# Univariate Regression

## Correlation and Regression

- The regression line summarizes the linear relationship between 2 variables
- Correlation coefficient, *r,* measures strength of relationship: the closer *r* is to +/- 1, the more closely the points of the scatterplot approach the regression line

## Squared Correlations

- $r^2$ is the proportion of the variance in the variable y which is accounted for by its relationship to x
  - i.e., how closely the dots cluster around the regression line
  - If r =.45 the two variables share ~20% of their variance

## Residuals

- Points usually don't all lie on the line.
  - The vertical difference between a real, observed y-value ($\boldsymbol{Y}$) and the point that the regression line predicts it should be ($\hat{\boldsymbol{Y}}$) is called the ***residual***.
- Regression involves an Independent (or explanatory) variable and a dependent (or response) variable.

## The Linear regression equation:

- It summarises / models real observations
- Allows us to try and make a prediction on the value of *y*, based on a given value of *x* beyond the values we have observed (between the limits of the observable data sample we have – i.e. max. and min. values of *x*).
- It describes a straight line which minimizes squared deviations of observed values of Y from those on the regression line, i.e., the ***squared residuals***

## Simple Linear Regression

$$\boldsymbol{y = \alpha + \beta x + \varepsilon}$$

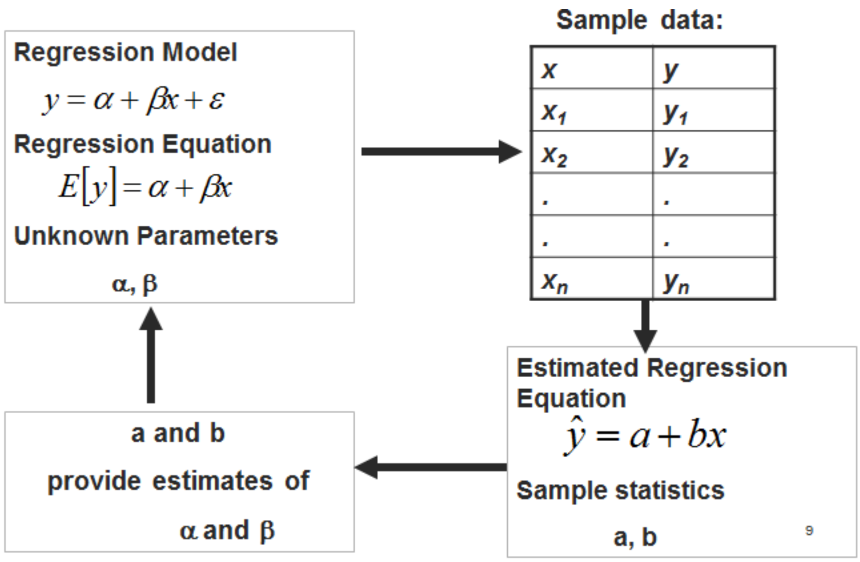
$\alpha$, $\beta$: model parameters;

$\boldsymbol{\varepsilon}$: the unpredictable random disturbance term

- $\alpha$,$\beta$ are unknown, and must be estimated using sample data
- We use the estimated regression equation

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

- Greek alphabet $\alpha$, $\beta$, $\gamma$…are used to denote parameters of a regression equation
- English alphabet a, b, c…are used to denote the estimates of these parameters




- You want to find a regression model which minimises the error term
  - Use the ***method of least squares*** to estimate $\alpha$ and $\beta$.

**Method of Least Squares**

- Provides the regression line in which the sum of squared differences between the observed values and the values predicted by the model is as small as possible

$$\Sigma(Y - Y_{pred})^2$$

$$\text{Deviation} = \Sigma(\text{observed} - \text{model})^2$$

  - Differences are squared to allow for positive / negative $(Y - Y_{pred})$

**Goodness of Fit**

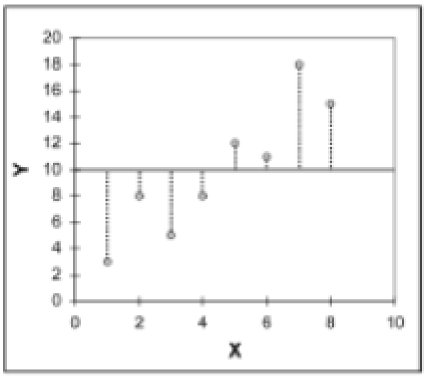
- How well does the model describe / reproduce the observed data?

→Use ***Sums of Squares***

**Sums of Squares (SS)**

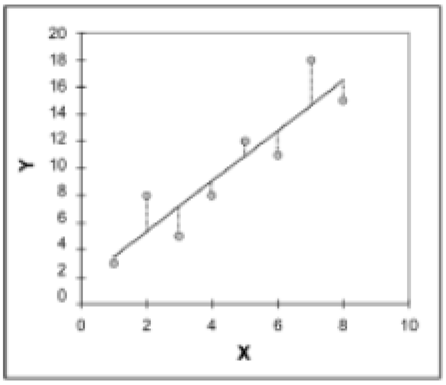
$$SS_{Total} = \sum (Y - Y_{Mean})^2$$

i.e. The sum of the squared differences between each observed value of $y$ and the mean of $y$.



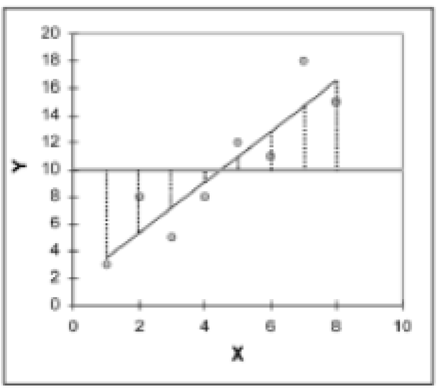

$$SS_{Residual} = \sum (Y - Y_{Predicted})^2$$

i.e. The sum of the squared differences between each observed value of $y$ and its corresponding predicted value of $y$.




$$SS_{Model} = \sum (Y_{Predicted} - Y_{Mean})^2$$

i.e. The sum of the squared differences between each predicted value of $y$ and the mean of $y$.



$$SS_{Model} = SS_{Total} - SS_{Residual}$$

- These are the same equations as $SS_{Between}$, $SS_{Total}$ and $SS_{Within}$, respectively, in ANOVA

**$R^2$**

$$R^2 = \frac{SS_M}{SS_T}$$

- An indication of how much better the model is at predicting Y than if only the mean of Y was used.
- **(Pearson Correlation Coefficient)$^2$**
- We want the ratio of $SS_M$:$SS_T$ to be LARGE –
- $R^2$ represents the proportion of variance in y that can be explained by the model
- $R^2$ * 100 = percentage of variance accounted for by the model
- In univariate regression, the correlation coefficient, *r*, is $\sqrt{R^2}$
    - Doesn't capture whether positive / negative, but this can be established by looking at a scatter plot or at *b* in the regression equation
- If the model is good at predicting, then $SS_M$ will be large compared to $SS_R$

**Testing the Model Using the F-Ratio**

$$F = \frac{MS_M}{MS_R}$$

- SS are totals, therefore affected by sample size
- ***Mean Squares*** (MS) can be used instead (as in ANOVA)

$$MS_M = \frac{SS_M}{df_M}$$
$$df_M = n_{predictors}$$

$$MS_R = \frac{SS_R}{df_R}$$
$$df_R = n_{participants} - n_{predictors} - 1$$

- F-ratio tells us how much better our model is at predicting values of Y than chance alone (the mean)
- As with ANOVA, we want our F to be LARGE
- Calculate critical value, or look up in table.
- Provide a *p*-value:
    - Generally speaking, when $p < .05$, the result is said to be significant.

**Important Values in a Regression output**

e.g.

$SS_M$

$SS_R$

$MS_M$

| Model | Sum of squares | df | Mean square | F | Sig |
|---|---|---|---|---|---|
| Reg'n | 344.49 | 1 | 344.49 | 21.17 | .000 |
| Resid'l | 2229.59 | 137 | 16.27 | | |
| Total | 2574.08 | 138 | | | |

$MS_R$

$SS_T$

| Mode | R | R Square | Adjusted R Square | Std. Error of he Estimate |
|---|---|---|---|---|
| 1 | .366a | .134 | .128 | 4.03416 |

- And you can write the regression equation:

| Model | | | | | |
|---|---|---|---|---|---|
| | B | Std err | Beta | t | Sig |
| (const) | 12.357 | 1.743 | | 7.088 | .000 |
| Health value | .237 | .051 | .366 | 4.601 | .000 |

$$Y_i = b_0 + b_1 X_i$$

***Health Value = 12.357 + (.237)*(whatever X is)***

- ***X is significantly positively correlated with Y***

- ***X explains approximately 13.4% of variance in Y***
- ***This is greater than the proportion expected by chance***
- ***We can ‘predict’ Y from X***