## Assumptions for Linear Regression

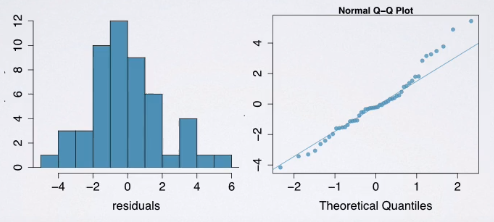
1. Linearity
2. Nearly normal residuals
3. Constant variability

### Linearity

* Relationship between the explanatory and the response variable should be linear
* Check using a scatterplot of the data, or a residuals plot (random scatter around zero)

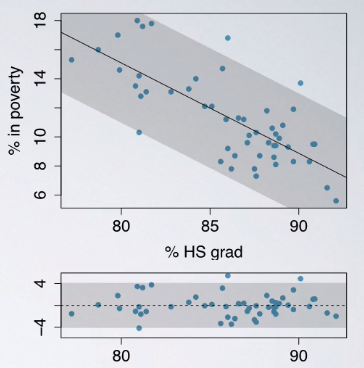
### Nearly normal residuals

* Residuals should be nearly normal distributed, centered at 0.
* May not be satisfied if there are unusual observations that don’t follow the trend of the rest of the data.
* Check using a histogram or normal probability plot of residuals (Q-Q plot)



### Constant variability (homoscedasticity)

* Variability of points around the least squares line should be roughly constant.
* Implies that the variability of residuals around the 0 line should be roughly constant as well
* Check using a residuals plot

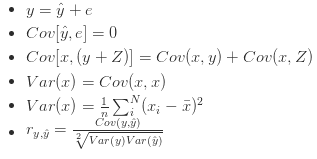
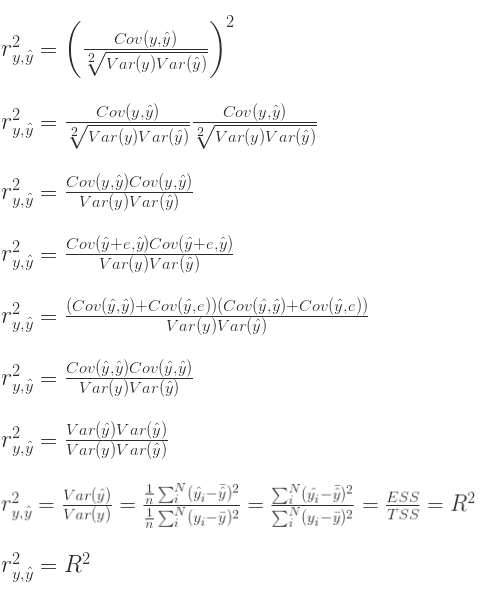


<https://gallery.shinyapps.io/slr_diag/>

## R squared

* Coefficient of determination
* Tells us what percent of variability in the response variable is explained by the model.
* Measures the strength of the fit of a linear model

Proof:



* Always between 0 to 1
* for is the same as for
* R-squared does not tell us about the percentage of observations that can be predicted accurately.
* Two ways to calculate
  + Using correlation: square of the correlation coefficient
  + From the definition: proportion of explained to total variability

## Correlation vs linear regression

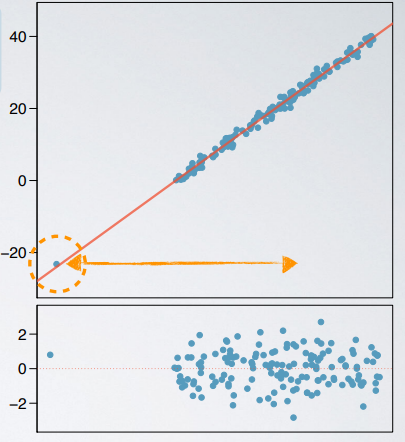
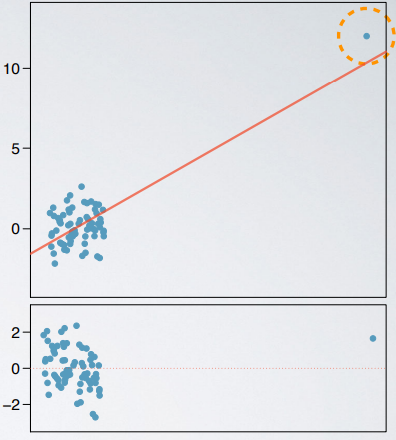


* Regression allows us to predict one variable from the other (not just say if there is an association).
* the magnitude (absolute value) of the correlation coefficient measures the strength of the **linear** association between two numerical variables
* the sign of the correlation coefficient indicates the direction of association
* the correlation coefficient is always between -1 and 1, -1 indicating perfect negative linear association, +1 indicating perfect positive linear association, and 0 indicating no linear relationship
* since the correlation coefficient is unitless, it is not affected by changes in the center or scale of either variable (such as unit conversions), but the slope in linear regression will be sensitive to the unit
* the correlation of X with Y is the same as of Y with X
* the correlation coefficient is sensitive to outliers
* If the correlation coefficient is 1, then the slope does not necessarily have to be 1. Imagine a dependent variable which increases by precisely 2 units for every 1 unit increase in the explanatory variable. These variables have correlation coefficient R*R* = 1 yet the slope of the regression line would be 2.

## Outliers

### Types of outliers

* outliers are points that fall away from the cloud of points
* outliers that fall horizontally away from the center of the cloud but don’t influence the slope of the regression line are called **leverage points** (lies around the trajectory of the regression line)
* outliers that actually influence the slope of the regression line are called **influential points**
  + usually high leverage points
  + to determine if a point is influential, visualize the regression line with and without the point, and ask: Does the slope of the line change considerably?

leverage point influential point

* We don’t exclude outliers without good reasons
* Influential points do not always reduce

They can reduce R-squared if the remainder of the data show a strong relationship and there is only one or few points that are outside the trajectory of the regression line.

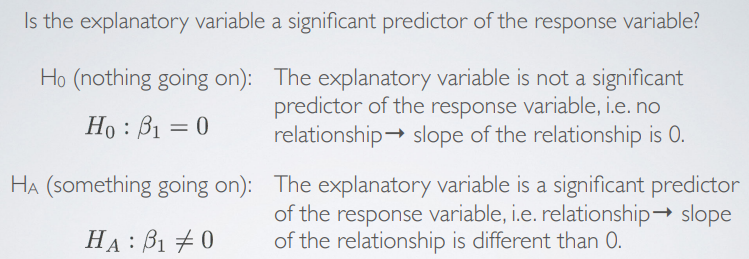
* Influential points change the slope (not necessarily the intercept).
* High leverage points (points farther from the center of the data) are more likely to be influential.
* An influential point does not necessarily change the form of relationship between the variables.

## Inference for Simple Linear Regression

### Test for the slope

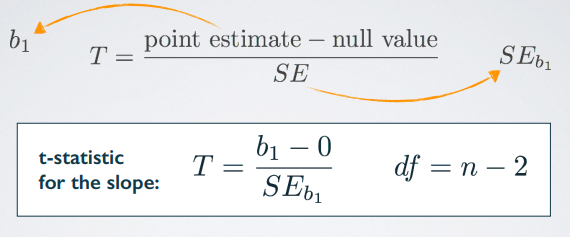
* Hypotheses

It’s about parameter (not point estimate), the null value is 0



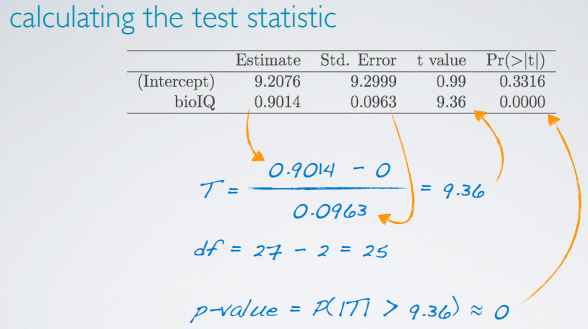
* Mechanics

Use point estimate and its standard error



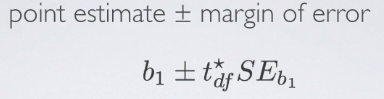
The reason for is because we have 2 unknown parameters here: and

Example:

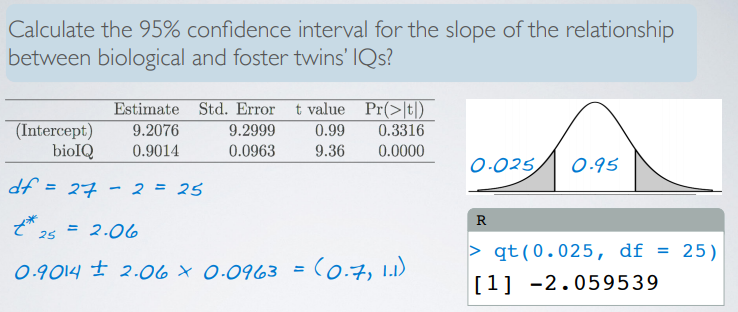


Use a two-sided t-test for the p-value.

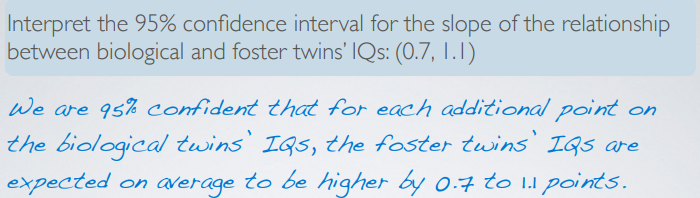
* Confidence interval is:



In R, use to find the critical t score.



**Interpretation on CI:**

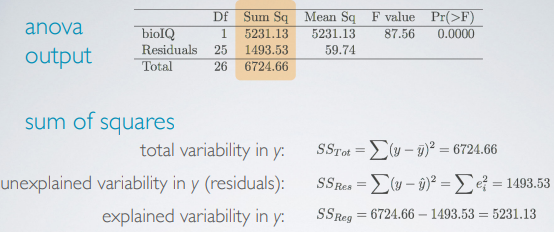


* If we have a sample that is non-random (biased), the results will be unreliable.

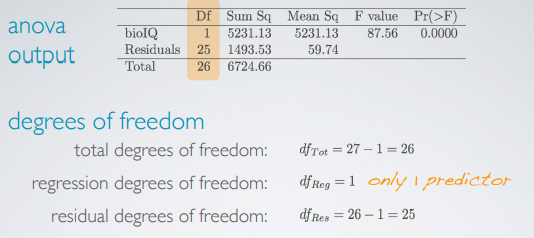
## Variability Partitioning

Partitioning the variability in y to explained and unexplained variability requires **analysis of variance (ANOVA)**.

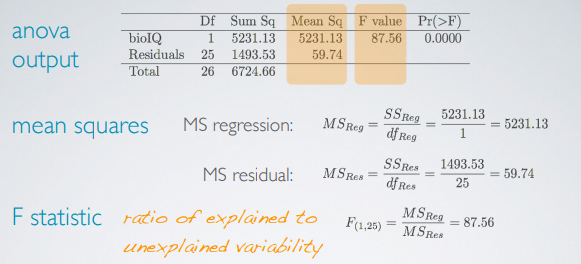
1. Sum of squares



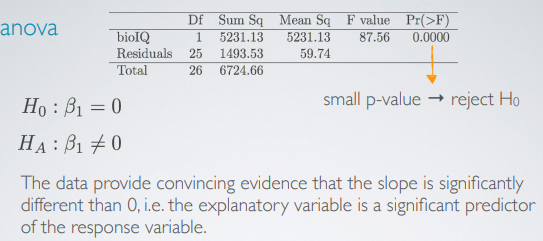
1. Degrees of freedom



1. Mean squares and F statistic



1. P-value



It also answers the questions ‘do we get a significantly better prediction of y from our regression equation than by just predicting the mean?’