

## Section 8: Stats refresher

Here are my Section 43-48 notes following the video stream to guide us.

### Basic Stats refresher

Four types of Variables:

Categorical (characteristic of data unit) subsets are:

- Nominal: gender (M/F) colors (red, green, black) names that can't be ordered.
- Ordinal: Small, Medium, Large or A, B, C (grades) names that can be ordered.

Verses...

Numeric: (basically whole numbers how much or how many) subsets are:

- Discrete: 1,2,3 businesses, 568 people
- Continuous: Age, Height (range)

Regressions analysis estimates the relationships amongst variables, with focus on relationship between a dependent and one or more independent variables.

~ Wikipedia

### Linear vs. Logistic regressions

Linear -> Simple vs Multiple

Logistical -> Simple vs Multiple

#### Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Y Dependent variable (DV)

X1 Independent Variable (IV)

B1 Coefficient

#### Multiple Linear Regression

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots$$

$$b_n * x_n$$

Y Dependent variable (DV)

X1...Xn Independent Variables (IV)

B1... Bn Coefficients

#### Simple Linear Regression chart example

Y axis Salary(\$)

X axis Experience

Salary =  $b_0 + b_1 \cdot \text{experience}$

Ordinary Least Squares:

Difference between the Observed versus the Model

$\sum (Y_i - \hat{Y}_i)^2 \rightarrow \min$

Looks for the minimum sum of squares

R Squared:

$SS_{\text{res}} = \sum (Y_i - \hat{Y}_i)^2$

$SS_{\text{tot}} = \sum (Y_i - \bar{Y})^2$

$R^2 = 1 - (SS_{\text{res}} / SS_{\text{tot}})$

Adjusted R Squared  $R^2_{\text{adj}}$

$SS_{\text{res}} = \sum (Y_i - \hat{Y}_i)^2$

$SS_{\text{tot}} = \sum (Y_i - \bar{Y})^2$

$R^2_{\text{adj}} = 1 - (SS_{\text{res}} / SS_{\text{tot}})$

$SS_{\text{res}} \rightarrow \min$

$R^2$  will never decrease if you add variables.

$R^2$  goodness of fit (greater the better)

Add a third variable (add last digit of phone number), did it get better or worse?

$\text{Adj. } R^2 = 1 - ((1 - R^2) (n-1/n-p-1))$

p = number of regressors (IV)

n = sample size

When P increases in the bottom of the denominator ( $n-p-1$ ), the entire denominator ( $n-p-1$ ) decreases, the division ratio increases ( $n-1/n-p-1$ ),  $(1-R^2)(n-1/n-p-1)$  increases and  $1-((1-R^2)(n-1/n-p-1))$  decreases.

When normal  $R^2$  increases, this part decreases ( $1-R^2$ ), then  $1-((1-R^2)(n-1/n-p-1))$  decreases.

If additional independent variables are not helping the model  $\text{Adj. } R^2$  will decrease further away from 1 versus if independent variables are helping the model via  $\text{Adj. } R^2$  increasing.

