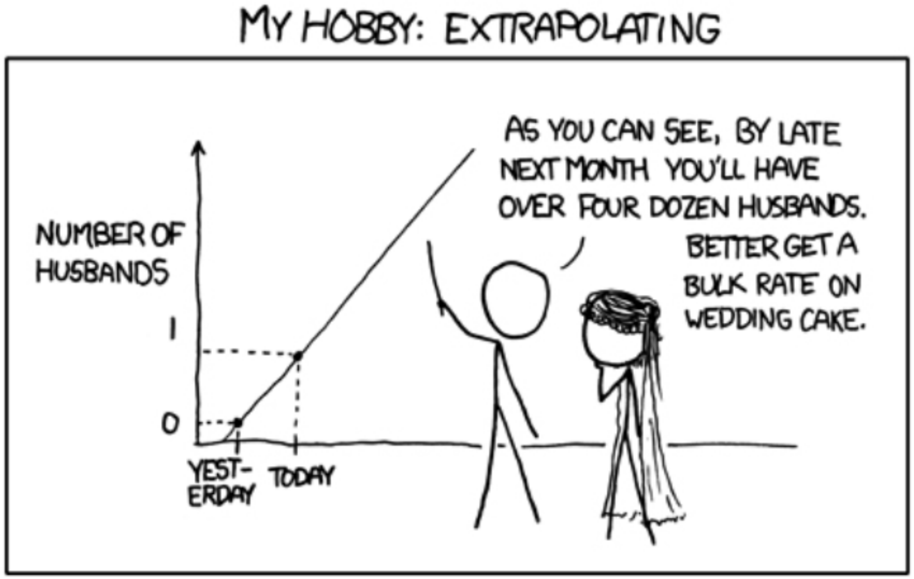
Linear Regression & ANOVA Concepts

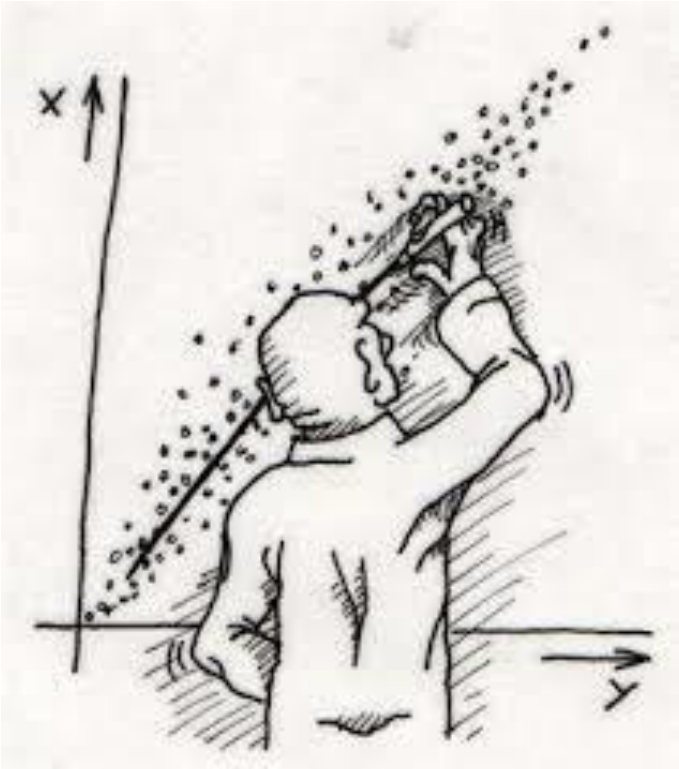
Presented by: Derek Kane

Overview of Topics

- Introduction to Regression Analysis
- Ordinary Least Squares
  - Assumptions
  - Detecting Violations
- Interaction Terms
- Log-Level & Log-Log Transformations
- ANOVA
- Practical Example
  - Real Estate
  - Supermarket Marketing


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS.
BETTER GET A BULK RATE ON WEDDING CAKE.

Introduction to Regression Analysis



- Regression Analysis is the art and science of fitting straight lines to patterns of data.

- Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning.

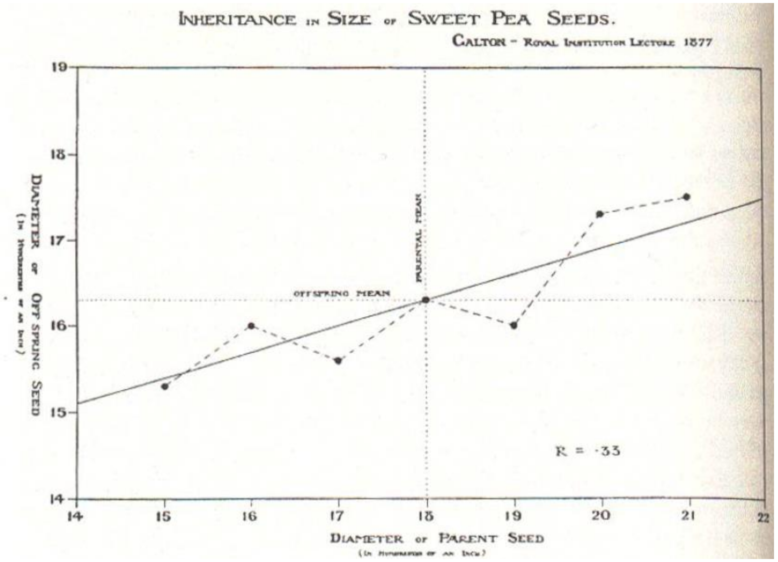
- In a linear regression model, the variable of interest is (dependent variable) is predicted from a single or multiple group of variables (independent variable) using a linear mathematical formula.

- Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.

Introduction to Regression Analysis

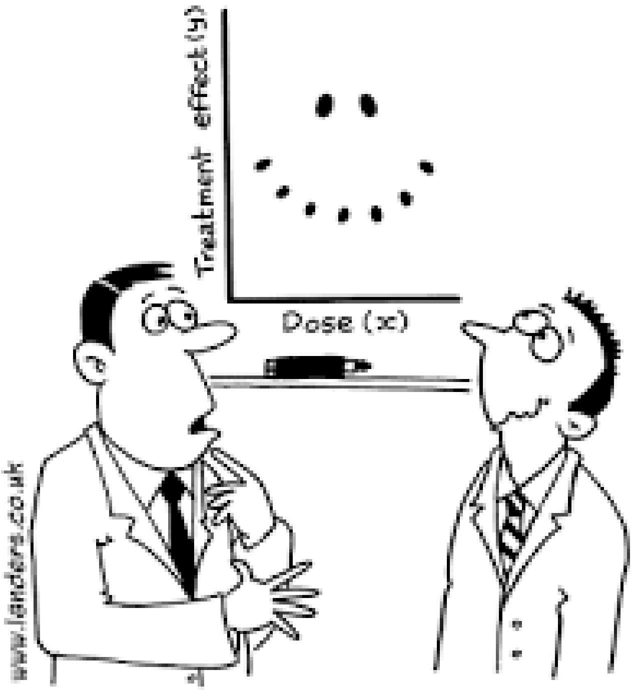
History:

- The earliest form of regression was the method of least squares, which was published by a French mathematician Adrien-Marie Legendre in 1805 and by German mathematician Gauss in 1809.
- Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun (mostly comets, but also later the then newly discovered minor planets).
- In the 1950s and 1960s, economists used electromechanical desk calculators to calculate regressions. Before 1970, it sometimes took up to 24 hours to receive the result from one regression.



The first published picture of a regression line by Francis Galton in 1877

Introduction to Regression Analysis



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

- Many techniques for carrying out regression analysis have been developed.
- Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data.
- Non-parametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.
- Our focus will be on ordinary least squares regression and parametric methods.

Introduction to Regression Analysis

Regression analysis may be used for a wide variety of business applications, such as:

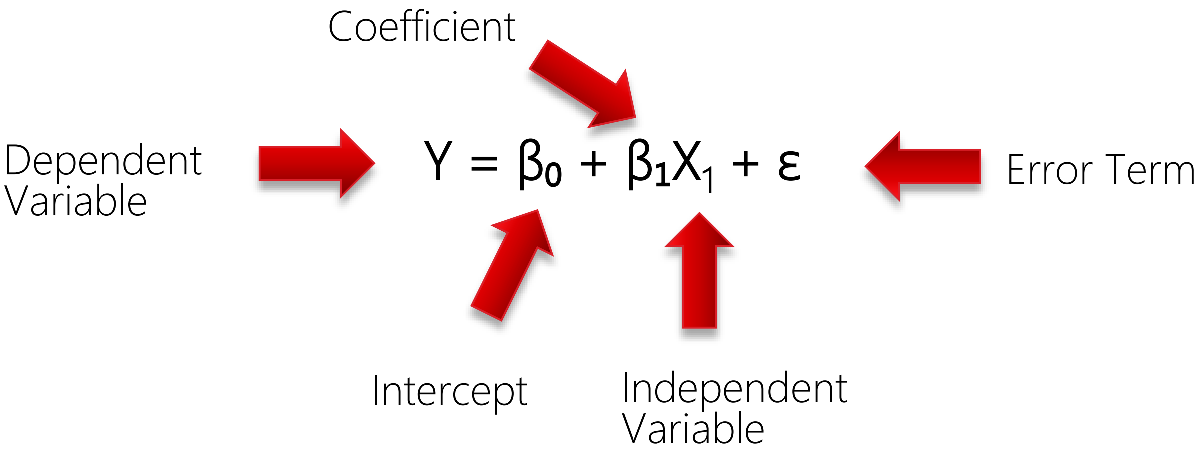
- Measuring the impact on a corporation's profits of an increase in profits.
- Understanding how sensitive a corporation's sales are to changes in advertising expenditures.
- Seeing how a stock price is affected by changes in interest rates.
- Calculating price elasticity for goods and services.
- Litigation and information discovery.
- Total Quality Control Analyses.
- Human Resource and talent evaluation.

- Regression analysis may also be used for forecasting purposes; for example, a regression equation may be used to forecast the future demand for a company's products.

Introduction to Regression Analysis

Simple Linear Regression Formula

- The simple regression model can be represented as follows:



Coefficient

Dependent Variable

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Error Term

Intercept

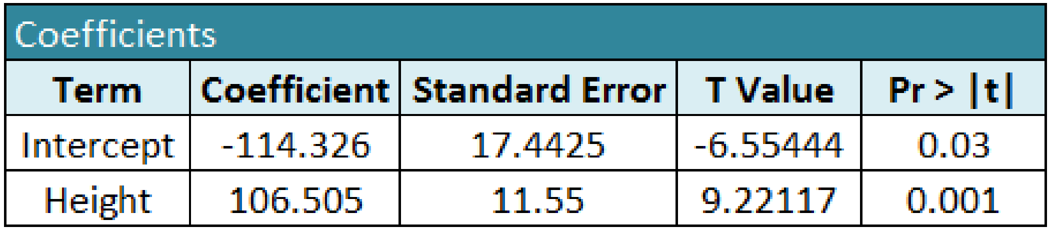
Independent Variable

- The β_0 represents the Y intercept value, the coefficient β_1 represents the slope of the line, the X_1 is an independent variable and ε is the error term. The error term is the value needed to correct for a prediction error between the observed and predicted value.

Introduction to Regression Analysis

Simple Linear Regression Formula

- The output of a regression analysis will produce a coefficient table similar to the one below.



Coefficients

Term	Coefficient	Standard Error	T Value	Pr > |t|
Intercept	-114.326	17.4425	-6.55444	0.03
Height	106.505	11.55	9.22117	0.001

- This table shows that the intercept is -114.326 and the Height coefficient is 106.505 +/- 11.55.
- This can be interpreted as for each unit increase in X, we can expect that Y will increase by 106.5
- Also, the T value and Pr > |t| indicate that these variables are statistically significant at the 0.05 level and can be included in the model.

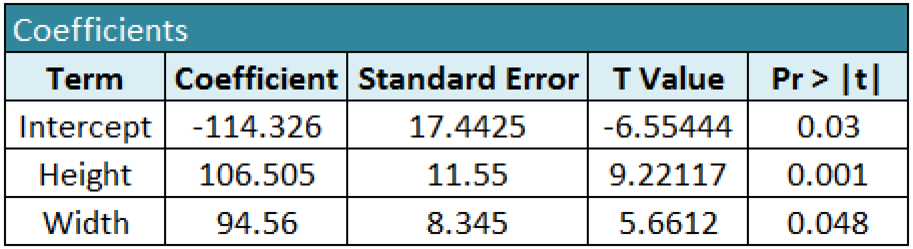
Introduction to Regression Analysis

Multiple Linear Regression Formula

- A multiple linear regression is essentially the same as a simple linear regression except that there can be multiple coefficients and independent variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

- The interpretation of the coefficient is slightly different than in a simple linear regression. Using the table below the interpretation can be thought of:



Coefficients

Term	Coefficient	Standard Error	T Value	Pr > |t|
Intercept	-114.326	17.4425	-6.55444	0.03
Height	106.505	11.55	9.22117	0.001
Width	94.56	8.345	5.6612	0.048

- For each 1 unit change in Width, increases Y by 94.56. This is while holding all other coefficients constant.

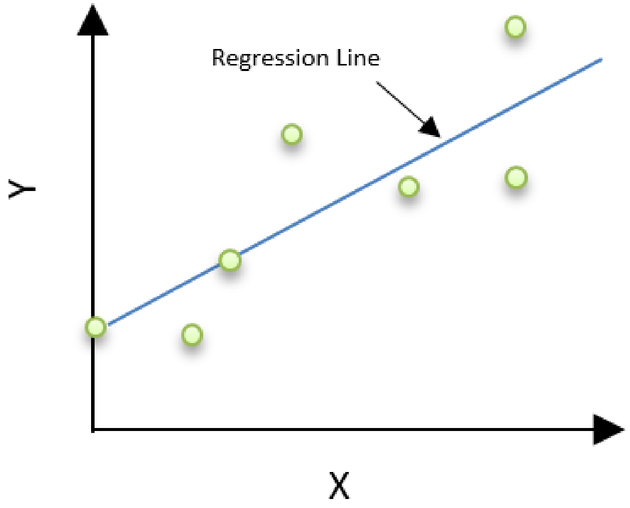
Ordinary Least Squares

What is Ordinary Least Squares or OLS?

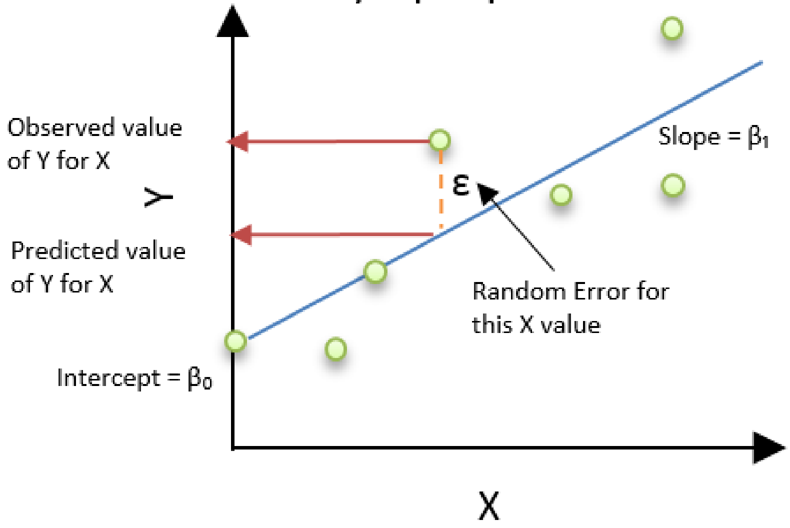
- In statistics, ordinary least squares (OLS) or linear least squares is a method for estimating the unknown parameters in a linear regression model.

- The goal of OLS is to minimize the differences between the observed responses in some arbitrary dataset and the responses predicted by the linear approximation of the data.

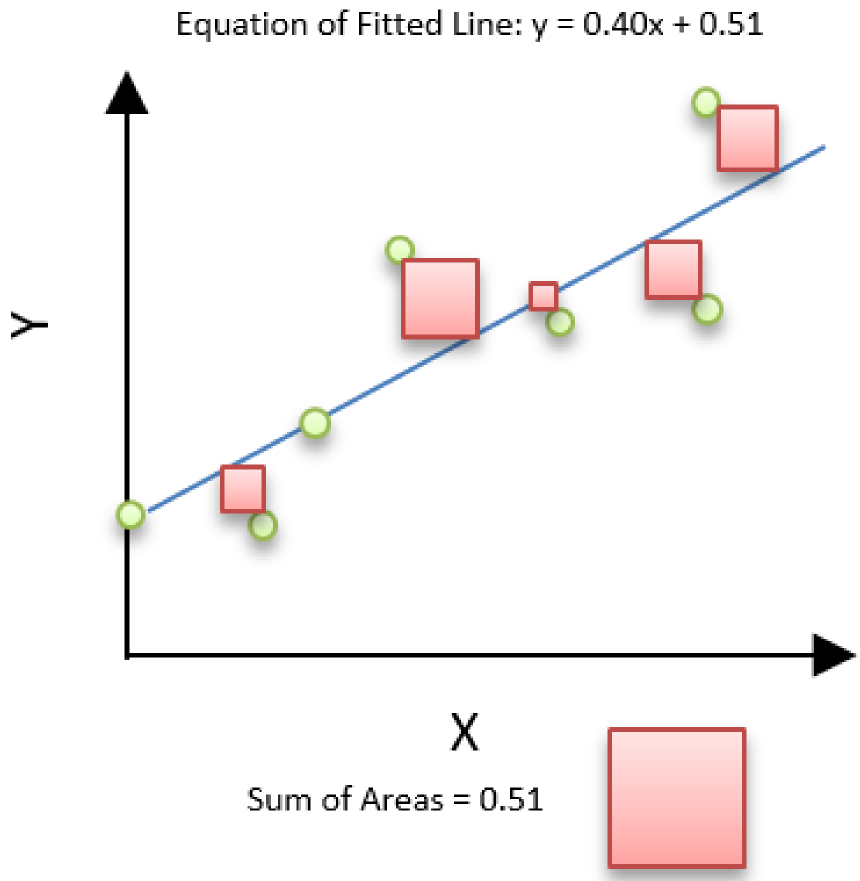
$$y_n = \sum_{i=0}^{k} \beta_i x_{ni} + \varepsilon_n$$



$$y = \beta_0 + \beta_1 x + \varepsilon$$


Observed value of Y for X
Y
Predicted value of Y for X
Intercept = β₀
ε
Slope = β₁
Random Error for this X value
X

Ordinary Least Squares



- Visually this is seen as the sum of the vertical distances between each data point in the set and the corresponding point on the regression line.

- The smaller the differences (square size), the better the model fits the data.

OLS Assumptions

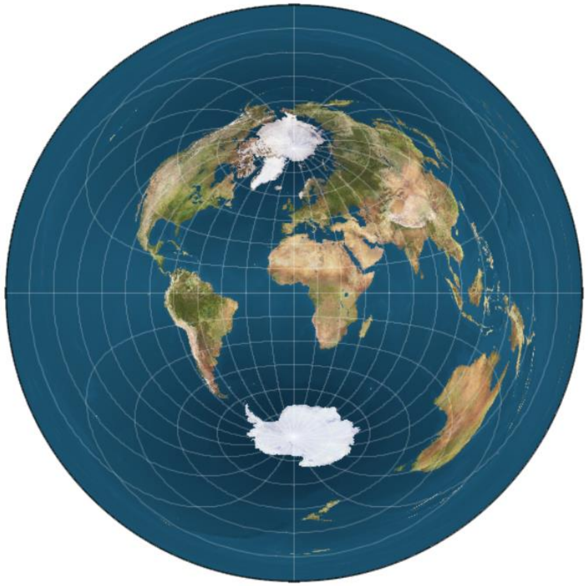
There are a number of classical assumptions which must hold true if we are to trust the outcome of the regression model.

- The sample is representative of the population for the inference prediction.
- The error is a random variable with a mean of zero conditional on the independent variables.
- The independent variables are measured with no error.
- The predictors are linearly independent, i.e. it is not possible to express any predictor as a linear combination of the others.
- The errors are uncorrelated, that is, the variance–covariance matrix of the errors is diagonal and each non-zero element is the variance of the error.
- The variance of the error is constant across observations (homoscedasticity).

OLS Violations of Assumptions

Consequences of using an invalid modeling procedure include:

- The consequences have a tremendous impact on the *theory* that formed the basis of investigating this aspect of human nature.
- A lack of linear association between independent and dependent variables, model misspecification, etc… insinuates that you have the *wrong theory*.
- Biased, inefficient coefficients due to poor reliability, collinearity, etc… lead to an incorrect interpretation regarding your *theory*.
- Outliers imply that you are not able to apply your *theory* to the entire population that you drew your data from.
- Over fitting implies that you are overconfident with your *theory*.



Regression Diagnostics

There are a number of statistics and diagnostic tests we can draw from to evaluate linear regression models beyond EDA.

- Coefficient of Determination
- Residual Plot
- Breusch-Pagan or White Test
- Variance Inflation Factor
- Influential Observations
- Leverage Points
- Cook's Distance
- Etc…

2-28 © LaughingStock International Inc., Dist. by Universal UClick for UFS, 2013

"Rapid pulse, sweating, shallow breathing ... According to the computer, you've got gallstones."

Regression Diagnostics

R^2 : Coefficient of Determination

- This is a measure of the goodness of fit for a linear regression model.
- It is the percentage of the dependent variable variation that is explained by a linear model

- R^2 = Explained variation / Total variation

- R^2 is always between 0 and 100%:
  - 0% indicates that the model explains none of the variability of the dependent data around its mean.
  - 100% indicates that the model explains all the variability of the dependent data around its mean.

Regression Diagnostics

Are Low R^2 Values Inherently Bad?

- No!

There are two major reasons why it can be just fine to have low R-squared values.

- In some fields, it is entirely expected that your R-squared values will be low. For example, any field that attempts to predict human behavior, such as psychology, typically has R-squared values lower than 50%. Humans are simply harder to predict than, say, physical processes.

- Furthermore, if your R-squared value is low but you have statistically significant predictors, you can still draw important conclusions about how changes in the predictor values are associated with changes in the response value.

- Regardless of the R-squared, the significant coefficients still represent the mean change in the response for one unit of change in the predictor while holding other predictors in the model constant. Obviously, this type of information can be extremely valuable.

Regression Diagnostics

- The number of independent variables in your model will increase the value of R-squared despite whether the variables offer an increase in explanatory power. To combat this issue, we should focus on utilizing the Adjusted R-Squared metric which penalizes a model for having too many variables.

- There is no generally accepted technique for relating the number of total observations to the number of independent variables in a model. One possible rule of thumb suggested by Good and Hardon is:

$$N = m^n$$

- Where N is the sample size, n is the number of independent variables, and m is the number of observations needed to reach the desired precision if the model had only one variable.

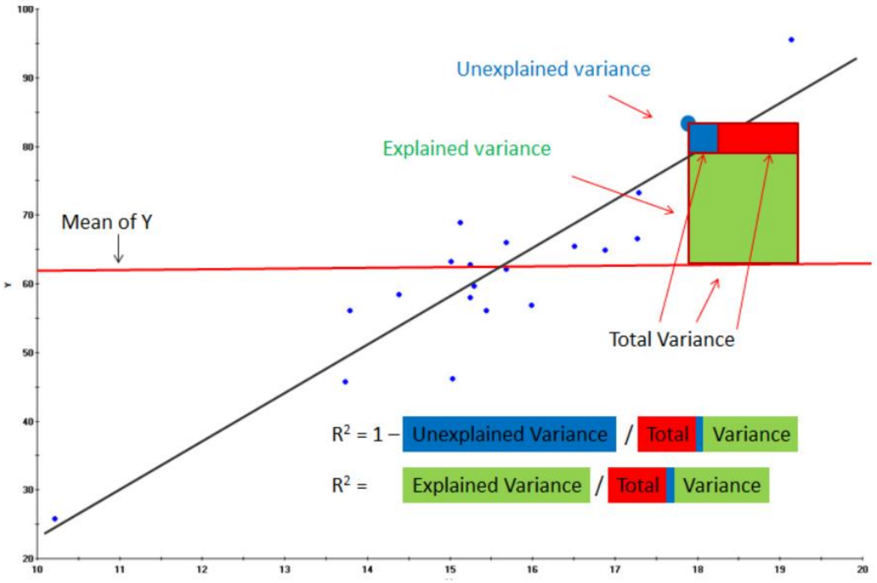
- For example, if the dataset contained 1000 observations and the researcher decided that 5 observations are needed to support a single variable, then the maximum number of independent variables the model can support is 4.

$$\frac{\log 1000}{\log 5} = 4.29$$

Regression Diagnostics

Key Limitations of R^2

- R-squared cannot determine whether the coefficient estimates and predictions are biased, which is why you must assess the residual plots.
- R-squared does not indicate whether a regression model is adequate. You can have a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data!
- The R-squared in your output is a biased estimate of the population R-squared.



$R^2 = 1 -$ Unexplained Variance / Total Variance

$R^2 =$ Explained Variance / Total Variance

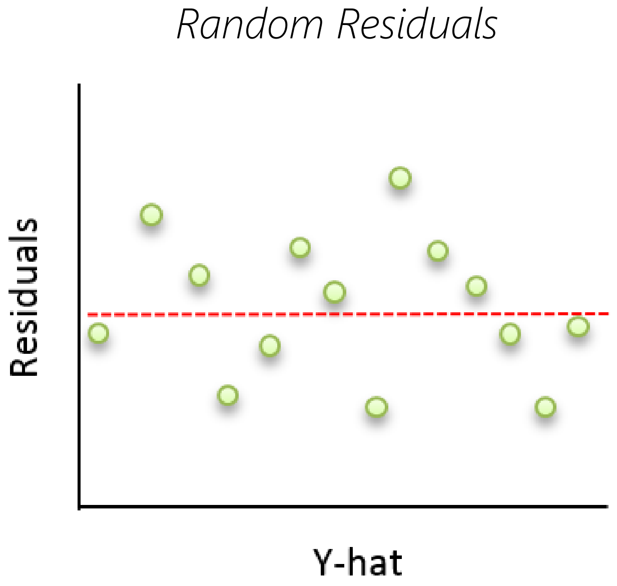
Residual Plots

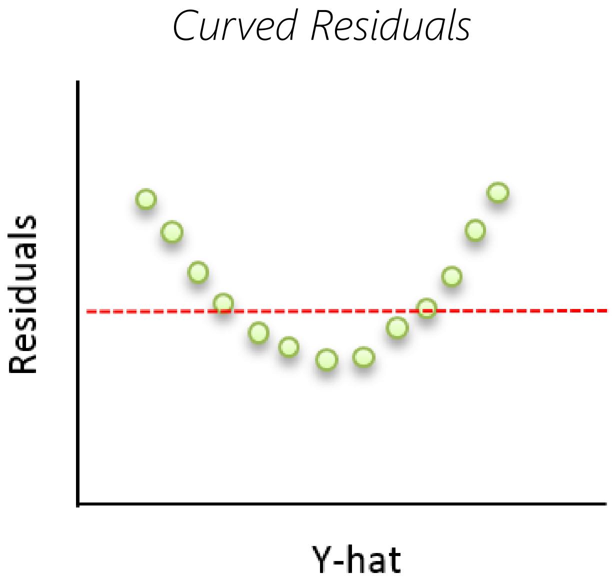
- A residual plot is a scatterplot of the residuals (difference between the actual and predicted value) against the predicted value.

- A proper model will exhibit a random pattern for the spread of the residuals with no discernable shape.

- Residual plots are used extensively in linear regression analysis for diagnostics and assumption testing.

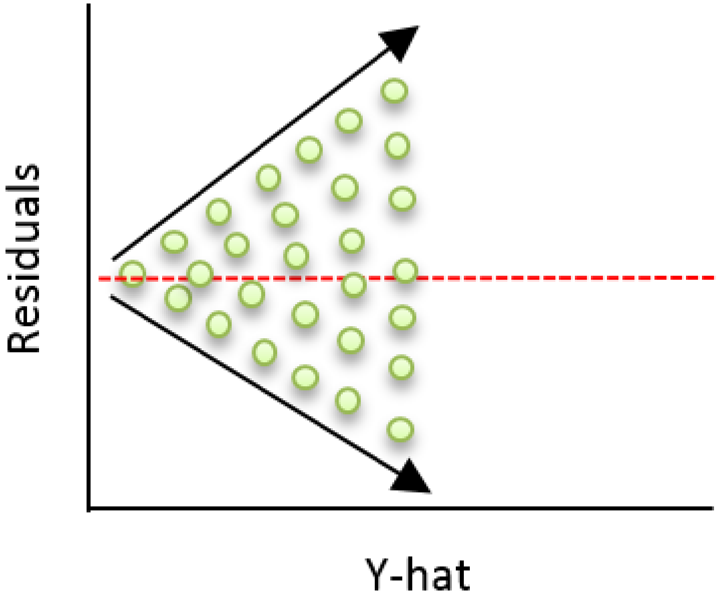
- If the residuals form a curvature like shape, then we know that a transformation will be necessary and can explore some methods like the Box-Cox.





Heteroskedasticity

Heteroskedasticity Pattern

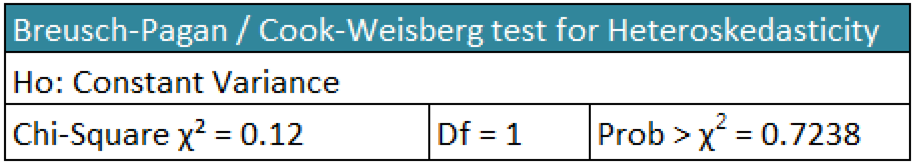


- Linear Regression Analysis using OLS contains an assumption that residuals are identically distributed across every X variable.
- When this condition holds, the error terms are homoskedastic, which means the errors have the same scatter regardless of the value of X.
- When the scatter of the errors is different, varying depending on the value of one or more of the independent variables, the error terms are heteroskedastic.
- A review of a scatterplot of the studentized residuals against the dependent variable can be used to detect if heteroskedasticity is present. The residuals will appear to fan outward in a distinct pattern.

Heteroskedasticity

- Heteroskedasticity has serious consequences for the OLS estimator. Although the OLS estimator remains unbiased, the estimated SE is wrong. Because of this, confidence intervals and hypotheses tests cannot be relied on.

- The Breusch-Pagan test (alt. White Test) is a method that can be employed to identify whether or not the error variances are all equal versus the alternative that the error variances are a multiplicative function of one or more variables.



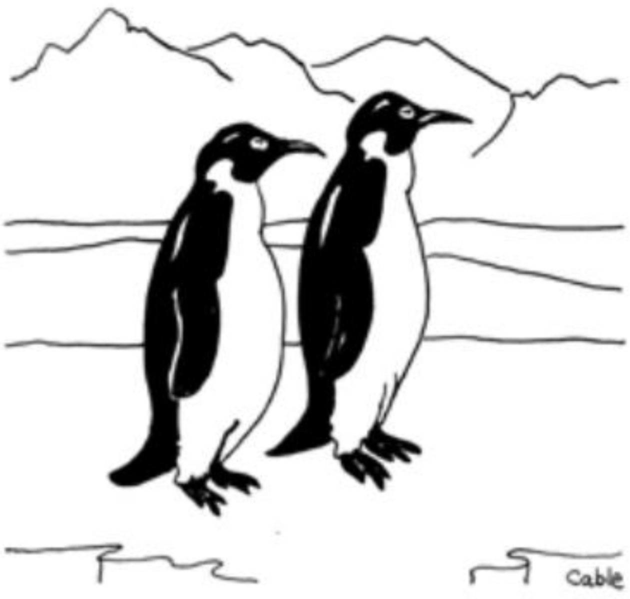
Breusch-Pagan / Cook-Weisberg test for Heteroskedasticity		
Ho: Constant Variance		
Chi-Square $\chi^2 = 0.12$	Df = 1	Prob > $\chi^2 = 0.7238$

- The results of this test show that the Chi-Square value is fairly low indicating that heteroskedasticity is probably not a problem.

Techniques to correct heteroskedasticity:

- Re-specify the model. (Include omitted variables)
- Transform the variables.
- Use Weighted Least Squares in place of OLS.

Multicollinearity



"Do you think all these film crews brought on global warming or did global warming bring on all these film crews?"

What is Multicollinearity?

- Collinearity (or multicollinearity) is the undesirable situation where the correlations among the independent variables are strong.

- In some cases, multiple regression results may seem paradoxical. For instance, the model may fit the data well (high F-Test), even though none of the X variables has a statistically significant impact on explaining Y.

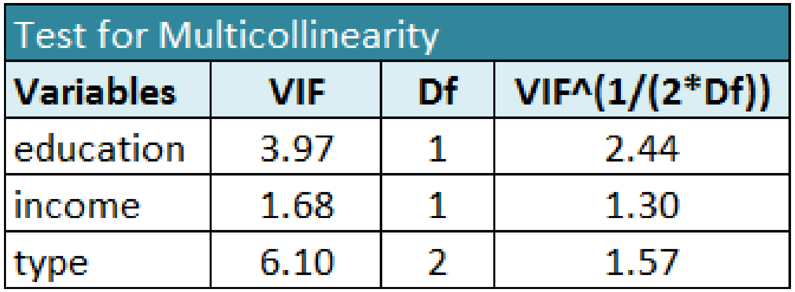
- How is this possible? When two X variables are highly correlated, they both convey essentially the same information. When this happens, the X variables are collinear and the results show multicollinearity.

Multicollinearity

Why is Multicollinearity a Problem?

- Multicollinearity misleadingly inflates the standard errors of the coefficients.
- Thus, it makes some variables statistically insignificant while they should be otherwise significant.
- It is like two or more people singing loudly at the same time. One cannot discern which is which. They offset each other.

Multicollinearity



Test for Multicollinearity

Variables	VIF	Df	VIF^(1/(2*Df))
education	3.97	1	2.44
income	1.68	1	1.30
type	6.10	2	1.57

VIF > 5,
Collinearity
Present

How to detect Multicollinearity?

- Formally, variance inflation factors (VIF) measure how much the variance of the estimated coefficients are increased over the case of no correlation among the X variables. If no two X variables are correlated, then all the VIFs will be 1.
- If VIF for one of the variables is around or greater than 5, there is collinearity associated with that variable.
- The easy solution is: If there are two or more variables that will have a VIF around or greater than 5, one of these variables must be removed from the regression model. To determine the best one to remove, remove each one individually. Select the regression equation that explains the most variance (R2 the highest).

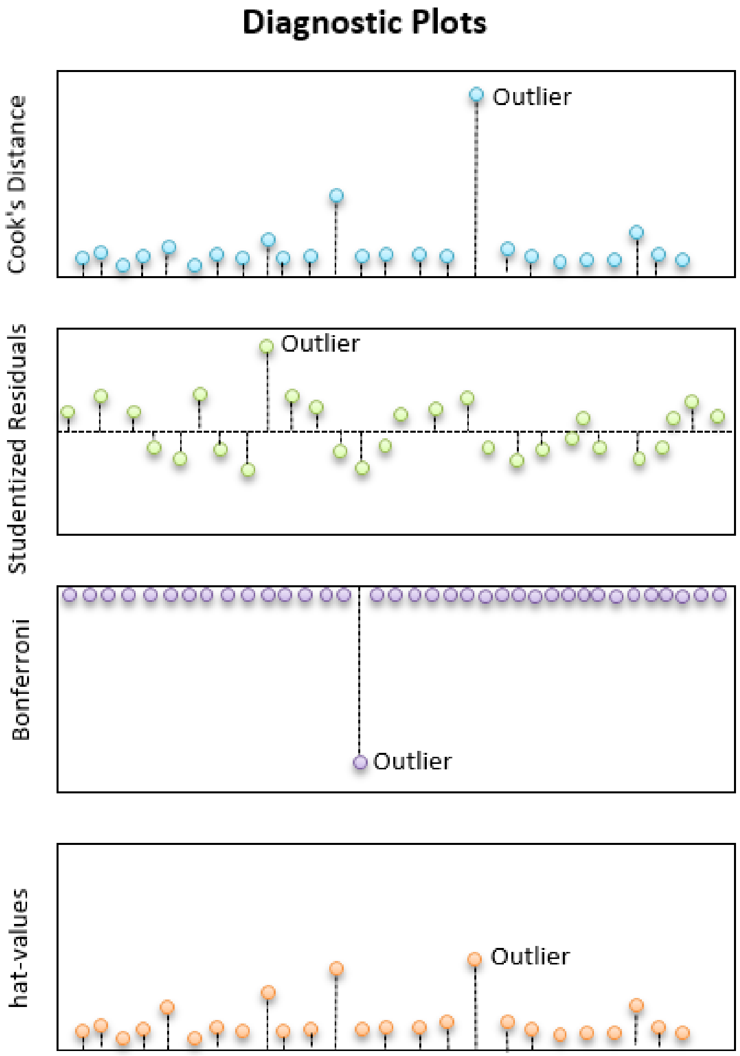
Diagnostic Plots for Outlier Detection

- Cook's distance or Cook's D is a commonly used estimate of the influence of a data point when performing OLS regression.

- Studentized residuals is a quotient resulting from the division of a residual by an estimate of its standard deviation.

- The Bonferroni method is a simple method that allows many comparison statements to be made (or confidence intervals to be constructed) while still assuring an overall confidence coefficient is maintained.

- The hat values represent the predicted Y values plotted against the actual Y values.



Interaction Terms in Model

- Adding interaction terms to a regression model can greatly expand understanding of the relationships among the variables in the model and allows more hypotheses to be tested.

Height = 42 + 2.3 * Bacteria + 11 * Sun

- Height = The height of a shrub. (cm)
- Bacteria = the amount of bacteria in the soil. (1000 per/ml)
- Sun = whether the shrub is located in partial or full sun. (Sun = 0 partial and Sun = 1 is full)

Interaction Terms in Model

- It would be useful to add an interaction term to the model if we wanted to test the hypothesis that the relationship between the amount of bacteria in the soil on the height of the shrub was different in full sun than in partial sun.

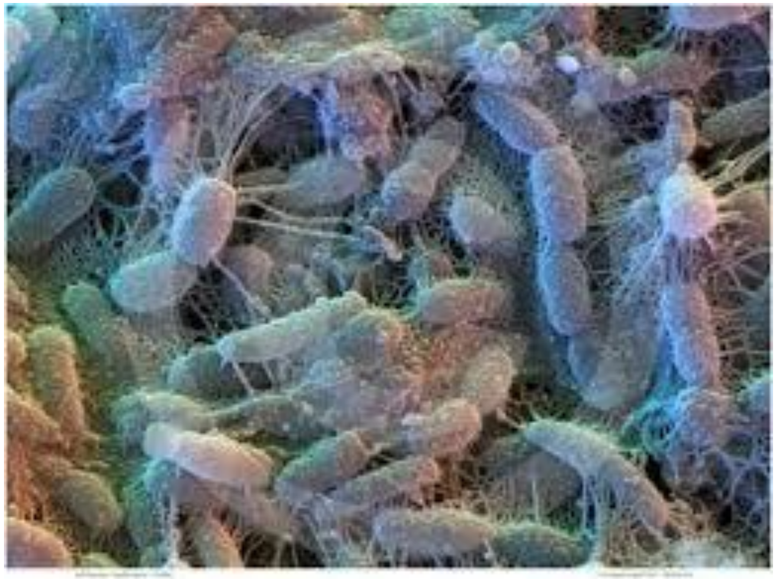
- One possibility is that in full sun, plants with more bacteria in the soil tend to be taller, whereas in partial sun, plants with more bacteria in the soil are shorter.

- Another possibility is that plants with more bacteria in the soil tend to be taller in both full and partial sun, but that the relationship is much more dramatic in full than in partial sun.

Interaction Terms in Model

- The presence of a significant interaction indicates that the effect of one predictor variable on the response variable is different at different values of the other predictor variable.

- It is tested by adding a term to the model in which the two predictor variables are multiplied.



The regression equation will look like this:

Height = B0 + B1 * Bacteria + B2 * Sun + B3 * (Bacteria * Sun)

Interaction Terms in Model

Height = B0 + B1 * Bacteria + B2 * Sun + B3 * (Bacteria * Sun)

- Adding an interaction term to a model drastically changes the interpretation of all of the coefficients.
- If there were no interaction term, B1 would be interpreted as the unique effect of Bacteria on Height. But the interaction means that the effect of Bacteria on Height is different for different values of Sun.
- So the unique effect of Bacteria on Height is not limited to B1, but also depends on the values of B3 and Sun.
- The unique effect of Bacteria is represented by everything that is multiplied by Bacteria in the model: B1 + B3*Sun. B1 is now interpreted as the unique effect of Bacteria on Height only when Sun = 0.

Interaction Terms in Model

- In our example, once we add the interaction term, our model looks like:
- Height = 35 + 4.2*Bacteria + 9*Sun + 3.2*Bacteria*Sun
- Adding the interaction term changed the values of B1 and B2.
- The effect of Bacteria on Height is now 4.2 + 3.2*Sun. For plants in partial sun, Sun = 0, so the effect of Bacteria is 4.2 + 3.2 * 0 = 4.2.
- So for two plants in partial sun, a plant with 1000 more bacteria/ml in the soil would be expected to be 4.2 cm taller than a plant with less bacteria.

Interaction Terms in Model

- For plants in full sun, however, the effect of Bacteria is 4.2 + 3.2*1 = 7.4.
- So for two plants in full sun, a plant with 1000 more bacteria/ml in the soil would be expected to be 7.4 cm taller than a plant with less bacteria.
- Because of the interaction, the effect of having more bacteria in the soil is different if a plant is in full or partial sun.
- Another way of saying this is that the slopes of the regression lines between height and bacteria count are different for the different categories of sun. B3 indicates how different those slopes are.

Height = 35 + 4.2 * Bacteria + 9 * Sun + 3.2 * (Bacteria*Sun)

Interaction Terms in Model

- Interpreting B2 is more difficult.

- B2 is the effect of Sun when Bacteria = 0. Since Bacteria is a continuous variable, it is unlikely that it equals 0 often, if ever, so B2 can be virtually meaningless by itself.

- Instead, it is more useful to understand the effect of Sun, but again, this can be difficult.

- The effect of Sun is B2 + B3*Bacteria, which is different at every one of the infinite values of Bacteria.

- For that reason, often the only way to get an intuitive understanding of the effect of Sun is to plug a few values of Bacteria into the equation to see how Height, the response variable, changes.

Height = 35 + 4.2 * Bacteria + 9 * Sun + 3.2 * (Bacteria*Sun)

Log Transformation Interpretations

- The presentation so far has only considered the following form of a linear regression equation:

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

- This is also considered a “level-level” specification because the raw values of y are being regressed against raw values of x

How do we interpret β_1?

- We differentiate X1 to find the marginal effect of x on y. In this case, β is the marginal effect.

$$\frac{dy}{dx} = \beta$$

Log Transformation Interpretations

- A log-level Regression specification:

$$\log(y) = \beta_0 + \beta_1 x_1 + \epsilon$$

- This is called a "log-level" specification because the natural log transformed values of Y are being regressed on the raw values of x.

- You might want to run this specification if you think that increases in x lead to a constant percentage increase in y.

- Ex. Wage on Education? Forest Lumber Volume on Years?

Log Transformation Interpretations

How do we interpret β$_1$?

- First solve for y:

$$\log(y) = \beta_0 + \beta_1 x_1 + \epsilon$$

$$y = e^{\beta_0 + \beta_1 x_1 + \epsilon}$$

- Then differentiate to get the marginal effect.

$$\frac{dy}{dx_1} = \beta e^{\beta_0 + \beta_1 x_1 + \epsilon} = \beta_1 y \qquad \beta_1 = \frac{dy}{dx_1}\frac{1}{y}$$

- So the marginal effect depends on the value of y, with β itself represents the growth rate.

- For example, if we estimate that β$_1$ is 0.04, we should say that for another year increases the volume of lumber by 4%.

Log Transformation Interpretations

- A log-log Regression specification:

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \epsilon$$

- This is called a "log-log" specification because the natural log transformed values of Y are being regressed on the log transformed values of x.

- You might want to run this specification if you think that percentage increases in x lead to a constant percentage changes in y. Ex. Constant Demand Elasticity

- To calculate marginal effects. Solve for y…

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \epsilon y = e^{\beta_0 + \beta_1 \log(x_1) + \epsilon}$$

- And differentiate x

$$\frac{dy}{dx_1} = \frac{\beta_1}{x_1} e^{\beta_0 + \beta_1 \log(x_1) + \epsilon} = \beta_1 \frac{y}{x_1}$$

Log Transformation Interpretations

- From the previous slide the marginal effect is:

$$\frac{dy}{dx_1} = \beta_1 \frac{y}{x_1}$$

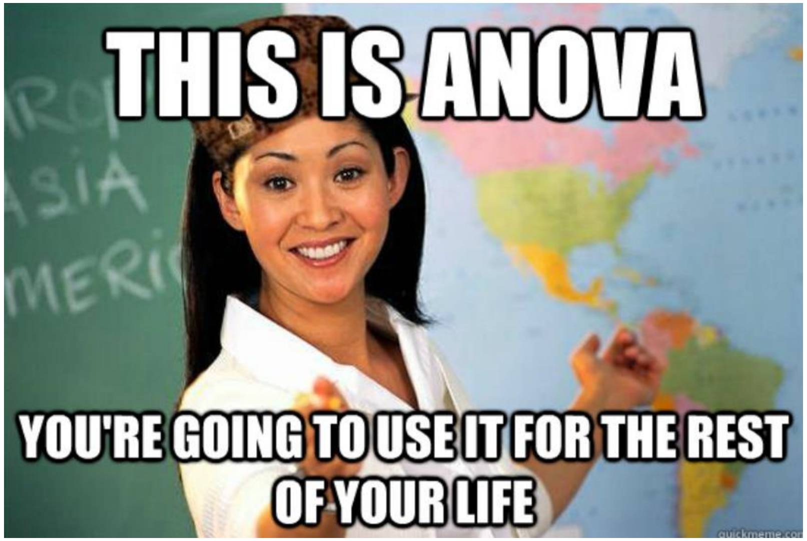
- Solving for β_1 we get:

$$\beta_1 = \frac{dy}{dx_1} \frac{x_1}{y}$$

- This makes β_1 an elasticity. If x_1 is a price and y is a demand and we estimate $\beta_1 = -0.6$, it means that a 1% increase in the price of a good would lead to a 6% decrease in demand.

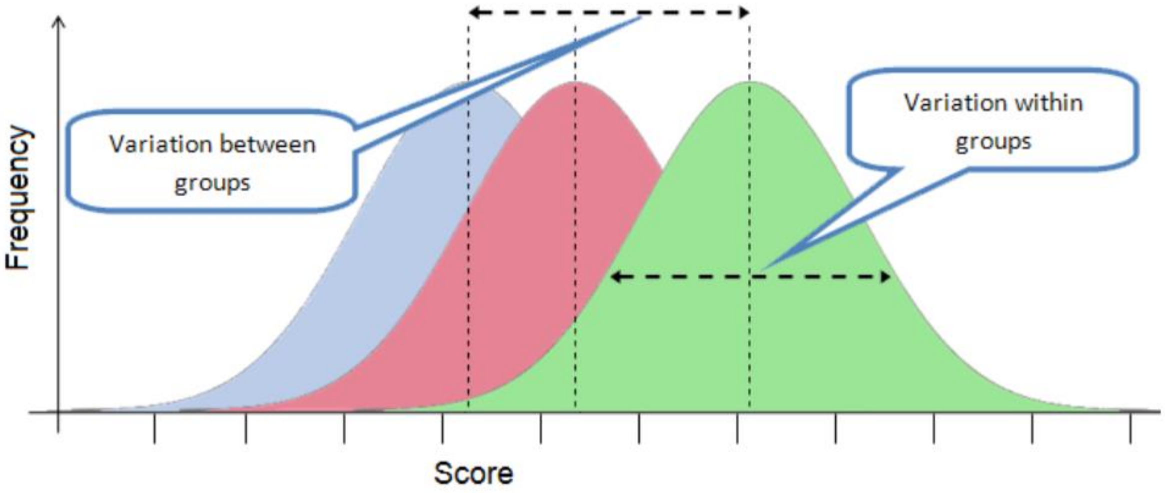
ANOVA

- Analysis of the variance or ANOVA is used to compare differences of means among more than 2 groups.
- It does this by looking at variation in the data and where that variation is found (hence its name).
- Specifically, ANOVA compares the amount of variation between groups with the amount of variation within groups.
- It can be used for both observational and experimental studies.



ANOVA

- When we take samples from a population, we expect each sample mean to differ simply because we are taking a sample rather than measuring the whole population; this is called sampling error but is often referred to more informally as the effects of “chance”.
- Thus, we always expect there to be some differences in means among different groups.
- The question is: is the difference among groups greater than that expected to be caused by chance? In other words, is there likely to be a true (real) difference in the population mean?



ANOVA

The ANOVA model

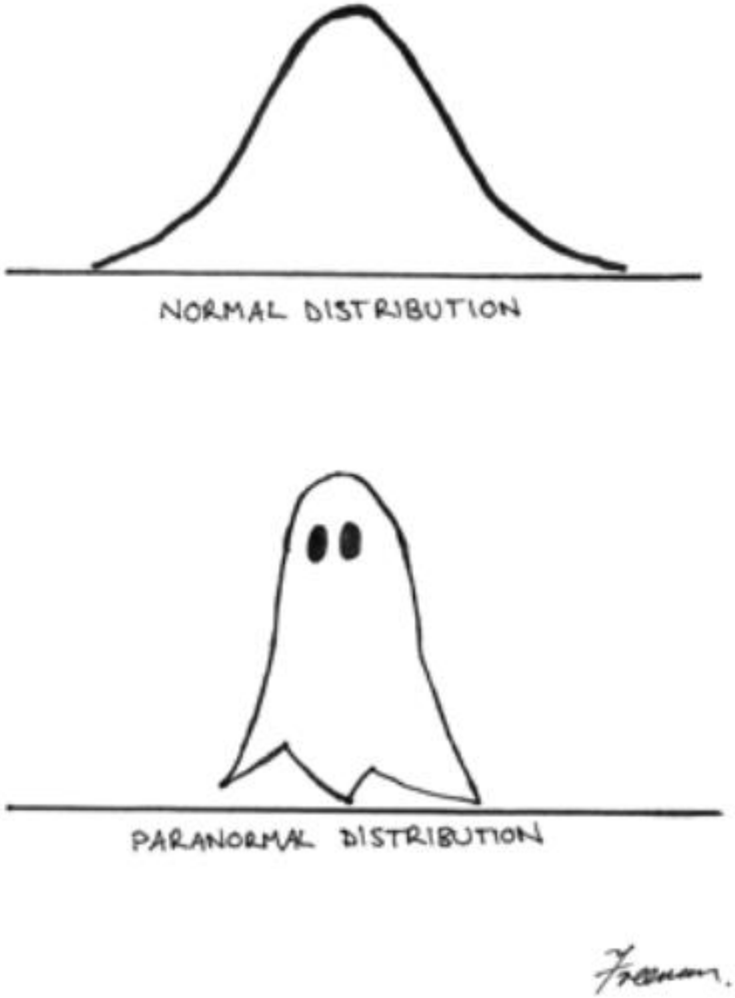
- Mathematically, ANOVA can be written as:

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

- where x are the individual data points (i and j denote the group and the individual observation), ε is the unexplained variation and the parameters of the model (μ) are the population means of each group. Thus, each data point (xij) is its group mean plus error.

Assumptions of ANOVA

- The response is normally distributed
- Variance is similar within different groups
- The data points are independent



ANOVA

Hypothesis testing

- Like other classical statistical tests, we use ANOVA to calculate a test statistic (the F-ratio) with which we can obtain the probability (the P-value) of obtaining the data assuming the null hypothesis.

- Null hypothesis: all population means are equal
- Alternative hypothesis: at least one population mean is different from the rest.

- A significant P-value (usually taken as $P<0.05$) suggests that at least one group mean is significantly different from the others. In other words, a variable with $p<0.05$ allows for us to consider including the variable within a predictive model.

- ANOVA separates the variation in the dataset into 2 parts: between-group and within-group. These variations are called the sums of squares, which can be seen in the following slides.

ANOVA

Calculation of the F ratio

Step 1) Variation between groups

- The between-group variation (or between-group sums of squares, SS) is calculated by comparing the mean of each group with the overall mean of the data.

- Specifically, this is:

$$Between\ SS = n_1(\bar{x}_1 - \bar{x})^2 + n_2\,(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2$$

- We then divide the BSS by the number of degrees of freedom [this is like sample size, except it is n-1, because the deviations must sum to zero, and once you know n-1, the last one is also known] to get our estimate of the mean variation between groups.

ANOVA

Step 2) Variation within groups

- The within-group variation (or the within-group sums of squares) is the variation of each observation from its group mean.

$$SS_r = s^2{}_{group1}\,(n_{group1\;\_\;1}) + s^2{}_{group2}\,(n_{group2} - 1) + s^2{}_{group3}\,(n_{group3} - 1)$$

- i.e., by adding up the variance of each group times by the degrees of freedom of each group. Note, you might also come across the total SS (sum of). Within SS is then Total SS minus Between SS.

- As before, we then divide by the total degrees of freedom to get the mean variation within groups.

ANOVA

Step 3) The F ratio

- The F ratio is then calculated as:

$$F\ Ratio = \frac{Mean\ Between\ Group\ SS}{Mean\ Within\ Group\ SS}$$

- If the average difference between groups is similar to that within groups, the F ratio is about 1. As the average difference between groups becomes greater than that within groups, the F ratio becomes larger than 1.

- Therefore, variables with higher F Ratio values provide greater explanatory power when utilized in predictive models.

- To obtain a P-value, it can be tested against the F-distribution of a random variable with the degrees of freedom associated with the numerator and denominator of the ratio. The P-value is the probably of getting that F ratio or a greater one. Larger F-ratios gives smaller P-values.

ANOVA

- Do you prefer ketchup or soy sauce?

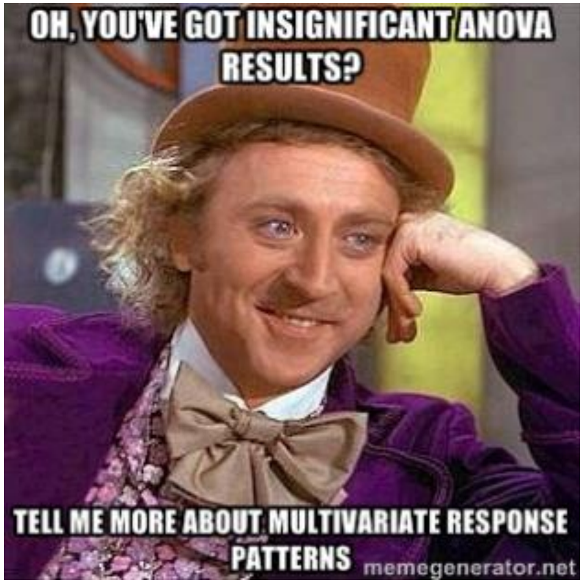
OR

- If someone asked you this question, your answer would likely depend upon what you were eating. You probably wouldn't dunk your spicy tuna roll in ketchup. And most people (pregnant moms-to-be excluded) don't seem to fancy eating soy sauce with hot French fries.

ANOVA

A Common Error When Using ANOVA to Assess Variables

- So you collect data about your variables of interest, and now you're ready to do your analysis. This is where many people make the unfortunate mistake of looking only at each variables individually.
- In addition to considering how each variable impacts your response variable, you also need to evaluate the interaction between those variables and determine if any of those are significant as well.
- And much like your preference for ketchup versus soy sauce depends upon what you're eating, optimum settings for a given variable will depend upon the settings of another variable when an interaction is present.

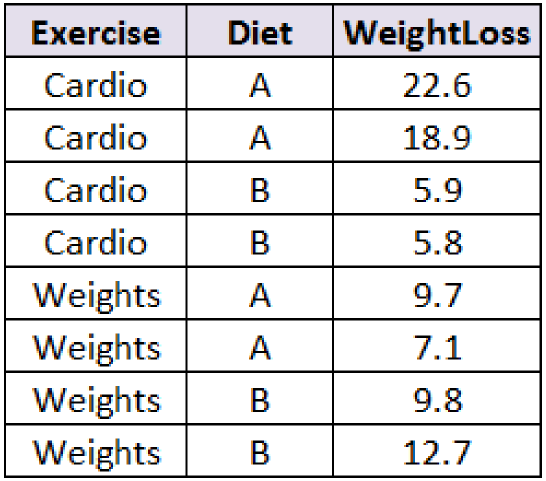

OH, YOU'VE GOT INSIGNIFICANT ANOVA RESULTS?
TELL ME MORE ABOUT MULTIVARIATE RESPONSE PATTERNS
memegenerator.net

ANOVA

How to Evaluate and Interpret an Interaction

- Let’s use a weight loss example to illustrate how we can evaluate an interaction between factors. We're evaluating 2 different diets and 2 different exercise programs: one focused on cardio and one focused on weight training. We want to determine which result in greater weight loss. We randomly assign participants to either diet A or B and either the cardio or weight training regimen, and then record the amount of weight they’ve lost after 1 month.

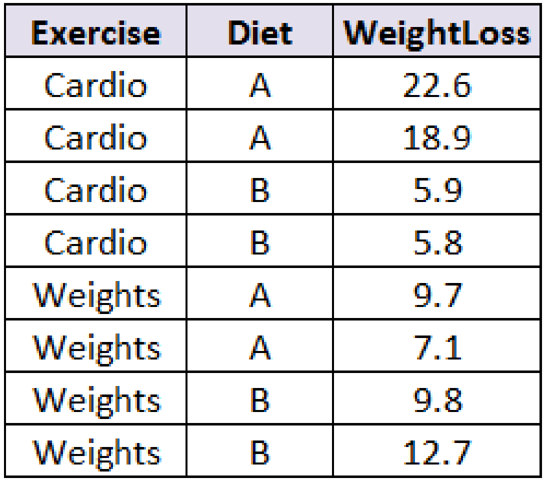
- Here is a snapshot of the data:



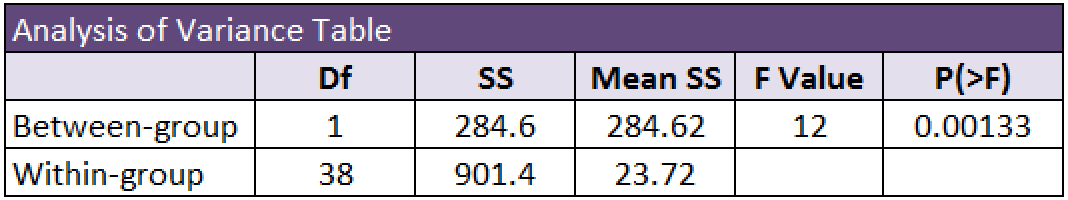
Exercise	Diet	WeightLoss
Cardio	A	22.6
Cardio	A	18.9
Cardio	B	5.9
Cardio	B	5.8
Weights	A	9.7
Weights	A	7.1
Weights	B	9.8
Weights	B	12.7

ANOVA

- **Example:** We are wanting to understand how to explain the WeightLoss variable from the diet variable.



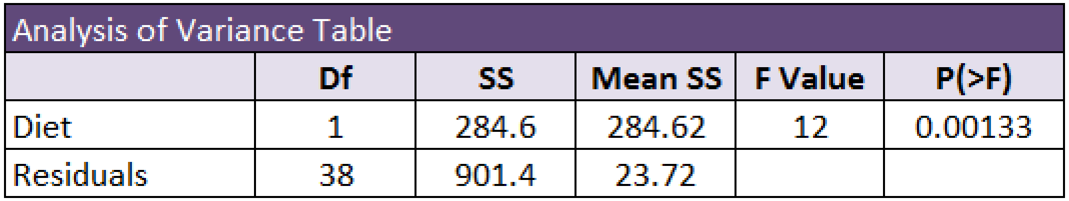
Exercise	Diet	WeightLoss
Cardio	A	22.6
Cardio	A	18.9
Cardio	B	5.9
Cardio	B	5.8
Weights	A	9.7
Weights	A	7.1
Weights	B	9.8
Weights	B	12.7



Analysis of Variance Table

	Df	SS	Mean SS	F Value	P(>F)
Between-group	1	284.6	284.62	12	0.00133
Within-group	38	901.4	23.72		

OR



Analysis of Variance Table

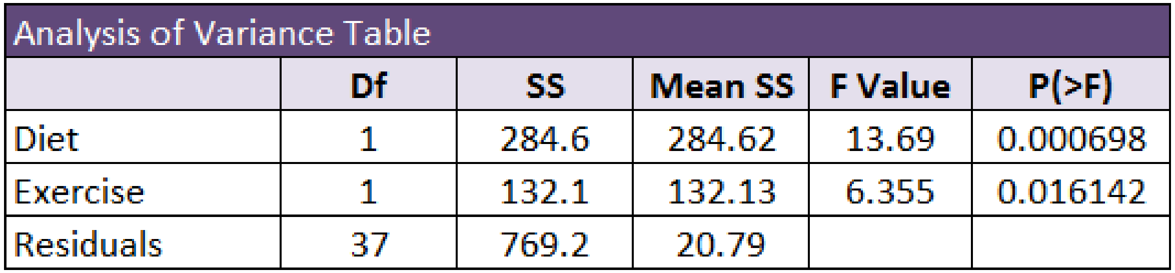
	Df	SS	Mean SS	F Value	P(>F)
Diet	1	284.6	284.62	12	0.00133
Residuals	38	901.4	23.72		

Observations:

- The F Value is well over 1 indicating that this variable has some explanatory value for WeightLoss.
- The P-Value is statistically significant at the 0.05 level.

ANOVA

- Let’s look at the ANOVA output for both the Diet and Exercise variables.



Analysis of Variance Table

	Df	SS	Mean SS	F Value	P(>F)
Diet	1	284.6	284.62	13.69	0.000698
Exercise	1	132.1	132.13	6.355	0.016142
Residuals	37	769.2	20.79		

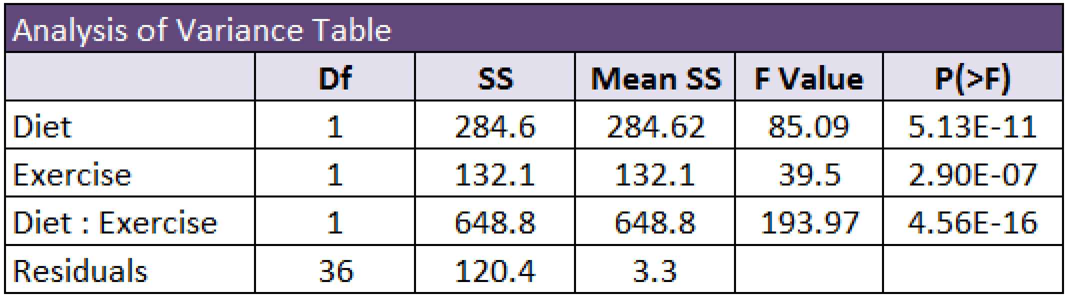
Between Group

Within Group

Observations:

- The Diet variable has a F Value of 13.69
- The Exercise variable has a F Value of 6.355
- Both variables are statistically significant at the 0.05 level

ANOVA



Analysis of Variance Table

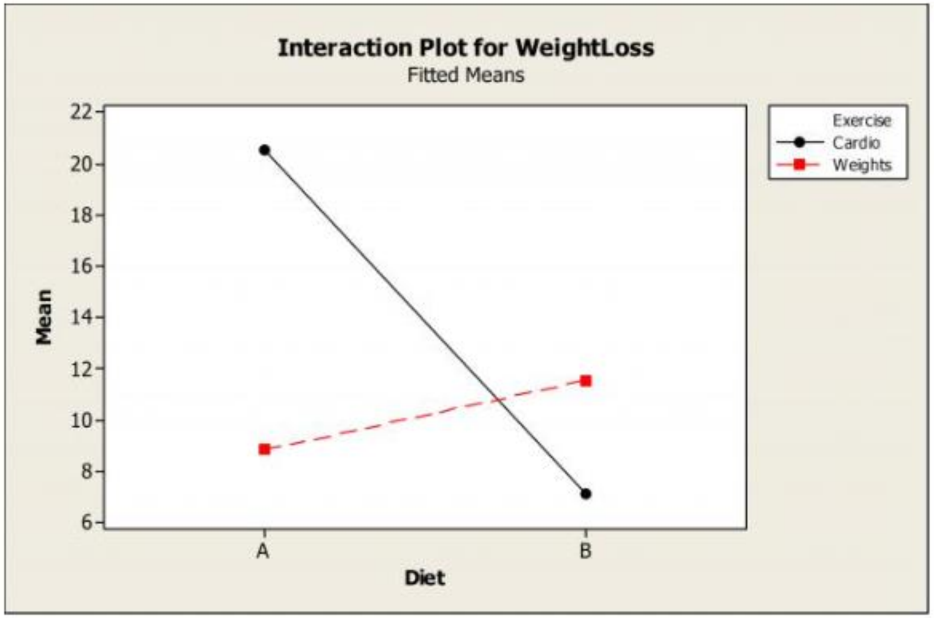
	Df	SS	Mean SS	F Value	P(>F)
Diet	1	284.6	284.62	85.09	5.13E-11
Exercise	1	132.1	132.1	39.5	2.90E-07
Diet : Exercise	1	648.8	648.8	193.97	4.56E-16
Residuals	36	120.4	3.3		

- We can see that the p-value for the Exercise*Diet interaction is 0.000. Because this p-value is so small, we can conclude that there is indeed a significant interaction between Exercise and Diet.

- So which diet is better? Our data suggest it’s like asking “ketchup or soy sauce?” The answer is, "It depends."

ANOVA

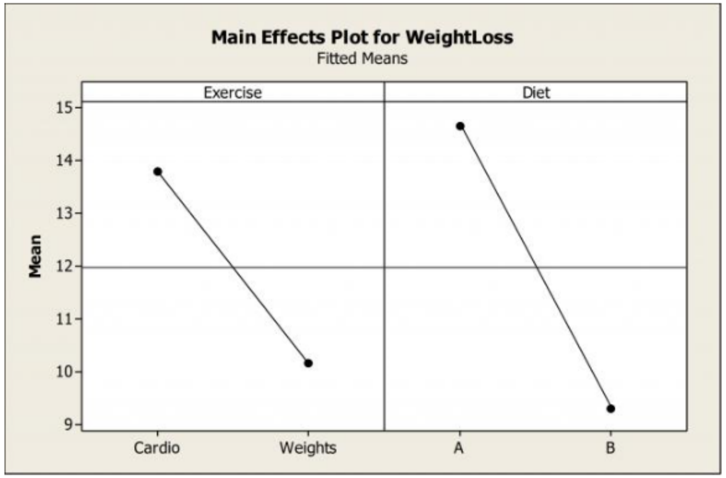
- Since the Exercise*Diet interaction is significant, let’s use an interaction plot to take a closer look:



- For participants using the cardio program (shown in black), we can see that diet A is best and results in greater weight loss. However, if you’re following the weight training regimen (shown in red), then diet B is results in greater weight loss than A.

ANOVA

- Suppose this interaction wasn't on our radar, and we instead focused only on the individual main effects and their impact on weight loss:



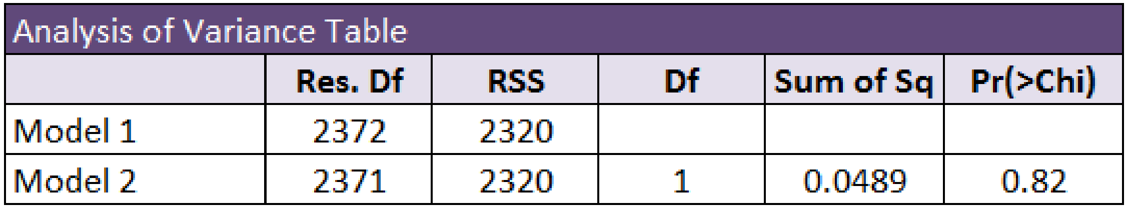
- Based on this plot, we would incorrectly conclude that diet A is better than B. As we saw from the interaction plot, that is only true IF we're looking at the cardio group.

- Clearly, we always need to evaluate interactions when analyzing multiple factors. If you don't, you run the risk of drawing incorrect conclusions…and you might just get ketchup with your sushi roll.

ANOVA

- ANOVA can also be used as a means to compare two linear regression models using the Chi-square measure.

- Here are two regression models we want to compare to each other. The order here is important so make sure you are applying the correct selection of the models.
  - Model 1: y = a
  - Model 2: y = b



Analysis of Variance Table

	Res. Df	RSS	Df	Sum of Sq	Pr(>Chi)
Model 1	2372	2320			
Model 2	2371	2320	1	0.0489	0.82

- The p-value of the test is 0.82. It means that the fitted model “Model 1" is not significantly different from Model 2 at the level of α=0.05 . Note that this test makes sense only if Model 1 and Model 2 are nested models. (i.e. it tests whether reduction in the residual sum of squares are statistically significant or not).

ANOVA and Linear Regression

- Linear regression is used to analyze continuous relationships; however, regression is essentially the same as ANOVA.

- In ANOVA, we calculate means and deviations of our data from the means.

- In linear regression, we calculate the best line through the data and calculate the deviations of the data from this line.

- The F ratio can be calculated in both.

"I can prove it or disprove it! What do you want me to do?"

Linear Regression Example
Boston Housing Market

Understanding the Data

- The dynamics and rapid change of the real estate market require business decision makers to seek advanced analytical solutions to maintain a competitive edge.

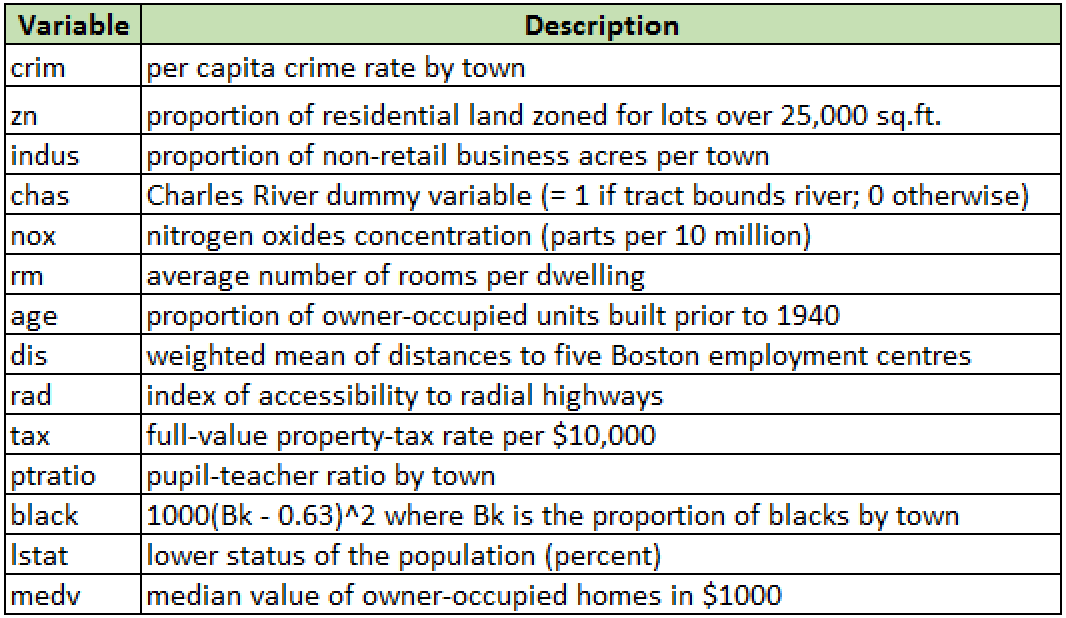
- Real estate pricing and home valuation can greatly benefit from predictive modeling techniques, in particular, linear regression.

- The dataset we will be working with reviews home values in Boston, Massachusetts and compiles a number of other statistics to help aid in determining property value.

- The goal for this exercise will be to provide a predictive model that can be leveraged to help real estate businesses in the Boston market.

Understanding the Data

- Here is a description of the variables within the dataset:

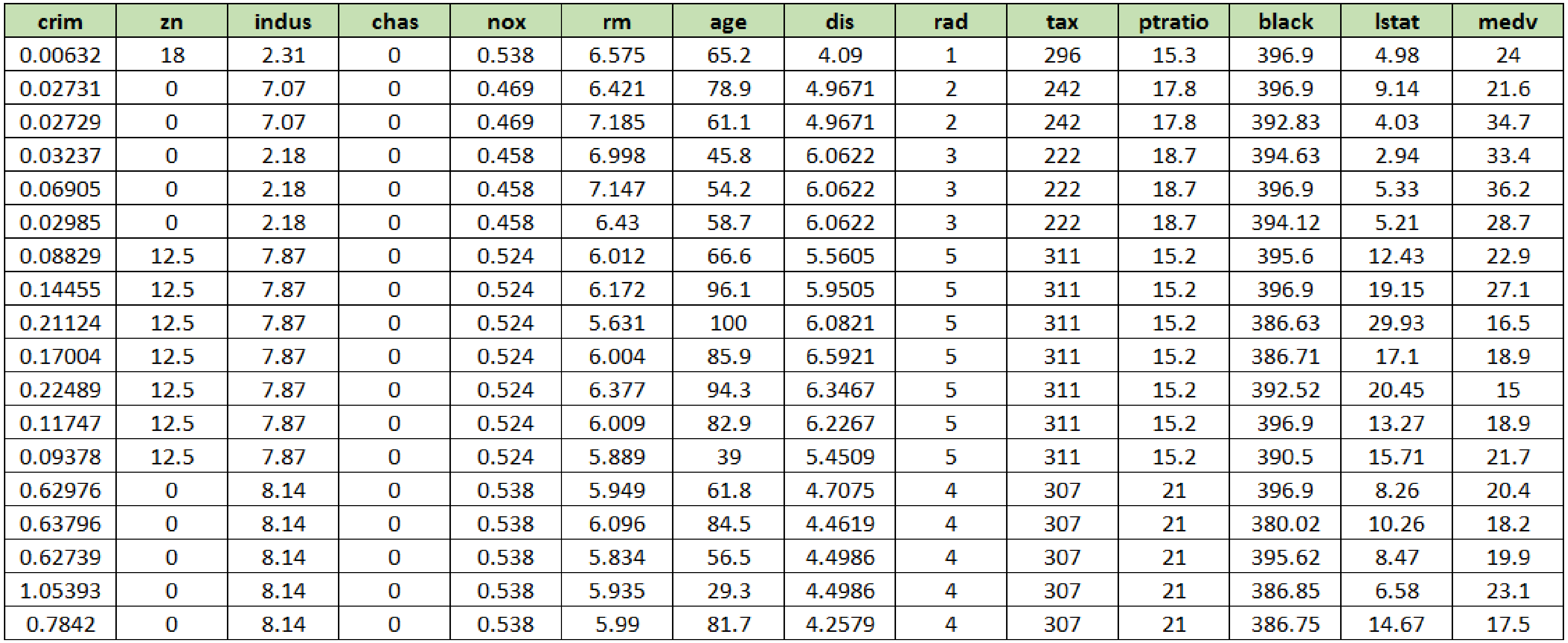


Variable	Description
crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft.
indus	proportion of non-retail business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
nox	nitrogen oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted mean of distances to five Boston employment centres
rad	index of accessibility to radial highways
tax	full-value property-tax rate per $10,000
ptratio	pupil-teacher ratio by town
black	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
lstat	lower status of the population (percent)
medv	median value of owner-occupied homes in $1000

- Our goal is to develop a multiple linear regression model for the median value of a home in Boston based upon the other variables.

Understanding the Data

- First, lets take a look at the raw data in the table.



crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9
0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15
0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9
0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7
0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4
0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2
0.62739	0	8.14	0	0.538	5.834	56.5	4.4986	4	307	21	395.62	8.47	19.9
1.05393	0	8.14	0	0.538	5.935	29.3	4.4986	4	307	21	386.85	6.58	23.1
0.7842	0	8.14	0	0.538	5.99	81.7	4.2579	4	307	21	386.75	14.67	17.5

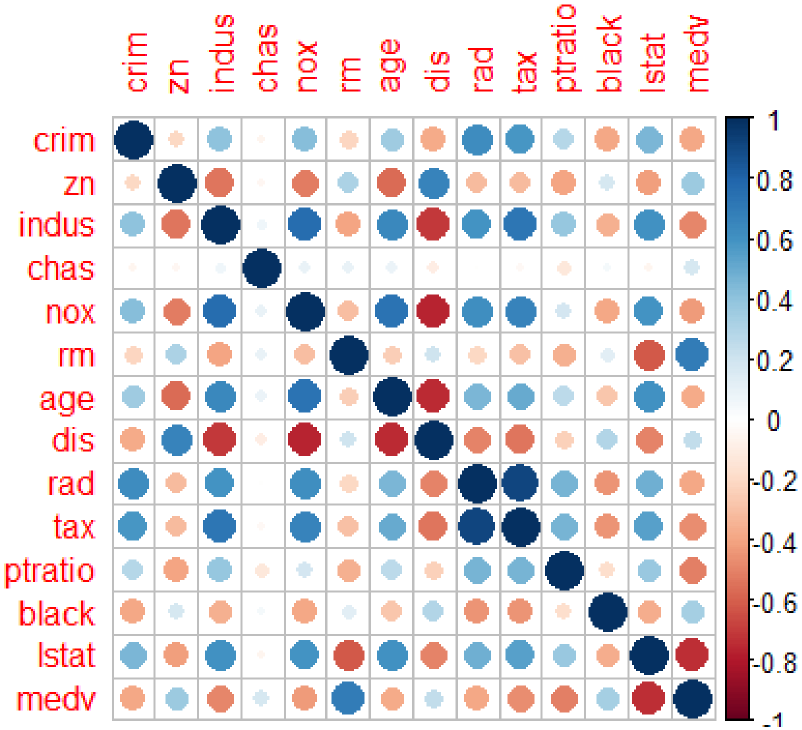
- With so many potential independent variables, we first need to reduce the field of variables to those which can help explain the model.

Understanding the Data

- A review of the correlation matrix indicates that there are a number of variables which we can use when building a model.

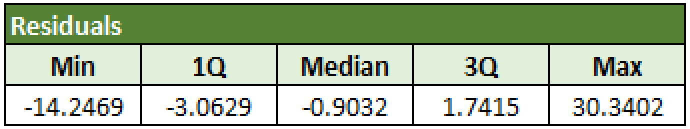
- Based upon the correlations, we will initially focus on utilizing the following variables: indus, rm, tax, ptratio, and lstat.

- Afterwards, we will assess quality of the models performance and utilize an alternative model approach.

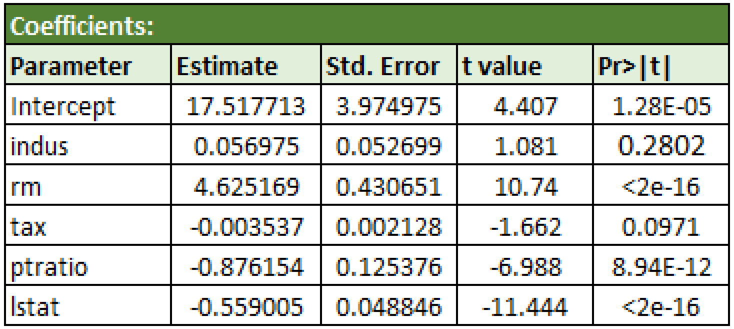


Correlation												
crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat
-0.388	0.360	-0.484	0.175	-0.427	0.695	-0.377	0.250	-0.382	-0.469	-0.508	0.333	-0.738

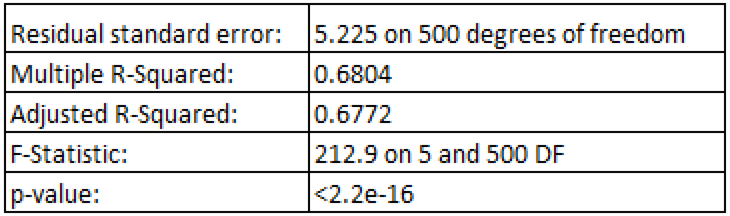
Preliminary Model Selection



Residuals				
Min	1Q	Median	3Q	Max
-14.2469	-3.0629	-0.9032	1.7415	30.3402



Coefficients:				
Parameter	Estimate	Std. Error	t value	Pr>|t|
Intercept	17.517713	3.974975	4.407	1.28E-05
indus	0.056975	0.052699	1.081	0.2802
rm	4.625169	0.430651	10.74	<2e-16
tax	-0.003537	0.002128	-1.662	0.0971
ptratio	-0.876154	0.125376	-6.988	8.94E-12
lstat	-0.559005	0.048846	-11.444	<2e-16



Residual standard error:	5.225 on 500 degrees of freedom
Multiple R-Squared:	0.6804
Adjusted R-Squared:	0.6772
F-Statistic:	212.9 on 5 and 500 DF
p-value:	<2.2e-16

- A tertiary review of the models output shows a couple of potential issues with the model.

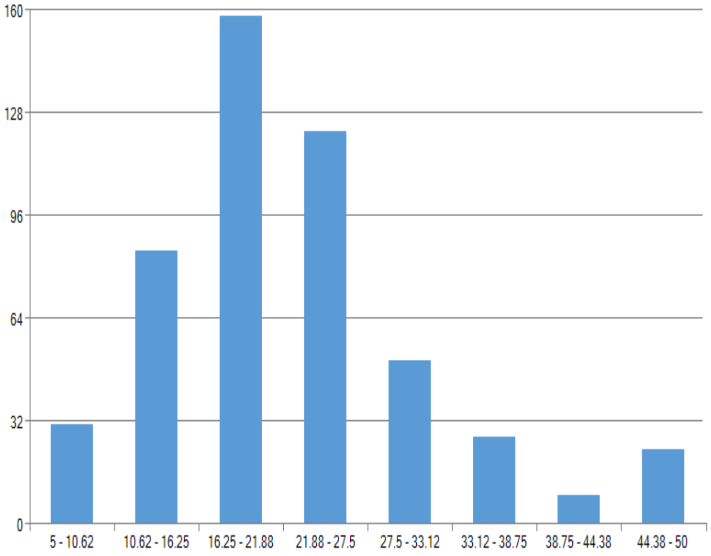
- Despite having a correlation of -0.484 to the median value, the indus variable is not statistically significant (0.2802) and should be dropped from the model. The tax variable is also statistically insignificant and should be removed from the model.

- The Adjusted R-squared is 0.6772, which indicates a reasonable goodness of fit and 67.7% of the variation in house prices can be explained by the five variables.

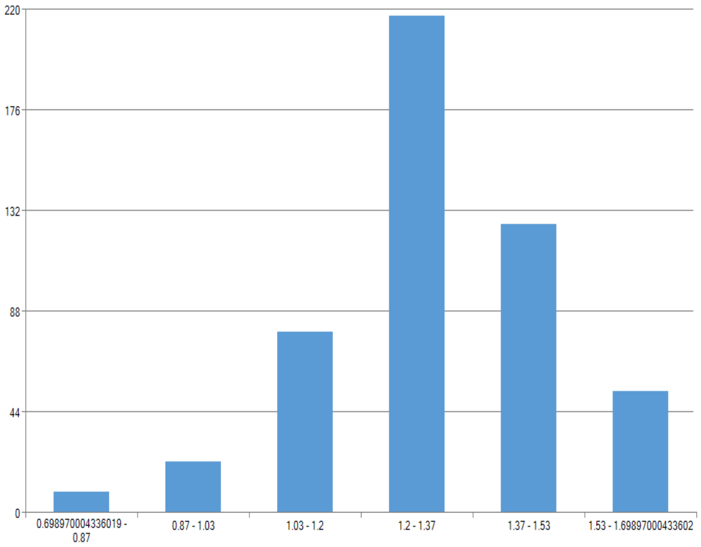
- There are some who would argue, based off of industry experience, that we could leave the model as is and the model performs reasonably well. Nevertheless, we will rebuild this model and improve its performance.

Preliminary Model Selection

- Lets double check that the dependent variable is normally distributed. It appears that the dataset is left skewed and could benefit from a log transformation.



Log Transformation

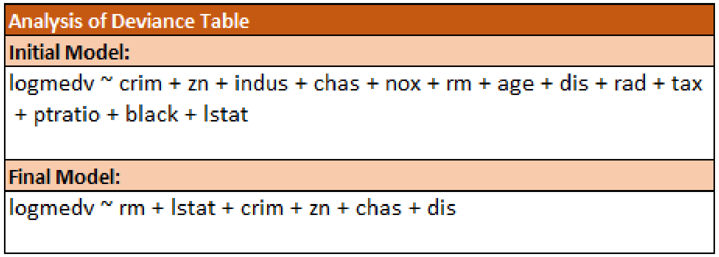


Preliminary Model Selection

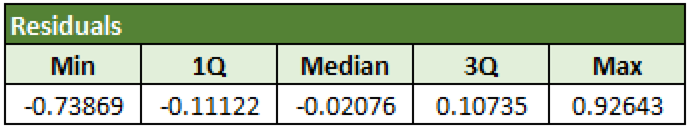
- Lets utilize an automated feature selection procedure called stepwise selection to identify those variables which are both statistically significant and add value to the regression model.

- This revised model now has all variables showing statistical significance at the p<0.05 level.

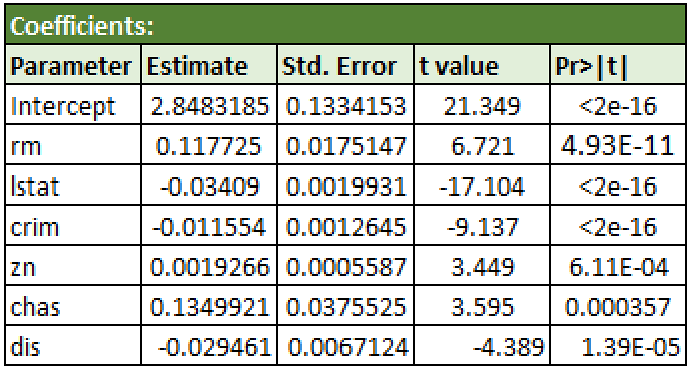
- Additionally, the model now has an Adjusted R-Square of 73.4% compared to 67.7% which is a notable improvement.



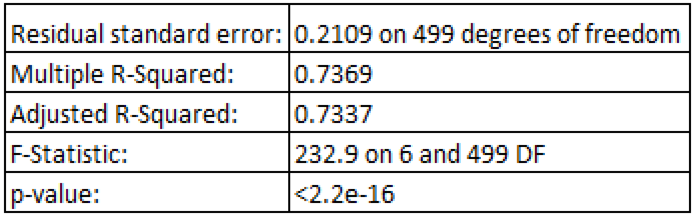
Analysis of Deviance Table
Initial Model:
logmedv ~ crim + zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio + black + lstat
Final Model:
logmedv ~ rm + lstat + crim + zn + chas + dis



Residuals				
Min	**1Q**	**Median**	**3Q**	**Max**
-0.73869	-0.11122	-0.02076	0.10735	0.92643



Coefficients:				
Parameter	**Estimate**	**Std. Error**	**t value**	**Pr>|t|**
Intercept	2.8483185	0.1334153	21.349	<2e-16
rm	0.117725	0.0175147	6.721	4.93E-11
lstat	-0.03409	0.0019931	-17.104	<2e-16
crim	-0.011554	0.0012645	-9.137	<2e-16
zn	0.0019266	0.0005587	3.449	6.11E-04
chas	0.1349921	0.0375525	3.595	0.000357
dis	-0.029461	0.0067124	-4.389	1.39E-05



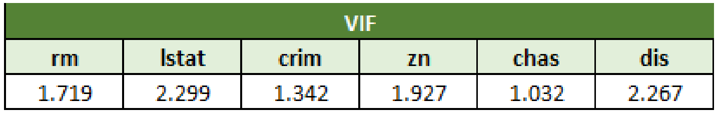
Residual standard error:	0.2109 on 499 degrees of freedom
Multiple R-Squared:	0.7369
Adjusted R-Squared:	0.7337
F-Statistic:	232.9 on 6 and 499 DF
p-value:	<2.2e-16

Model Performance

- We checked the VIF to determine whether multicollinearity is an issue. All of the values are below 3 which indicates that this is not an issue.

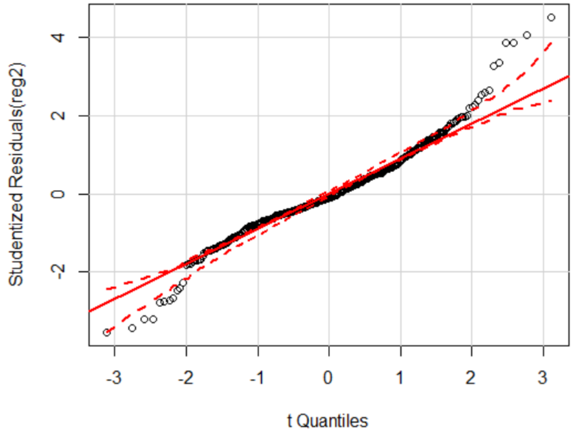
- A review of the QQ Plot indicates that the data generally agrees with a normal distribution, however, there are longer tails at the ends of the distribution.

- A review of the residual plots indicates the potential need to apply some transformation of the independent variable to further improve the model.

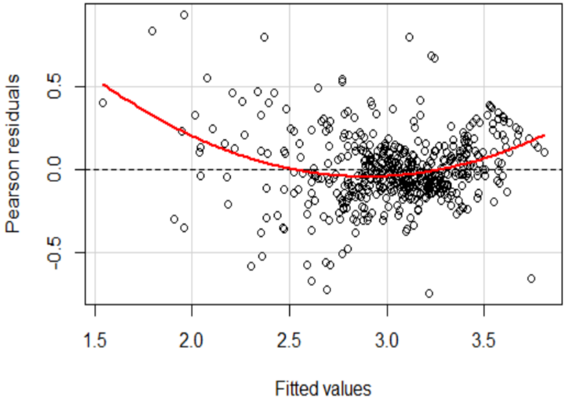


VIF					
rm	lstat	crim	zn	chas	dis
1.719	2.299	1.342	1.927	1.032	2.267

QQ Plot



Residual Plot



Model Interpretation

Here is the final model that we had produced:

Log(Price) = 2.8 + .12 Rooms - .03 Income - .01 Crime + .01 Zone + .13 River - .03 Employ

- When we performed a log-level transformation of the data, we now must interpret the change in x as a constant percentage increase in y.

Interpretation:

- Therefore, each additional room a house has leads to an increase of the price by 12%, holding other variables constant.
- If the home is near the river, the price increases by 13%, holding other variables constant.
- When a house is close to the main employment centers, the price decreases by 3% per unit.

Practical Example - Price Elasticity and Optimization

Understanding the Data

- A supermarket is selling a new type of grape juice in some of its stores for pilot testing.
- The senior management wants to understand the relationship between the grape juice and its impact on apple juice, cookie sales, and profitability.
- We will showcase how it is possible to build off of linear OLS regression models and econometric methodologies to solve a series of advanced business problems.
- The goal will be to provide tangible recommendations from our analyses to help this business manage their portfolio.

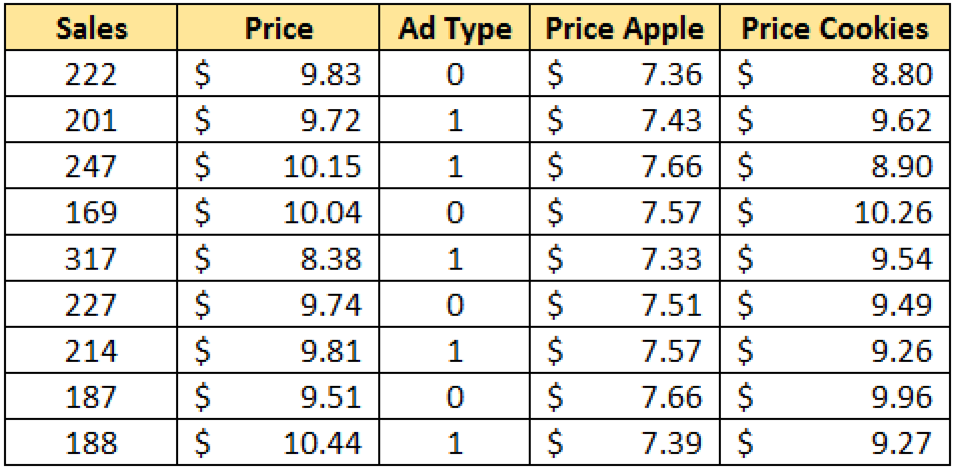
Understanding the Data

Our goal is to setup an experiments to analyze:

- Which type of in-store advertisement is more effective? The marketing group has placed two types of ads in stores for testing, one theme is natural production of the juice, the other theme is family health caring.
- The Price Elasticity – the reactions of sales quantity of the grape juice to its price change.
- The Cross-Price Elasticity – the reactions of sales quantity of the grape juice to the price changes of other products such as apple juice and cookies in the same store.
- How to find the best unit price of the grape juice which can maximize the profit and the forecast of sales with that price?

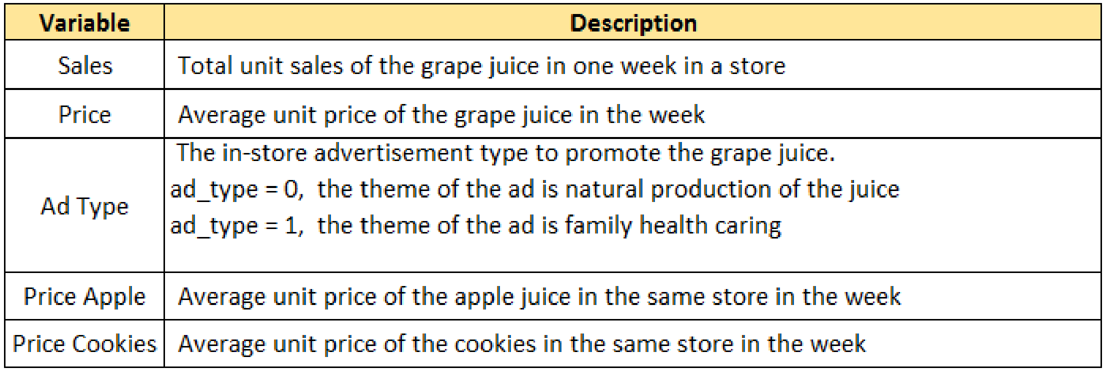
Understanding the Data

- First, lets take a look at the raw data in the table.



Sales	Price	Ad Type	Price Apple	Price Cookies
222	$ 9.83	0	$ 7.36	$ 8.80
201	$ 9.72	1	$ 7.43	$ 9.62
247	$ 10.15	1	$ 7.66	$ 8.90
169	$ 10.04	0	$ 7.57	$ 10.26
317	$ 8.38	1	$ 7.33	$ 9.54
227	$ 9.74	0	$ 7.51	$ 9.49
214	$ 9.81	1	$ 7.57	$ 9.26
187	$ 9.51	0	$ 7.66	$ 9.96
188	$ 10.44	1	$ 7.39	$ 9.27

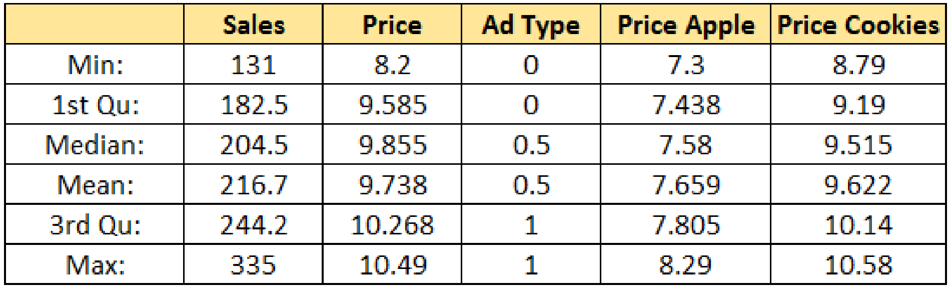
- Here is a description of the variables within the dataset:



Variable	Description
Sales	Total unit sales of the grape juice in one week in a store
Price	Average unit price of the grape juice in the week
Ad Type	The in-store advertisement type to promote the grape juice. ad_type = 0, the theme of the ad is natural production of the juice ad_type = 1, the theme of the ad is family health caring
Price Apple	Average unit price of the apple juice in the same store in the week
Price Cookies	Average unit price of the cookies in the same store in the week

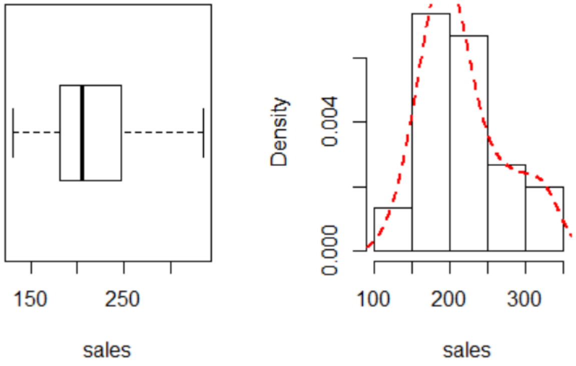
Understanding the Data

- From the summary table, we can roughly know the basic statistics of each numeric variable. For example, the mean value of sales is 216.7 units, the min value is 131, and the max value is 335.



	Sales	Price	Ad Type	Price Apple	Price Cookies
Min:	131	8.2	0	7.3	8.79
1st Qu:	182.5	9.585	0	7.438	9.19
Median:	204.5	9.855	0.5	7.58	9.515
Mean:	216.7	9.738	0.5	7.659	9.622
3rd Qu:	244.2	10.268	1	7.805	10.14
Max:	335	10.49	1	8.29	10.58

- We can further explore the distribution of the data of sales by visualizing the data in graphical form as follows. We don't find outliers in the above box plot graph and the sales data distribution is roughly normal. It is not necessary to apply further data cleaning and treatment to the data set.



Analysis of Ad Effectiveness

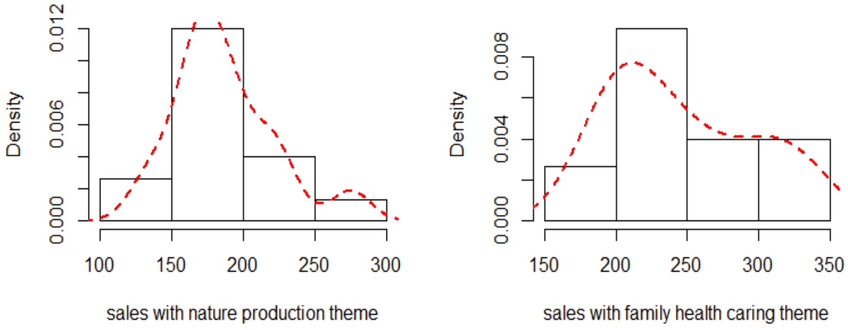
- The marketing team wants to find out the ad with better effectiveness for sales between the two types of ads:
  - a natural production theme
  - with family health caring theme.
- To find out the better ad, we can calculate and compare the mean of sales with the two different ad types at the first step.
- The mean of sales with nature product theme is about 187; the mean of sales with family health caring theme is about 247.
- It looks like that the latter one is better.

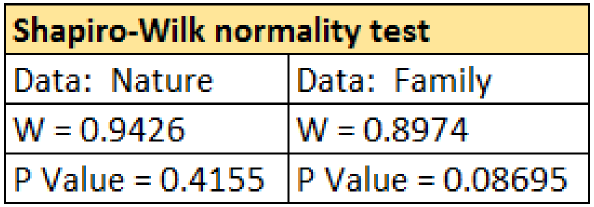
Analysis of Ad Effectiveness

- To find out how likely the conclusion is correct for the whole population, it is necessary to do statistical testing – two-sample t-test.

- We can see that both datasets are normally distributed and to be certain we can run the Shapiro-Wilk test.

- The p-values of the Shapiro-Wilk tests are larger than 0.05, so there is no strong evidence to reject the null hypothesis that the two groups of sales data are normally distributed.

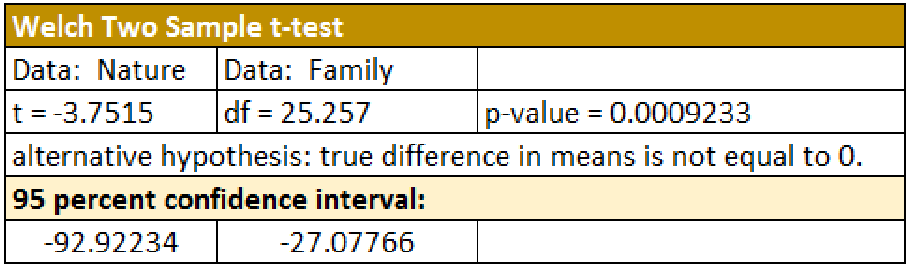




Shapiro-Wilk normality test	
Data: Nature	Data: Family
W = 0.9426	W = 0.8974
P Value = 0.4155	P Value = 0.08695

Analysis of Ad Effectiveness

- Now we can conduct the Welch two sample t-test since the t-test assumptions are met.



Welch Two Sample t-test		
Data: Nature	Data: Family	
t = -3.7515	df = 25.257	p-value = 0.0009233
alternative hypothesis: true difference in means is not equal to 0.		
95 percent confidence interval:		
-92.92234	-27.07766	

From the output of t-test above, we can say that:

- We have strong evidence to say that the population means of the sales with the two different ad types are different because the p-value of the t-test is very small;
- With 95% confidence, we can estimate that the mean of the sales with natural production theme ad is somewhere in 27 to 93 units less than that of the sales with family health caring theme ad.
- So the conclusion is that the ad with the theme of family health caring is BETTER.

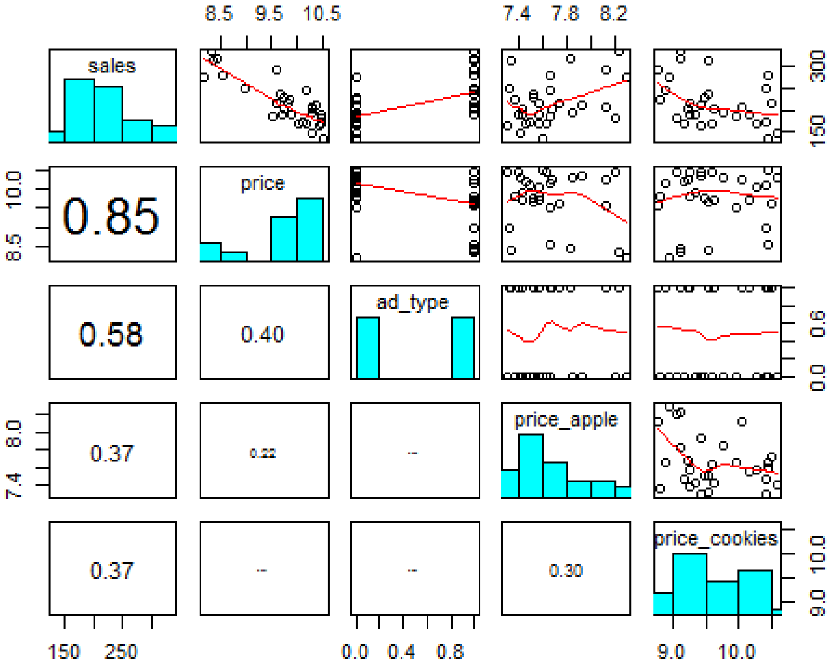
Sales Driver and Price Elasticity Analysis

- With the information given in the data set, we can explore how grape juice price, ad type, apple juice price, cookies price influence the sales of grape juice in a store by multiple linear regression analysis.

- Here, "sales" is the dependent variable and the others are independent variables.

- Let's investigate the correlation between the sales and other variables by displaying the correlation coefficients in pairs.

- The correlation coefficients between sales and price, ad type, price apple, and price cookies are 0.85, 0.58, 0.37, and 0.37 respectively, that means they all might have some influences to the sales.



Preliminary Model Selection

Residuals				
Min	1Q	Median	3Q	Max
-36.29	-10.488	0.884	10.483	29.471

Coefficients:				
Parameter	**Estimate**	**Std. Error**	**t value**	**Pr>|t|**
Intercept	774.81	145.349	5.331	1.59E-05
Price	-51.239	5.321	-9.63	6.83E-10
Ad Type	29.742	7.249	4.103	0.00038
Price Apple	22.089	12.512	1.765	0.08971
Price Cookies	-25.277	6.296	-4.015	0.00048

Residual standard error:	18.2 on 25 degrees of freedom
Multiple R-Squared:	0.8974
Adjusted R-Squared:	0.881
F-Statistic:	54.67 on 4 and 25 DF
p-value:	5.32E-12

- We can try to add all of the independent variables into the regression model:

- The p-value for Price, Ad Type, and Price Cookies is much less than 0.05. They are significant in explaining the sales. We are confident to include these variables into the model.

- The p-value of Price Apple is a bit larger than 0.05, seems there are no strong evidence for apple juice price to explain the sales. However, according to our real-life experience, we know when apple juice price is lower, consumers likely to buy more apple juice, and then the sales of other fruit juice will decrease.

- So we can also add it into the model to explain the grape juice sales.

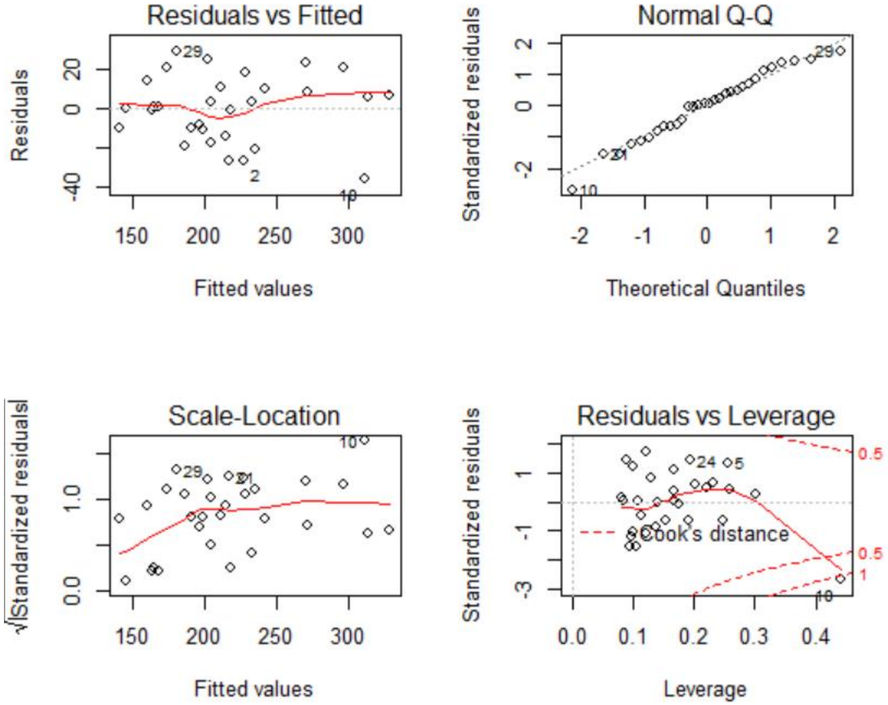
- The Adjusted R-squared is 0.881, which indicates a reasonable goodness of fit and 88% of the variation in sales can be explained by the four variables.

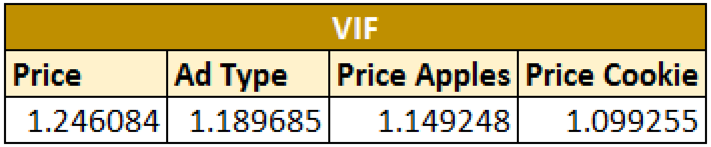
Preliminary Model Selection

- The assumptions for the regression to be true are that data are random and independent; residuals are normally distributed and have constant variance. Let's check the residuals assumptions visually.

- The Residuals vs Fitted graph above shows that the residuals scatter around the fitted line with no obvious pattern, and the Normal Q-Q graph shows that basically the residuals are normally distributed. The assumptions are met.

- The VIF test value for each variable is close to 1, which means the multicollinearity is very low among these variables.





VIF			
Price	Ad Type	Price Apples	Price Cookie
1.246084	1.189685	1.149248	1.099255

Price Elasticity Analysis

Linear Regression Model:

Sales = 774.81 – 51.24 * Price + 29.74 * Ad Type + 22.1 * Price Apple – 25.28 * Price Cookies

- With model established, we can analysis the Price Elasticity(PE) and Cross-price Elasticity(CPE) to predict the reactions of sales quantity to price.

Price Elasticity

- PE = (ΔQ/Q) / (ΔP/P) = (ΔQ/ΔP) * (P/Q) = -51.24 * 0.045 = -2.3
- P is price, Q is sales quantity
- ΔQ/ΔP = -51.24 , the parameter before the variable “price” in the above model
- P/Q = 9.738 / 216.7 = 0.045
- P is the mean of prices in the dataset, Q is the mean of the Sales variable.

Interpretation: The PE indicates that 10% decrease in grape juice price will increase the grape juice sales by 23%, and vice versa.

Cross Price Elasticity Analysis

Let's further calculate the CPE on apple juice and cookies to analyze the how the change of apple juice price and cookies price influence the sales of grape juice.

Cross Price Elasticity

- $CPE_{apple} = (\Delta Q / \Delta P_{apple}) * (P_{apple} / Q) = 22.1 * (7.659 / 216.7) = 0.78$
- $CPE_{cookies} = (\Delta Q / \Delta P_{cookies}) * (P_{cookies} / Q) = -25.28 * (9.622 / 216.7) = -1.12$

Interpretation:

- The CPE_{apple} indicates that 10% decrease in apple juice price will DECREASE the sales of grape juice by 7.8%, and vice verse. So the grape juice and apple juice are substitutes.
- The $CPE_{cookies}$ indicates that 10% decrease in cookies price will INCREASE the grape juice sales by 11.2%, and vice verse. So the grape juice and cookies are compliments. Place the two products together will likely increase the sales for both.
- We can also know that the grape juice sales increase 29.74 units when using the ad with the family health caring theme (ad_type = 1).

Optimization of the Price

- Usually companies want to get higher profit rather than just higher sales quantity.

- So, how to set the optimal price for the new grape juice to get the maximum profit based on the dataset collected in the pilot period and our regression model?

- To simplify the question, we can let the Ad Type = 1, the Price Apple = 7.659 (mean value), and the Price Cookies = 9.738 (mean value).

The model is simplified as follows:

- Sales = 774.81 – 51.24 * price + 29.74 * 1 + 22.1 * 7.659 – 25.28 * 9.738

- Sales = 772.64 – 51.24*price

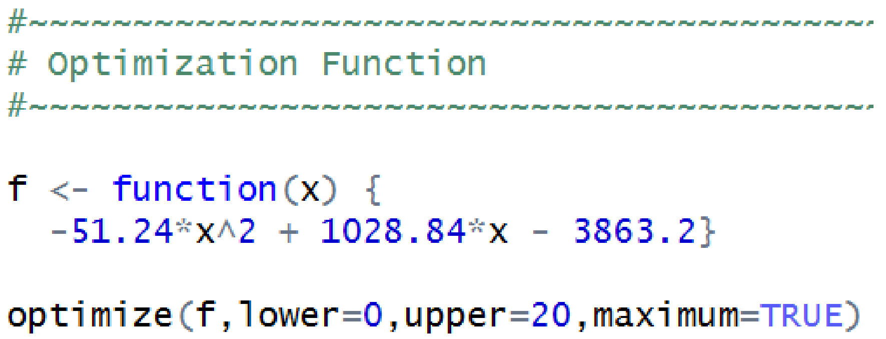
Optimization of the Price

- Assume the marginal cost (C) per unit of grape juice is $5.00. We can calculate the profit (Y) by the following formula:

- Y = (price – C) * Sales Quantity = (price – 5) * (772.64 – 51.24*price)

- Y = – 51.24 * price^2 + 1028.84 * price – 3863.2

- To get the optimal price to maximize Y, we can use the following R function.



```r
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# Optimization Function
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

f <- function(x) {
  -51.24*x^2 + 1028.84*x - 3863.2}

optimize(f,lower=0,upper=20,maximum=TRUE)
```

Optimization of the Price

- The optimal price is $10.04; the maximum profit will be $1301 according to the above output. In reality, we can reasonably set the price to be $10.00 or $9.99.

- We can further use the model to predict the sales while the price is $10.00.
  - Additionally, the ad type = 1
  - Mean Price Apple = 7.659
  - Mean Price Cookies = 9.738

Linear Regression Model:

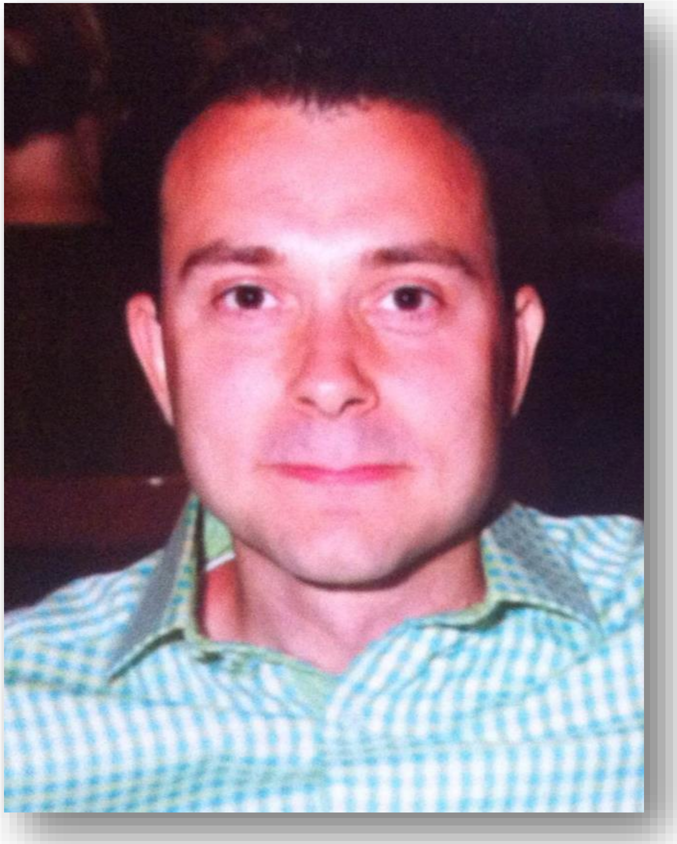
Sales = 774.81 – 51.24 * Price + 29.74 * Ad Type + 22.08 * Price Apple – 25.27 * Price Cookies

Sales = 774.81 – 51.24 * (10) + 29.74 * (1)+ 22.08 * (7.659) – 25.27 *(9.738)

- The sales forecast will be 215 units with a variable range of 176 ~ 254 with 95% confidence in a store in one week on average.
- Based on the forecast and other factors, the supermarket can prepare the inventory for all of its stores after the pilot period.

About Me

- Reside in Wayne, Illinois
- Active Semi-Professional Classical Musician (Bassoon).
- Married my wife on 10/10/10 and been together for 10 years.
- Pet Yorkshire Terrier / Toy Poodle named Brunzie.
- Pet Maine Coons’ named Maximus Power and Nemesis Gul du Cat.
- Enjoy Cooking, Hiking, Cycling, Kayaking, and Astronomy.
- Self proclaimed Data Nerd and Technology Lover.



Fine

Acknowledgements

- http://en.wikipedia.org/wiki/Regression_analysis
- http://www.ftpress.com/articles/article.aspx?p=2133374
- http://people.duke.edu/~rnau/Notes_on_linear_regression_analysis--Robert_Nau.pdf
- http://www.theanalysisfactor.com/interpreting-interactions-in-regression/
- http://www.edanzediting.com/blog/statistics_anova_explained#.VIdeEo0tBSM
- http://en.wikipedia.org/wiki/Ordinary_least_squares
- http://www.unt.edu/rss/class/mike/6810/OLS%20Regression%20Summary.pdf
- http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
- http://www.chsbs.cmich.edu/fattah/courses/empirical/multicollinearity.html
- http://home.wlu.edu/~gusej/econ398/notes/logRegressions.pdf
- http://www.dataapple.net/?p=19