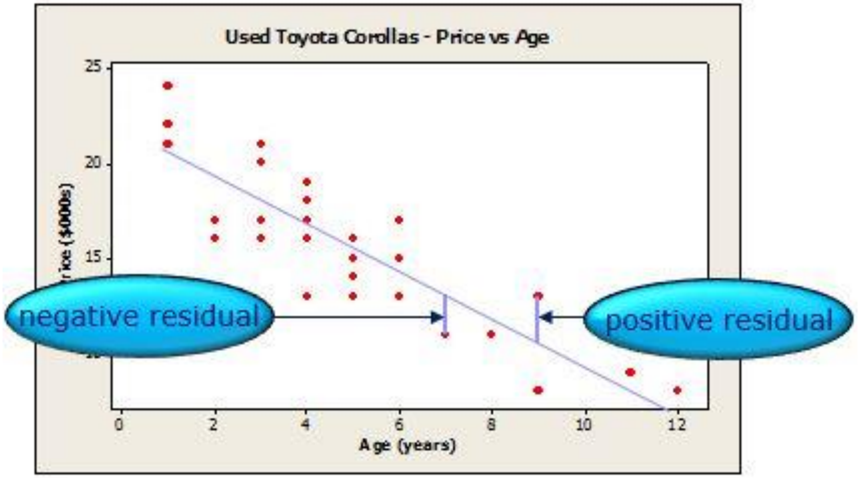
- The regression line summarises the relation between X and Y. But how good a summary is it? That is, ***how well does the line fit the data?***
- Fits well: Not such a good fit:





- Since residuals measure the discrepancy between the data and the fitted line, we base our **goodness-of-fit** measure, **r^2**, on the residuals.

Goodness-of-fit



Residuals measure the discrepancy between the data and the fitted line.

The *residuals* in a scatter plot are obtained by subtracting the predicted values from the responses, that is,

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

Residuals can be *positive* (for data points *above* the line) or *negative* (for points *below* the line).

Goodness-of-fit

To obtain this goodness-of-fit measure, we first need to consider 3 measures of variation in the response, Y:

- ○ the total variation

  which is made up of:

  - ○ the variation explained by the model

    and

  - ○ the variation left unexplained.

For the car data, we are considering the total variation in price around the mean which is made up of the variation explained by looking at the age of the car and the other variation that can't be explained by the age of the car.

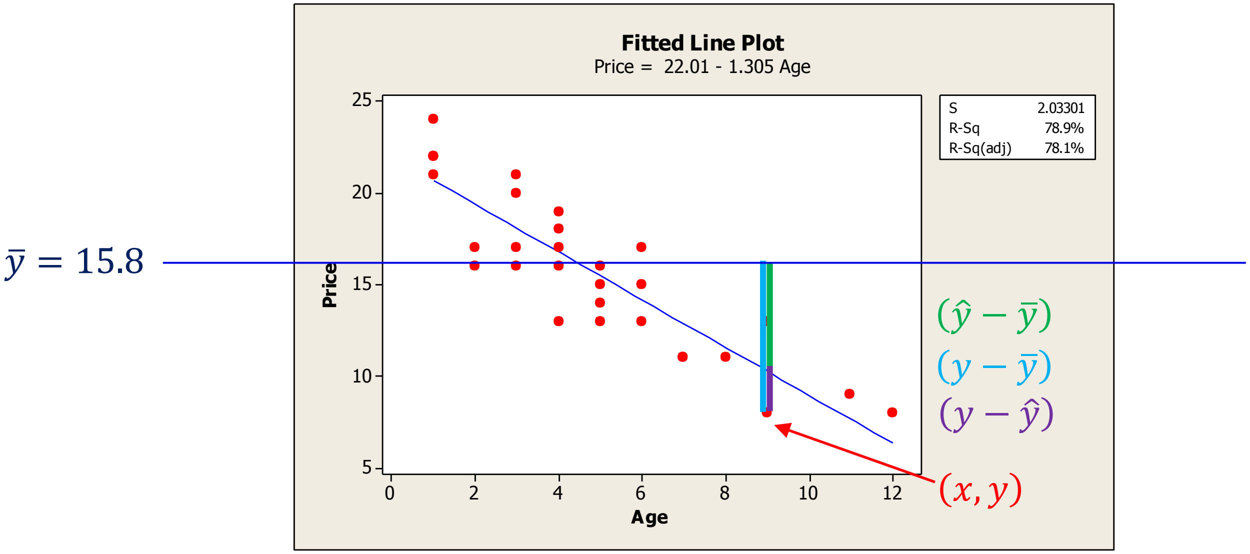
Goodness-of-fit

For the cars, the total variation is measured by looking at the difference in price between each car’s actual price and the mean price: $(y_i - \bar{y})$

The variation which can be explained by the model is measured by looking at the difference between each car’s predicted price and the mean: $(\hat{y}_i - \bar{y})$

The variation which can’t be explained by the model is measured by looking at the difference each car’s actual price and the predicted price: $(y_i - \hat{y}_i)$ ie. the residuals.

Minitab calculates these variations for each point (x_i, y_i) and uses these to calculate the goodness of fit statistic, r^2, which gives the proportion of variation in the response, Y, which can be explained by X, the determinant.



Interpreting the Goodness-of-fit Statistic

- The goodness-of-fit statistic is the variation in the response variable, y, that can be explained by the model, expressed as a proportion of the total variation in the y values.
- *This is just the proportion of the variation in the y values which is accounted for by the x values.*
- From the Minitab output for the car data we had:

```
Model Summary

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

Regression Equation

Price = 22.011 - 1.305 Age
```

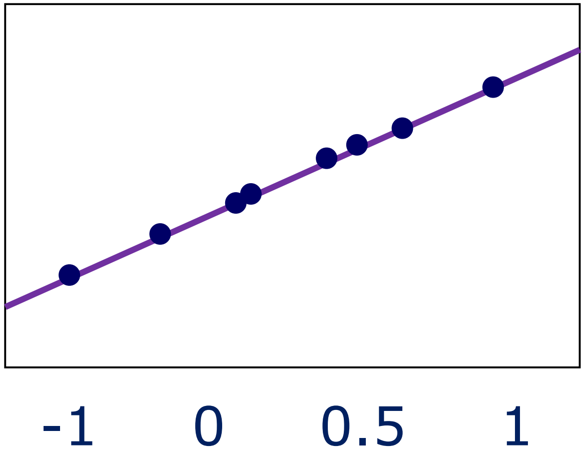
This tells us that nearly 79% of the variation in the prices of used Toyota Corollas can be explained by their age. Not all cars of the same age would have the same price - we would have to infer that the other 21% of variation was explained by other factors eg. odometer reading, condition etc. etc........

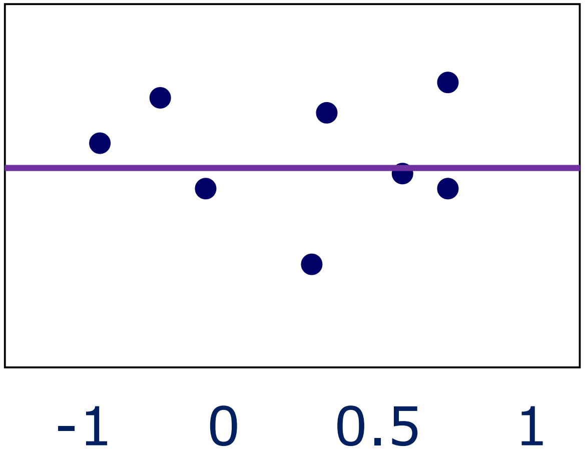
Quiz 4

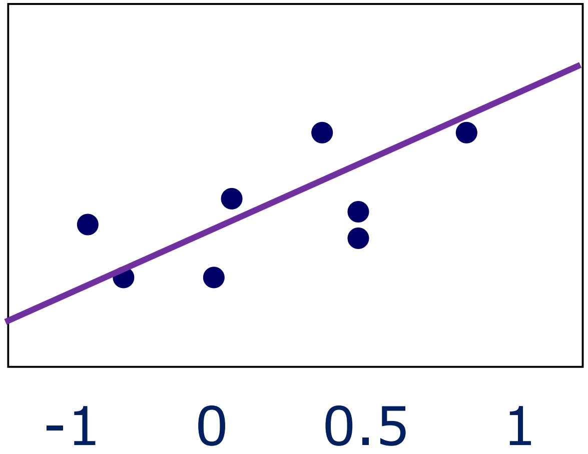
For the four scatter plots displayed on the next slide, indicate the approximate value of the goodness-of-fit statistic.

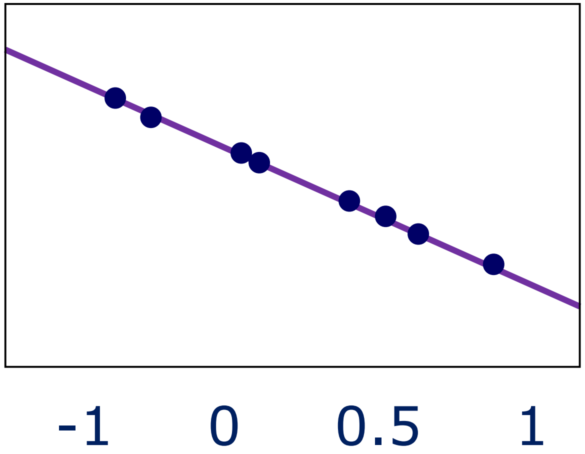
Remember the goodness-of-fit statistic always lies between 0 and 1 (ie. between 0% and 100%).

Quiz 4



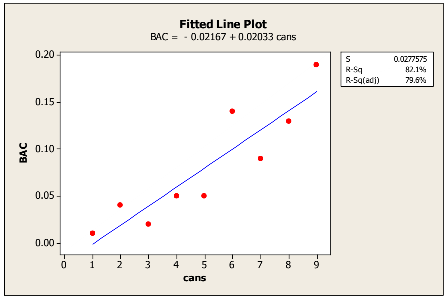






Quiz 5

a. Use the Minitab output from the beer drinkers exercise to report on how well a linear model fits these data?



```
Model Summary

        S    R-sq  R-sq(adj)  R-sq(pred)
0.0277575  82.14%     79.59%      71.00%

Coefficients

Term          Coef  SE Coef  T-Value  P-Value   VIF
Constant   -0.0217   0.0202    -1.07    0.318
Cans       0.02033  0.00358     5.67    0.001  1.00

Regression Equation

BAC = -0.0217 + 0.02033 Cans
```

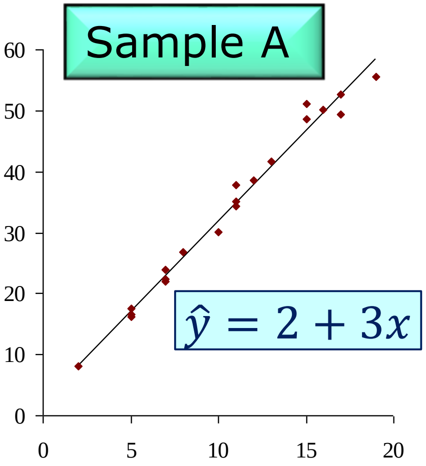
b. When fitting a linear model: weight = a + b × height for a group of students, the goodness-of-fit statistic is found to be 45%. What proportion of the variation in students' weights is explained by factors other than height? What might these factors be?

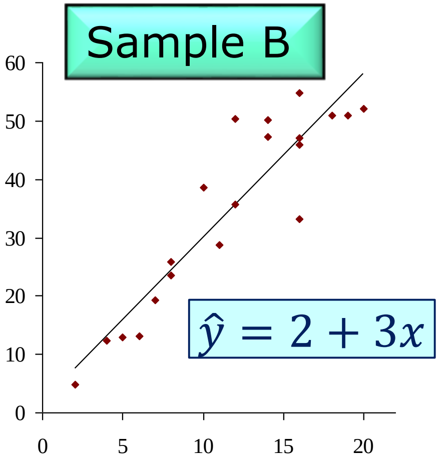
Using Goodness-of-fit to Compare Models

How well does a fitted line describe the relation in a scatter plot?

It depends on the spread of residuals around the fitted line.

We might find that the same regression line summarises two relations, but with different goodness-of-fit statistics (r^2).





small residuals

very good fit

r^2 = 98.5%

larger residuals

moderate fit

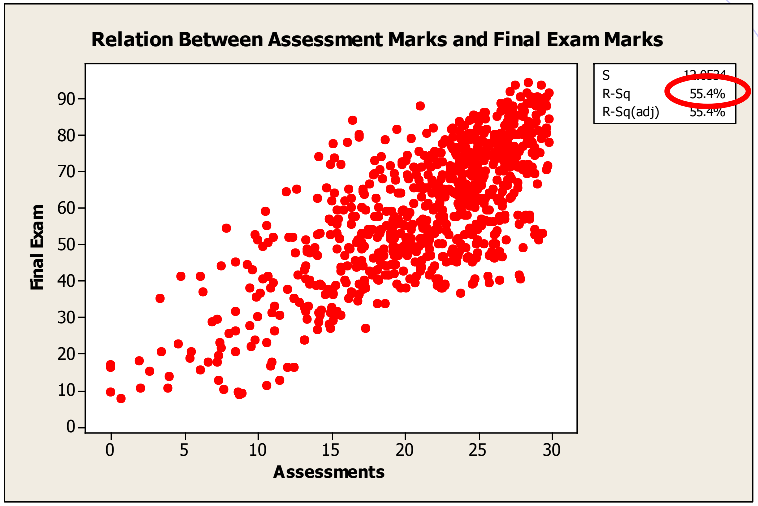
r^2 = 84.2%

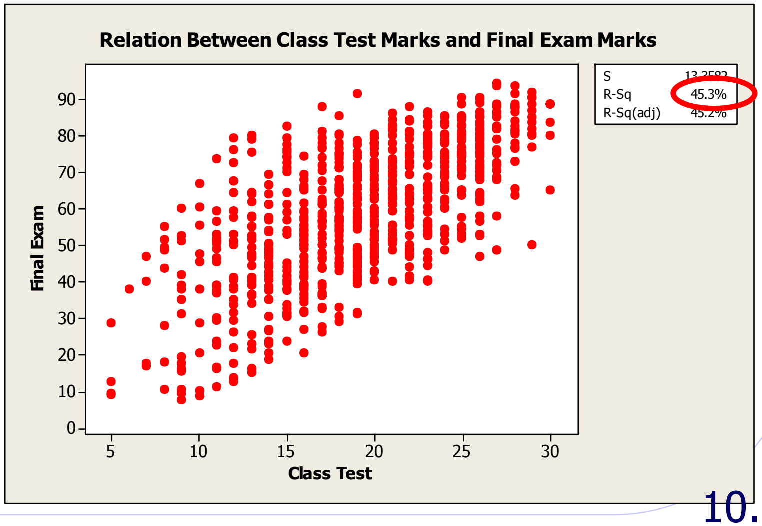
Comparing Goodness-of-fit

We used some of the results for STAT170 students in 2013 to compare assessment marks and class test marks as predictors of exam marks.

We found that r^2 was higher for the relation between assessment marks and exam marks.

Assessment marks are, therefore, better predictors than class test marks.





Correlation coefficient

The Correlation Coefficient

- The correlation coefficient, **r**, measures the *strength* and the *direction* of the *linear* relation between two variables, X and Y.

- This value of **r** always lies between −1 and 1

  ie. $-1 \leq r \leq 1$

- It can be shown that the goodness-of-fit statistic, $\mathbf{r^2}$, is the square of the correlation coefficient, **r**.

The Correlation Coefficient

For the car data we had the regression output from Minitab:

```
The regression equation is
Price = 22.0 - 1.31 Age
Predictor      Coef  SE Coef       T      P
Constant    22.0110   0.7089   31.05  0.000
Age         -1.3051   0.1276  -10.23  0.000
S = 2.03301   R-Sq = 78.9%   R-Sq(adj) = 78.1%
```

To calculate the correlation coefficient from this output:

Since $r^2 = 0.789$, and the slope, $b = -1.31$ is negative,

→ the correlation coefficient $r = -\sqrt{0.789} = -0.888$

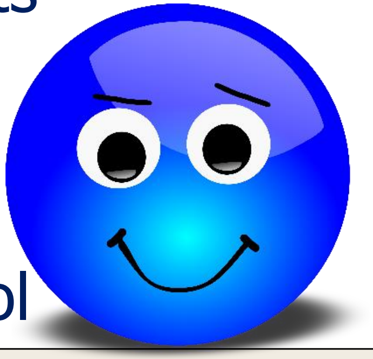
indicating that there is a strong negative linear relation between the price and the age of used Toyota Corollas

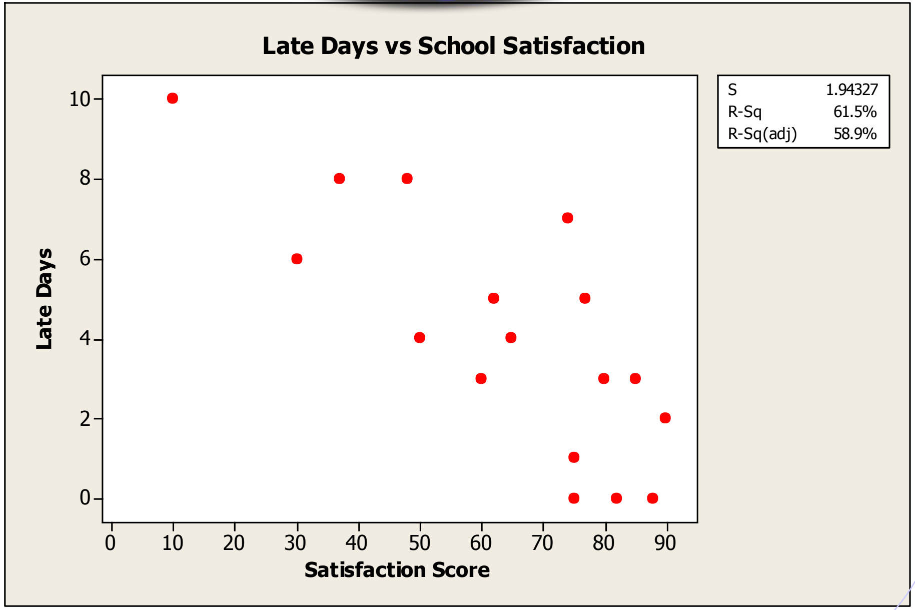
Alternatively Minitab can produce the following:

```
Correlations: Price, Age
Pearson correlation of Price and Age = -0.888
```

Quiz 6

A random sample of Year 9 students are selected. The number of days each child has been late for school in the past six months is recorded. Each child is asked to give a "school satisfaction score". These scores have been plotted against late days on the right. Use any relevant information to calculate a correlation coefficient and interpret this coefficient.

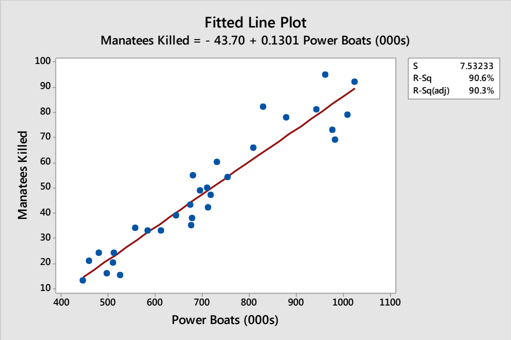




Goodness-of-fit and Correlation

Manatees:

$\hat{y} = -43.7 + 0.13x$



goodness-of-fit:

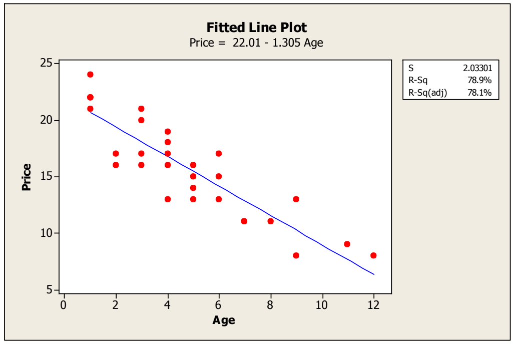
$r^2 = 90.6\%$

correlation coefficient:

$r = 0.95$

Toyotas: **price/age**

$\hat{y} = 22.0 - 1.31x$



goodness-of-fit:

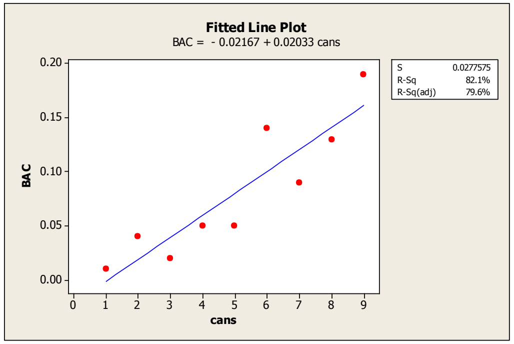
$r^2 = 78.9\%$

correlation coefficient:

$r = -0.89$

Beer:

$\hat{y} = -0.02 + 0.02x$



goodness-of-fit:

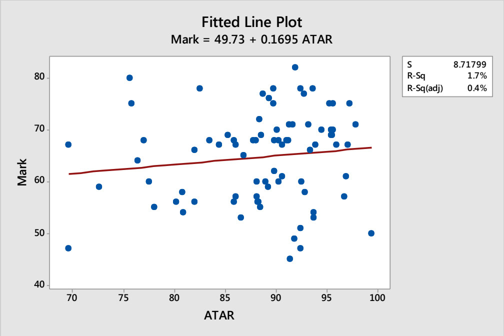
$r^2 = 82.1\%$

correlation coefficient:

$r = 0.906$

Students: Mark/ATAR

$\hat{y} = 49.7 + 0.17x$



goodness-of-fit:

$r^2 = 1.7\%\%$

correlation coefficient:

$r = 0.130$

More on Scatter Plots

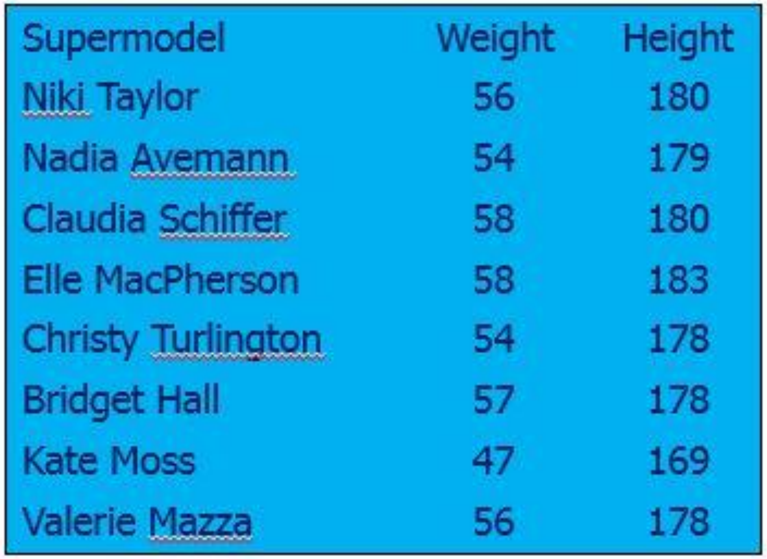
Look at the Scatter Plot First!

- When considering the relation between two numerical variables, we should always look at a scatter plot before any further analysis.
- The scatter plot should indicate whether a linear model is appropriate and may also indicate whether there are any other problems with the model.

For example:

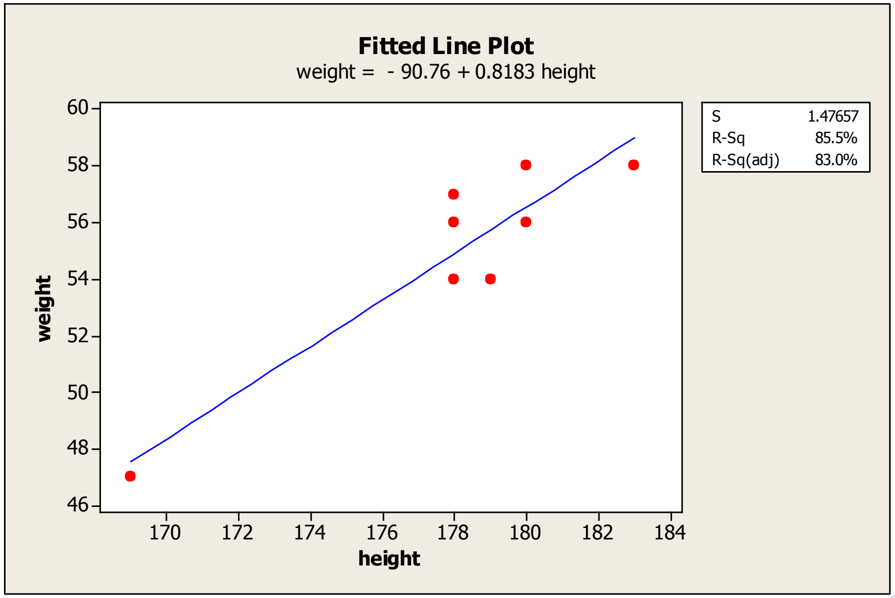
- a point far away from the majority of points can have an 'undue influence' on the regression line.
- if the plot indicates 'groups of points' there may be some reason for fitting separate models for different 'groups'

Quiz 7



Supermodel	Weight	Height
Niki Taylor	56	180
Nadia Avemann	54	179
Claudia Schiffer	58	180
Elle MacPherson	58	183
Christy Turlington	54	178
Bridget Hall	57	178
Kate Moss	47	169
Valerie Mazza	56	178

The data on the left of weights (kg) and heights (cm) were used to fit the linear model below.

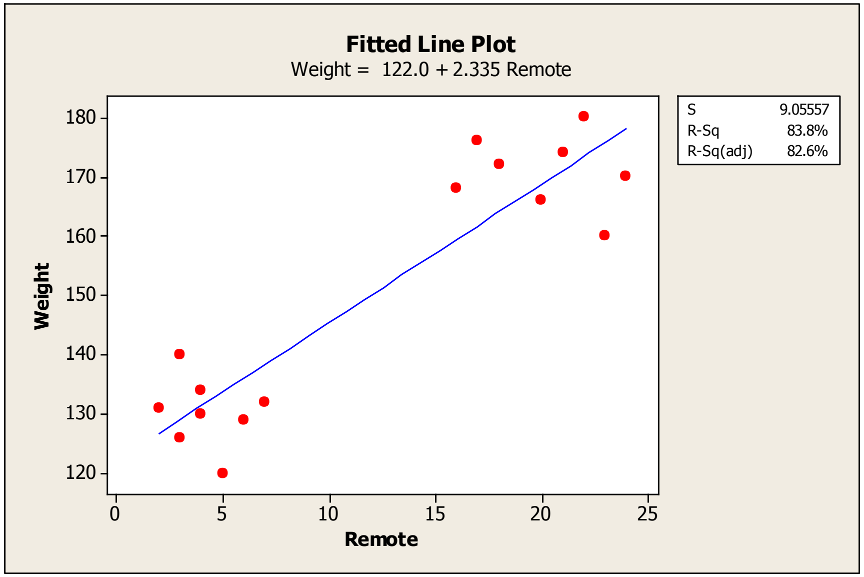


Does the scatter plot suggest there are any problems with the fitted model?

Source: Triola, M.F. (2005) *Elementary Statistics*, Pearson

Quiz 8

Weight	Remote	Sex
120	5	female
126	3	female
129	6	female
130	4	female
131	2	female
132	7	female
134	4	female
140	3	female
160	23	male
166	20	male
168	16	male
170	24	male
172	18	male
174	21	male
176	17	male
180	22	male

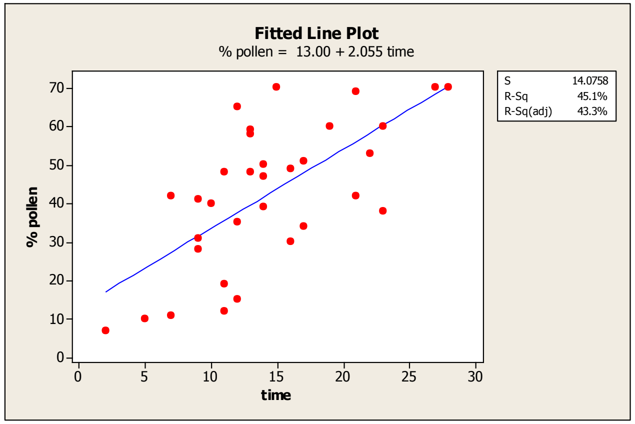


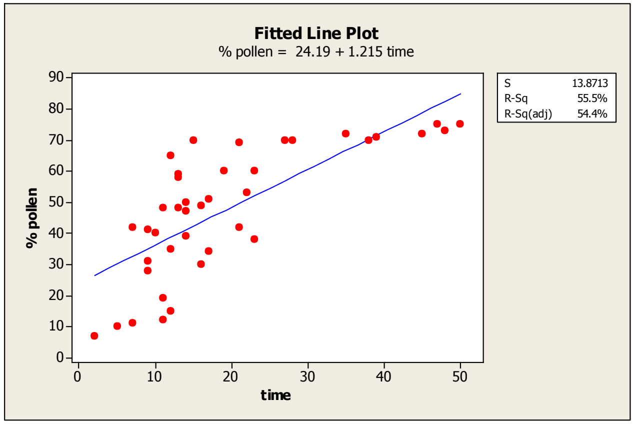
The data on the left were recorded on a sample of 16 people. The weight of each subject and the number of times he/she used a remote control in one hour were recorded and used to fit the model above. Any obvious problems with fitting a linear model here?

Source: Triola, M.F. (2005) *Elementary Statistics,* Pearson .

Quiz 9

Data were collected from a sample of bumblebees to determine whether the percentage of **pollen** that a removed by a bee from a flower is dependent on the **time** (seconds) it spends sitting on the flower? The plot on the left has a linear model fitted to the scatter plot only for the 33 bees which were observed sitting on the flowers for less than 30 seconds. The plot on the right has a linear model fitted to the scatter plot for all 40 bees actually observed in the study. Comment on the appropriateness of these models.





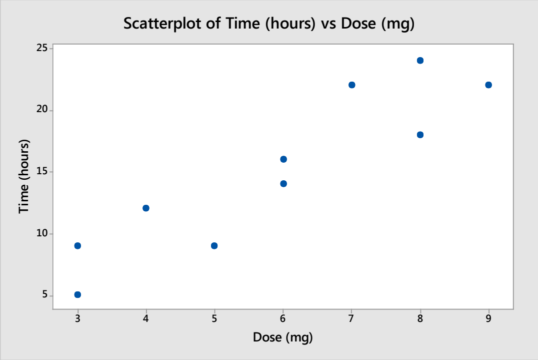
Homework Questions

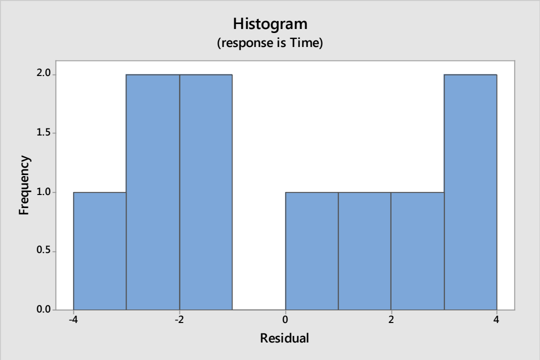
Homework Question 1

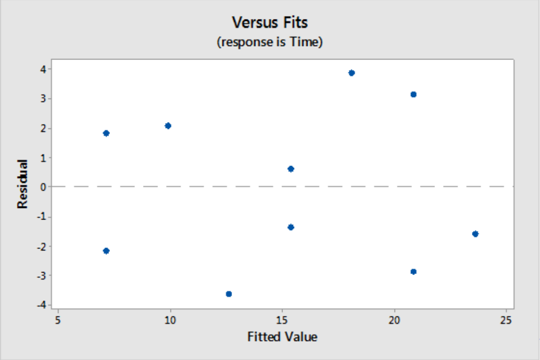
A trial was undertaken to test a new drug for allergies. Ten allergy sufferers were given various doses of the drug to determine the relation between the dose (milligrams) and the duration of relief (hours) from allergic symptoms. Each patient received a specified dosage and was asked to report back as soon as the protection of the drug appeared to have worn off. Use the output on the following slide which was obtained from the study to answer the questions which follow.

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

Homework Question 1







```
Regression Analysis: Time versus Dose

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
2.82074  82.84%     80.69%      73.65%

Coefficients

Term        Coef  SE Coef  T-Value  P-Value
Constant  -1.07     2.75    -0.39    0.707
Dose      2.741    0.441     ****    *****
```

✓Solution to Homework Question 1

○ *The dependent variable is:*______________________________

○ *The independent variable is:*____________________________

○ *Comment on the scatter plot:*___________________________

____

○ *Write down the least squares regression line:*

____

Solution to Homework Question 1

○ *Interpret the least squares regression line:*

____

____

○ *The least squares regression line has its sum of residuals equal to__________and its sum of squared residuals is a ______________________________*

Solution to Homework Question 1

- *The goodness-of-fit statistic is:*______________________________

- *Interpret the goodness of fit statistic:*______________________

____

- *The correlation coefficient is:*______________________________

- *Interpret the correlation coefficient:*________________________

____

Solution to Homework Question 1

- *Write down a null and an alternative hypothesis to test for a significant linear relation between dosage and duration of relief:*_______________________________

- *Write down the three assumptions of a linear model:*_______________________________________

____

- *Comment on whether each assumption is satisfied by referring to the appropriate plot for each assumption:*___________________________________

____

Solution to Homework Question 1

- *The test statistic is:____________and the df is:___________*
- *The p-value is:____*
- *The correct decision is:_________________________________*
- *Write an appropriate conclusion:________________________*

____

____

____

Solution to Homework Question 1

- *If appropriate make the following predictions:*
  - *Predict the duration of protection for a patient who is given a dosage of 4.5 mg:*______________________________
  - *Predict the duration of protection for a patient who is given a dosage of 10 mg:*________________________________
  - *Predict the dosage given to a patient who has protection from allergies for a duration of 16 hours:*

____

Summary of Lectures 9 and 10

- A scatter plot is used to graph the relation between two variables X and Y. The determinant (x) is plotted on the horizontal axis and the outcome (y) on the vertical axis.
- If the relation is linear, a straight line $\hat{y}$ = a+bx, which estimates $\widehat{y}$ = α+βx, summarises the relation and may be fitted by minimising the sum of the squares of the residuals.
- A t-test may be used to test the statistical significance of the linear relation. The slope can be interpreted as the expected amount of increase in Y for a unit increase in X.
- The model enables us to predict the outcome within the range of the determinant.

Summary of Lectures 9 and 10 continued

- The goodness-of-fit statistic is measures the proportion of the variation in Y accounted for by the fitted li

$$0 \leq r^2 \leq 1 \quad \text{or} \quad 0\% \leq r^2 \leq 100\%$$

- The correlation coefficient measures the strength and direction of the linear relation.

$$-1 \leq r \leq 1$$

- The model assumes (1) a linear relation, and (2) errors having constant spread with (3) a normal distribution.

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