UCD School of Mathematical Sciences ---- Introduction to Data Analytics

Lecture 10 – Correlation and Simple Linear Regression

Overview

In this section we will:

- Use a scatter plot to examine the relationship between two variables.
- Understand how a correlation coefficient measures the strength of relationship.
- Discuss possible kinds of relationships
- Examine how a regression line can provide quantitative information about a relationship
- Assess the fit of a proposed regression line, graphically and numerically
- Make inferences and predictions based on a regression line fit to data, including whether a linear relationship exists in a population.

Two methods for exploring linear relationships

1. Correlation: uses a single number between —1 and +1
2. Regression: uses the equation of the line that best fits the data

1. Correlation

Correlation is a statistical method used to determine whether a linear relationship between variables exists; calculates a value r (between -1 and 1) measuring the strength of relationship.
i.e. Does one variable tend to occur with large(or small) values of the other?
These are called positive and negative relationships respectively.

Examples:

1. Minimum daily temperature and heating costs -- Negative
2. Interest rates and number of loan applications -- Negative
3. Incomes of employed men and employed partners -- Positive
4. Height and IQ -- None
5. Maths LC score and Irish LC score -- Positive

2. Regression

Regression is a statistical method used to algebraically describe the nature of the relationship between variables—that is, positive or negative, linear or nonlinear.

The regression equation
e.g. $E(y) = 3 - 0.1x$
can provide quantitative information.

e.g. if minimum temperature falls by $1^{o}C$, by how much, on average, do heating costs increase?

Statistical Questions of Interest

- Are two or more variables related?
- If so, what is the strength of the relationship?
- What type of relationship exists?
- What kind of predictions can be made from the relationship?

10.1 Scatter Plots & Correlation

Scatter Plots are used to visualise a relationship and Pearson's correlation coefficient is used to measure the strength of a linear relationship.

Scatter Plots

- A scatter plot is a graph of the ordered pairs (x, y) of numbers consisting of the independent variable, x, and the dependent variable, y.

- A scatter plot is a visual way to describe the nature of the relationship between the independent and dependent variables.

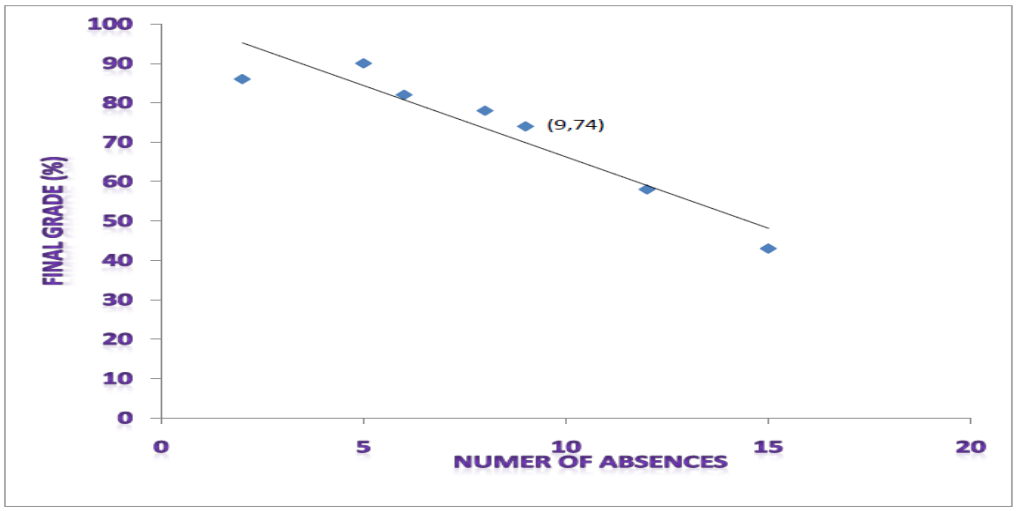
Example:

Student	**No of absences, *x***	**Grade (%), y**
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

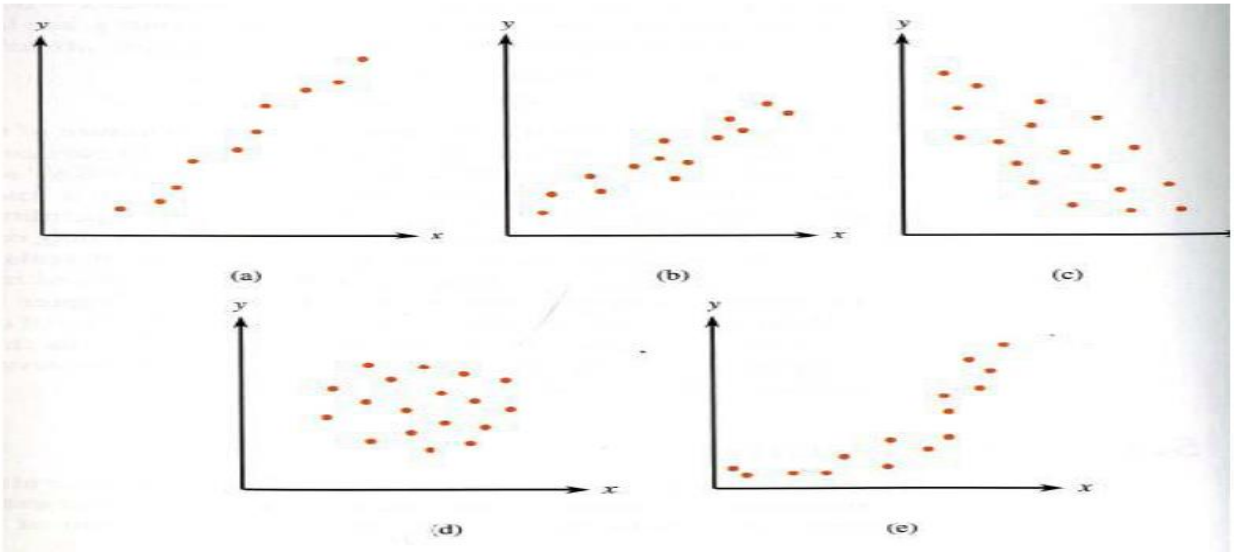
Order the data by number of absences:

Student	**No of absences, *x***	**Grade (%), y**
B	2	86
F	5	90
A	6	82
G	8	78
D	9	74
E	12	58
C	15	43

Scatterplot of the Data:



Scatter plots - various relationships:



(a) Positive linear; (b) Positive linear; (c) Negative linear; (d) No relation; (e) Curved relation

Correlation Coefficient

- The correlation coefficient, r, computed from data measures the strength and direction of a linear relationship between two variables.

- r does not depend on the units of measurement.

i.e. if you calculate r for your data when it is measured in cm and when it is measured in m, the value of r will be the same. The relationship between the variables is not affected by the unit of measurement

- r does not depend on which variable is x, y.
  i.e. Suppose we have data measuring the height and weight of 10 subjects. Our (x, y) bivariate data would be of the form (Height, Weight) or (Weight, Height). While our scatterplots would look slightly different depending on which form we used, the calculated correlation coefficient, r, will be the same in both cases.

- The symbol for the sample correlation coefficient is r, and for the population correlation coefficient is ρ(rho).

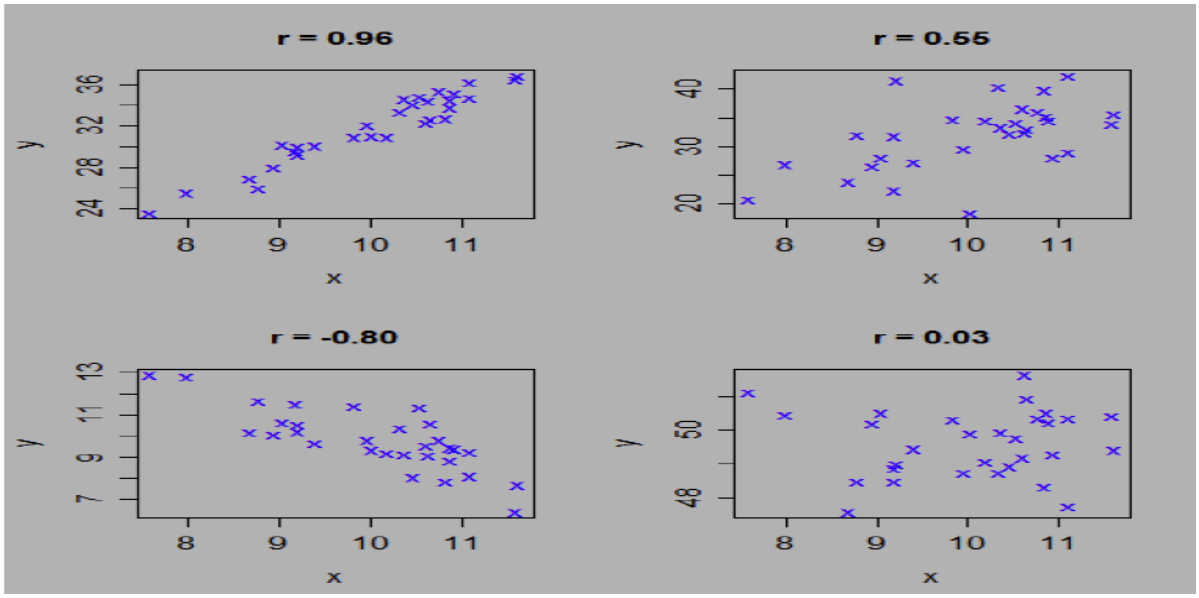
- The range of the correlation coefficient is from -1 to +1.

- If there is a strong positive linear relationship between the variables, the value of r will be close to +1.

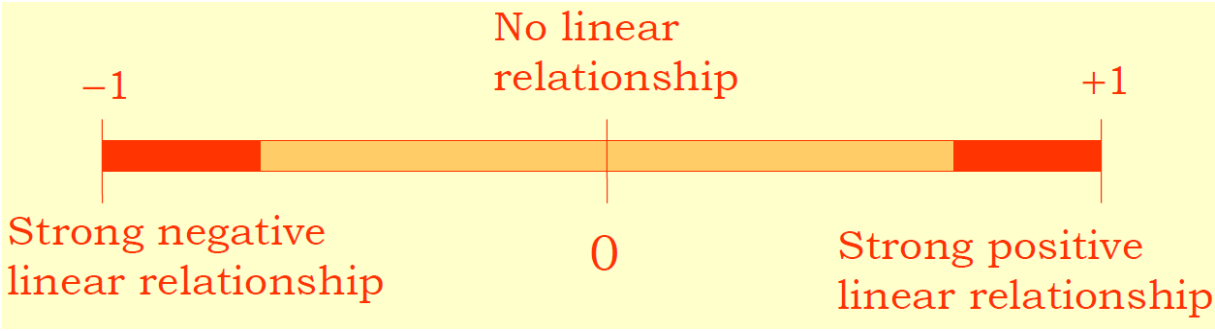
- If there is a strong negative linear relationship between the variables, the value of r will be close to -1.

<u>r and Scatter Plots</u>

A high value of |r| is seen when the points are closely scattered around a line.



When there is no linear relationship between the variables or only a weak relationship, the value of r will be close to 0.



Calculating r

The formula for r:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

$$S_{xy} = \sum(x - \bar{x})(y - \bar{y}) = \sum xy - \frac{\sum x \sum y}{n}$$

$$S_{xx} = \sum(x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$
$$S_{yy} = \sum(y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

Therefore

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Example - Calculating r

Country	Sugar(x)	Depression(y)	x^2	y^2	xy
Korea	150	2.3	22500	5.29	345
US	300	3	90000	9	900
France	350	4.4	122500	19.36	1540
Germany	375	5	140625	25	1875
Canada	390	5.2	152100	27.04	2028
New Zealand	480	5.7	230400	32.49	2736
Totals	2045	25.6	758125	118.18	9424

$\sum x = 2045$ $\sum y = 25.6$ $\sum x^2 = 758{,}125$ $\sum y^2 = 118.18$ $\sum xy = 9294$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

$$= \frac{9424 - \frac{(2045)(25.6)}{6}}{\sqrt{\left(75815 - \frac{(2045)^2}{6}\right)\left(118.18 - \frac{(25.6)^2}{6}\right)}}$$

$$= \frac{698.6667}{\sqrt{61120.83 \; x \; 9.95333}}$$

$$= \frac{698.6667}{\sqrt{739.7535}} = 0.944 \text{ (correct to 3 decimal places)}$$

Conclusion: There is a strong positive linear association between national sugar consumption and depression rate.

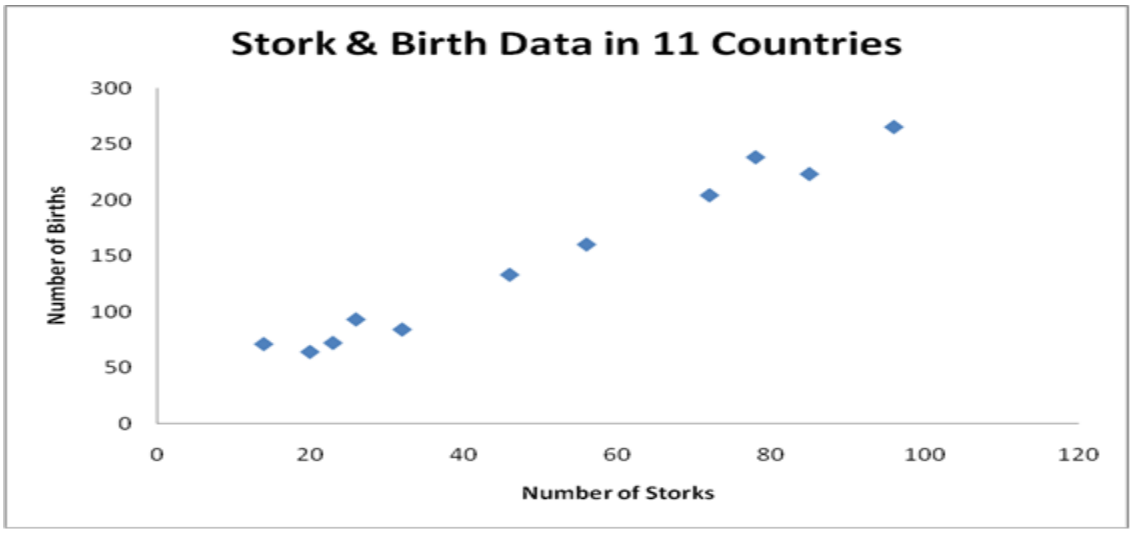
Population Correlation

Formally defined, the population correlation coefficient, ρ, is the correlation computed by using all possible pairs of data values (x, y) taken from a population.

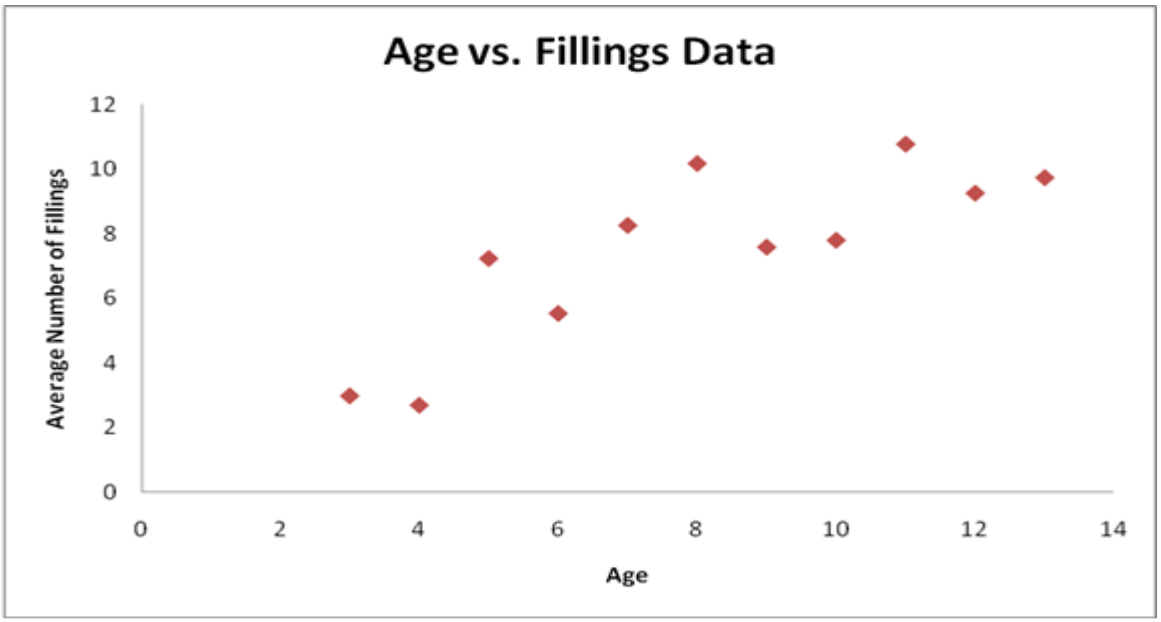
Correlation and Causation

What conclusions are valid when we find a strong correlation? e.g. |r| > 0.8?

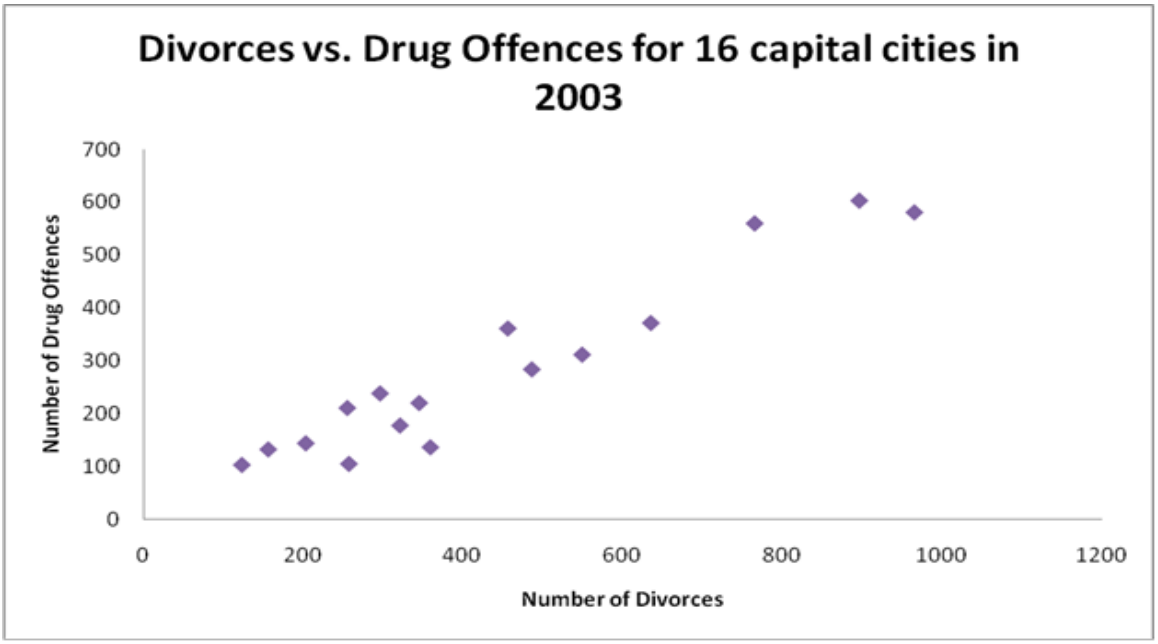
1. There is a strong correlation between the numbers of storks in a country and the number of births in that country. Countries with many storks have a high number of births and countries with low stork counts have low numbers of births.



2. There is a high correlation among primary school children between vocabulary and numbers of tooth fillings. Children with many fillings have a larger vocabulary than children with only a small number or with no fillings.



3. Divorce rates are correlated with numbers of drug offences.



What should we conclude from these facts?

1. That storks really are responsible for bringing babies.
2. That eating Mars bars will increase your vocabulary.
3. That drugs are responsible for marriage breakdown.

No, these examples illustrate a very important point: *Correlation is not the same as causation.*

Explanations

1. Larger countries have larger stork populations and usually have higher human populations as well and so there will be higher numbers of babies born than in smaller countries.
2. Young children have very few fillings because they have only been around for a few years whereas older children have had time to eat lots of sweets, get a lot of bad teeth and learn a lot of new words.
3. ???

Possible Relationships

- There is a direct cause-and-effect relationship between the variables: that is, x causes y.
- There is a reverse cause-and-effect relationship between the variables: that is, y causes x.
- The relationship between the variables may be caused by a third variable: that is, y may appear to cause x but in reality z causes x.
- There may be a complexity of interrelationships among many variables; that is, x may cause y but w, t, and z fit into the picture as well.
- The relationship may be coincidental: although a researcher may find a relationship between x and y, common sense may prove otherwise.

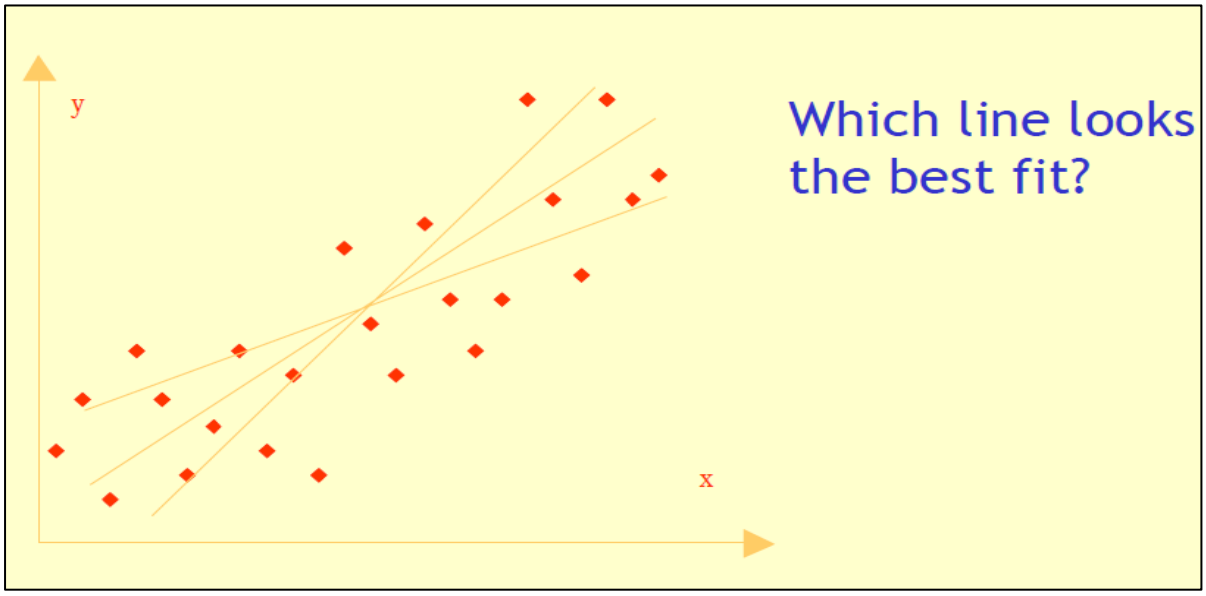
Interpreting Relationships

- A researcher must consider all possibilities giving rise to apparent associations.
- Remember, even when a genuine association exists as indicated by correlation, this does not necessarily imply causation.

10.2 Regression Line: Fitting the Model

If a scatter plot and the value of the correlation coefficient indicate a linear relationship, the next step is to determine the equation of the regression line which is the data's line of best fit.
Best fit means that the sum of the squares of the vertical distance from each point to the line is at a minimum.

A Scatterplot with three lines:



Equation of a Line

The equation of a line is often written as

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

where **b** is the **slope** of the line and **a** is the **intercept** (i.e. where the line cuts the **y** axis).

Explanatory and Response Variables
In many experiments, one of the variables is fixed or controlled and the point of the experiment is to determine how the other variable varies with the first. The fixed/controlled variable is known as the explanatory or independent variable and the other variable is known as the response or dependent variable.

We shall use "x" for the explanatory variable and "y" for the response variable, but we could have used any letters.

In a cause and effect relationship, the independent variable is the cause, and the dependent variable is the effect. Least squares linear regression is a method for predicting the value of a dependent variable *y*, based on the value of an independent variable x.

Fitting the Model: The Least Squares Approach

For the best-fit line:

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

where

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

and $\bar{y} = a + b\bar{x}$ (giving a)

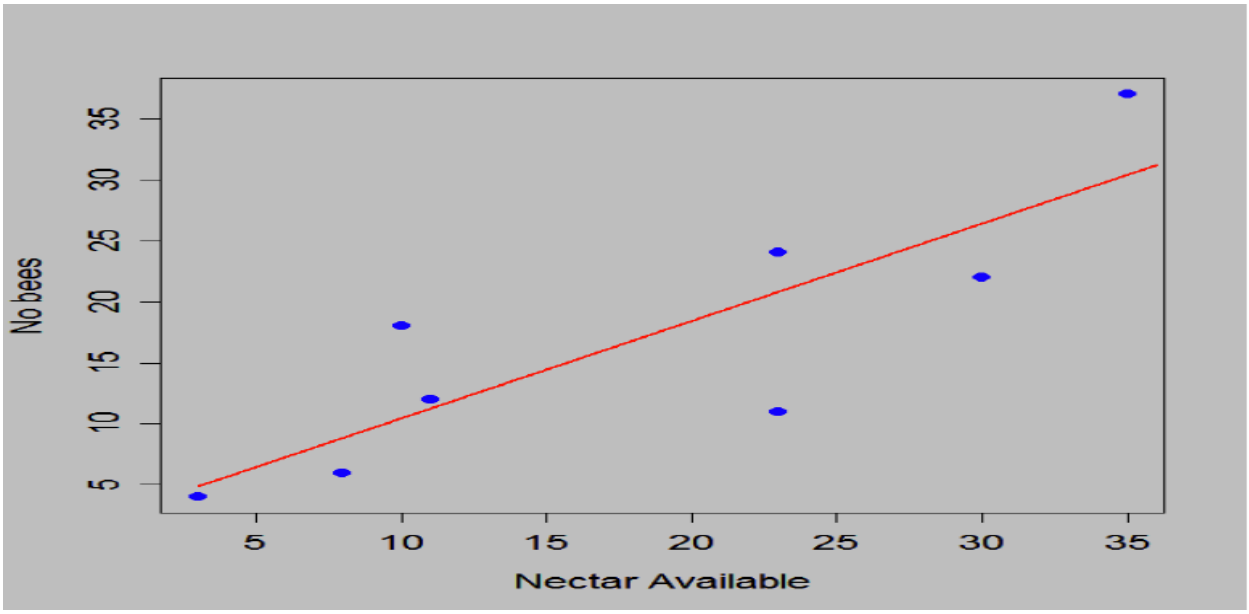
Calculations by hand

- Step 1 - Make a table with columns for *x*, *y*, *xy*, x^2, and y^2.
- Step 2 - Find the values of *xy*, x^2, and y^2. Place them in the appropriate columns.
- Step 3 - Substitute in the formula to find the value of *r, a* and *b*.

Example: Nectar availability and abundance of bumblebees

The following table shows data from 8 sites on the abundance of bumblebees (y) and the availability of nectar (x):

Nectar Availability	3	8	11	10	23	23	30	35
Bumblebee Abundance	4	6	12	18	11	24	22	37



Calculations:

	x	y	xx (x^2)	yy (y^2)	xy
	3	4	9	16	12
	8	6	64	36	48
	11	12	121	144	132
	10	18	100	324	180
	23	11	529	121	253
	23	24	529	576	552
	30	22	900	484	660
	35	37	1225	1369	1295
Sum (i.e. totals)	143	134	3477	3070	3132

$$\sum x = 143 \qquad \sum y = 134 \qquad \sum x^2 = 3477 \qquad \sum y^2 = 3070 \qquad \sum xy = 3132$$

$$\bar{x} = 17.875 \qquad \bar{y} = 16.75$$

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{3132 - \frac{(143)(134)}{8}}{3477 - \frac{(143)^2}{8}} = \frac{736.75}{920.93} = 0.8$$

$$a = \bar{y} - b\bar{x} = 16.75 - (0.8)(17.875) = 2.45$$

The best fit (least squares) line is:

$$\hat{y} = 2.45 + 0.8x$$

where $\hat{y}$ is a value calculated from the least squares line, and not a value of y occurring in the data set.

10.3 Assessing the Fit of a Line

- Residuals
- Coefficient of Determination, r^2
- Standard deviation about the Least Squares (LS) Line

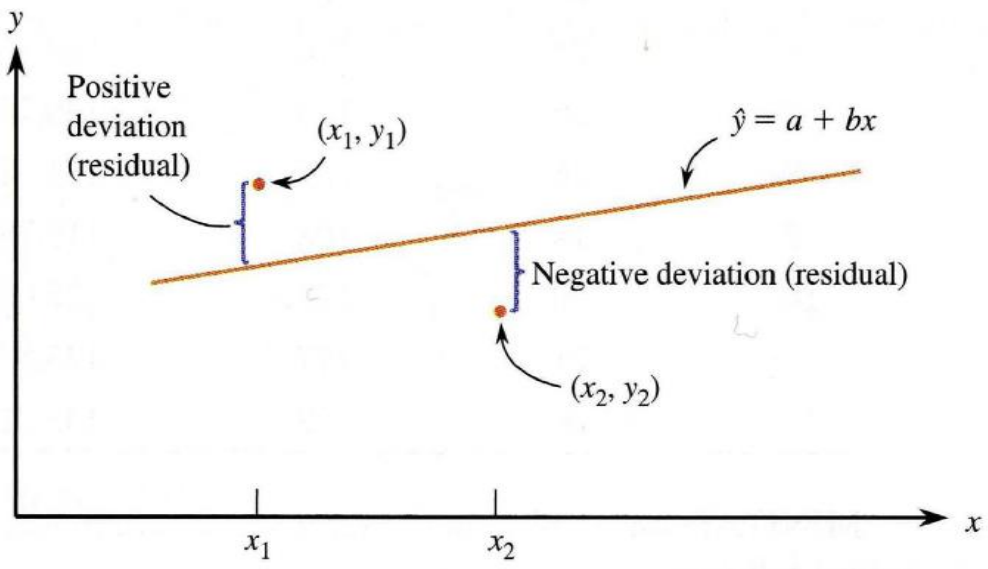
<u>Residuals</u>

- For each point (x, y) in the dataset:
- y - (a + bx) measures the vertical deviation (vertical distance) from the point to the line.
- Some points are above the line and y -(a + bx) will be positive for these points. For points below the line y-(a+ bx) will be negative.
- Now if we calculate [y-(a+ bx)]^^{2} for each point (x,y) and add them all up we get the sum of the squared vertical distances of all the points from the line.
- The least squares line minimises this sum of squared values.

If we have data values (x_1, y_1), (x_2, y_2), ……, (x_n, y_n) then the point (x_i, $\hat{y}_i$), where $\hat{y}_i$ = a + bx_i lies on the least squares line.

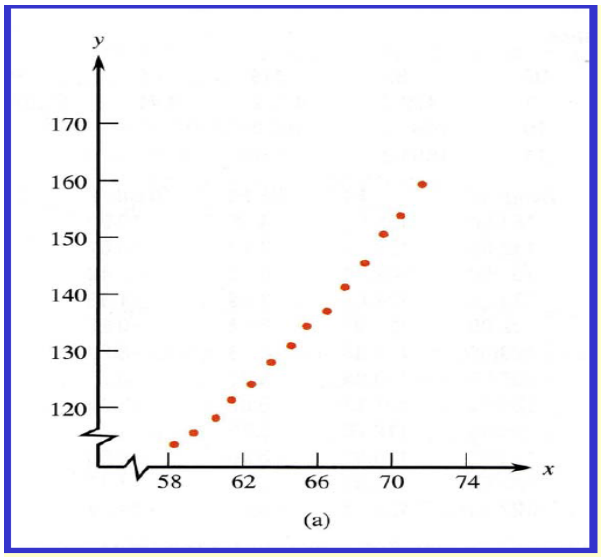
$\hat{y}_i$ is called the predicted value for x_i.

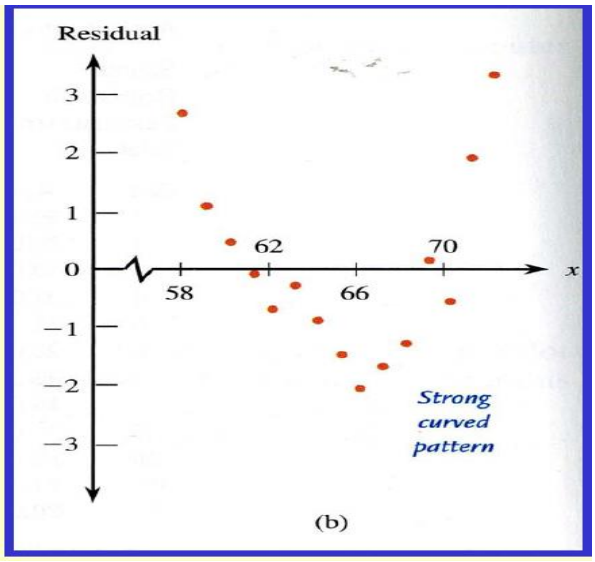
The residual for this point is $y_i - \hat{y}_i = y_i - (a + b\,x_i)$



Example

The following shows a scatterplot of data on *x*= height (ins) and *y* = average weight (lb) of American women is given in *The World Almanac and Book of Facts*





For a good fit no patterns or unusual points should appear in a residual plot.

Hence in the example:

- Even though the scatter plot looks straight, the curvature in the residuals indicates a straight regression line does not fit the data.
- In fact average weight increases more rapidly for large heights than it does for small heights.

Total Variation

- The total variation in the y values $\sum(y-\bar{y})^2$ is the sum of the squares of the vertical distance each point is from the mean.
- The total variation can be divided into two parts: that which is attributed to the regression relationship of x and y, and that which is due to chance (residual).

Two Parts of Total Variation

- The explained variation is SSReg = $\sum(\hat{y}-\bar{y})^2$
- Variation due to chance, found by SSResid = $\sum(y-\hat{y})^2$ is called the unexplained variation. This variation cannot be attributed to the relationships.

The total variation, SSTo is equal to the sum of the explained variation, SSreg, and the unexplained variation, SSresid.

SSTo = SSReg + SSResid

Algebraically:

$$\sum(y-\bar{y})^2 = \sum(\hat{y}-\bar{y})^2 + \sum(y-\hat{y})^2$$

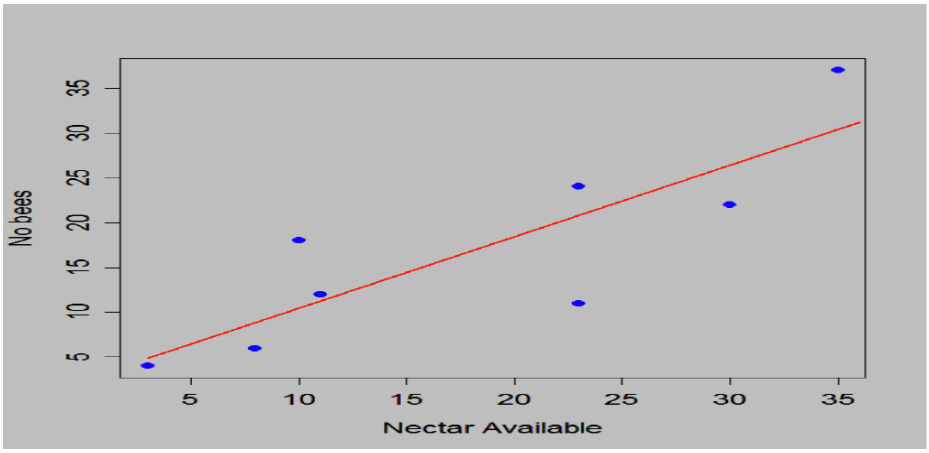
Calculations:

$$\text{SSTot} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$\text{SSResid} = \sum y^2 - a\sum y - b\sum xy$$

SSReg = SSTo - SSResid

Bees & Nectar Example continued



a	=	2.4490	b	=	0.8001
$\sum x$	=	143	$\sum y$	=	134
$\sum x^2$	=	3477	$\sum y^2$	=	3070
$\sum xy$	=	3132	n	=	8

Hence

SSTot = 3070 – $(134)^2/8$ = 825.5

SSResid = 3070 – (2.4490)(134) – (0.8001)(3132) = 235.9208

SSReg = 825.500 – 235.9208 = 589.5792

r^2 = 589.5792/825.500 = 0.71421

r = 0.84 (since b>0)

Interpretation of r^2

71% of variation in number of bees explained by nectar.
This is helpful in interpreting the strength of the relationship.
Coefficient of Determination

How good is this model at explaining the behaviour of the y values?
The coefficient of determination is the proportion of the variation in y that is explained by the regression line and the independent variable.
i.e. coefficient of determination

SSReg / SSTo = 1 - (SSResid / SSTo)

Mathematically it can be proved that the coefficient of determination = r^2

Standard Error about LS line

The standard error about the LS line, s_e, estimates the typical amount by which an observation deviates from the LS

$$s_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}}$$

This is referred to as the estimated standard error of the regression line.
Interpreting s: We expect most (approximately 95%) of the observed y values to lie within $2s_e$ of their respective least squares predicted values, $\hat{y}$

10.4 Summary

- The strength and direction of the linear relationship between variables is measured by the value of the correlation coefficient r.
- r can assume values between and including -1 and +1.
- The closer the value of the correlation coefficient is to -1 or +1, the stronger the linear relationship is between the variables.
- A value of -1 or +1 indicates a perfect linear relationship.
- Relationships can be linear or curvilinear.
- To determine the shape, one draws a scatter plot of the variables.

- If the relationship is linear, the data can be approximated by a straight line, called the regression line or the line of best fit.
- In addition, relationships can be multiple. That is, there can be two or more independent variables and one dependent variable.
- A coefficient of correlation and a regression equation can be found for multiple relationships, just as they can be found for simple relationships.
- The coefficient of determination is a better indicator of the strength of a linear relationship than the correlation coefficient.
- It is better because it identifies the percentage of variation of the dependent variable that is directly attributable to the variation of the independent variable.
- The coefficient of determination is obtained by squaring the correlation coefficient and converting the result to a percentage.
- Another statistic used in correlation and regression is the standard error of estimate, which is an estimate of the standard deviation of the y values about the predicted y' values.
- The standard error of estimate can be used to construct a prediction interval about a specific value point estimate $\hat{y}$ of the mean or the y values for a given x.

Conclusion

Many relationships among variables exist in the real world. One way to determine whether a relationship exists is to use the statistical techniques known as correlation and regression.

10.5 Example: Mother's age and baby's birth weight

Researchers claim that adolescent females are much more likely to deliver low-birth-weight babies than are adult females.

The following data on

x = maternal age (years)

y = birth weight of baby (grams)

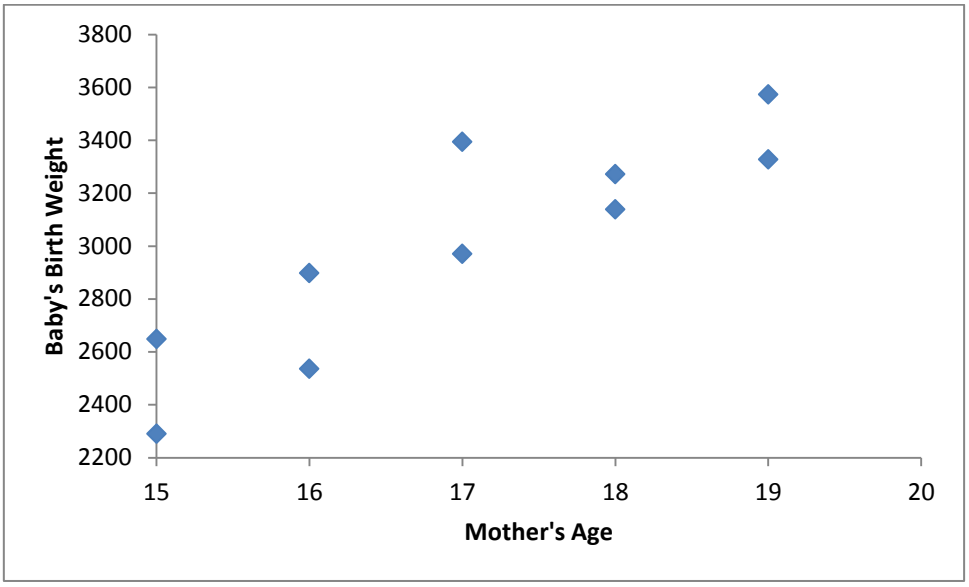
Mother's Age	15	17	18	15	16	19	17	16	18	19
Baby's Birth Weight	2289	3393	3271	2648	2897	3327	2970	2535	3138	3573

$$n = 10 \qquad \sum x = 170 \qquad \sum y = 30{,}041$$

$$\sum x^2 = 2910 \qquad \sum y^2 = 91{,}785{,}351 \qquad \sum xy = 515{,}600$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 20 \qquad S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 4903.0$$

Least squares line: $\hat{y} = -1163.45 + 245.15\ x$



Interpreting the LS line

- For x = 17.5 yrs the LS lines predicts a birth weight $\hat{y}$ = 3126.6 grams.
- However, birth weight is not completely determined by mother's age. The actual y value will differ for different 17.5 yr olds.
- If the actual weight is y then

  y = -1163.45 + 245.15 x + e

  where e may be positive or negative.

10.6 Using the Model for Estimation and Prediction

Assumptions about e

- e is Normally distributed.
- The values of e associated with any two x's are independent.
- e is a random variable with a mean of 0, i.e. E(e) = 0.
- The variance of e is a constant σ^2 for all values of x.

Estimating α and β

In the population E(y) = α + βx

From a sample data set the least squares line is $\hat{y}$ = a + b x.
a and b provide estimates of α and β respectively.

The LS line fitted to the birth weight data is: $\hat{y}$ = -1163.45 + 245.15 x

- The value 245.15 is an estimate of the slope β of the population regression line.
- β is the amount by which the average birth weight differs when comparing mothers with a one year age difference.
- Hence we estimate the average birth weight increases by 245.15g for each year of mother's age.

Properties of the estimate b

Recall:

$$b = \frac{S_{xy}}{S_{xx}}$$

Where

$$S_{xy} = \sum (x - \bar{x})(y - \bar{y}) \quad \text{and} \quad S_{xx} = \sum (x - \bar{x})^2$$

Provided $y_1, y_2, ..., y_n$ are independent:

1. The mean of b is β i.e. E(b) = β
2. The variance of b is $\mathrm{var(b)} = \frac{\sigma^2}{S_{xx}}$
3. If for each value of x the variable y has a normal distribution, then b is normally distributed so

$$b \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$

Estimating σ^2

The result in 3 (above) depends on a value for σ^2 which is usually unknown.

An estimate for $\sigma^2$is

$$s_e^2 = \frac{SSResid}{n-2}$$

Therefore standard errors (estimated standard deviations) for e and b are

$$s_e = \sqrt{\frac{SSResid}{n-2}}$$

and

$$s_b = \frac{s_e}{\sqrt{S_{xx}}} = \frac{\sqrt{\frac{SSResid}{n-2}}}{\sqrt{S_{xx}}}$$

We will need to use this when calculating confidence and prediction intervals.

Example: Mother's age and baby's birth weight (continued)

For the birth weight data:

SSResid = $\sum y^2 - a\sum y - b\sum xy$
$= 91785 - 1163.45(30041) - (245.15)(515600)$
$= 337212.5$

$$s_e{}^2 = \frac{SSResid}{n-2} = \frac{337212.5}{10-2} = 42151.56$$

$$s_e = \sqrt{42151.56}$$

$$s_b = \frac{s_e}{\sqrt{S_{xx}}} = \frac{\sqrt{42151.56}}{\sqrt{20}} = 45.90837$$

Confidence interval for β

A 95% confidence interval for β the slope of the population regression line, is of the form:

$$b \pm (\text{t critical value})\, s_b$$

where the t critical value is based on df = n - 2 and the required confidence level (e.g. 95% or 90%).

Example: Mother's age and baby's birth weight (continued)
A 95% confidence interval for β is:

$$245.15 \pm (2.306)(45.90837)$$
$$245.15 \pm 105.8647$$
$$(139.28, 351.01)$$

This confidence interval is the interval estimate of the slope parameter β. In terms of this example we can be 95% confident that the true mean increase in baby's weight for each additional year age difference in mother's age is between 139.28 and 351.01 grams.

Interpreting $\hat{y}$

Let x* be a specific value of the predictor x
e.g. x = 17.5 yr in the birth-weight example.

1. When x = x*, E(y) = α+ βx*. Thus

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

   is a point estimate of the mean y when x = x*.

   For the example:

   $$\hat{y} = -1163.45 + 245.15(17.5) = 3126.675$$

2. $\hat{y}$ would also be our prediction for an observation that would be observed when x = x*.

The precision of the estimate in 1 will be better than in 2.
In 2 we allow for the lack of precision in estimating E(y), and also of the amount by which an individual may differ from the mean.

Standard errors and confidence intervals

The standard error for a + bx* is

$$s_{a + bx^*} = s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

A confidence interval for α + βx*, the mean of y when x = x* is

$$\text{a + bx* ± (t-critical value) } s_{\text{a + bx}^*}$$

Example Mother's age and baby's birth weight (continued)

Computer output gives the following ANOVA table with values for SSReg, SSResid and SSTo and the associated df.

Source	df	SS	MS
Regression	1	1201970.45 (SSReg)	1201970.45
Residual Error	8	337212.45 (SSResid)	42152.45 (s_e^2)
Total	9	1539182.9 (SSTot)	

From this $s_e = \sqrt{42151.56}$ = 205.30844 (as seen previously).

$$s_{\text{a + b}(17.5)} = (205.30844)\sqrt{\frac{1}{10} + \frac{(17.5 - 17)^2}{20}} = (205.30844)\sqrt{0.1125} = 68.86251$$

For 95% confidence with 8df, t-critical value = 2.306.
Therefore a 95% confidence interval for the mean birth weight of babies born to 17.5 year old mothers is:

3126.675 ± (2.306)(68.86254)
3126.675 ± 158.7970
(2968, 3285)

WE are 95% confident that the mean weight of babies born to 17.5 year old mothers lies between 2968 and 3285 grams.

Prediction Interval

A prediction interval for a single y value made when x = x* is

$$a + bx \pm \text{(t-critical value)}\sqrt{s_e^2 + s_{\text{a + bx}^*}^2}$$

Example Mother's age and baby's birth weight (continued)

$$s_e^2 + s_{\text{a + bx}^*}^2 = 42152.45 + 68.86254^2 = 46894.5$$

$$\sqrt{s_e^2 + s^2_{\text{a + bx}^*}} = \sqrt{46894.5} = 216.5514$$

Hence the required 95% prediction interval is:

$$3126.675 \pm (2.306)(216.5514)$$
$$3126.675 \pm 499.3675$$
$$(2627, 3626)$$

When mother's age is equal to 17.5 years we predict with 95% confidence that the babies weight for this new individual will lie between 2627 and 3626 grams.

Note that the prediction interval is wider that the confidence interval for the mean.

Note:

Prediction intervals:

$$a + bx \pm \text{(t-critical value)}\sqrt{s_e^2 + s^2_{\text{a + bx}^*}}$$

Now $s_{\text{a + bx}^*} = s_e\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$ which is also equal to $\sqrt{\frac{s_e^2}{n} + \frac{s_e^2(x^* - \bar{x})^2}{S_{xx}}}$

Therefore $s^2_{\text{a + bx}^*} = \frac{s_e^2}{n} + \frac{s_e^2(x^* - \bar{x})^2}{S_{xx}}$

Now $s_e^2 + s^2_{\text{a + bx}^*} = s_e^2 + \frac{s_e^2}{n} + \frac{s_e^2(x^* - \bar{x})^2}{S_{xx}} = (s_e^2)\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)$

Now $\sqrt{s_e^2 + s^2_{\text{a + bx}^*}} = \sqrt{s_e^2 + \frac{s_e^2}{n} + \frac{s_e^2(x^* - \bar{x})^2}{S_{xx}}} = s_e\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$

10.7 Assessing the Utility of the Model: Hypothesis tests concerning β

Suppose there is a linear relationship of the form: $E(y) = \alpha + \beta x$

If $\beta = 0$:

- $E(y) = \alpha$ a constant, so E(y) does not depend on x
- knowing x does not give information about y
- in particular x does not explain the variation in y

Tests about β(General)

Null Hypothesis:

H_0: β = hypothesized value (β_0)

Test statistic:

$$t = \frac{b - \beta_0}{s_b}$$

based on n - 2 df

Alternative hypothesis H_A

H_A	**p-value**
$\beta > \beta_0$	Area to right of t under the t curve
$\beta < \beta_0$	Area to left of t under the t curve
$\beta \neq \beta_0$	2(Area to right of t) if t > 0 2(Area to left of t) if t < 0

Model Utility Test (Specific Test: H_0: β = 0)

This is the previous test with hypothesized value = 0, so:

H_0: β = 0

H_A: β ≠ 0

with test statistic

$$t = \frac{b}{s_b}$$

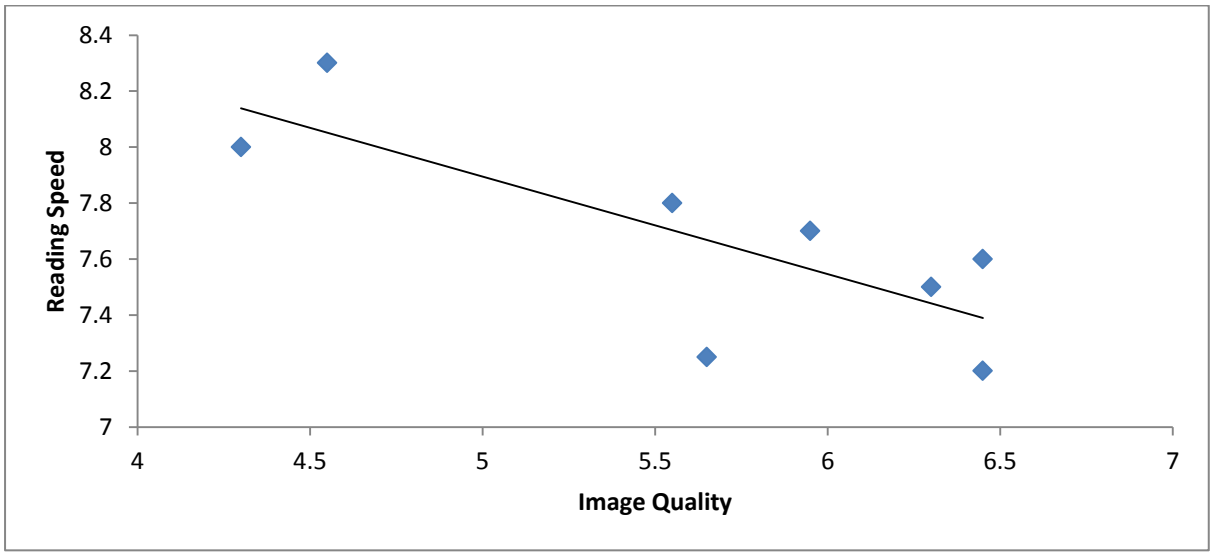
This tests whether a simple linear model is useful.

Example: Image Quality and Reading Speed

An experiment studied CRT image quality (x) and average time for a group of subjects to read certain passages (y, sec).

The following data were obtained:

x	4.30	4.55	5.55	5.65	5.95	6.30	6.45	6.45
y	8.00	8.30	7.80	7.25	7.70	7.50	7.60	7.20



Is a simple linear model is useful?

H_0: $\beta = 0$

H_A: $\beta \neq 0$

For this example:

n = 8, b = -0.348501 SSResid = 0.367626, Sxx = 4.835

$$s_e^2 = \frac{0.367626}{6} = 0.061271 \qquad s_b = \sqrt{\frac{0.061271}{4.835}} = 0.1125717$$

$$t = \frac{b}{s_b} = \frac{-0.348501}{0.1125717} = -3.10$$

Table 9, page 43 NCST

TABLE 9. THE t-DISTRIBUTION FUNCTION

$v =$	...	6
t = 0.0	...	0.5000
⋮	⋮	⋮
3.0	...	0.9880
.1	...	**0.9894**
⋮	⋮	⋮

p = 2(1 - 0.9894) = 0.02 (2 dec. pls.)

Conclusion: Taking α= 0.05, H_0: β = 0 is rejected if p < 0.05.

Hence, the data rejects H_0 (at the 5% significance level) and we conclude that $\beta \neq 0$ (i.e. x is useful in explaining the variation in y).

Computer output - Regression Analysis

The regression equation is: reading time = 9.63 - 0.35 image quality

Predictor	**Coefficient**	**Standard Deviation**	**T**	**P-value**
Intercept	9.6378	0.6419	15.015	0.000
Image quality	-0.3485	0.1125	-3.096	0.0212