

please get a handout from the back!

Stats 101/101G/108 Workshop

Regression and Correlation [RC] *Ch10!*

2020

by Leila Boyle

-we'll start @ 9.35am...



Stats 101/101G/108 Workshops

The Statistics Department offers workshops and one-to-one/small group assistance for Stats 101/101G/108 students wanting to improve their statistics skills and understanding of core concepts and topics.

Leila's website for Stats 101/101G/108 workshop hand-outs and information is here: www.tinyURL.com/stats-10x *

Resources for this workshop, including pdfs of this hand-out and Leila's scanned slides showing her working for each problem are available here: www.tinyURL.com/stats-RC *

Want to get in touch with Leila?

Leila Boyle

Undergraduate Statistics Assistance, Department of Statistics
Room 303S.288 (second floor of the Science Centre, Building 303S)

l.boyle@auckland.ac.nz; (09) 923-9045; 021 447-018

Want help with Stats?

Stats 101/101G/108 appointments

Book your preferred time with Leila here: www.tinyurl.com/appt-stats, or contact her directly (see above for her contact details).

in person or online! or via phone.

Stats 101/101G/108 Workshops

One computing workshop, four exam prep workshops and four drop-in sessions are held during the second half of the semester.

Workshops are run in a relaxed environment and allow plenty of time for questions. In fact, this is encouraged! ☺

Please make sure you bring your calculator with you to all of these workshops!

No booking is required – just turn up to any workshop! You are also welcome to come along virtually on Zoom if you prefer. Search your emails for "Leila" to find the link – email Leila at l.boyle@auckland.ac.nz if you can't find it.

• Computer workshop: Hypothesis Tests in SPSS

www.tinyURL.com/stats-HTS

Computing for Assignment 3 – covers the computing you need to do for Questions 3 and 4 (in Zight plots & SPSS output). There are six **identical sessions**.

- Friday 16 October, 3-4pm
- Monday 19 October, 10-11am
- Monday 19 October, 2-3pm
- Tuesday 20 October, 4-5pm
- Wednesday 21 October, 11am-midday
- Wednesday 21 October, 3-4pm

• Exam prep workshops

Chi-Square Tests

last week! www.tinyURL.com/stats-CS
Exam revision for **Chapter 9** – Saturday 24 October, 1-4pm, LibB15 (useful exam prep and also useful for the **Chapter 9 Quiz** due at 11pm on Wednesday 28 October!)

Regression and Correlation

today! www.tinyURL.com/stats-RC
Exam revision for **Chapter 10** – Saturday 31 October, 9.30am-12.30pm, LibB10 (useful exam prep and also useful for the **Chapter 10 Quiz** due at 11pm on Wednesday 4 November!)

Hypothesis Tests: Proportions

www.tinyURL.com/stats-HTP
Exam revision for **Chapters 6 & 7** (with a focus on proportions) – Tuesday 3 November, 9.30am-12.30pm, LibB10 (useful exam prep)

Hypothesis Tests: Means

www.tinyURL.com/stats-HTM
Exam revision for **Chapter 6, 7 & 8** – Tuesday 3 November, 1-4pm, LibB10 (useful exam prep)

(with a focus on means)

• Drop-in sessions

- Saturday 17 October, 9.30am-4pm, LibB10
- Saturday 24 October, 9.30am-12.30pm, LibB15
- Monday 26 October, 9.30am-4pm, LibB10
- Saturday 31 October, 1-4pm, LibB10

CS

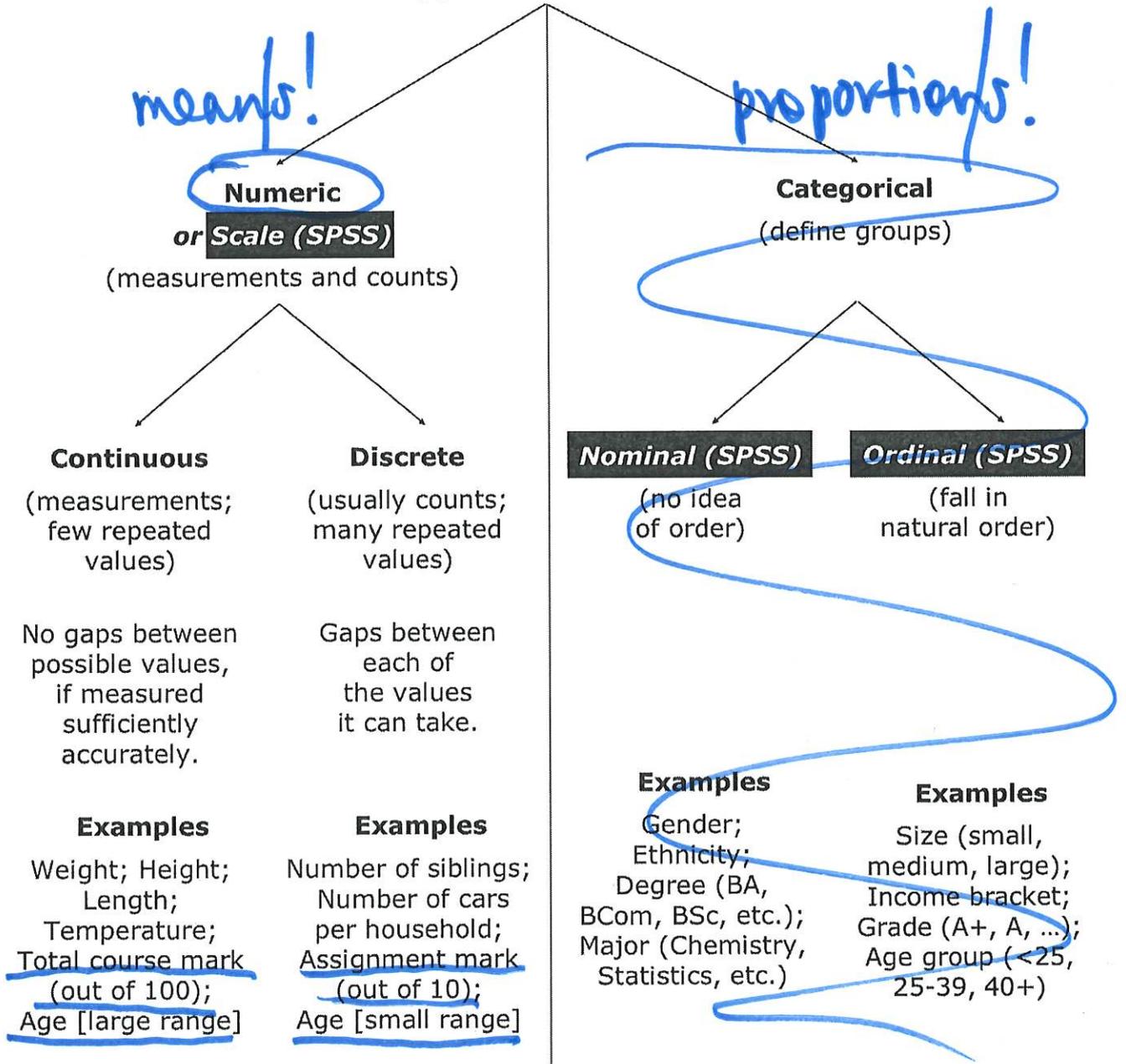
RC

HTP

HTM

Regression and Correlation

Types of Variables

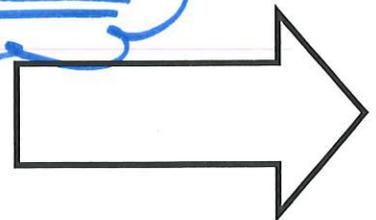


Useful reference: Chance Encounters, pages 40 – 42

The main tool for comparing two numeric variables is the scatter plot.

What to look for in a scatter plot:

- Trend (pattern)
- Scatter
- Strength of the relationship
- Association
- Outliers
- Groupings



Exploring relationships between two variables

Ch 10

Ch 6-8 (means)

Ch 9

Two Numeric

Numeric versus Categorical

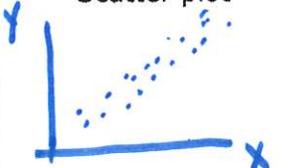
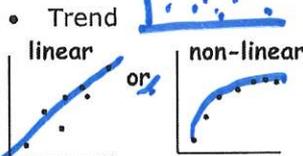
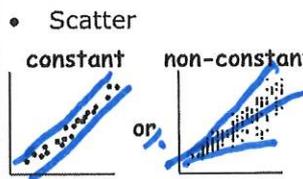
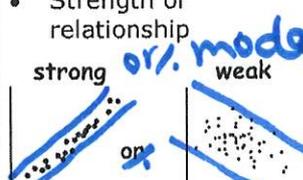
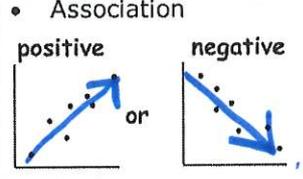
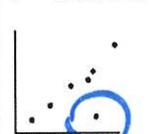
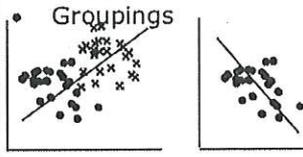
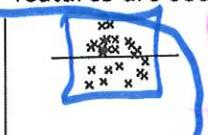
Two Categorical

RC

HTM (1&2)

CST/HTP

Exploratory tools to use to explore relationship/s

<p>Scatter plot</p> 	<p>Side-by-side plots on the same scale:</p> <ul style="list-style-type: none"> any n { • Dot plots $n \geq 20$ { • Stem-and-leaf plots { • Box plots $n \geq 50$ { • Histograms 	<p>Two-way table of counts and/or Bar graphs of proportions (on rows/columns)</p>
<p>Features to look for:</p>		
<ul style="list-style-type: none"> • Trend <ul style="list-style-type: none"> linear or non-linear • Scatter <ul style="list-style-type: none"> constant or non-constant • Strength of relationship <ul style="list-style-type: none"> strong or moderate! or weak • Association <ul style="list-style-type: none"> positive or negative • Outliers • Groupings      	<p>Compare the groups by looking at:</p> <ul style="list-style-type: none"> • Any group differences: <ul style="list-style-type: none"> ◦ averages (centres) <ul style="list-style-type: none"> medians means ◦ variability (spread) <ul style="list-style-type: none"> IQRs ranges ◦ shapes <ul style="list-style-type: none"> left/negative right/positive symmetric/skewed modes <ul style="list-style-type: none"> unimodal bimodal trimodal • Details of individual groups: <ul style="list-style-type: none"> ◦ outliers, gaps, clusters, groupings <p>Think about reasons why these differences, similarities and features are seen</p> 	<ul style="list-style-type: none"> • Most common and least common combinations • Differences in distributions (e.g. row and/or column bar graphs) <p>Chi-sq test for indep.</p>

regression & correlation

t-test or F-test

$$y = mx + c \rightarrow y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Regression

Regression looks at the relationship between two numeric variables where the two variables take on special roles:

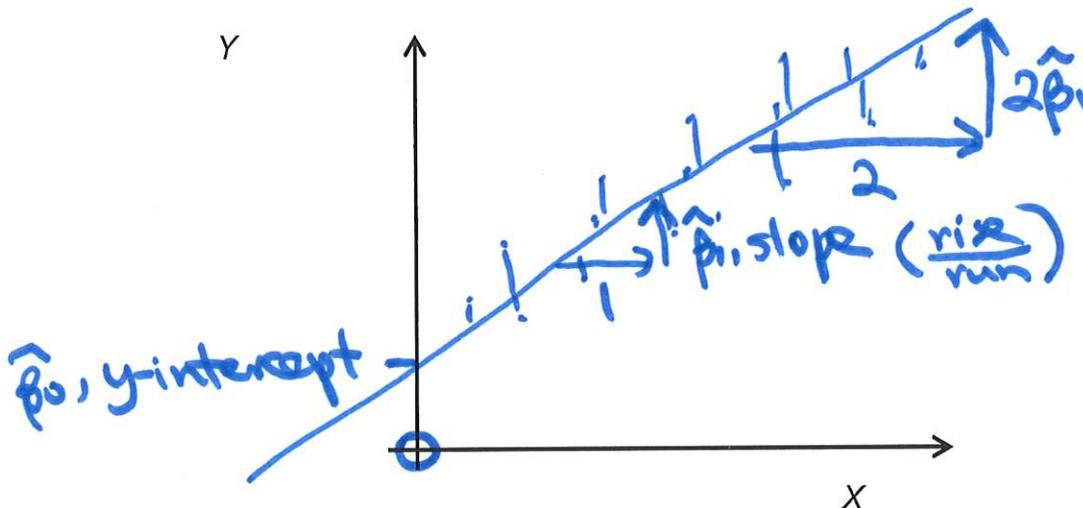
- X is used to **explain** or **predict** the behaviour of Y
- X is the **explanatory** or **independent** variable
- Y is the **dependent** or **response** variable

The two main components of the regression model are:

- **trend** and
- **scatter**.

see back page
for Formulae Sheet

We use a **least squares regression line** $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ fitted by the computer / calculator to estimate the unknown population parameters β_0 and β_1 .



"sum", "total" (Sigma)

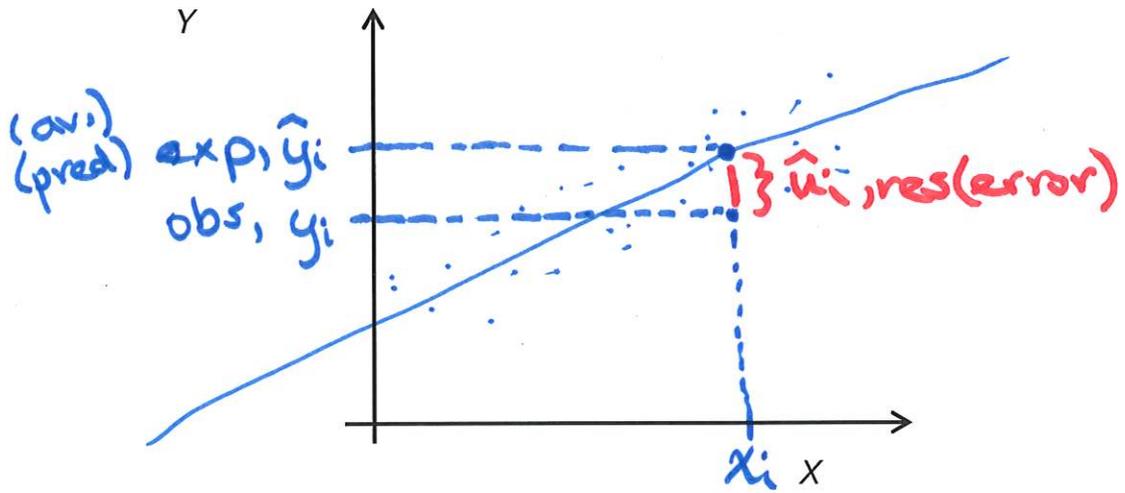
The single **least squares regression line** for each linear regression model:

- minimises the sum of the squared residuals/prediction errors
- has \sum residuals = 0 (but so do many other lines)
- has (\bar{x}, \bar{y}) lying on it

• **Residuals**

- Errors, residuals or prediction errors are all terms for the same thing.
- A residual is the (vertical) distance between the **actual observed value** y_i and the **expected estimated value** \hat{y}_i , i.e.:

Errors = observed - expected $(\hat{u}_i = y_i - \hat{y}_i)$ *memorise!*



• **Hypotheses**

Recall the *t*-test?

see back page
for Formulae Sheet

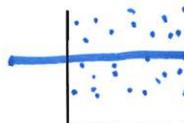
Use: $t_0 = \frac{\text{estimate} - \text{hypothesised value}}{\text{std error}}$

where in this case the degrees of freedom are:

$df = n - 2$

$H_0: \beta_1 = 0$ (there **is no** linear relationship)

↑
Slope



$H_1: \beta_1 \neq 0$ (there **is a** linear relationship)

↓



In SPSS, the output always comes out in the same way:

predicted, average, expected

Regression

Coefficients(a)

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	$\hat{\beta}_0$	$se(\hat{\beta}_0)$		t_0	p-value
X-axis_Variable	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	r	t_0	p-value

a. Dependent Variable: **Y-axis_Variable**

$df = n - 2$

y-intercept
slope, β_1

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

$t_0 = \frac{\text{est-hyp val}}{\text{std err}} = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}$

Correlations

		X-axis_Variable	Y-axis_Variable
X-axis_Variable	Pearson Correlation	1	r
	Sig. (2-tailed)		p-value
	N	n	n
Y-axis_Variable	Pearson Correlation	r	1
	Sig. (2-tailed)	p-value	
	N	n	n

Saying the same thing!

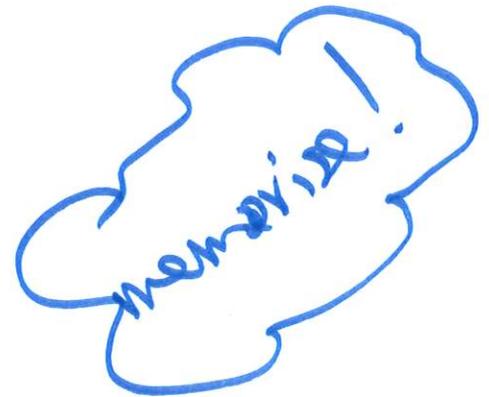
Recall: The P-value:

- In regression, we are carrying out *t*-tests, just like in Chapter 7 and 8!
- Therefore, the **P-value:**
 - ✓ is the **conditional** probability of observing a test statistic as extreme as that observed or more so, **given that** the null hypothesis, H_0 , is true.
 - ✓ is the probability that sampling variation would produce an estimate that is at least as far from the hypothesised value than the estimate we obtained from our data, **assuming that** the null hypothesis, H_0 , is true.
 - ✓ measures the strength of evidence **against** H_0 .

memorise!

- We interpret the *P-value* as a **description** of the **strength of evidence against the null hypothesis, H_0** . The **smaller** the *P-value*, the **stronger** the evidence against H_0 :

<i>P-value</i>	Evidence against H_0
> 0.10	None
≈ 0.07	Weak
≈ 0.05	Some
≈ 0.01	Strong
≤ 0.001	Very Strong



- An alternative approach often found in research articles and news items is to describe the test result as (statistically) significant or not significant. A test result is said to be significant when the *P-value* is "small enough"; usually people say a *P-value* is "small enough" if it is less than 0.05 (5%):

Testing at a 5% level of significance:

<i>P-value</i>	Test result	Action
< 0.05	Significant	Reject H_0 in favour of H_1
> 0.05	Nonsignificant	Do not reject H_0

✓ .049
X .051

Testing can be done at any level of significance; 1% is common but 5% is what most researchers use.

The level of significance can be thought of as a false alarm error rate, i.e. it is the proportion of times that the null hypothesis will be rejected when it is actually true (which can result in action being taken when really no action should be taken).

Thus, a statistically significant result means that a study has produced a "small" *P-value* (usually $< 5\%$).

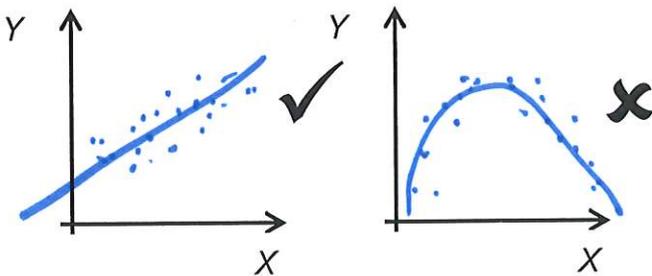
• **Assumptions** of simple linear regression are:

1. There is a **linear** relationship between X and Y .
2. **Errors** are all **independent**. *can't be checked!*
3. **Errors** are **Normally** distributed (with $\mu = 0$).
4. **Errors** all have the **same std deviation**, σ , regardless of the value of x .

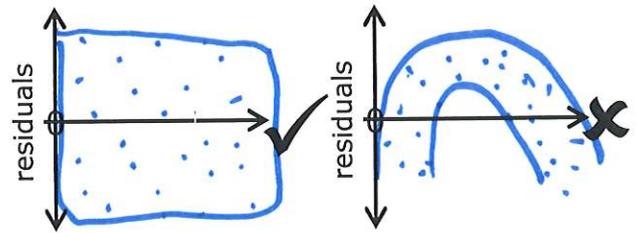
• **Assumption checking using plots of the data and residual plots**

1. There is a **linear** relationship between X and Y .

Scatterplot of data:

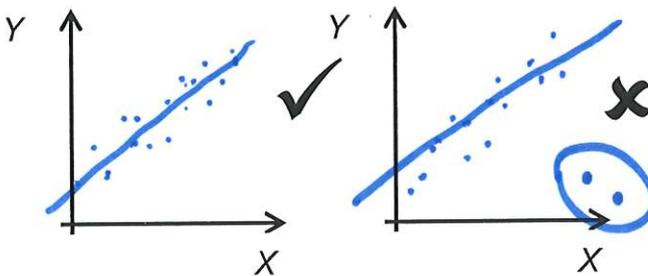


Residual plot:

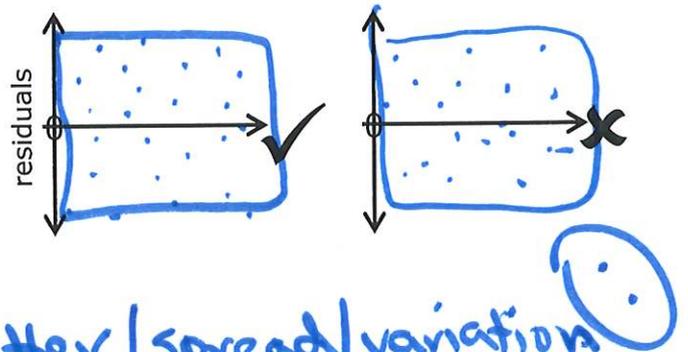


3. Errors are **Normally** distributed (with $\mu = 0$).

Scatterplot of data:

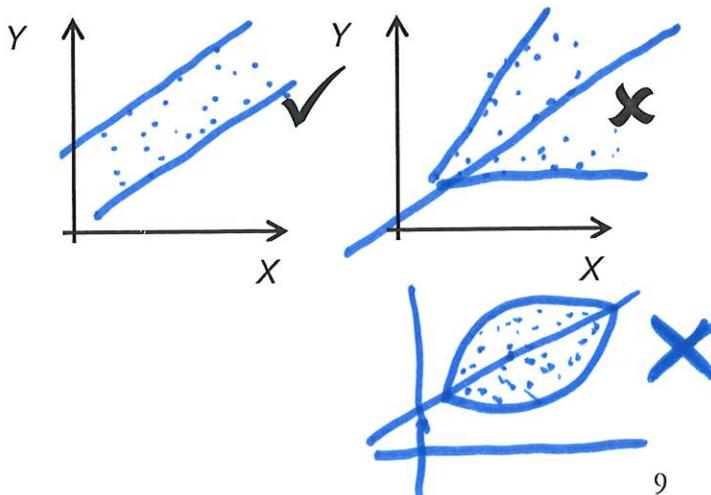


Residual plot:

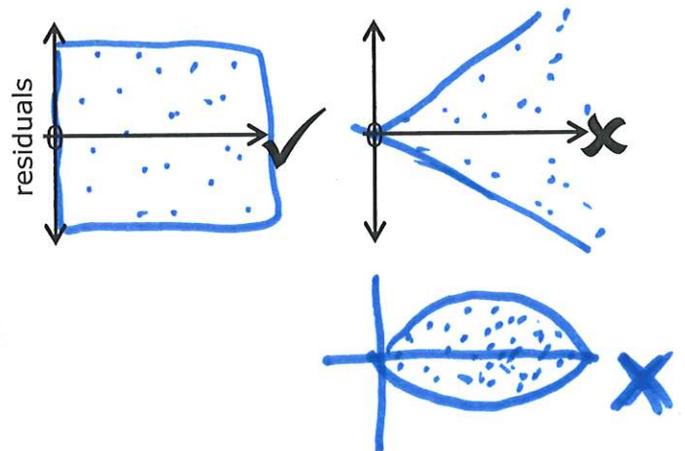


4. Errors all have the **same std deviation**, σ , regardless of the value of x .

Scatterplot of data:



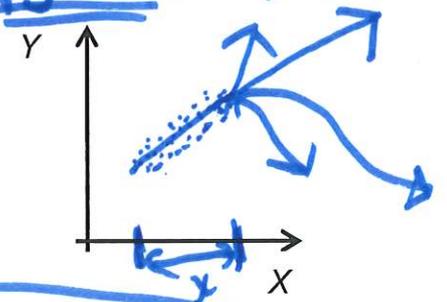
Residual plot:



11.20-11.35am - have a break!
 restart on pg 10 @ 11.35...

• **Estimating / Predicting**

- ✓ Within the range of our observed X-values this can be done with confidence. Predicting outside the range of our observed X-values is dangerous. A relationship that fits the data well may not extend outside that range.

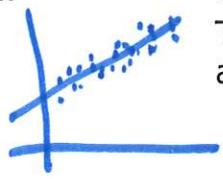


✓ **Confidence Interval (for the mean)**

This estimates the **mean** Y-value at a specified value of x. The width of the interval allows for:

- uncertainty about the values of β_0 and β_1 .

$$\text{estimate} \pm t \times \text{se}(\text{estimate})$$



✓ **Prediction Interval**

This predicts the Y-value for **an individual** with a specified value of x.

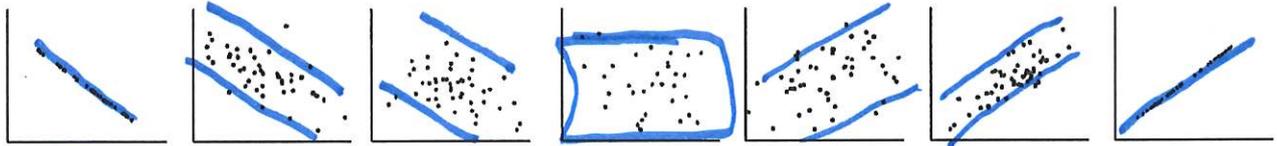
The width of the interval allows for:

- uncertainty about the values of β_0 and β_1 **and**
- uncertainty due to the random scatter about the line.

- ✓ For a given value of x, the **95% prediction interval** is **always wider** than the **95% confidence interval for the mean**.

✓ **The Sample Correlation Coefficient, r**

- ✓ r has a value between -1 and +1:



$r = -1$ $r = -0.7$ $r = -0.4$ $r = 0$ $r = 0.3$ $r = 0.8$ $r = 1$

- $r = -1$, then X and Y have a **perfect negative linear relationship**
- $r = 0$, then X and Y have **no linear** relationship but they may have **some other non-linear relationship**
- $r = 1$, then X and Y have a **perfect positive linear relationship**
- ✓ r measures the **strength** and **direction** of the **linear** association between **two numeric variables**
- ✓ r measures how close the points **come to lying on a straight line**
- ✓ The value of r is the same if the axes are swapped around - it doesn't matter which variable is X and which one is Y
- ✓ r has no units → a computer / calculator can give you the value of r

✓ **Correlation DOES NOT imply causation**