MATH 124 Spring 2005

Lecture: Regression

Skills you should acquire from this lecture:

- These are lecture notes about regression. It is unlikely that we will get time to fully discuss these in class, but they will be useful for doing the final Excel assignment.

Related readings in the textbook:

- Sections 2.3, 2.4, 10.1

This part introduces the concept of simple linear regression. The goal is to find the equation of the line which best fits the scatterplot.

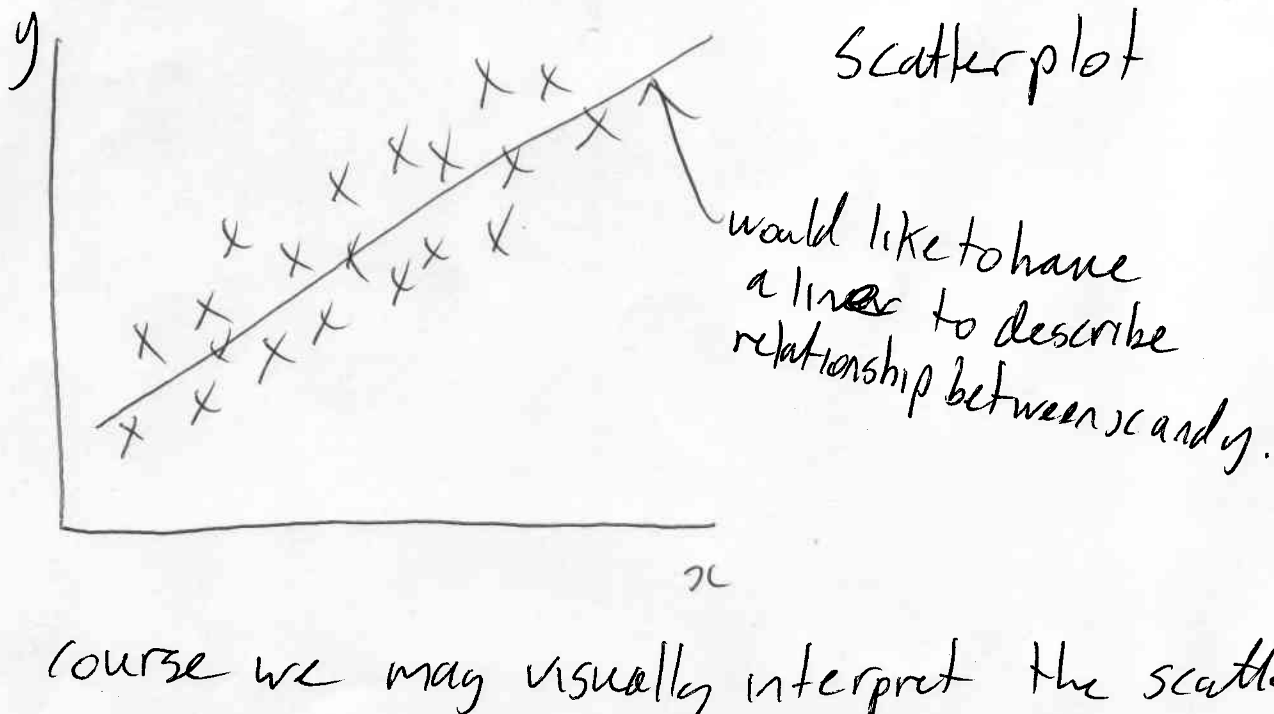
Basic Principles

n observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
on x, y paired variables

x is a explanatory variable
y is a response variable

A visual way to examine the data

Scatterplot



Of course we may visually interpret the scatterplot as we have discussed in earlier lectures. However, It would be nice to have a mathematical description of the relationship also. Often, the relationship observed in the scatterplot appears to be linear so it makes sense to describe the relationship in the form

$$\underbrace{\mu_y}_{\text{the mean of y for a given x}} = \underbrace{\beta_0}_{\text{intercept}} + \underbrace{\beta_1}_{\text{slope}} \underbrace{x}_{x \text{ variable}}$$

Note β_0, β_1 are called parameters.

However we do not observe the regression line when we draw a scatterplot because the observed y values vary about there means (μ_y). So a stricter statistical description of the regression relationship should also include a description of the variability. We do this by including a term in the model which we call the <u>residual</u>. It is the difference between the regression line and the observed y. We represent it with an ϵ (greek symbol epsilion). In addition we describe the distribution of each ϵ by assuming that it is $N(0,\sigma)$. This leads to the following formal description know as the "simple linear regression model"

Intercept ~~Slope~~ parameter

slope parameter

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

observed response

observed explanatory

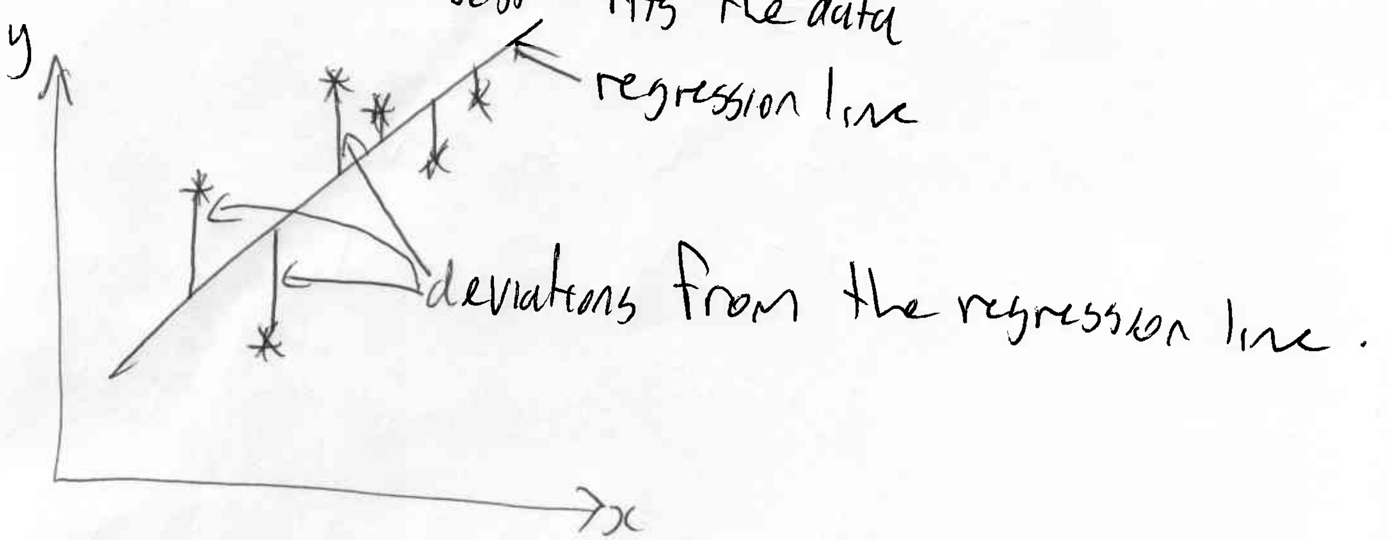
the residual

with the assumption that the ϵ_i are independent and distributed $N(0,\sigma)$

also a parameter.

Estimating β_0, β_1, σ (aka fitting the regression line)

Basic idea Find the equation of the line which best fits the data



The line which best fits the data will be the one which minimizes the deviations. What is conventionally done is to minimize the sum of the deviations squared. This is called the method of least squares. With lots of mathematics it can be shown that the estimates of β_0 and β_1 which we will call b_0 and b_1 respectively given by this method are

standard deviation of y

$$b_1 = r \frac{s_y}{s_x}$$

$\leftarrow$ standard deviation of x

correlation between x and y

and

this value is from above

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

mean of x

this is the mean of y

We call

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

the predicted/fitted regression line. And given a value of x say x^* we can predict a value of y (ie $\hat{y} = b_0 + b_1 x^*$)

We estimate the residuals using e_i (in place of ϵ_i)

$$\begin{aligned} e_i &= \text{observed response variable} - \text{predicted response variable.} \\ &= y_i - \hat{y}_i \\ &= y_i - (b_0 + b_1 x_i) \end{aligned}$$

and use these to estimate σ.

In particular,

$$s = \hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-2}} \quad \left(= \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} \right)$$

This part shows an example and then discusses how you can use a regression line to predict the value of the response variable given a value for the explanatory variable. It then talks about the topic of transformation. The goal of transformation is to make the relationship between the two variables more linear so that you can fit a simple linear regression line. Finally it mentions R-squared which can be used to judge how well the regression line fit the observed data. Higher R-squared values means better fit

Research was conducted to measure the strethability of Mozzarella cheese. In particular, the researchers were interested in how temperature related to elongation (measured as percentage increase) at breakage. They gathered the following data

x (temperature)	y (elongation)
59	118
60	143
63	182
65	210
68	247
70	243

Find the equation of the regression line.

Note that

$\sum x_i = 385$ $\quad \sum x_i^2 = 24799$ $\quad \bar{x} = 385/6 = 64.16$

$\sum y_i = 1143$ $\quad \sum y_i^2 = 231655$ $\quad \bar{y} = 1143/6 = 190.5$

$\sum x_i y_i = 74464$ $\quad n = 6$

So $s_x = \sqrt{\dfrac{24799 - 6(64.16)^2}{5}} = 4.3550$

$s_y = \sqrt{\dfrac{231655 - 6(190.5)^2}{5}} = 52.7513$

So $r = \dfrac{1}{5}\,\dfrac{1}{4.355}\,\dfrac{1}{52.7513}\left(74464 - 6(64.16)(190.5)\right)$

$= 0.9763$ $\qquad$ recall $b_1 = r\dfrac{s_y}{s_x}$ $\quad b_0 = \bar{y} - b_1\bar{x}$

So slope and intercept estimates are

$b_1 = (0.9763)\left(\dfrac{4.3550}{52.7513}\right) = 11.83$

$b_0 = 190.5 - 11.83(64.16) = -568.34$

In otherwords the fitted regression line is

$$\text{Elongation} = -568.34 + 11.83\,\text{Temperature}.$$

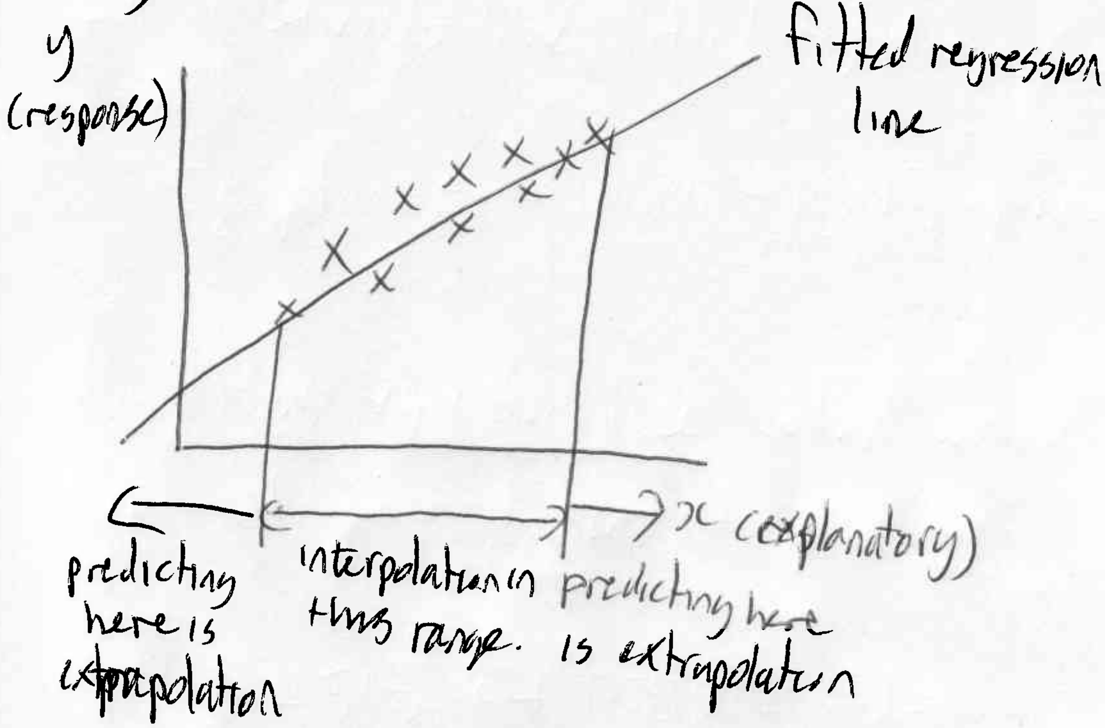
Using this regression line what would we predict the elongation to be at temperature 66.5?

$$-568.34 + 11.83(66.5) = \underline{218.355}$$

<u>Can we always use regression line to predict response given explanatory?</u>

In general **NO!!** It is usually only safe to predict at values of the explanatory within the range of the data observed. We call this <u>interpolation</u>. Predicting at values outside the range of the observed explanatory variables is called <u>extrapolation</u> it is dangerous.

eg

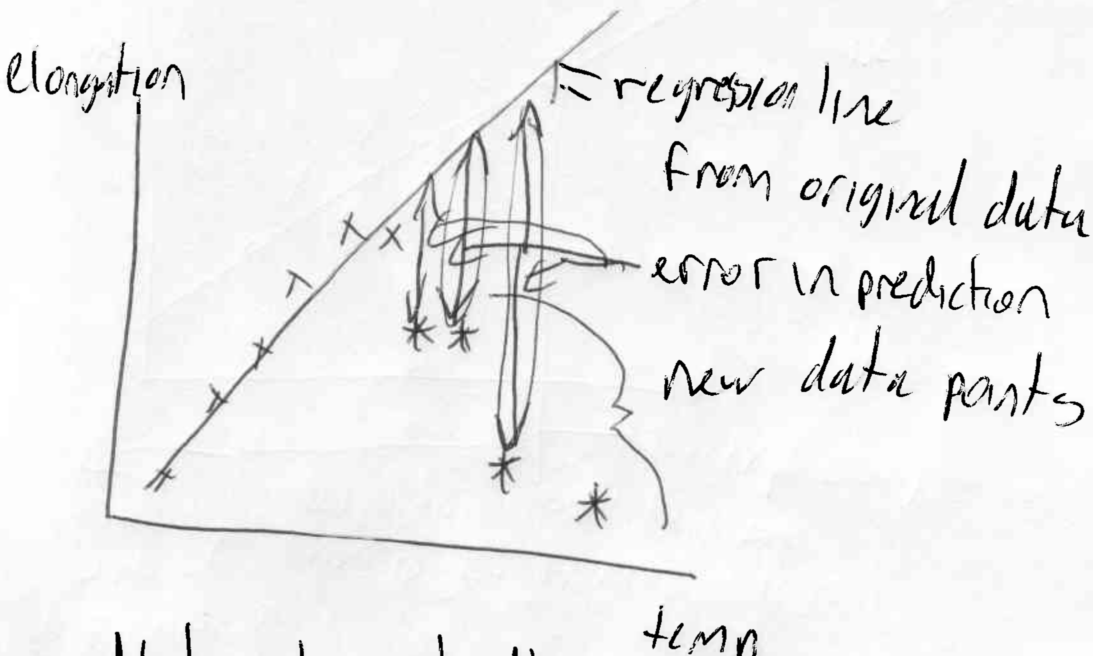


Why is it dangerous? Because we cannot be sure that the linear pattern we saw over the observed range continues outside this range

For instance further data on the mozzerella cheese study came in and was

x (temp)	y (elongation)
72	208
74	197
78	135
83	132

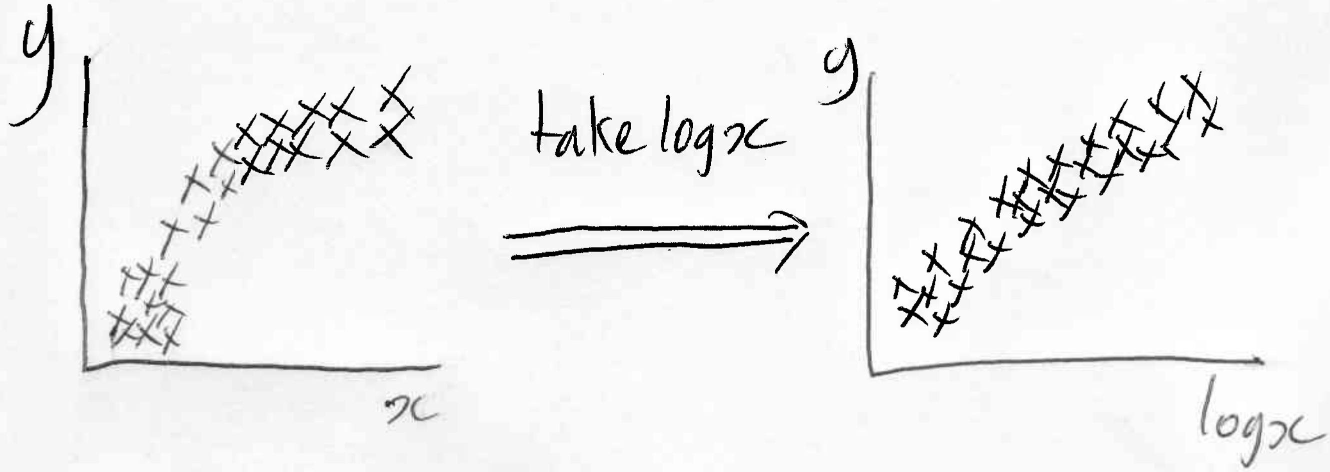
Rough sketch of situation



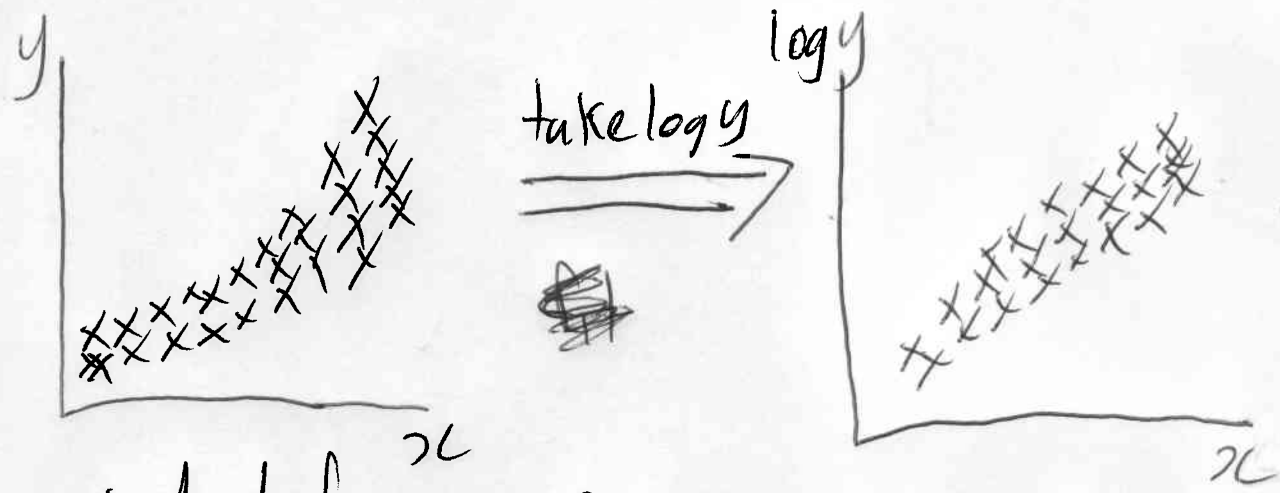
Notice how badly the regression line would have done at predicting the new points.

Transformation

Sometimes a scatterplot will not initially show what looks like a reasonable linear relationship between the two variables. This might lead you to think that a linear regression line would be inappropriate. Some of the time it is possible to transform the data and produce a more linear relationship between the two variables. We discuss 3 cases here (your textbook discusses somemore)

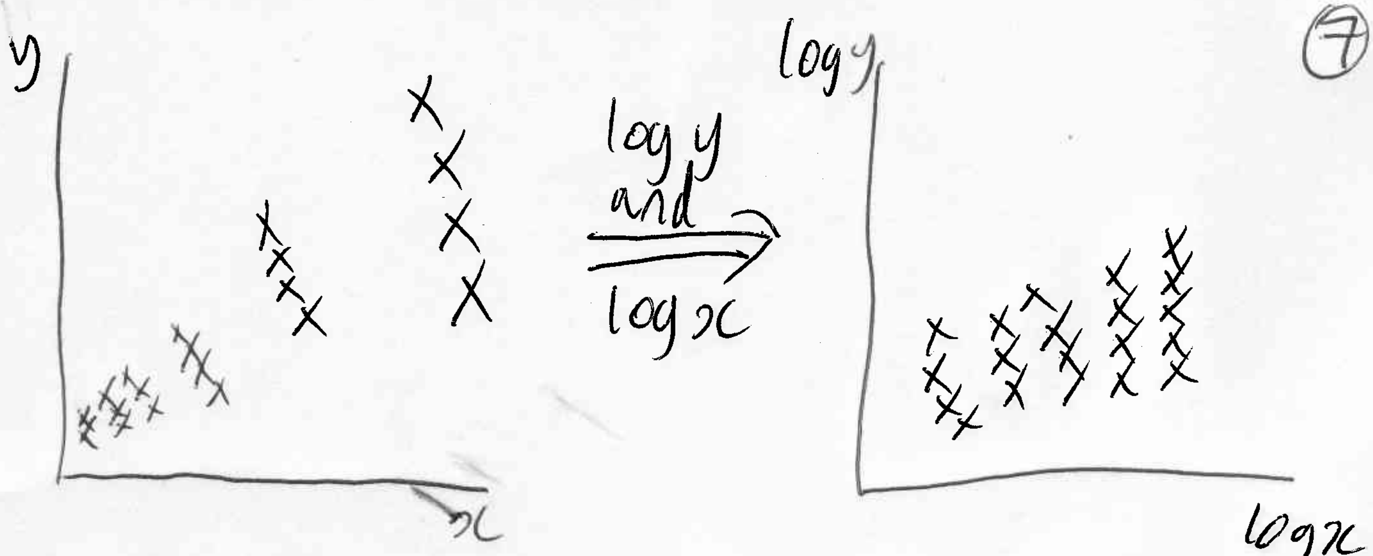
1. 

Instead of $y = b_1 x + b_0$ fit $y = b_1 \log x + b_0$

2 

instead of $y = b_1 x + b_0$ fit $\log y = b_1 x + b_0$

3.



instead of

$y = b_1 x + b_0$

fit $\log y = b_1 \log x + b_0$

<u>R-sq (aka R^2)</u>

R^2 is the square of the correlation. It can be used to judge how well the linear regression line fit the observed data.

Note

$$0 \leq R^2 \leq 1$$

no linear relationship

linear relationship models data perfectly.

This part is very advanced and talks about how you can carry out confidence intervals and hypothesis tests for the regression parameters.

Confidence intervals for regression parameters

A level C confidence interval for β_0 ← intercept is

$$b_0 \pm t^* SE_{b_0}$$

where $P(-t^* < T < t^*) = C$ and T has $t(n-2)$ distribution.

A level C confidence interval for β_1 ← slope is given by

$$b_1 \pm t^* SE_{b_1}$$

where $P(-t^* < T < t^*) = C$ and T has $t(n-2)$ distribution.

Example

Suppose I fit a linear regression to some data using excel and part of the output is

$b_0 = -2.2761$, $b_1 = 2.0680$, $SE_{b_0} = 0.1647$, $SE_{b_1} = 0.1054$
with $n=52$

What are the 95% CI for β_0, β_1?

for β_0 the 95% CI is

$$-2.2761 \pm (2.009)(0.1647)$$

$$-2.2761 \pm 0.3309$$

$$(-2.6070, -1.9452)$$

A 95% CI for β_1 is

$$2.0680 \pm (2.009)(.1054)$$

$$2.0680 \pm 0.2117$$

$$(1.8562, 2.2798)$$

<u>Note</u> I do not expect you to know how SE_{b_0}, SE_{b_1} are calculated. Only how to use them for CI and testing questions.

Hypothesis test for the slope

to test

$$H_0: \beta_1 = 0$$

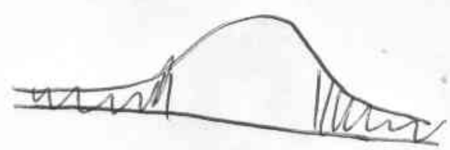
vs

$$H_A: \beta_1 \neq 0$$

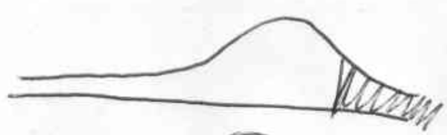
use $t = \dfrac{b_1}{SE_{b_1}}$ ← test statistic, which has t distribution with $n-2$ df.

As discussed previously the alternative tells you which tail to use for your P-value

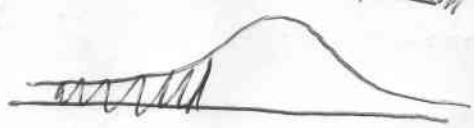
$H_A: \beta_1 \neq 0$ $\qquad 2P(T > |t|)$



$H_A: \beta_1 > 0$ $\qquad P(T > t)$



$H_A: \beta_1 < 0$ $\qquad P(T < t)$



__Note__ This test can be used to test whether the "x" variable is useful or not at predicting the "y" variable. If $\beta_1 = 0$ is true then the "x" variable is of no use (ie if your prefer H_0). Otherwise it is useful (ie if you prefer H_A).

<u>Example</u> Computer output for the regression line fitted to a particular dataset gave the following results,

$b_0 = 2.436 \qquad SE_{b_0} = 2.973$

$b_1 = -10.431 \qquad SE_{b_1} = 6.473$

$n = 18$

test whether or not the "x" variable is useful for determining the "y" value.

$H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

$t = \dfrac{-10.431}{6.473} = -1.611 \qquad df = 18 - 2 = 16$

so $\text{Pvalue} = 2P(T > |-1.611|)$

from table

$$1.337 < 1.611 < 1.746$$

$$.1 > P(T > 1.611) > .05$$

$$\Rightarrow .1 < 2P(T > 1.611) < .2$$

Since Pvalue is large cannot reject H_0.
Therefore conclude "x" variable is of no use in predicting "y" variable.

Residuals

Recall that the residual is the difference between the observed value of the response variable and that predicted by the regression line ie

$$e_i = y_i - \hat{y}_i$$

plots of residuals against the explanatory variables can be useful for assessing whether the

Linear regression line fit or not.

Plots of residuals can also be highly informative visually about the regression line and how well it fit the observed data.

This last part discusses residuals and residuals plots. These can be used to visually assess how well the linear regression line that you have fitted models the observed data.

Residuals

Recall that the residual is the difference between the observed value of the response variable and the predicted value u

$$e_i = y_i - \hat{y}_i$$

It turns out that because residuals show how far the observed data falls from the regression line they are useful for assessing how well a fitted regression line modelled the observed data. Note also that the mean of the residuals is always zero.

Specifically, the tools we use are called a <u>residual plots</u>.

Residuals plot

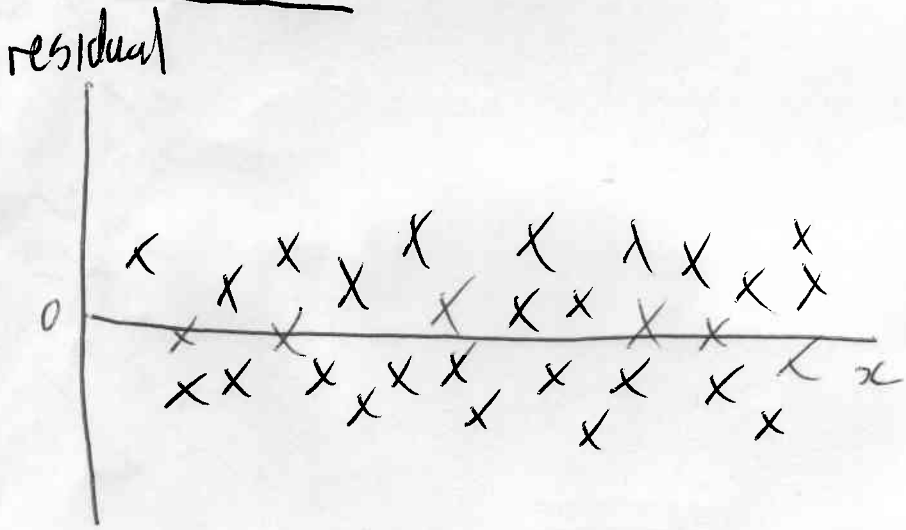
These are plots of residuals against ______

- residuals vs explanatory variable
- residuals vs predicted values

<u>Note</u> never plot residuals versus response variable.

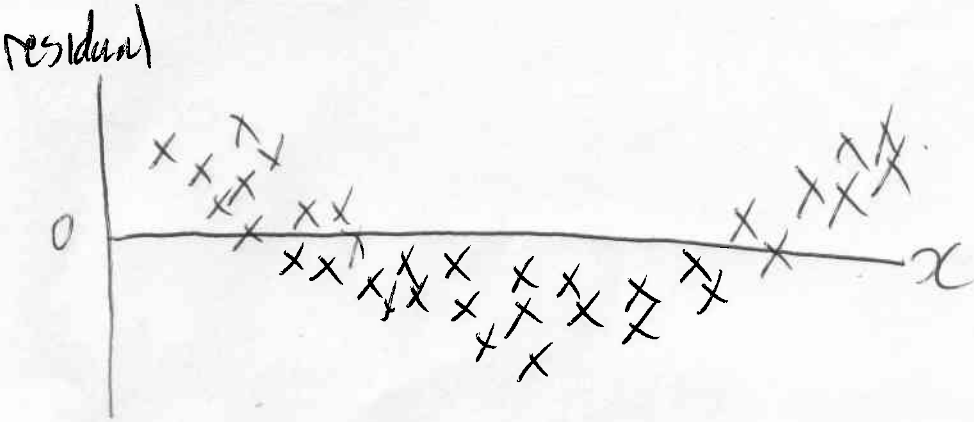
They are used to visually assess the fit of the regression line.

<u>Ideal plot</u>

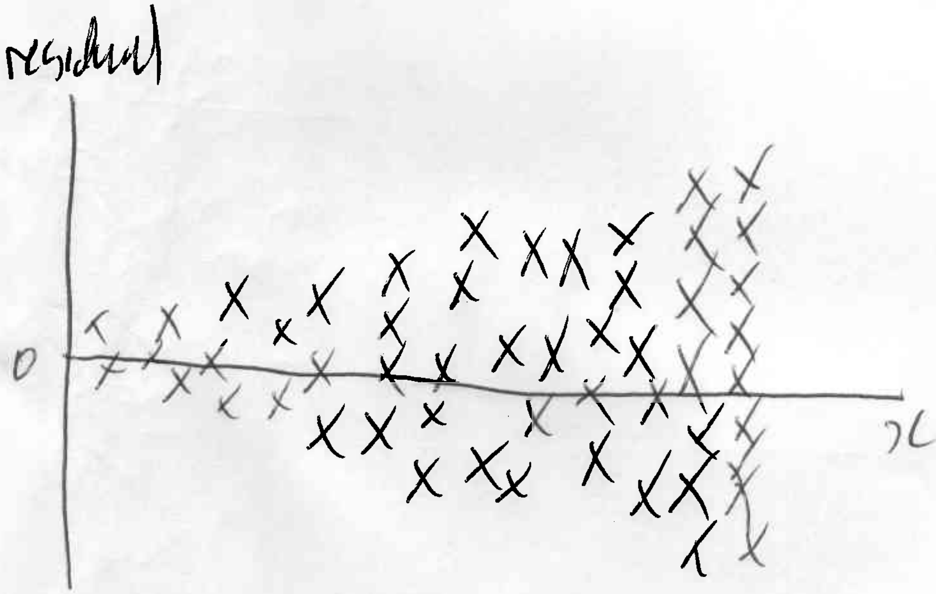


- Shows irregular pattern ie no clear pattern.
- Even vertical spread across all values on x axis

Non ideal plots



- pattern (this particular pattern indicates a curved relationship exists between y and x that is not modelled by the linear regression line)



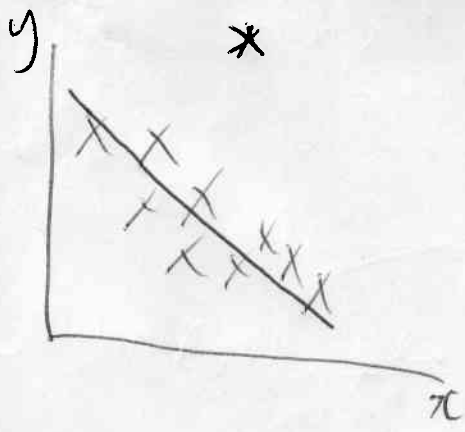
- Uneven vertical spread (this particular relationship indicates a transformation of y variable might ~~be~~ useful)

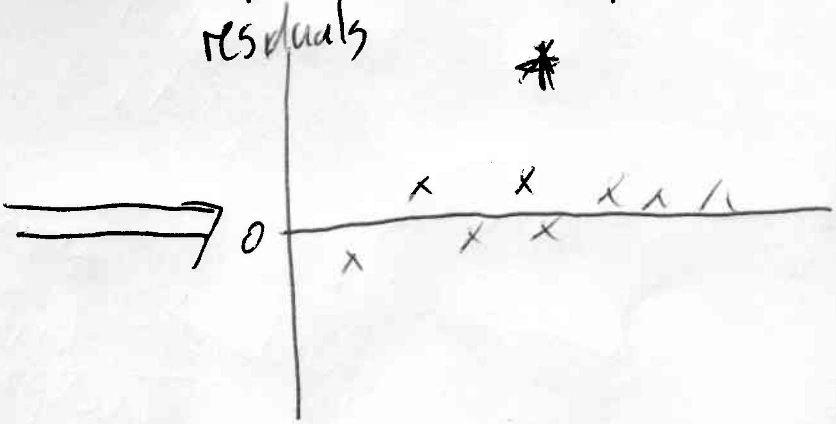
Outliers/Influential observations

An outlier is an observation that lies outside the overall pattern of the other observations.

An observation is influential if removing it would have large effect on the fitted regression line (perhaps even completely changing the equation of the regression line)

- Outliers in y tend to show up in residuals plots as





- outliers in x are often (but not always) influential

eg