Lecture 29 (MWF) Linear regression: The linear model and statistical inference

Suhasini Subba Rao
Suppose I randomly pick a pick an adult and I ask you to guess their height. You would probably give me an interval, of say, 4.5 to 6.5 feet.

Suppose I gave you the additional information that they have size 5 feet, would you reassess your previous estimate?

Your would probably change your estimate. In this case you may say their height would be between 5-5.5.

The size 5 gives us additional information about that person. It allows us to narrow down our estimate and make a more ‘precise’ estimate of her height.

Put into statistical terms, without knowledge of their shoe size the standard deviation is quite large. Recall that standard deviation is a
measure of error. Once we know their shoe size the standard deviation (amount of error) decreases.

- Often we believe that one variable may have an influence on another variable.

- For example the variable $X$ (shoe size of person) may influence the variable $Y$ (the height of that person).
  - We call $X$ the independent variable.
  - We call $Y$ the dependent variable.

- To see if $X$ has an influence on $Y$ we often plot a scatter plot with $X$ on the $x$-axis and $Y$ on the $Y$-axis. We look for a relationship between the two.
• Sometimes it is not clear what influences what (for example does shoe size have an influence on height or height have an influence on shoe size), in which case, you let the dependent variable $Y$ be the variable of interest.

• In many cases, the dependency between the independent and dependent variable can be modelled using a linear relationship, which we describe below.
Review: The equation of a line $y = \beta_0 + \beta_1 x$

- Given any value $x$, we can use the equation $y = \beta_0 + \beta_1 x$ to determine the corresponding $y$-value.

- $\beta_0$ corresponds to the line of the $y$-axis when $x = 0$ (often called the intercept).

- $\beta_1$ the slope and ‘determines’ how $x$ is related to $y$. For one unit change in $x$, $\beta_1$ corresponds to the change in $y$. So if $\beta_1 = 2$, and $x$ increases by
one, then $y$ will increase by two. On the other hand, if \( \beta_1 = -2 \), and $x$ increases by one, then $y$ will decrease by two. When \( \beta_1 = 0 \), the slope is flat. This means that $x$ exerts no linear influence on $y$. 
Data: size of a person and their shoe size

Data rarely follows an exact linear line. But in several examples, a linear-type of relationship is plausible. Here is one example:

<table>
<thead>
<tr>
<th>Height $y_i$</th>
<th>6</th>
<th>14</th>
<th>10</th>
<th>14</th>
<th>26</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feet size $x_i$</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

Bivariate Fit of Height By Foot
Least squares - the line of best fit

This is the estimator which leads to the smallest squared distance between each observation pair \((x_i, y_i)\) and the corresponding point on the line, \(\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i\) (this difference turns out to be the residual). This distance is known as the residual sum of squares:

\[
\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.
\]

\(\hat{\beta}_0\) and \(\hat{\beta}_1\) are selected to minimise it.
Expressions for $\hat{\beta}_1$ and $\hat{\beta}_0$

- We need to calculate:
  
  $S_{xy} = \sum_i (y_i - \bar{y})(x_i - \bar{x})$ and $S_{xx} = \sum_i (x_i - \bar{x})^2$

  - $S_{xy}$: relationship
  - $S_{xx}$: spread of x-axis

- $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

- You never have to calculate these (that is the purpose of statistical software).

- Equation of best fitting line: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ (which we can use for prediction).
What $S_{xy}$ and $S_{xx}$ mean

• What the slope, $\hat{\beta}_1$, means:

  – A positive $\hat{\beta}_1$ means a positive relationship. In other words, if $(X_i - \bar{X})$ is positive then $(Y_i - \bar{Y})$ tends to be positive. Thus a positive $S_{xy}$ means that $\hat{\beta}_1$ will be positive.
  – On the other hand a negative $\hat{\beta}_1$ means a negative relationship. If $(X_i - \bar{X})$ is positive then $(Y_i - \bar{Y})$ tends to be negative. Thus a negative $S_{xy}$ means that $\hat{\beta}_1$ will be negative. $\hat{\beta}_1$ means a negative relationship.
  – If $X_i$ does not exert any linear influence on $Y_i$, then the product $(X_i - \bar{X})(Y_i - \bar{Y})$ can be either negative or positive and the sum $S_{xy} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$ will cancel out the negative and positive and is likely to be close to zero. In turn this means that $\hat{\beta}_1$ will be ‘close’ to zero.
• $S_{xx}$ is simply the sample standard deviation before dividing by $n - 1$, and measure the amount of variation of the independent variables.

• The value of the coefficient $\hat{\beta}_1$ will vary according to the units you use. For example, suppose you want to measure the temperature has on the volume of ice on a lake, if you measure the temperature in Celcius, the slope will be different to if you measure the temperature in Fahrenheit. Thus the slope (like the mean) is sensitive to the units used.
Factors which may effect the regression: Leverage and outliers

• The least squares estimator $\hat{\beta}_1$ of $\beta_1$ is heavily influenced by outliers and what are known as leverage points. This is why we need to look out for these points when we perform a linear regression and look for their effect on the estimators.

  – An outlier is a $y$ value (value on the y-axis) that lies outside most of the $y$-values.
  – A leverage point is a $x$ value (value on the $x$-axis) that lies outside most of the $x$-values.

• We shall now see how the influence the parameter estimates.
The effect of Leverage points

- Leverage points lie quite far from the main cluster of $x$-values (they can be considered as outliers on the $x$-axis).

- Because they are far from the main cluster, they can dramatically
influence the gradient of the line \((\hat{\beta}_1)\).

• There are three ways to check for the influence of leverage points:
  
  – Leverage plots. This is basically a qqplot for the x-values.
  – Cooks distance. This measures how much the fitted line changes if the x-value is omitted.
  – See Ott and Longnecker, Chapter 13, for details.
The effect of outliers

- The picture illustrates how the outlier point can shift and change the shape of the linear term.
Height example

| Height $y_i$ feet size $x_i$ | 6   | 14 | 10 | 14 | 26 | $\bar{y} = 14$ |
|                             | 1   | 3  | 4  | 5  | 7  | $\bar{x} = 4$ |

| Height $y_i$ $y_i - \bar{y}$ | 6   | 14 | 10 | 14 | 26 |
| feet size $x_i$ $x_i - \bar{x}$ | 1   | 3  | 4  | 5  | 7  |
| $x_i - \bar{x}$ $y_i - \bar{y}$ | -8  | 0  | -4 | 0  | 12 |
| $(x_i - \bar{x})^2$ | -3  | -1 | 0  | 1  | 3  |
| $(y_i - \bar{y})(x_i - \bar{x})$ | 9   | 1  | 0  | 1  | 9  |
| $(-8) \times (-3) = 24$ | 0   | 0  | 0  | 0  | 12 $\times$ 3 $= 36$ |

- $S_{xy} = \sum_{i=1}^{4} (y_i - \bar{y})(x_i - \bar{x}) = 24 + 0 + 0 + 0 + 36 = 60.$

- $S_{xx} = \sum_{i=1}^{4} (x_i - \bar{x})^2 = 9 + 1 + 0 + 1 + 9 = 20.$
• Then we have $\hat{\beta}_1 = \frac{60}{20} = 3$ and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \times \bar{x} = 14 - 3 \times 4 = 2.$$ 

• The line of best fit is $\hat{Y} = 2 + 3x$. 

Height example in JMP

Bivariate Fit of height By footsize

Linear Fit
height = 2 + 3*footsize

Summary of Fit
RSquare 0.803571
RSquare Adj 0.738095
Root Mean Square Error 3.829708
Mean of Response 14
Observations (or Sum Wgts) 5

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>180.000000</td>
<td>180.000</td>
<td>12.2727</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>3</td>
<td>44.000000</td>
<td>14.667</td>
<td></td>
<td>0.0394*</td>
</tr>
<tr>
<td>C. Total</td>
<td>4</td>
<td>224.000000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parameter Estimates

| Term   | Estimate | Std Error | t Ratio | Prob>|H|
|--------|----------|-----------|---------|-----|
| Intercept | 2 3.829708 | 0.52 | 0.6376 |
| footsize | 3 0.856349 | 3.50 | 0.0394* |
Is the slope really there?

- What does the slope $\hat{y} = 2 + 3x$ (where $x$ is the shoe size and $\hat{y}$ is the predictive length) tell us about the relationship between shoe size and height?

- The slope looks quite convincing, the value $\hat{\beta}_1 = 3$ is ‘large’ and positive. Looking at the plot, you can easily believe there is linear a relationship between these two variables.
• However, we can put a line through any set of points and see a relationship which may not be there.

Consider the following example, the height of people and the number of bananas they have eaten that day:

It appears that the number of bananas one eats has a profound influence on their height. However, when you look closer you see the sample size is three and it is quite easy to random choose three people which can give you such data when there is no underlying relationship between bananas and height.
Therefore our objectives are to determine:

(i) If the slope (estimator) is statistically significant? I.e. is there really evidence of a relationship. Here we need to use statistical techniques since as usual we do not observe the entire population (in this example it is just 5 people!).

(ii) If there is a relationship, how strong is this relationship, the strength of a relationship is determined by how ‘well’ the line fits the points (this is commonly measured using the $R^2$ or correlation).
Comparing bananas and feet

The main part of the inference is that we are trying to see if the global slope (to be defined precisely later) $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$. JMP tells us that the p-value for the banana example is 33%, which means that there is a 33% chance of regenerating a slope like that when there isn’t a linear relationship. Whereas, the p-value for the shoe size example is 3.94%.

Before understanding the JMP output, we first introduce the notion of a linear model.
The sample and true slope

• For every random sample we draw, we will get new values of \((Y_i, x_i)\) which gives us different values on the X-Y graph and also different lines of best fit. In other words, each new sample leads to a new \(\hat{\beta}_0\) and \(\hat{\beta}_1\), hence \(\hat{\beta}_0\) and \(\hat{\beta}_1\) are random variables and have a distribution.

• What do we mean by random sample, and what are \(\hat{\beta}_0\) and \(\hat{\beta}_1\) estimating?

• We start by understanding what \((Y, x)\) means.
Returning to the shoe size/height example

- Recall if we randomly select someone from the population of all people there is wide variation in what their height could be.

- Suppose we restrict ourselves to the subpopulation of all people with size $x = 12$ feet. Then we narrow the amount of variation in the height. Size 12 is quite a large shoe size, hence it is reasonable to suppose that people with size 12 feet are also quite tall.

- However, even if you knew that someone has size 12 feet, you will not know their exact height. Their shoe size will only give an indication of their height. You could say, that the mean height of a person with size 12 feet will be 6 feet give or take a bit. That give or take a bit, is the random variation due to differences between people.
• But it seems reasonable to assume that the mean height of a person varies linearly according to shoe sizes, ie.

\[ \bar{y} = \beta_0 + \beta_1 x \]

where \( x \) denotes the shoe size and \( \bar{y} \) the mean height of a person with size \( x \) feet. (for now ignore what \( \beta_0 \) and \( \beta_1 \) are).

Why this formula?

- Remember before we had a population of people, the average height is taken over the entire population.
  Now we have restricted ourselves to a subpopulation of people with size 12 people. The average is take over this subpopulation. It is quite reasonable to suppose that the average of this subpopulation should be quite large.
- Compare this average with the subpopulation of people with size 2
feet. This is a small size, so it is quite likely the average height of people with size 4 feet should be small, say 4.11 feet.

– Compare this average with the subpopulation of people with size 6 feet. This size is somehow ‘normal’ hence we would expect the average height of people with this shoes size to be about 5.4.

• Now let us look at the general average $\beta_0 + \beta_1 x$ again.

• If $\beta_1$ is positive, it tells us that the average height grows with shoe size. This fits with out intuition that height increases with shoe size. Hence Modelling the mean height as $\beta_0 + \beta_1 x$ seems quite reasonable.

• But remember there is a lot of variation about the mean. That is height of a randomly selected size 12 may be quite far from the mean $\beta_0 + 12 \beta_1$.

To model the variation we add error to the mean.
• Hence the linear regression model is $Y = \beta_0 + \beta_1 x + \varepsilon$ (the error is due to random variation due to differences in individuals).
The linear regression model

• The height of a randomly selection person with size $x$ feet is

$$ Y = \beta_0 + \beta_1 x + \varepsilon. $$

• $\beta_0 + \beta_1 x$ the mean height of a person who is $x$ feet tall. In general $\beta_0 + \beta_1 x$ is the mean in a linear regression.

• Examples. Suppose for now it is known that $\beta_0 = 1$ and $\beta_1 = 2$

We model the height of a person who has size two feet as

$$ Y = 1 + 2 \times 2 + \varepsilon. $$

due to random variation
Observe that 5 is the average height of a person with size two feet. The residual $\varepsilon$ is the variation about the mean because every individual is different (this is ‘individual effect’).

We model the height of a person who has size 12 feet as

$$Y = 1 + 2 \times 12 + \varepsilon.$$ 

- $23 = 1 + 2 \times 12$ is the average height of a person with size 12 feet.
The linear model and the least squares estimators

• In reality we will not know $\beta_0$ and $\beta_1$ (just like the mean is unknown), therefore the mean equation $\beta_0 + \beta_1 x$ is also unknown.

• However, we do observe $(y_i, x_i)$ (recall the shoe size/height example we had 5 observations).

• The least squares estimator $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators of $\beta_0$ and $\beta_1$ based on the sample, and $\hat{y}(x) = \hat{\beta}_1 x + \hat{\beta}_0$ is an estimators of the mean $\beta_1 x + \beta_0$.

• Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators of $\beta_0$ and $\beta_1$ we will want to do the obvious things like construct confidence intervals for $\beta_0$ and $\beta_1$ and do statistical tests.
The Prediction equation (predicting the mean)

• Once we have estimated \( \beta_0 \) and \( \beta_1 \) we can use the prediction equation 
\[
\bar{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x
\]
 to predict the mean of \( Y \) given any value of \( x \).

At this point there are a few issues that we need to keep in mind when we do this.

• As budding statisticians, we know that \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are estimators of \( \beta_0 \) and \( \beta_1 \). Thus, there will be errors in the prediction, \( \bar{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \), and it would be wiser to construct a confidence interval.

• However, we need to decide, what exactly are we trying to predict?

(i) If we are predicting the the mean \( \beta_0 + \beta_1 x \), and construct a confidence interval for the mean, \( \beta_0 + \beta_1 x \). The confidence interval will get narrower as the estimators become more reliable.
(ii) On the other hand, we are trying to predict $Y = \beta_1 x + \beta_0$ itself, the confidence interval for $Y$ will not become too narrow, even if the estimators are extremely reliable. This confidence interval not only has to take into account the variation of the estimator, but also the variability in the population. It can be wide, and will not get narrower as the sample size grow. Eg. Suppose I am interested in a CI for the height of someone with size 8 feet, even if I have huge sample size, that give me accurate estimates of $\beta_0$ and $\beta_1$, I still I will not be able to make my interval narrow, as it is impossible to predict the inter-person variation $\varepsilon$. Details are not given here but can be found in Ott and Longnecker, Chapter 11.
**Prediction and extrapolation**

If $x$ lies outside the interval of observations that we have used to calculate $\beta_0$ and $\beta_1$ (the interval where we do have information) then $\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ is extrapolating. It can be done, but extreme caution must be used, because we have not have data on how $x$ influences $Y$ in this interval. To see how it can go terribly wrong, look at the Nature article on Olympic times on my website.
Inference for the slope $\beta_1$

- In the one sample t-test that we did several weeks ago (though we never directly specified the model), the model is

$$Y = \mu + \varepsilon,$$

$\mu$ is the population mean (which we do not know), $Y$ is the observation (eg. height or a randomly chosen person) and $\varepsilon$ is the perturbation or the error about the mean (eg. of the mean height of a person is 170cm, I am 160cm, in this case $\varepsilon = -10cm$ - of course in practice we do not know what the mean height is and this has to be estimated from the a sample of heights). The sample average $\bar{X}$, calculated form $n$ observations is an estimator of $\mu$. 
• This can be extended to the linear model:

\[ Y = \beta_0 + \beta_1 X + \varepsilon. \]

Like the population mean \( \mu \), \( \beta_1 \) and \( \beta_0 \) are unknown.

• We observe the sample \((y_1, x_1), \ldots, (y_n, x_n)\), from this we can calculate \( \hat{\beta}_1 \) and \( \hat{\beta}_0 \). \( \hat{\beta}_1 \) and \( \hat{\beta}_0 \) are estimators of \( \beta_1 \) and \( \beta_0 \).

• To make inference about the unknown \( \beta_1 \) we need to know the distribution of its estimator \( \hat{\beta}_1 \).

• Since \( \hat{Y}(x) = \beta_0 + \beta_1 x \), understanding the distribution behind \( \hat{\beta}_1 \) will help us to make inference about the mean value of \( Y \) and construct CIs for \( \beta_0 + \beta_1 x \). This is what the computer is very useful for.
The distribution of $\hat{\beta}_1$

- Because for every sample drawn we will get new $(Y_1, X_1), \ldots, (Y_n, X_n)$, the estimator of the slope $\hat{\beta}_1$ will be different for each sample. Hence $\hat{\beta}_1$ is a random variable and has a distribution.

- The variance of $\hat{\beta}_1$ is $\frac{\sigma^2_\varepsilon}{S_{XX}}$ ($\sigma^2_\varepsilon$ is the variance of $\varepsilon$).

- If all the assumptions are satisfied (and the variance of $\varepsilon$ is known) it has the normal distribution

$$\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2_\varepsilon}{\sum_{i=1}^{n} (x_i - \bar{x})^2}) = \mathcal{N}(\beta_1, \frac{\sigma^2_\varepsilon}{S_{XX}}),$$

(aside: $X$ here is nonrandom).
• This result holds true, if the sample size is relatively small and the residuals $\varepsilon_t$ are normally distributed or the sample size is large (in which case we do not need normality of the residuals).
Estimating the variance and standard error

- To estimate the variance of the residuals $\sigma^2_\varepsilon$ we use the estimated residuals;

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i,$$

and use as an estimator of $\sigma^2_\varepsilon$, $s^2_\varepsilon$, where

$$s^2_\varepsilon = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

- The standard error for $\hat{\beta}_1$ is

$$\sqrt{\frac{\sigma^2_\varepsilon}{S_{XX}}} = \sqrt{\frac{\text{var}(\varepsilon_i)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$
Factors that influence the standard error of $\beta_1$

- Of course we want a small standard error.

- We see to get a small variance (thus a better estimator of $\beta_1$), we require;
  - The sample size $n$ to be large. This means more $(X_i - \bar{X})^2$ terms, the larger the denominator $\sum_{i=1}^{n}(x_i - \bar{x})^2$, therefore $\frac{\sigma^2}{S_{XX}}$ gets smaller.
  - More variation in $X_i$. This means the more different values $X_i$ takes the larger $(x_i - \bar{x})^2$, hence $S_{XX}$ will be larger and $\frac{\sigma^2}{S_{XX}}$ will be smaller.

- More different values of $X_i$ mean the more sure we are of the gradient.

- If you are able to design your experiment (this means select $X_i$ - eg. people with certain shoe size), choose $X_i$'s which are very different.
For example, don’t choose people with size 3 feet. Choose people with a wide range of feet sizes, this gives a lot of information about the slope $\beta_1$.

- More spread out $X_i$ means we get more information about the entire linear function $Y = \beta_0 + \beta_1 X + \varepsilon$.

- It is worth noting the distribution of $\hat{\beta}_0$ is also known. See page 561 of Ott and Longnecker.
The distribution of $\hat{\beta}_1$

- Since $\sigma^2_\varepsilon$ is unknown and we use instead the sample variance based on the residuals $s^2_\varepsilon$, we do not use the normal distribution but the t-distribution.

\[
\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{s^2_\varepsilon}{S_{XX}}} \sim t_{n-2}}.
\]

- It is a $t$-distribution with $(n-2)$-degrees of freedom (the reason for -2 rather than -1, is because we have now estimated two parameters $\beta_0$ and $\beta_1$). Before we had estimated only one parameter, the mean $\mu$.

- We can use this result to:
  - Construct confidence intervals for the true slope $\beta_1$. 


Test whether the x-variable exerts an influence on the y-variable, in other words test for the true slope \( \beta_1 \) (often \( H_0 : \beta_1 = 0 \) against \( H_A : \beta_1 \neq 0 \), or \( H_0 : \beta_1 \leq 0 \) against \( H_A : \beta_1 > 0 \) or \( H_0 : \beta_1 \geq 0 \) against \( H_A : \beta_1 < 0 \)).

Construct CI for the mean \( \beta_0 + \hat{\beta}_1 x \) and also the individual observation \( Y \) (usually done in JMP).
Confidence intervals for $\beta_1$

- To make a 95% CI for $\beta_1$ use

$$[\hat{\beta}_1 - t_{n-2}(0.025) \sqrt{\frac{s^2_i}{S_{XX}}}, \hat{\beta}_1 + t_{n-2}(0.025) \sqrt{\frac{s^2_i}{S_{XX}}}],$$

where $S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2$. 
Testing $\beta_1 = 0$

We can apply the above result to test $H_0 : \beta_1 = 0$ against the alternative $H_A : \beta_1 \neq 0$.

- Make Z-transformation:

$$t = \frac{\hat{\beta}_1}{\sqrt{\frac{s_{\epsilon}^2}{S_{XX}}} \sqrt{S_X}}.$$

- Now look calculate the p-value of this $P(t(n-2) \geq |t|)$ or construct the non-rejection region.

- We can also do a one-sided test, but by calculating the area to the left or right of $t$, depending on which side the alternative is ‘pointing’.
Example: Height and Shoe size

The output tells us that the prediction equation (for the mean) is \( \hat{y} = 3x + 2 \). To test \( H_0 : \beta \leq 0 \) against \( H_A : \beta > 0 \) (noting that our hunch is that larger feet increase the chance of a taller person) the
t-transform is $t = \frac{3}{0.85} = 3.5$. The number of degrees of freedom is given next to Error and is $df=5 - 2 = 3$. The area to the right of 3.5 (using the t-tables with 3df is between 1-2%). As it is a one-sided test, the p-value is between 1-2%. If you look at the output it gives the p-value as 3.94% (this is because Statcrunch is doing a two-sided test). From this p-value we can deduce that the exact p-value in the one-sided test is $1.96\% = 3.94/2\%$.

- This means there is a 1.96% chance of seeing a positive slope like this when there is no linear dependence between shoe size and height. If we use 5% as the significance level, there is evidence in the data to suggest there is a linear dependence between the shoe size and height.

- A 95% confidence interval for the slope is

$$[3 \pm 3.182 \times 0.85].$$
Example: Height and bananas

Let us return to the bananas example:

- The prediction equations (for the mean) is $\hat{y} = 0.375x + 4.875$.

- To test $H_0 : \beta = 0$ against $H_A : \beta \neq 0$ (we have no real reason to
believe that the hypothesis should be pointing one way or the other) the
t-transform is $t = \frac{0.375}{0.21} = 1.74$. The number of degrees of freedom is
df=3 – 2 = 1. The area to the right of 1.73 (using the t-tables with 1df)
is more than 15%). As it is a two-sided test, the p-value is more than
30%. If you look at the output the precise p-value is 33%.

- This means that there that the chance of having a curve that fits like it
does through the data given that there is no linear association between
bananas and height is 33%. As this p-value is so large, there isn’t any
evidence in the data of linear dependence between height and bananas.

Remember, this result does not tell us that bananas do not influence
height, but there is no evidence in the data that there is a linear influence.
The reason for out getting this result:

- There really is NO influence.
- There is a linear influence, but we do not have enough data to detect
it (may be the influence is too small for a sample size of three to detect).
– The influence could be nonlinear - which a linear model cannot account for.
Example: Banks and Businesses

As one part of a study of commercial bank branches, data are obtained on the number of independent businesses \((x)\) and number of banks in 11 different areas. The commercial centers of cities are excluded. The data is given below:

<table>
<thead>
<tr>
<th>busi.</th>
<th>92</th>
<th>116</th>
<th>124</th>
<th>210</th>
<th>216</th>
<th>267</th>
<th>306</th>
<th>378</th>
<th>415</th>
<th>502</th>
<th>615</th>
<th>703</th>
</tr>
</thead>
<tbody>
<tr>
<td>banks</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

Plot the data, does a linear line seem plausible?
Solution: Banks and Businesses

Bivariate Fit of Bank By Business

Linear Fit
Bank = 1.7668461 + 0.0111049*Business

Summary of Fit

| Term            | Estimate | Std Error | t Ratio | Prob>|t|
|-----------------|----------|-----------|---------|-----|
| Intercept       | 1.766846 | 0.321175  | 5.50    | 0.003*|
| Business        | 0.0111049| 0.000845  | 13.14   | <0.001*|
Analysing the output

• It is clear from the output that the parameter estimators are $\hat{\beta}_1 = 0.011$ and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 5.4 - 0.011 \times 328.7 = 1.87$. Thus the prediction equation is $\hat{y} = 0.011x + 1.87$.

• Our hypothesis is this case is that the number of businesses positively influence the businesses, which means $H_0 : \beta_1 \leq 0$ against $H_A : \beta_1 > 0$. The t-value is $t = 0.011/0.000845 = 13.14$, the t-distribution has 10df and the area to the right of 13.14 is tiny (less than $(0.01/2)\%$).

• The data suggests that the number businesses in a region has a linear influence on the number of banks.
Example: Global Temperatures over time

We want to see global temperature anomalies have increased over time:

![Bivariate Fit of global By year graph]

\[
\text{global} = -7.291542 + 0.0036852 \times \text{year}
\]

Summary of Fit
- \( R^2 = 0.551408 \)
- \( R^2 \text{ Adjusted} = 0.548476 \)
- Root Mean Square Error = 0.149692
- Mean of Response = -0.19023
- Observations (or Sum Wgts) = 155

Analysis of Variance
- Source: Model
  - DF = 1
  - Sum of Squares = 4.2141500
  - Mean Square = 4.21415
  - F Ratio = 188.0671
  - Prob > F = <.0001*
- Source: Error
  - DF = 153
  - Sum of Squares = 3.4283771
  - Mean Square = 0.02241
  - Prob > F = <.0001*
- Source: C. Total
  - DF = 154
  - Sum of Squares = 7.6425271
  - Mean Square = 0.00001
  - Prob > F = <.0001*

Parameter Estimates
- Term: Intercept
  - Estimate = -7.291542
  - Std Error = 0.317964
  - t Ratio = -14.08
  - Prob > |t| = <.0001*
  - year
  - Estimate = 0.0036852
  - Std Error = 0.000269
  - t Ratio = 13.71
  - Prob > |t| = <.0001*
Discussion: Global Temperatures over time

• The prediction equation is $\hat{Y} = 0.0036x - 7.29$.

• Here we want to test $H_0 : \beta_1 \leq 0$ against $H_A : \beta_1 > 0$.

• The t-value is $t = 0.0036/0.00026 = 13.71$. As this is a one-sided test the p-value is the area to the right of 13.71, which is less than $(0.01/2)\%$. Therefore there is evidence that temperatures are increasing.

• We can examine how much it has increased by constructing confidence intervals for the slope. The 95\% confidence interval is

$$[0.00368 \pm 1.96 \times 0.000269] = [0.0032, 0.0042].$$
This suggests a yearly increase of between 0.0032 to 0.0042 degrees. Therefore over the past 150 years this suggests an increase of between $150 \times 0.0032$ to $150 \times 0.0042$ celsius, ie between $[0.48, 0.63]$ degrees.
Example: Global Temperatures and CO2

Next we want to investigate whether man-made emission CO2 has impacted the global temperatures.
Discussion: Global Temperatures and CO2

• The prediction equation is $\hat{Y} = 0.079x - 0.35$.

• Here we want to test $H_0 : \beta_1 \leq 0$ against $H_A : \beta_1 > 0$.

• The t-value is $t = 0.079/0.0043 = 16.44$. The area to the right of 16.44 is less than 0.01/2%. Therefore there is evidence that increasing CO2 levels are corresponding with increasing temperature levels (notice I have avoided the use of a causality here).

• Comparing CO2 with time as explanatory variables for temperature, it appears that CO2 better fits the data.

We will describe a numerical comparison for the fit in a later lecture.
Residuals

- Recall that when we did an ANOVA, and to check the assumptions we calculated the residuals.

  In this case, the residuals were the difference between the sample mean in each group and each of the observations in each group.

- Residuals are an important component in many statistical analysis. They help in determining whether such modelling assumptions are realistic or not.

- In the case of regression the residuals are the difference between each \( Y_i \) and corresponding point on the line of best fit \( \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \). The residuals are calculated with \( \hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \). The residuals are basically estimates of the deviation from the mean in the ‘true’ model \( Y = \beta_1 x + \beta_0 + \varepsilon \).
• We came across residuals above when obtaining an estimator of $\sigma^2_{\varepsilon}$ (which was used in the expression for the standard error).

• Usually, you do not have to calculate the residuals, the statistical software should give it.

• Recall the residuals are used to calculate the standard error of $\hat{\beta}_1$ (see slide 37).
The residuals in shoe size/height example

The predictor of $y_i$ given $x_i$ is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Hence the predictive line is $\hat{y} = 2 + 3x$. We use this to obtain the residuals:

| Height $y_i$ | 6 | 14 | 10 | 14 | 26 |
| Feet size $x_i$ | 1 | 3 | 4 | 5 | 7 |
| $\hat{y}_i = 3x_i + 2$ | 5 | 11 | 14 | 17 | 24 |
| Error in prediction: $\hat{\varepsilon}_i = y_i - \hat{y}_i$ | 1 | 3 | 4 | 3 | 2 |

- Residuals are very important. If they residuals are ‘small’ the line fits the data very well. If the residuals are ‘large’, the line does not fit so well. Of course, in many situations, the line fit will not be particularly good, because the association between the x and y variable is weak. However, even if the association is weak it can still be important. For example, the amount of milk a child drink may have an influence on his or her height,
but it is unlikely the link is strong. However, it is still an important association.

A strong link can be seen when the cloud of points cling closely to the line.

A weak link can be seen when the cloud of points do not cling to the cloud, but there appears to be a trend nevertheless.
The residuals in JMP
Example 2

As one part of a study of commercial bank branches, data are obtained on the number of independent businesses ($x$) and number of banks in 11 different areas. The commercial centers of cities are excluded. The data is given below:

<table>
<thead>
<tr>
<th>busi.</th>
<th>92</th>
<th>116</th>
<th>124</th>
<th>210</th>
<th>216</th>
<th>267</th>
<th>306</th>
<th>378</th>
<th>415</th>
<th>502</th>
<th>615</th>
<th>703</th>
</tr>
</thead>
<tbody>
<tr>
<td>banks</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

What are the residuals?
Solution 2

- We know from previously that the predictive equation is $\hat{Y} = 1.87 + 0.011X$.

<table>
<thead>
<tr>
<th>$x$</th>
<th>92</th>
<th>116</th>
<th>124</th>
<th>210</th>
<th>216</th>
<th>267</th>
<th>306</th>
<th>378</th>
<th>415</th>
<th>502</th>
<th>615</th>
<th>703</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>$\hat{y}$</td>
<td>2.8</td>
<td>3.0</td>
<td>3.1</td>
<td>4.1</td>
<td>4.1</td>
<td>4.7</td>
<td>5.1</td>
<td>5.9</td>
<td>6.3</td>
<td>7.3</td>
<td>8.5</td>
<td>9.5</td>
</tr>
<tr>
<td>$\hat{\epsilon}_i$</td>
<td>0.22</td>
<td>-1.04</td>
<td>-0.13</td>
<td>0.92</td>
<td>-0.14</td>
<td>0.30</td>
<td>-0.13</td>
<td>0.08</td>
<td>0.67</td>
<td>-0.29</td>
<td>0.47</td>
<td>-0.50</td>
</tr>
</tbody>
</table>

$\hat{\epsilon}_i = y_i - \hat{y}_i$ are the estimated residuals, they look ‘small’ with respect to $x$ and $y$ values, suggesting the fit is good.
Assumptions required for linear regression

In the analysis we did above we made the assumption that the estimated slope $\hat{\beta}_1$ was close to normally distributed (and thus $t$, since we had to estimate the standard deviation of the residuals). Therefore, the confidence intervals and p-values were derived under this assumption. How accurate they are depend on the following assumptions being satisfied.

(1) A straight line fits the data:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

In reality, this means if we plot the data, we see several a cloud of dots about a line.

(2) If the sample size is relatively small, the errors $\varepsilon$ are approximately normal).
(3) The variance of the errors $\varepsilon$ is the same for all samples (checking this can be a bit tricky and is not discussed here). In other words, there isn’t any heteroscedasticity.

(4) The observations are independent of each other.

If there is heteroscedasticity or dependence in the data then the standard errors are incorrect, which can give rise to the incorrect t-transform and p-values/CIs. There are ways to take these into account, but they require a more complex analysis.
1. Checking to see whether the linear model is correct

To see check whether the linear model is appropriate.

• We should fit the linear models to the data \((x_1, Y_1), \ldots, (x_n, Y_n)\).

• Obtain the parameter estimators \(\hat{\beta}_1, \hat{\beta}_0\).

• Estimate the residuals \(\hat{\epsilon}_i\).

• Make a scatter plot of \(\hat{\epsilon}_i\) against \(\{x_i\}\) or make a scatter plot of prediction \(\hat{\beta}_1 x_i + \hat{\beta}_0\) against the residuals (they are similar).

• If you still see a ‘pattern’ in the scatter plot (what looks like a relationship), then the linear model may not be the most appropriate.
• If the scatter plots looks ‘random’, then the linear plot may be appropriate.

• A pattern would suggest there is still some probably nonlinear dependence between the $Y_i$ and $x_i$ which has not been taken into account through the linear model.

To overcome this problem we could make a transformation of $x_i$ (such as looking at the relationship between $Y_i$ and $x_i^2$ etc.).
2. Checking for normality of the residuals

- It is very easy to check for normality.

- Effectively we just make a QQplot of the residuals.

  We recall we calculate the residuals by fitting a line of best fit to the data, \( \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x \), and estimate the residuals with \( \hat{\varepsilon}_t = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \).

- After estimating the residuals (there are the same number as the number of observations), we can make a QQplot of the residuals to check for normality and also a boxplot to check for outliers.

A large deviation from normality and many outliers will give inaccurate CIs and misleading results in statistical tests.
Example: Banks/Businesses and assumptions

Lecture 29 (MWF) Linear regression: The linear model and statistical inference
Discussion: Assumptions

• From the first plot of the residuals against the predicted banks, we don’t see any clear pattern. This suggests that there is no dependence between the residuals and the prediction equation. In other words, the residuals do not contain information about the x-variable that the linear model cannot model. Therefore, the linear model appears to be appropriate.

• From the second plot we see that the prediction equation closely matches the $Y$-values. This means that the linear equation appears to capture most the variability in the data (in other words the linear equation models well the number of banks).

• The right hand plot is the QQplot of the residuals. We recall that our inference on the slope was based on the normality of $\hat{\beta}_1$. Since the sample size is small, this means we require that the residuals are close
to normal. Looking at the QQplot there does not appear to be a much deviation from the line. Therefore the residuals do not appear to deviate much from normality.

- From how the data has been collected it appears that the observations are independent of each other and there is no real reason to think the data is heteroscedastic.

In summary the data appears to satisfy the assumptions for doing a simple linear regression,
Example: Temperatures over time and assumptions

**Bivariate Fit of global By year**

**Linear Fit**
global = -7.291542 + 0.00038652*year

**Diagnostics Plots**

**Residual by Predicted Plot**

**Actual by Predicted Plot**

**Distributions**

**Residuals global**

**Summary Statistics**
- Mean: 1.433e-18
- Std Dev: 0.1492052
- Std Err Mean: 0.0119644
- Upper 95% Mean: 0.0236751
- Lower 95% Mean: -0.0236756
- N: 155
Discussion: Global temperatures and assumptions

• A plot of the residuals against the predicted values, shows some sort of pattern, this suggests that the mean temperatures have not evolved in a precisely linear way over time. Therefore, if possible a more complex model (rather than just a linear one) may be more appropriate.

• The QQplot of the residuals does not show huge deviation from normality (and the sample size is quite large), which suggests that our normal/t approximation is accurate.

• However, one assumption which probably does not hold is independence of the data. As the data are temperatures over time, they are highly likely to be dependent, and this as not been accounted for in the analysis. Therefore the standard errors obtained may not be accurate.