Correlation and Regression

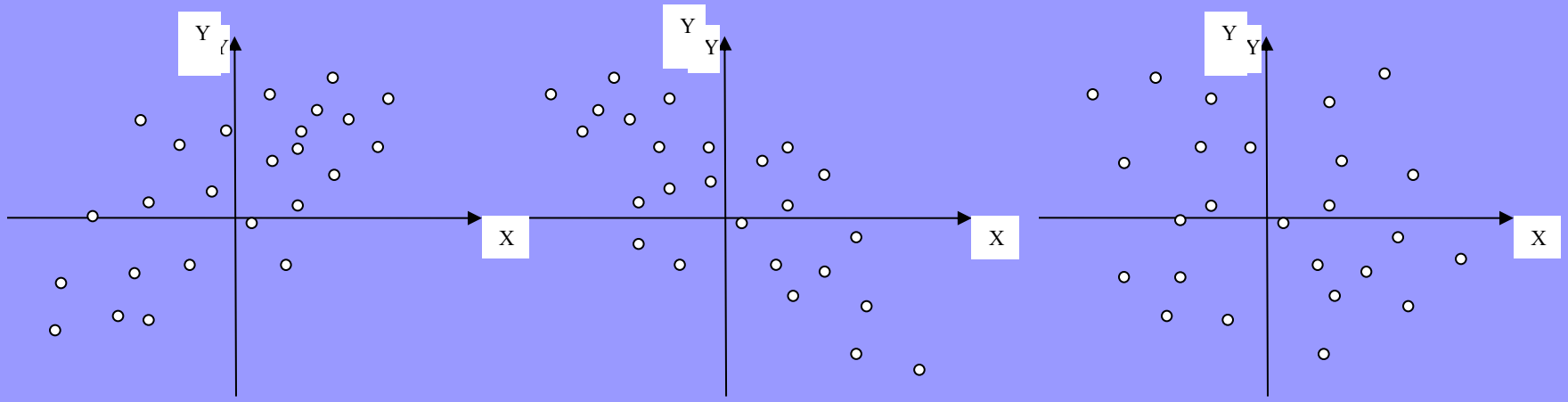
Topics Covered:

- Is there a relationship between *x* and *y*?
- What is the strength of this relationship
  - Pearson's r
- Can we describe this relationship and use this to predict *y* from *x*?
  - Regression
- Is the relationship we have described statistically significant?
  - t test
- Relevance to SPM
  - GLM

The relationship between *x* and *y*

- Correlation: is there a relationship between 2 variables?
- Regression: how well a certain independent variable predict dependent variable?
- CORRELATION ≠ CAUSATION
  - In order to infer causality: manipulate independent variable and observe effect on dependent variable

Scattergrams



Positive correlation

Negative correlation

No correlation

Variance vs Covariance

- *First, a note on your sample:*
  - *If you're wishing to assume that your sample is representative of the general population (RANDOM EFFECTS MODEL), use the degrees of freedom (n – 1) in your calculations of variance or covariance.*
  - *But if you're simply wanting to assess your current sample (FIXED EFFECTS MODEL), substitute n for the degrees of freedom.*

Variance vs Covariance

- Do two variables change together?

Variance:

• Gives information on variability of a single variable.

$$S_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Covariance:

• Gives information on the degree to which two variables vary together.

• Note how similar the covariance is to variance: the equation simply multiplies x's error scores by y's error scores as opposed to squaring x's error scores.

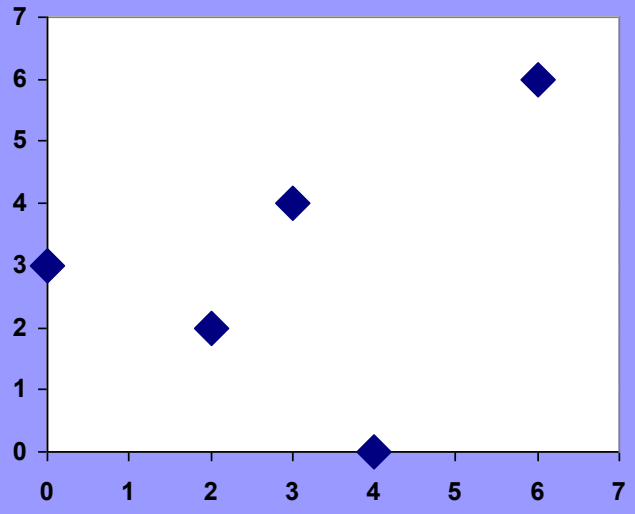
$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Covariance

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

- When X ↑ and Y ↑: cov (x,y) = pos.
- When X ↓ and Y ↑: cov (x,y) = neg.
- When no constant relationship: cov (x,y) = 0

Example Covariance



x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
0	3	-3	0	0
2	2	-1	-1	1
3	4	0	1	0
4	0	1	-3	-3
6	6	3	3	9
$\bar{x} = 3$	$\bar{y} = 3$			$\sum = 7$

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}))}{n-1} = \frac{7}{4} = 1.75$$

What does this number tell us?

Problem with Covariance:

- The value obtained by covariance is dependent on the size of the data's standard deviations: if large, the value will be greater than if small… *even if the relationship between x and y is exactly the same in the large versus small standard deviation datasets.*

Example of how covariance value relies on variance

	High variance data				Low variance data		
Subject	**x**	**y**	**x error * y error**		**x**	**y**	**X error * y error**
1	101	100	2500		54	53	9
2	81	80	900		53	52	4
3	61	60	100		52	51	1
4	51	50	0		51	50	0
5	41	40	100		50	49	1
6	21	20	900		49	48	4
7	1	0	2500		48	47	9
Mean	51	50			51	50	
Sum of x error * y error :			7000		**Sum of x error * y error :**		28
Covariance:			**1166.67**		**Covariance:**		**4.67**

Solution: Pearson's r

- **Covariance does not really tell us anything**
  - ***Solution: standardise this measure***
- **Pearson's R: standardises the covariance value.**
- **Divides the covariance by the multiplied standard deviations of X and Y:**

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{s_x s_y}$$

Pearson's R continued

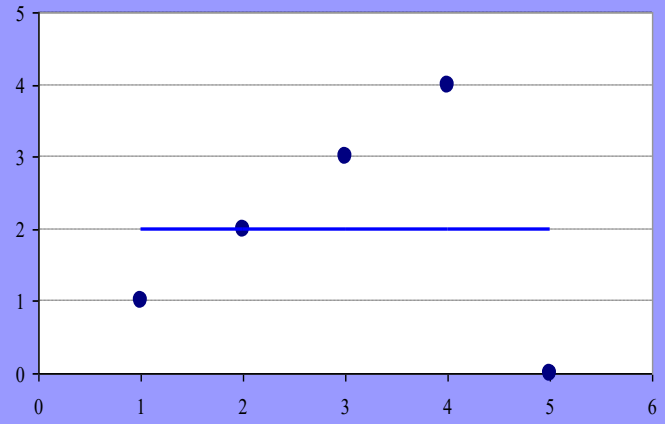
$$\mathrm{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \rightarrow r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

$$\downarrow$$

$$r_{xy} = \frac{\sum_{i=1}^{n} Z_{x_i} * Z_{y_i}}{n-1}$$

Limitations of r

- When r = 1 or r = -1:
  - We can predict y from x with certainty
  - all data points are on a straight line: y = ax + b
- r is actually $\hat{r}$
  - r = true r of whole population
  - $\hat{r}$ = estimate of r based on data
- r is very sensitive to extreme values:



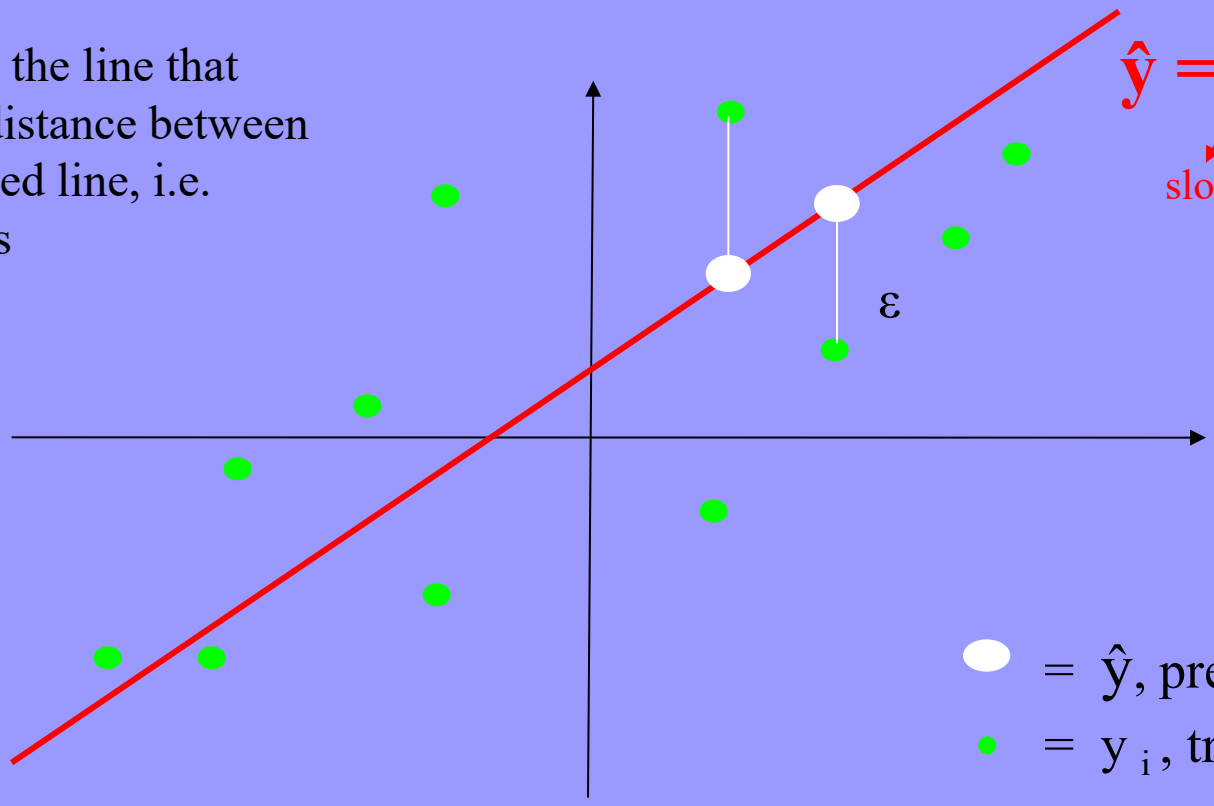
Regression

- Correlation tells you if there is an association between x and y but it doesn't describe the relationship or allow you to predict one variable from the other.

- To do this we need REGRESSION!

Best-fit Line

- Aim of linear regression is to fit a straight line, $\hat{y} = ax + b$, to data that gives best prediction of y for any value of x
- This will be the line that minimises distance between data and fitted line, i.e. the residuals



$\hat{y} = ax + b$

slope intercept

= $\hat{y}$, predicted value

= y_i, true value

ε = residual error

Least Squares Regression

- To find the best line we must minimise the sum of the squares of the residuals (the vertical distances from the data points to our line)

  Model line: $\hat{y} = ax + b$ a = slope, b = intercept

  Residual (ε) = $y - \hat{y}$

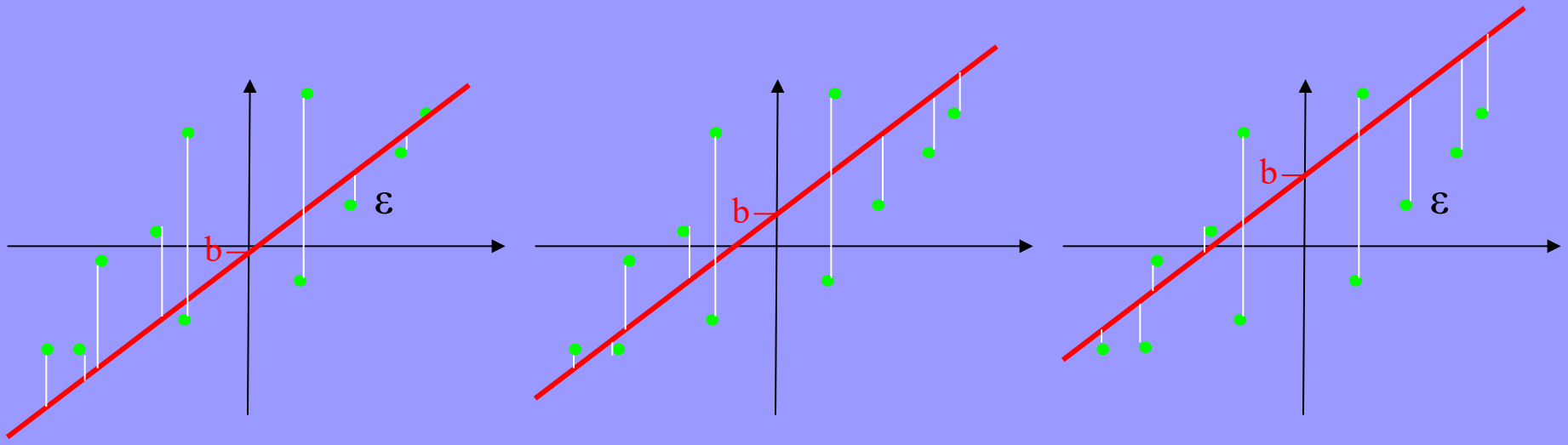
  Sum of squares of residuals = $\Sigma (y - \hat{y})^2$

  - we must find values of a and b that minimise

    $\Sigma (y - \hat{y})^2$

Finding b

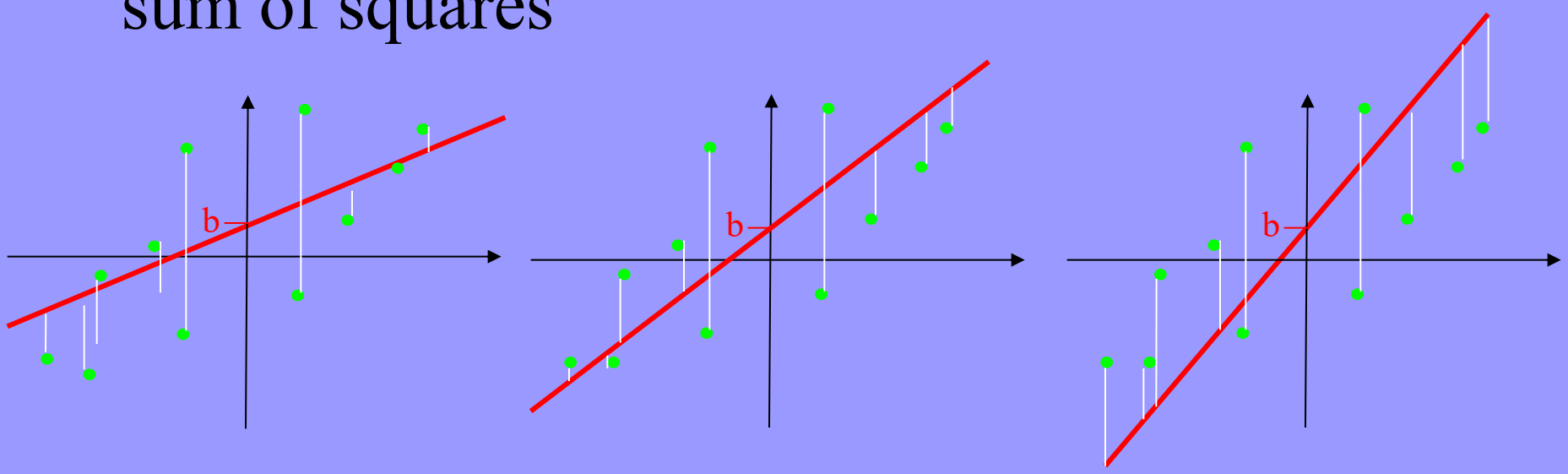
- First we find the value of b that gives the min sum of squares



- Trying different values of b is equivalent to shifting the line up and down the scatter plot

Finding a

- Now we find the value of a that gives the min sum of squares



- Trying out different values of a is equivalent to changing the slope of the line, while b stays constant

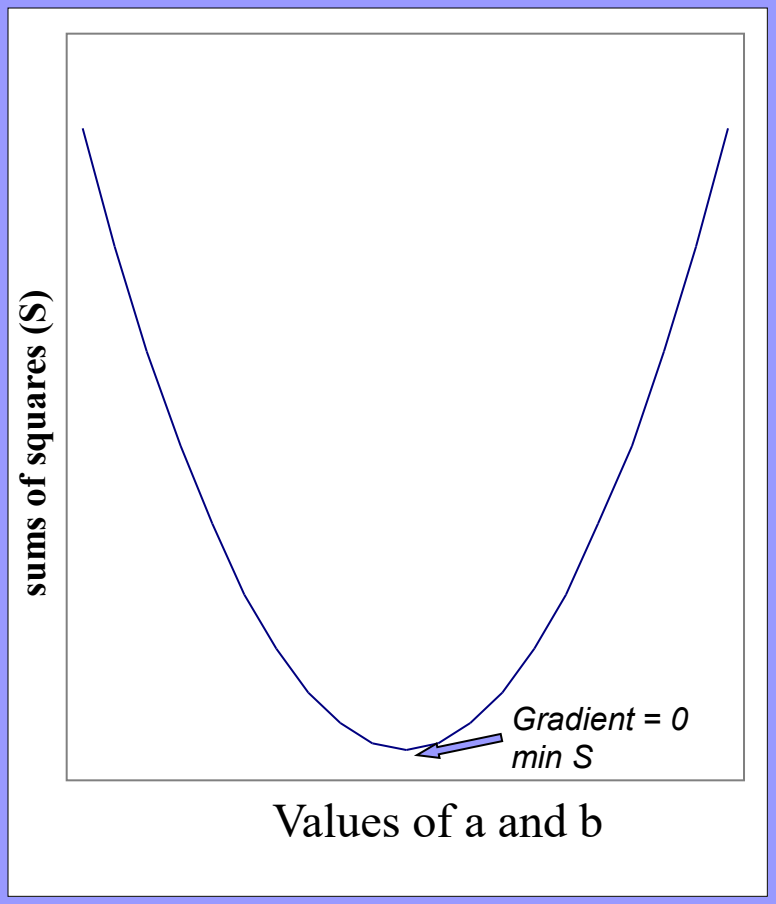
Minimising sums of squares

- Need to minimise $\Sigma(y-\hat{y})^2$
- $\hat{y} = ax + b$
- so need to minimise:

$$\Sigma(y - ax - b)^2$$

- If we plot the sums of squares for all different values of a and b we get a parabola, because it is a squared term

- So the min sum of squares is at the bottom of the curve, where the gradient is zero.



The maths bit

- The min sum of squares is at the bottom of the curve where the gradient = 0
- So we can find a and b that give min sum of squares by taking partial derivatives of $\Sigma(y - ax - b)^2$ with respect to a and b separately
- Then we solve these for 0 to give us the values of a and b that give the min sum of squares

The solution

- Doing this gives the following equations for a and b:

$$a = \frac{r\, s_y}{s_x}$$

r = correlation coefficient of x and y
s_y = standard deviation of y
s_x = standard deviation of x

- From you can see that:
  - A low correlation coefficient gives a flatter slope (small value of a)
  - Large spread of y, i.e. high standard deviation, results in a steeper slope (high value of a)
  - Large spread of x, i.e. high standard deviation, results in a flatter slope (high value of a)

The solution cont.

- Our model equation is $\hat{y} = ax + b$
- This line must pass through the mean so:

$$\bar{y} = a\bar{x} + b \quad \Rightarrow \quad b = \bar{y} - a\bar{x}$$

- We can put our equation for a into this giving:

$$b = \bar{y} - \frac{r\, s_y}{s_x}\, \bar{x}$$

r = correlation coefficient of x and y
s_y = standard deviation of y
s_x = standard deviation of x

- The smaller the correlation, the closer the intercept is to the mean of y

Back to the model

$$\hat{y} = ax + b = \frac{r\,s_y}{s_x}\,x + \bar{y} - \frac{r\,s_y}{s_x}\,\bar{x}$$

(a: $\frac{r\,s_y}{s_x}$; b: $\bar{y} - \frac{r\,s_y}{s_x}\,\bar{x}$; a: $\frac{r\,s_y}{s_x}$)

Rearranges to: $$\hat{y} = \frac{r\,s_y}{s_x}\,(x - \bar{x}) + \bar{y}$$

- If the correlation is zero, we will simply predict the mean of y for every value of x, and our regression line is just a flat straight line crossing the x-axis at y
- But this isn't very useful.
- We can calculate the regression line for any data, but the important question is how well does this line fit the data, or how good is it at predicting y from x

How good is our model?

- Total variance of y: $s_y^2 = \frac{\sum(y - \bar{y})^2}{n - 1} = \frac{SS_y}{df_y}$

- Variance of predicted y values ($\hat{y}$):

$$s_{\hat{y}}^2 = \frac{\sum(\hat{y} - \bar{y})^2}{n - 1} = \frac{SS_{pred}}{df_{\hat{y}}}$$

This is the variance explained by our regression model

- Error variance:

$$s_{error}^2 = \frac{\sum(y - \hat{y})^2}{n - 2} = \frac{SS_{er}}{df_{er}}$$

This is the variance of the error between our predicted y values and the actual y values, and thus is the variance in y that is NOT explained by the regression model

How good is our model cont.

- Total variance = predicted variance + error variance

$$s_y^2 = s_{\hat{y}}^2 + s_{er}^2$$

- Conveniently, via some complicated rearranging

$$s_{\hat{y}}^2 = r^2 s_y^2$$

$$r^2 = s_{\hat{y}}^2 / s_y^2$$

- so r^2 is the proportion of the variance in y that is explained by our regression model

How good is our model cont.

- Insert $r^2 s_y^2$ into $s_y^2 = s_{\hat{y}}^2 + s_{er}^2$ and rearrange to get:

$$s_{er}^2 = s_y^2 - r^2 s_y^2$$
$$= s_y^2 (1 - r^2)$$

- From this we can see that the greater the correlation the smaller the error variance, so the better our prediction

Is the model significant?

- i.e. do we get a significantly better prediction of y from our regression equation than by just predicting the mean?

- F-statistic:

$$F_{(df_{\hat{y}}, df_{er})} = \frac{s_{\hat{y}}^2}{s_{er}^2} \quad \overset{\textit{complicated rearranging}}{=\ldots\ldots=} \quad \frac{r^2 (n-2)^2}{1 - r^2}$$

- And it follows that:

(because $F = t^{2)}$)

$$t_{(n-2)} = \frac{r (n-2)}{\sqrt{1 - r^2}}$$

So all we need to know are r and n

General Linear Model

- Linear regression is actually a form of the General Linear Model where the parameters are a, the slope of the line, and b, the intercept.

$$y = ax + b + \varepsilon$$

- A General Linear Model is just any model that describes the data in terms of a straight line

Multiple regression

- Multiple regression is used to determine the effect of a number of independent variables, x_1, x_2, x_3 etc, on a single dependent variable, y
- The different x variables are combined in a linear way and each has its own regression coefficient:

$$y = a_1x_1 + a_2x_2 + \ldots\ldots + a_nx_n + b + \varepsilon$$

- The a parameters reflect the independent contribution of each independent variable, x, to the value of the dependent variable, y.
- i.e. the amount of variance in y that is accounted for by each x variable after all the other x variables have been accounted for

SPM

- Linear regression is a GLM that models the effect of one independent variable, x, on ONE dependent variable, y
- Multiple Regression models the effect of several independent variables, x_1, x_2 etc, on ONE dependent variable, y
- Both are types of General Linear Model
- GLM can also allow you to analyse the effects of several independent x variables on several dependent variables, y_1, y_2, y_3 etc, in a linear combination
- This is what SPM does and all will be explained next week!