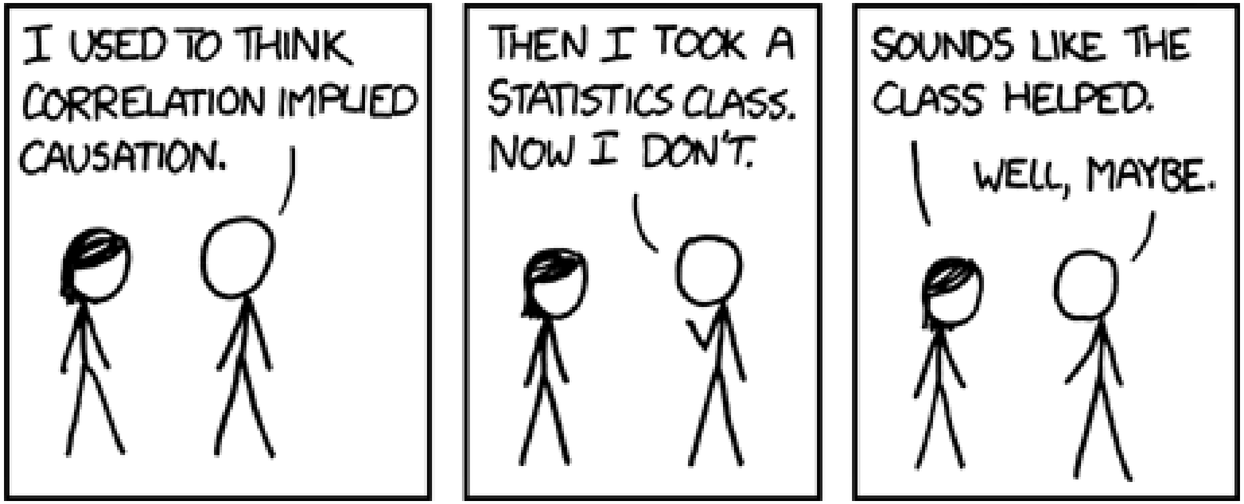
Correlation & Linear Regression


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

"Definition of Statistics:
The science of producing unreliable facts from reliable figures."

Evan Esar (Humorist & Writer)

Correlation & Regression Analyses

When do we use these?

- Predictor and response variables **must** be continuous

  Continuous: values can fall anywhere on an unbroken scale of measurements with real limits

  E.g. temperature, height, volume of fertilizer, etc.

- **Regression Analysis** –

  PART 1: find a relationship between response variable (Y) and a predictor variable (X) (e.g. Y~X)

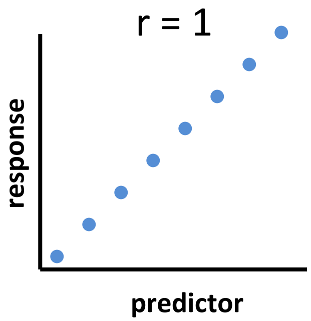
  PART 2: use relationship to predict Y from X

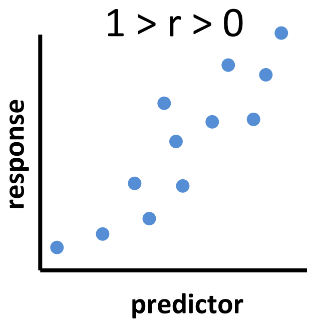
- **Correlation Analysis –** investigating the strength and direction of a relationship

Correlation Coefficients

r = correlation coefficient
range -1 to 1

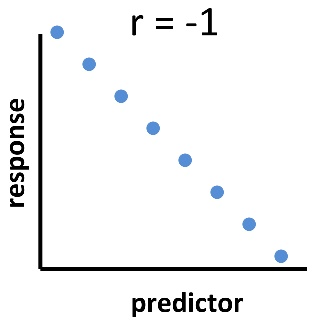
Positive relationship

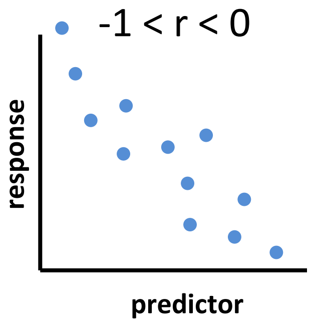




- Increase in X = increase in Y
- r = 1 doesn't have to be a one-to-one relationship

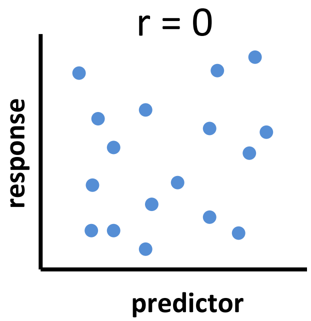
Negative relationship

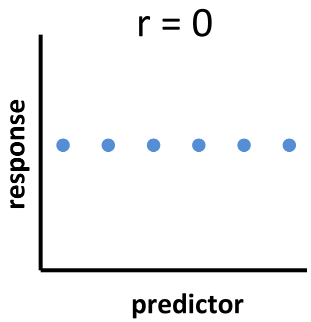




- Increase in X = decrease in Y
- r = -1 doesn't have to be a one-to-one relationship

No relationship





- Increase in X has none or no consistent effect on Y

Correlation Assumptions

1. The experimental errors of your data are normally distributed

2. Equal variances between treatments

   Homogeneity of variances
   Homoscedasticity

3. Independence of samples

   Each sample is randomly selected and independent

Pearson’s Correlation Coefficient

r = correlation coefficient

range -1 to 1

Standard correlation coefficient if assumptions are met

Pearson’s Correlation Coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Calculates relationship based on raw data

Pearson’s Correlation in R:

```
cor(predictor,response,method=”pearson”)
```

Kendall’s Correlation Coefficient

τ = correlation coefficient
range -1 to 1

If your data is non-normal and/or doesn’t have equal variances

Calculated on ranks of data rather than the raw values

1. Rank all of your observations for X (1:N) and Y(1:N)

   - Each row does not necessarily get the same rank for column X and Y

2. Compare the ranks between columns for each row

Kendall’s Correlation Coefficient:

$$\tau = \frac{(\#of\ concordant\ pair) - (\#of\ discordant\ pairs)}{{}^{1}/_{2}\, n(n-1)}$$

- Concordant pair – when the rankings between two rows match

  e.g. $x_i < x_j$ and $y_i < y_j$

Kendall’s Correlation in R:

```
cor(predictor,response,method=”kendall”)
```

Spearman's Rank Correlation Coefficient

If your data is highly non-normal or has significant outliers

ρ = correlation coefficient
range -1 to 1

Calculated on ranks of data rather than the raw values

1. Rank all of your observations for X (1:N) and Y(1:N)

   - Each row does not necessarily get the same rank for column X and Y

2. Compare the ranks between columns for each row

Spearman's Correlation Coefficient:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i}{n(n-1)}$$

- d_i is the difference between x_i and y_i ranks

  e.g. $d_i = x_i - y_j$

Spearman's Correlation in R:

```
cor(predictor,response,method=”spearman”)
```

Correlation Methods

Comparison between methods

- Pearson’s Correlation:
  - relationship **order (direction) and magnitude** of the data values is determined

- Kendall’s & Spearman’s Correlation:
  - relationship **order (direction)** of the data values is determined magnitude cannot be taken from this value because it is based on ranks not raw data
  - Be careful with inferences made with these
  - Order is OK (positive vs negative) – but the magnitude is misleading

- Kendall and Spearman coefficients will likely be larger than Pearson coefficients for the same data because coefficients are calculated on ranks rather then the raw data

Testing the significance of correlation coefficients

“What is the probability I would observe this or a more extreme correlation coefficient by random chance.”

- For Pearson’s *r* :
  - p-values reference the normal distribution
- For Kendall’s τ and Spearman’s ρ :
  - p-values reference the respective distribution of ranks

Pearson’s Correlation in R:

```
cor.test(predictor,response,method=”pearson”)
```

Kendall’s Correlation in R:

```
cor.test(predictor,response,method=”kendall”)
```

Spearman’s Correlation in R:

```
cor.test(predictor,response,method=”spearman”)
```

Dealing with Multiple Inferences

Making inferences from tables of correlation coefficients and p-values

- If we want to use multiple correlation coefficients and p-values to make general conclusions we need to be cautious about inflating our Type I Error due to the multiple test/comparisons

Climate variable	Correlation w/ growth (r^2)	p-value
Temp Jan	0.03	0.4700
Temp Feb	0.24	0.2631
Temp Mar	0.38	0.1235
Temp Apr	0.66	0.0063
Temp May	0.57	0.0236
Temp Jun	0.46	0.1465
Temp Jul	0.86	0.0001
Temp Aug	0.81	0.0036
Temp Sep	0.62	0.0669
Temp Oct	0.43	0.1801
Temp Nov	0.46	0.1465
Temp Dec	0.07	0.4282

Research Question: *Does tree growth dependent on climate?*

Answer (based on a cursory examination of this table)**:** *Yes, there are significant relationships with temperature in April, May, July, and August at α=0.05*

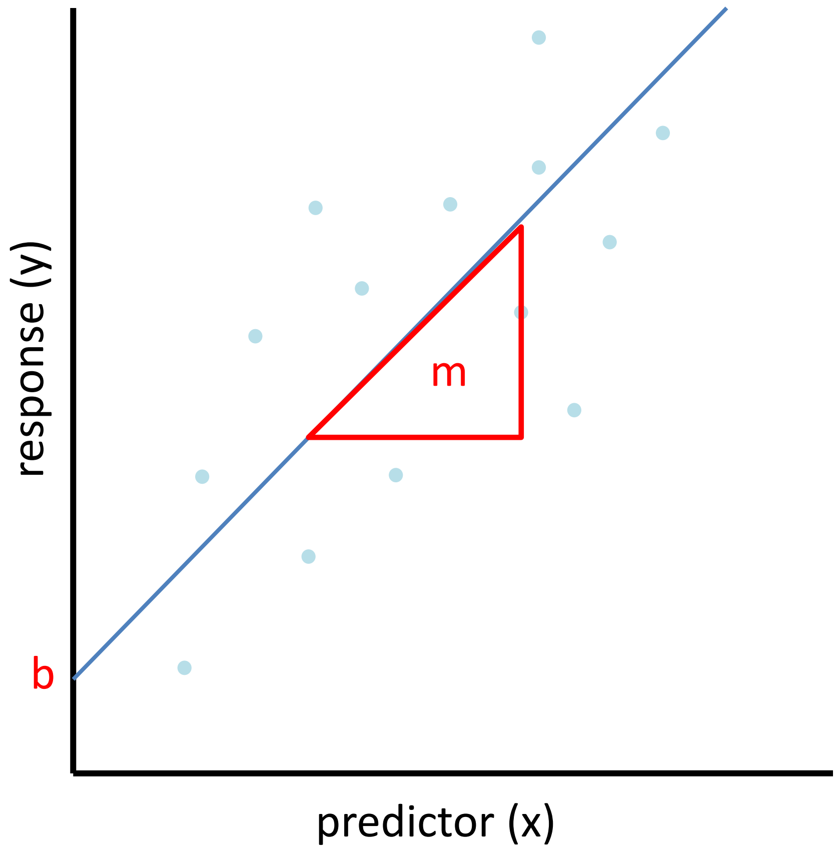
But this is not quite right – we need to adjust p-values for multiple inferences

Adjusting p-values in R:

```
p.adjust(originalP-value,method="bonferroni",n=numberOfComparisons)
```

Linear Regression

Linear relationships



Regression Analysis

PART 1: find a relationship between response variable (Y) and a predictor variable (X)
(e.g. Y~X)

PART 2: use relationship to predict Y from X

Equation of a line: $y = mx + b$

m = slope of the line $\left(\frac{RISE}{RUN}\right)$

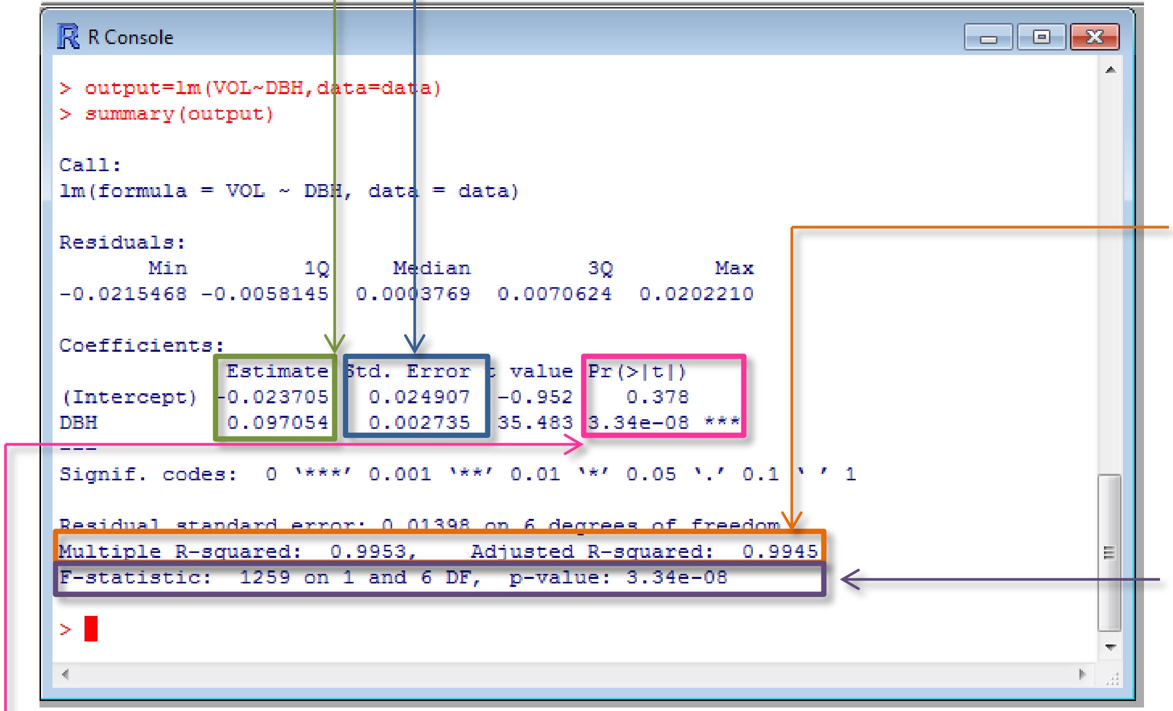
b = y-intercept

Linear Regression in R:

```
lm(response~predictor)
summary(lm(response~predictor))
```

Linear Regression

Output from R



```
> output=lm(VOL~DBH,data=data)
> summary(output)

Call:
lm(formula = VOL ~ DBH, data = data)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0215468 -0.0058145  0.0003769  0.0070624  0.0202210

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.023705   0.024907  -0.952    0.378
DBH          0.097054   0.002735  35.483 3.34e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01398 on 6 degrees of freedom
Multiple R-squared:  0.9953,    Adjusted R-squared:  0.9945
F-statistic:  1259 on 1 and 6 DF,  p-value: 3.34e-08

>
```

Estimate of model parameters (intercept and slope)

Standard error of estimates

Coefficient of determination a.k.a "Goodness of fit"

Measure of how close the data are to the fitted regression line

The significance of the overall relationship described by the model

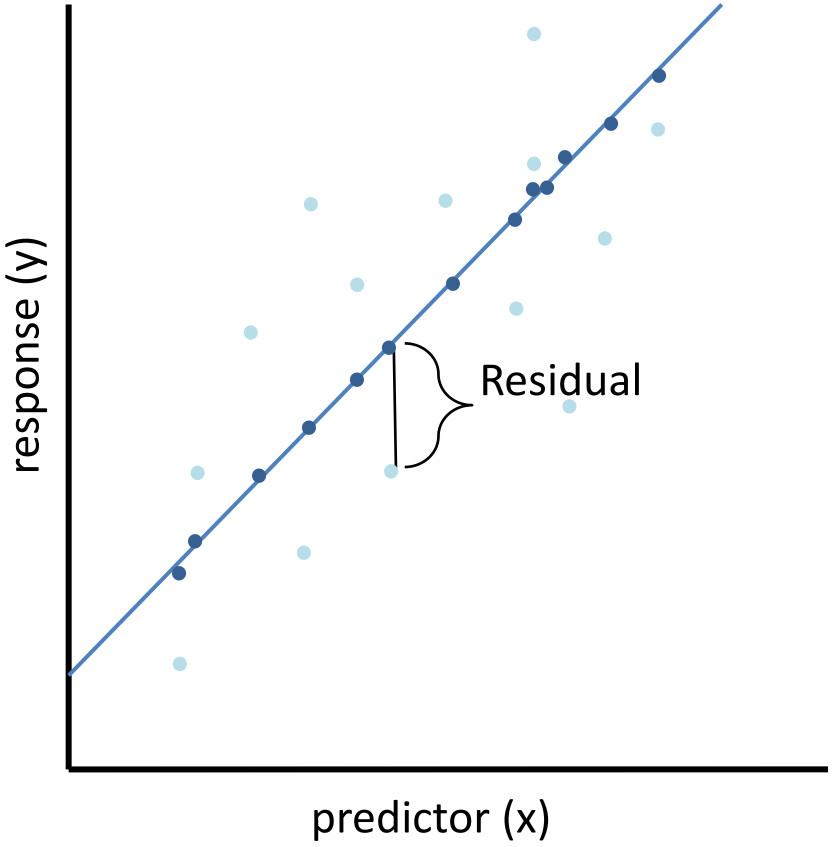
Tests the null hypothesis that the coefficient is equal to zero (no effect)

A predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable

A large p-value suggests that changes in the predictor are not associated with changes in the response

Linear Regression

Method of Least Squares



- For every value along our x-axis we get a predicted value of y ($\hat{y}$) which falls along our regression line

- The difference between the observed y and the predicted y (e.g. $y_i - \widehat{y_i}$) is the residual

- The method of least squares finds the values of m and b that minimize the sum of the squares of all the deviations

Estimation of linear regression coefficients

$$b = \bar{y} - b\bar{x}$$

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

Linear Regression

Relation to correlation coefficient

- Slope of regression equation (m) describes the direction of association between x and y, *but…*
  - The magnitude of the slope depends on the units of the variables
  - The correlation is a *standardized* slope that does not depend on units

$$Standardize = \frac{original\ data\ value\ - mean}{standard\ deviation}$$

- Values now represent units of standard deviations away from the mean

- Correlation r relates to slope m of prediction equation by:

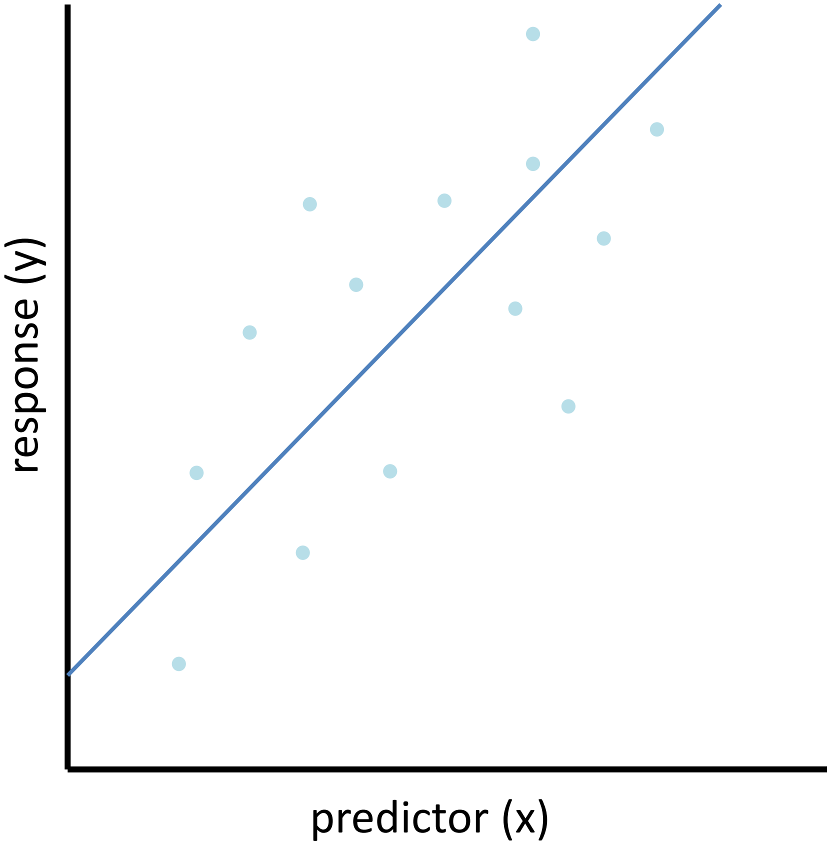
$$\boxed{r = m\left(\frac{s_x}{s_y}\right)}$$

where s_x and s_y are sample standard deviations of x and y.

The direction of your correlation coefficient and the slope of your regression line will be the same (positive or negative)

Linear Regression

Test the how strong the relationship between your variables is



- If we assume there is no significant relationship we test, *Is the slope of my line significantly different than zero?*

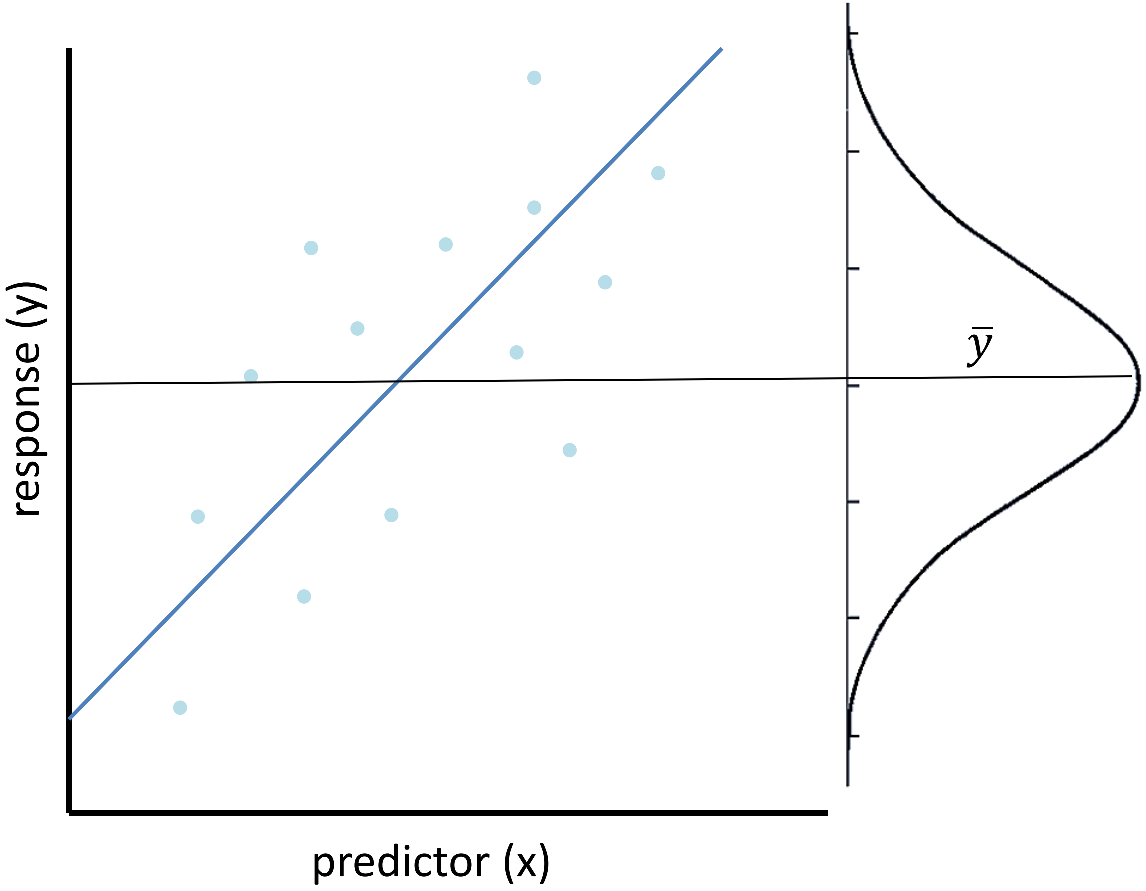
Test statistic:

$$r = \frac{signal}{noise}$$

$$r = \frac{\begin{matrix} varaince\ explained \\ by\ the\ regression\ equation \end{matrix}}{total\ variance\ in\ y}$$

Linear Regression

Total variance in y

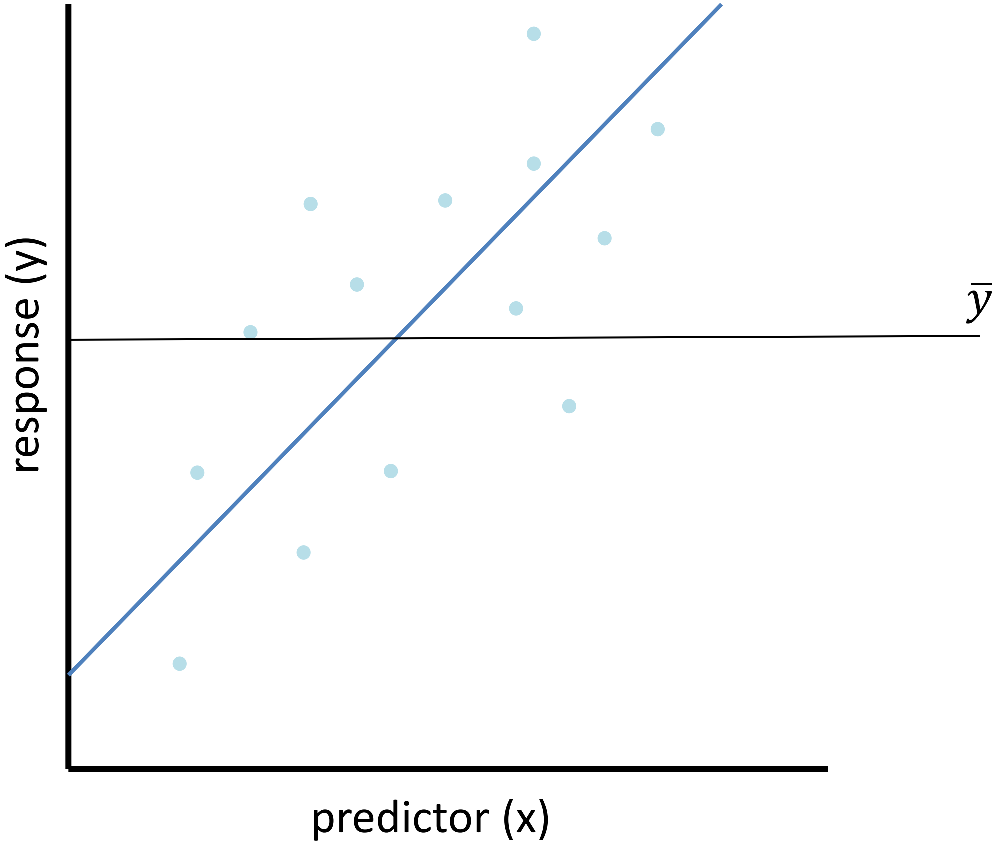


- **Total variance in y** is just the variance in your response variable y

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

Linear Regression

Variance explained by the model

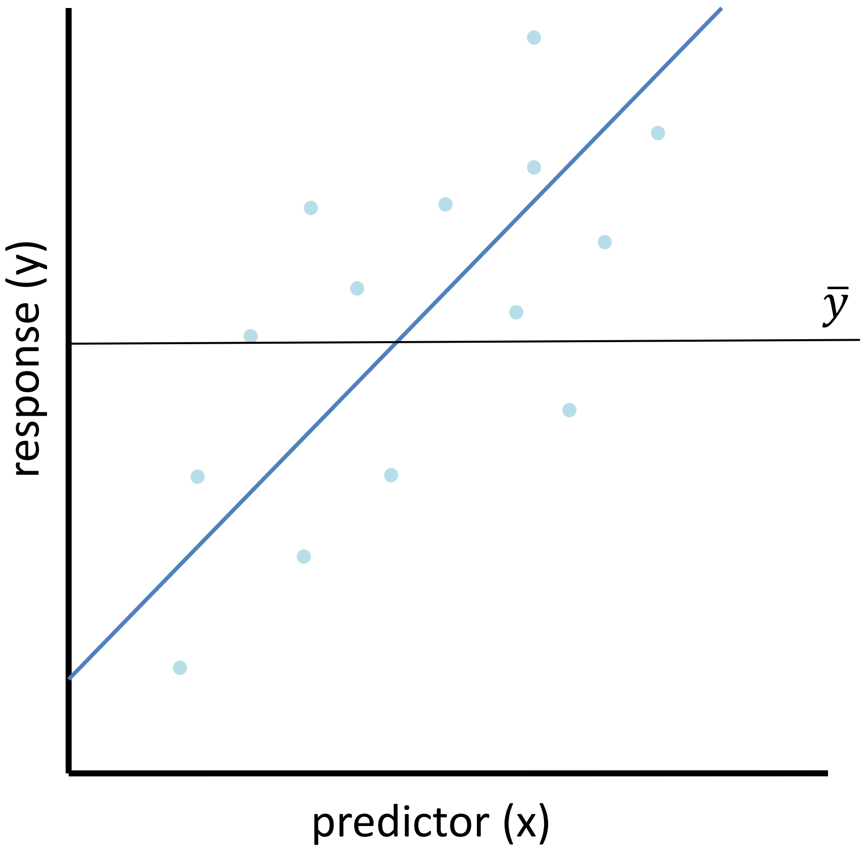


- **Variance explained by the regression model** is simply the amount of variation that occurs when you apply the relationship Y~X of which $\hat{y}$ is the result

$$s_{\hat{y}} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{n-1}}$$

Linear Regression

Test the how strong the relationship between your variables is



Test statistic:

$$r = \frac{signal}{noise}$$

$$r = \frac{\begin{matrix} varaince\ explained \\ by\ the\ regression\ equation \end{matrix}}{total\ variance\ in\ y}$$

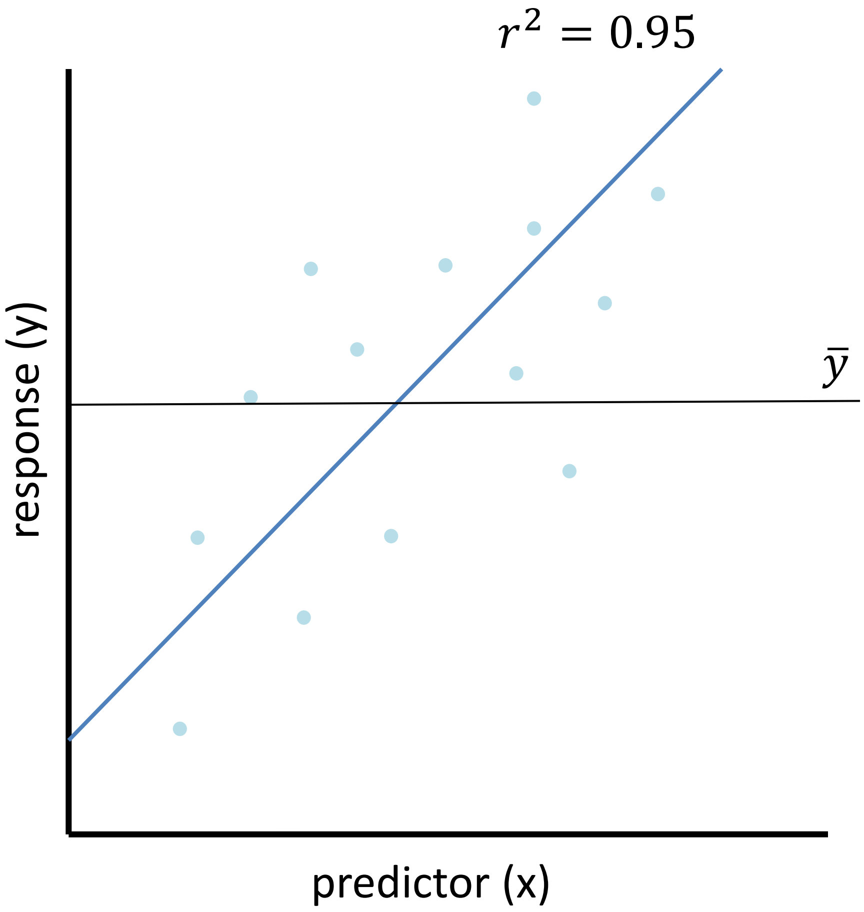
$$r = \frac{\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{n-1}}}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}$$

Apply rules of square root:

$$r = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Linear Regression

R-squared



Test statistic:

$$r = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

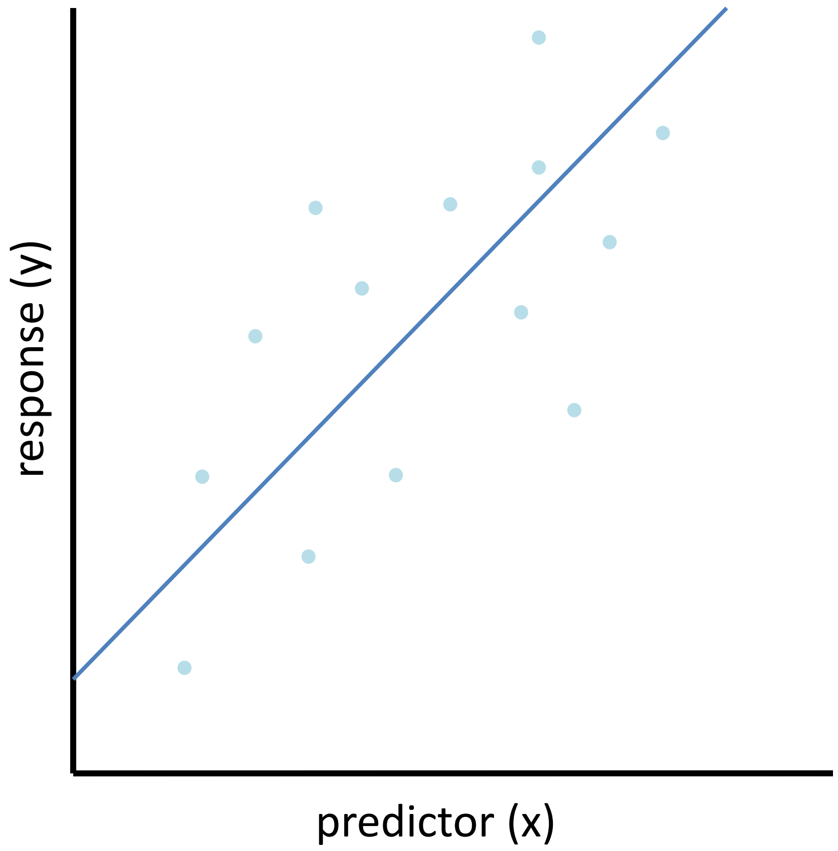
$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$R^2 = \frac{SS_{regression}}{SS_{total}}$$

- R^2 is always positive
- Ranges from 0 to 1 with values closer to 1 indicating a stronger relationship
- R will also export an adjusted R^2

Linear Regression

Unexplained variance



- Unless your regression line is a perfect fit (very rare) there is always part of the variance that cannot be explained

Unexplained variance =
total variance-explained variance

Multiple Linear Regression Assumptions

1. For any given value of X, the distribution of Y must be normal
    - BUT Y does not have to be normally distributed as a whole

2. For any given value of X, of Y must have equal variances

You can again check this by using the Shaprio Test, Bartlett Test, and residual plots on the residuals of your model

What we have all ready been doing!

No assumptions for X – but be conscious of your data

The relationship you detect is obviously reflective of the data you include in your study

Important to Remember

Correlation **DOES NOT** imply causation!

A linear relationship **DOES NOT** imply causation!

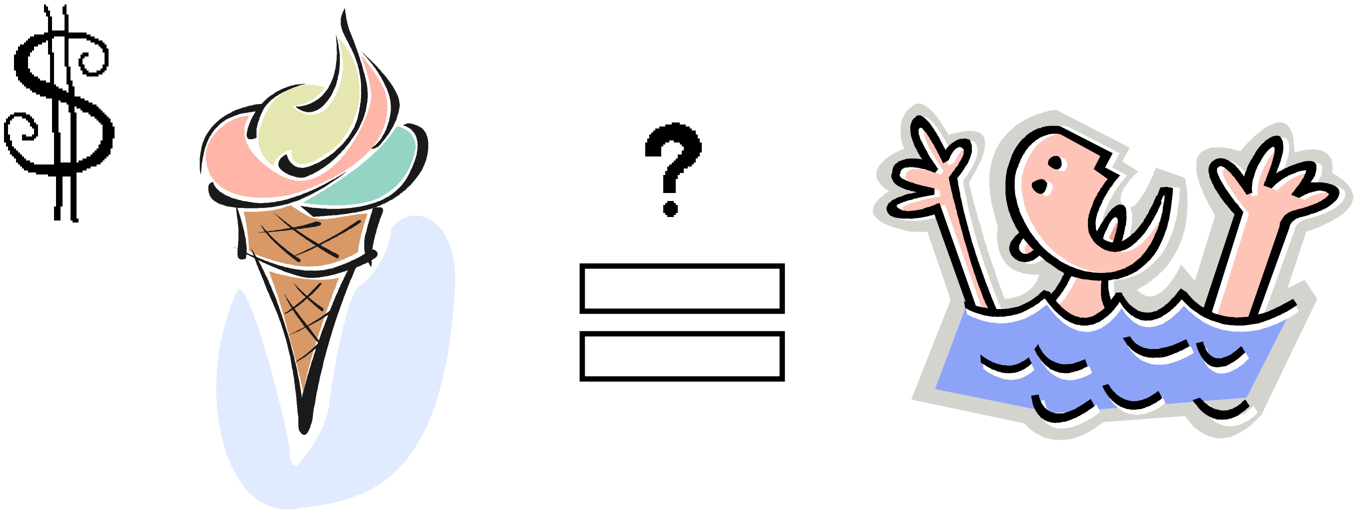
Both of these values imply a relationship rather than one factor causing another factor value

Be careful of your interpretations!

Correlation vs Causation

Example:

If you look at historic records there is a highly significant positive correlation between ice cream sales and the number of drowning deaths



Do you think drowning deaths cause ice cream sales to increase?
Of course NOT!

Both occur in the summer months – therefore there is another mechanism responsible for the observed relationship