Inspire…Educate…Transform.

Statistics and Probability in Decision Modeling

Multiple Linear Regression

Dr. Sridhar Pappu
Executive VP – Academics, INSOFE

October 07, 2018

Degrees of Freedom

Degrees of freedom, ν: # of independent observations for a source of variation minus the number of independent parameters estimated in computing the variation.*

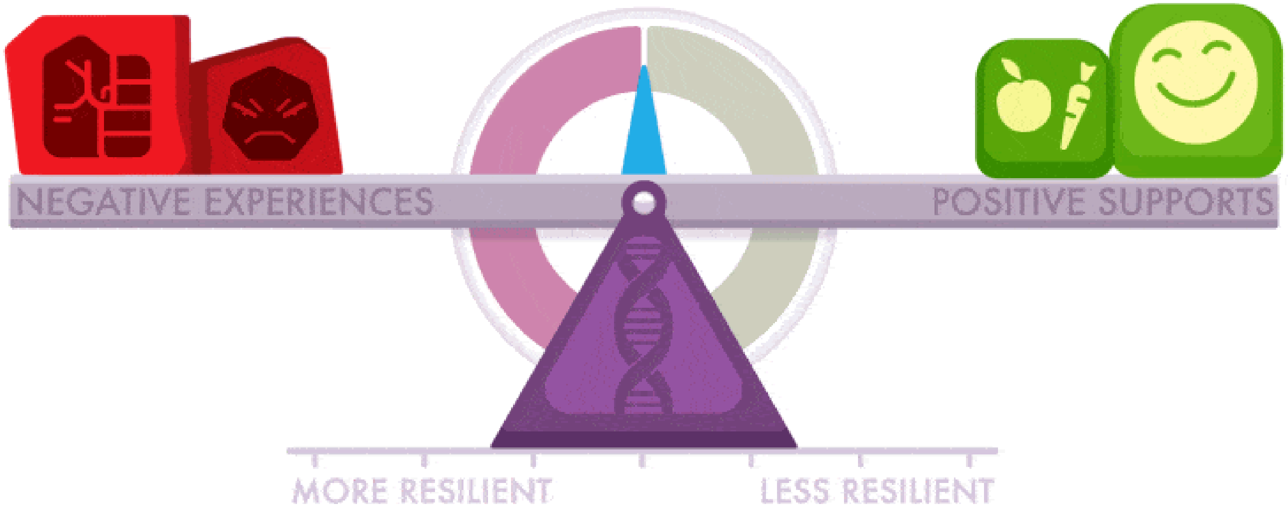
When sample size is considered, degrees of freedom are *n-1*.

* Roger E. Kirk, *Experimental Design: Procedures for the Behavioral Sciences*. Belmont, California: Brooks/Cole, 1968.

Influential Observations - Leverage

How much the observation's value on the **predictor variable** differs from the mean of the **predictor variable**.

That is, it tells us about extreme x values, which have the potential to highly influence the regression in certain conditions. *Remember Eric McCoo.*



CSE 7302c

Influential Observations - Leverage

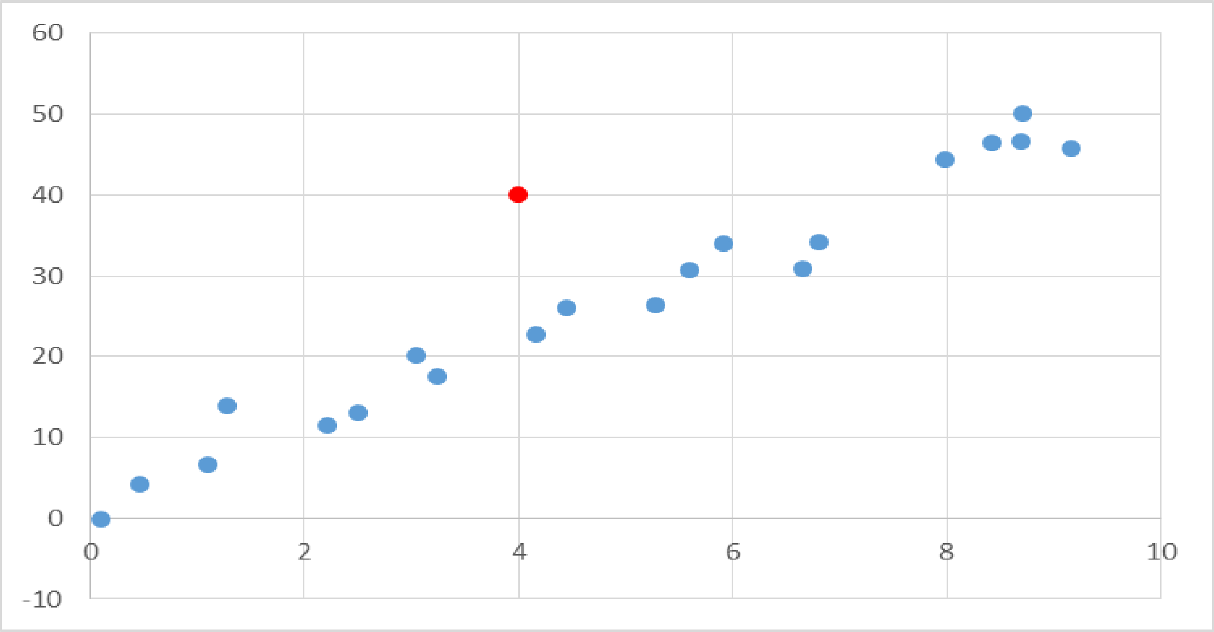
Leverage of the i^{th} data point is given by:

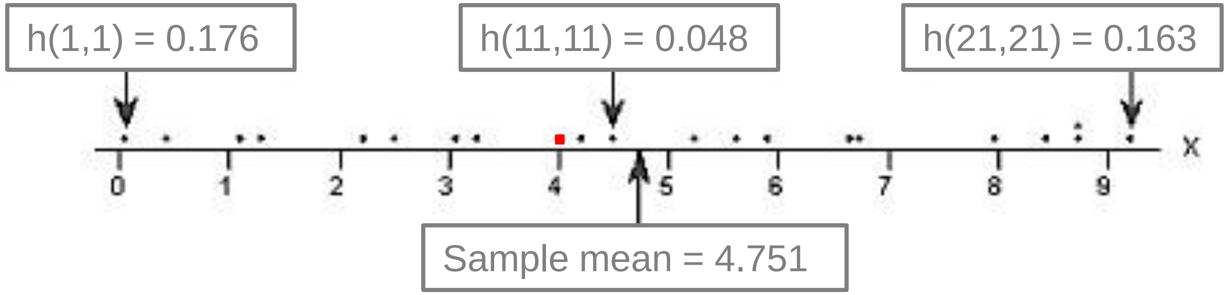
$$h_i = \frac{1+z^2}{n}$$

The sum of leverages = # of parameters, *p* (regression coefficients **including intercept**).

EXCEL ACTIVITY

Influential Observations - Leverage





Flag observations whose h > 3* avg(h) or h > 2* avg(h)

$$Avg(h) = \frac{sum(h)}{n} = \frac{p}{n}$$

Influential Observations - Distance

Based on error of prediction and is measured by Studentized Residual. This is calculated on the **dependent** variable and is a measure of 'outlierness'.

Recall Student's t-test. So, Studentizing is related to calculating the t-statistic of the metric in question, i.e., it is related to error of prediction of that observation divided by the standard deviation of the errors of prediction.

Influential Observations - Distance

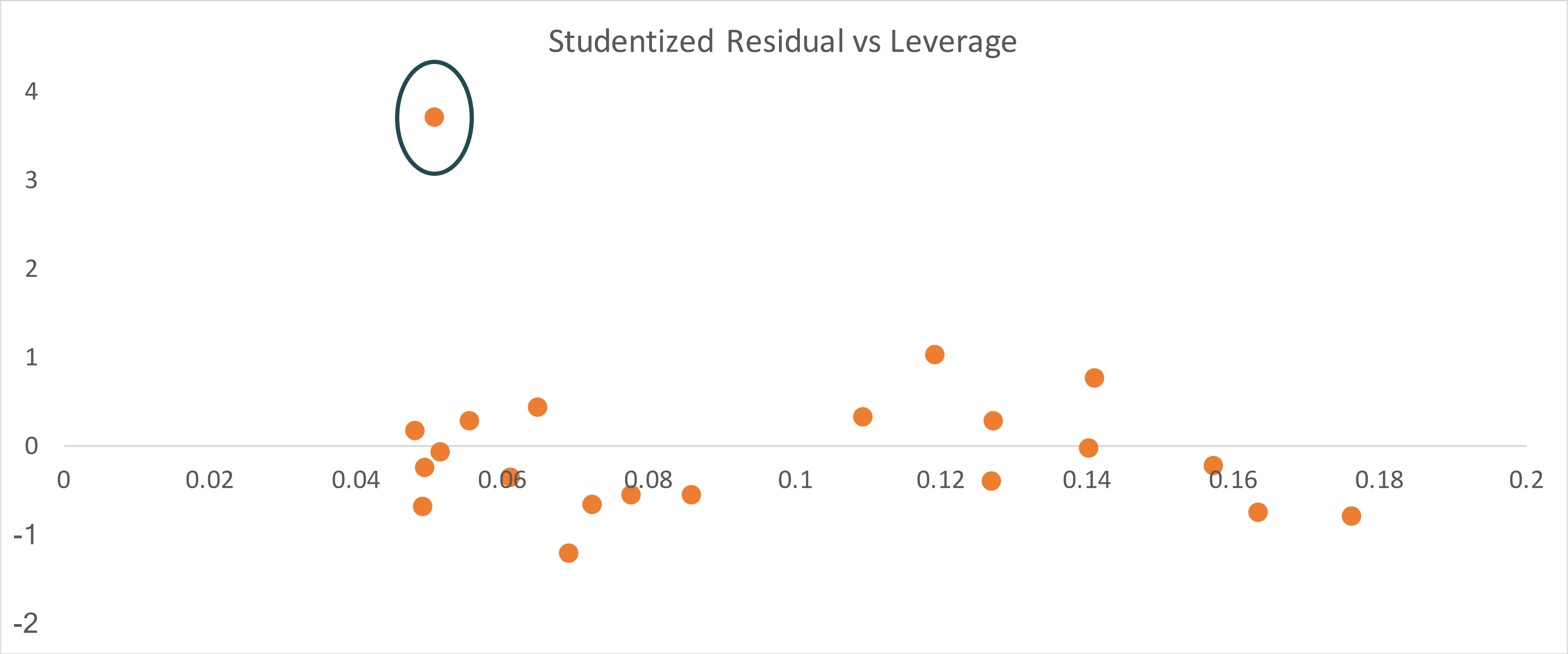
$$stdres_i = \frac{e_i}{\sqrt{MSE(1 - h_i)}}$$

Investigate observations with internally studentized residuals smaller than -2 or larger than 2.

Recall the empirical rule for normal distribution and the assumption that residuals follow normal distribution.

EXCEL ACTIVITY

Influential Observations - Distance

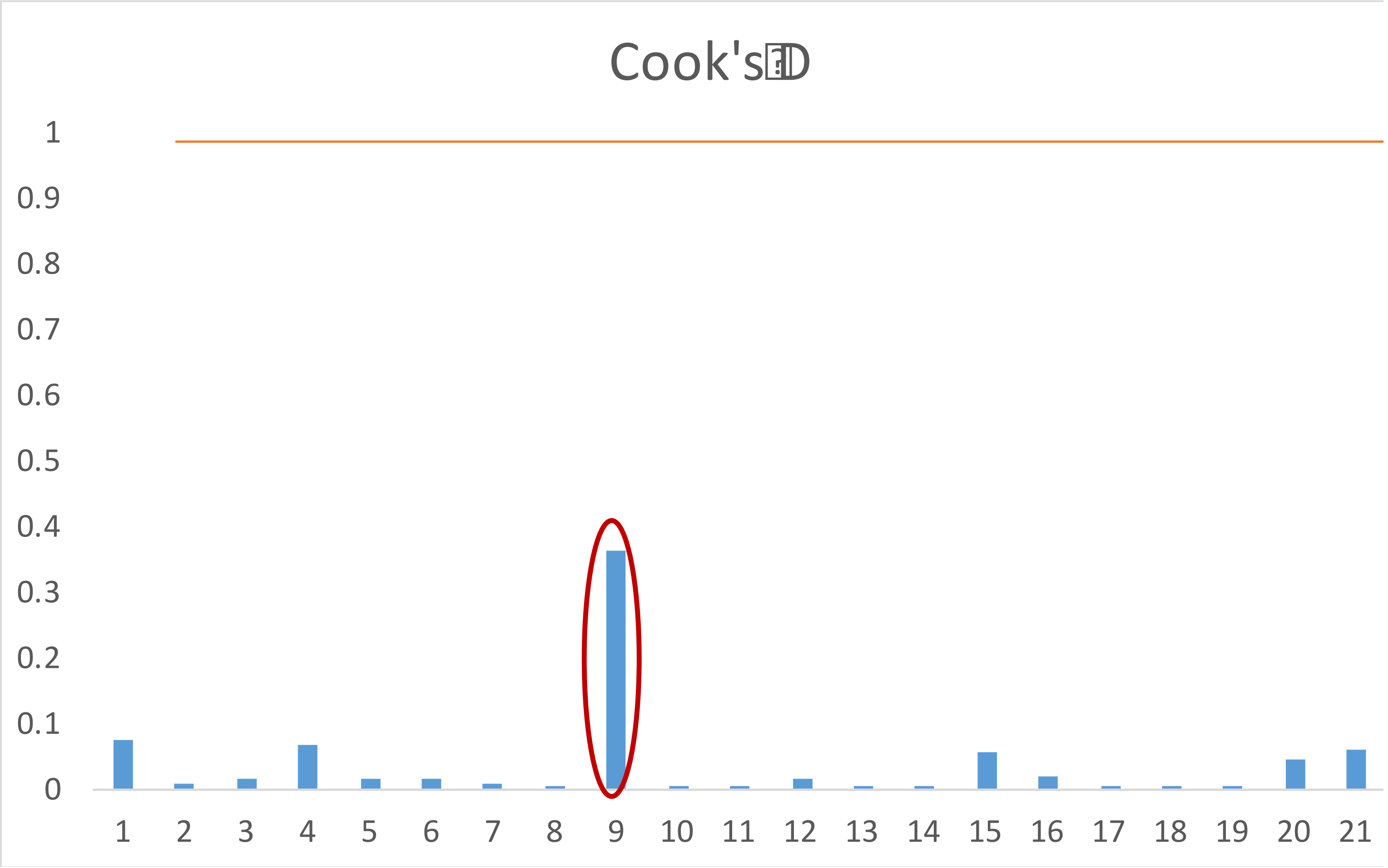


Influential Observations – Cook's D

Measures overall influence of an observation by seeing the impact on the regression coefficients when this observation is omitted. It accounts both for **leverage** and **residual**.

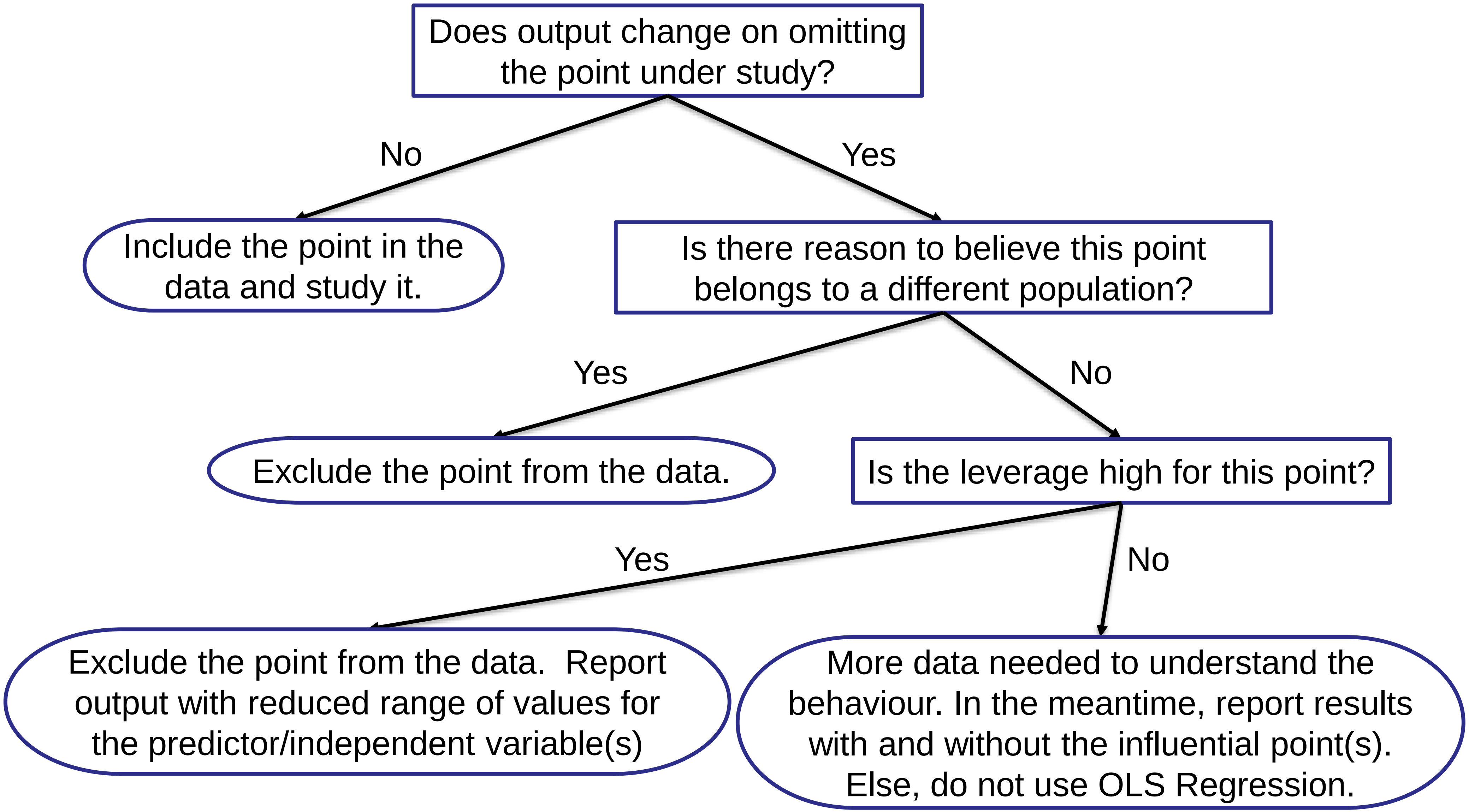
$$D_i = \frac{1}{p}(stdres_i)^2\left(\frac{h_i}{1-h_i}\right)$$

Influential Observations – Cook's D



CSE 7302c

Handling Influential Observations



Caution – R^2 in Regression Through Origin

- Temptation to drop intercept if it is **not significant**.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.717055011							
R Square	0.514167888							
Adjusted R Square	0.494734604							
Standard Error	4.21319131							
Observations	27							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	469.6573265	469.6573265	26.4581054	2.57053E-05			
Residual	25	443.7745253	17.75098101					
Total	26	913.4318519						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	-4.154014573	2.447784673	-1.697050651	0.102104456	-9.195321476	0.88729233	-10.97705723	2.669028089
Big Mac Price ($)	3.547427488	0.689658599	5.143744297	2.57053E-05	2.127049014	4.967805962	1.625048409	5.469806567

Caution – R^2 in Regression Through Origin

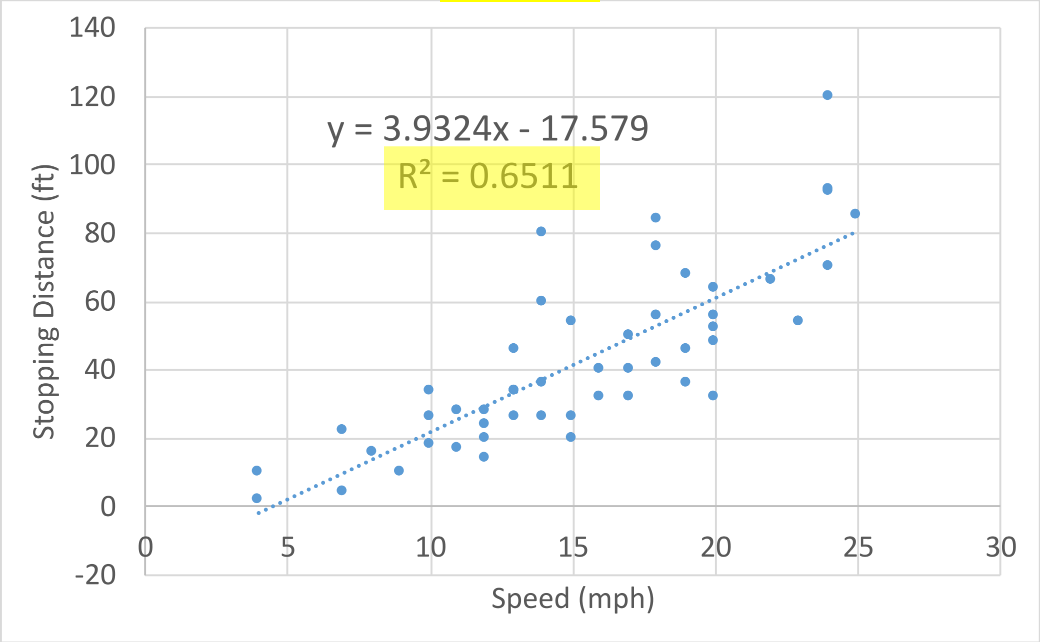
- Physical process makes intuitive sense for *y*=0 when *x*=0. For example, if the speed of the car = 0 mph, the distance travelled before it comes to a stop = 0 ft.

Caution – R^2 in Regression Through Origin

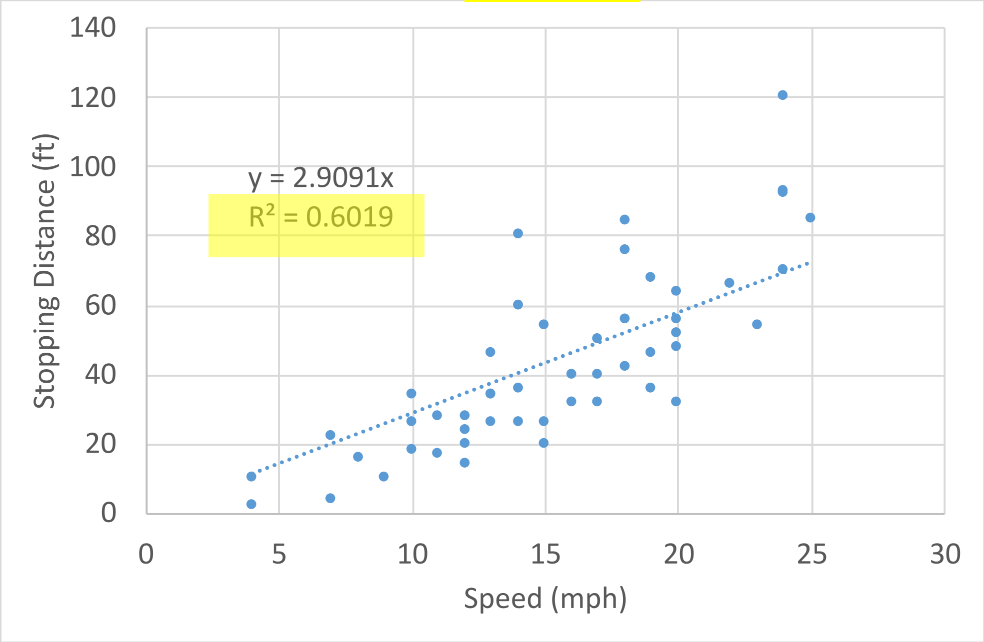
- However, you should evaluate other metrics before deciding on Regression Through Origin (RTO).
- Also, if (0,0) or values close to the origin are not in the dataset, you should be cautious about using RTO, as it is quite possible that data are not linear in the vicinity of the origin.

Caution – R^2 in Regression Through Origin

OLS



RTO



Excel gives a lower R^2 when the trendline method is used. This is using the R^2 formula we learned earlier (1-SSE/SST).

CSE 7302c

Caution – R^2 in Regression Through Origin

RTO output from Excel

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.94672557					
R Square	0.89628931					
Adjusted R Sq	0.87588114					
Standard Erro	16.2592371					
Observations	50					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	111949.223	111949.223	423.468152	1.8549E-25	
Residual	49	12953.7768	264.362793			
Total	50	124903				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
speed	2.90913214	0.14136864	20.5783418	9.2278E-26	2.62504123	3.19322306

RTO output from R

```
Call:
lm(formula = dist ~ speed + 0, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-26.183 -12.637  -5.455   4.590  50.181

Coefficients:
      Estimate Std. Error t value
speed   2.9091     0.1414   20.58
      Pr(>|t|)
speed   <2e-16 ***
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
  0.1 ' ' 1

Residual standard error: 16.26 on 49 degrees of freedom
Multiple R-squared:  0.8963,    Adjusted R-squared:  0.8942
F-statistic: 423.5 on 1 and 49 DF,  p-value: < 2.2e-16
```

Caution – R^2 in Regression Through Origin

The high R^2 values output by the software in models without intercepts is misleading, although mathematically sound. The SST calculation is used with respect to the origin instead of the mean.

$$\left((y_i - 0)^2 \text{ instead of } (y_i - \bar{y})^2\right)$$

This gives a much higher SST resulting in higher R^2. It is, therefore, best to check for other performance criteria when doing RTO. There needs to be a VERY STRONG reason for doing RTO, though.

Multiple Linear Regression

THE OUTPUT

Multiple Linear Regression

- Simple Linear Regression models the effect of one independent variable, *x*, on one dependent variable, *y*
- Multiple Regression models the effect of several independent variables, x_1, x_2 etc., on one dependent variable, *y*
- The different x variables are combined in a linear way and each has its own regression coefficient:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots .. + \beta_n x_n + \varepsilon$$

- The β parameters reflect the **independent contribution** of each independent variable, *x*, to the value of the dependent variable, *y*.

Interpreting Regression Coefficients

A coefficient is the slope of the linear relationship between the dependent variable (DV) and the **independent contribution** of the independent variable (IV), i.e., that part of the IV that is independent of (or uncorrelated with) all other IVs.

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.89666084					
R Square	0.804000661					
Adjusted R Square	0.750546296					
Standard Error	2.90902388					
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	381.8467141	127.282238	15.04087945	0.00033002	
Residual	11	93.08661926	8.462419933			
Total	14	474.9333333				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286

Assumptions of Multiple Linear Regression

- Same as simple linear regression
  - Linearity
  - Independence of errors
  - Homoscedasticity (constant variance)
  - Normality of errors

- Methods of checking assumptions are also the same

Determining the Multiple Regression Equation

- *k+1* equations to solve for *k* independent variables and the intercept.
- In solving for intercept and slope in a simple linear regression model, we needed $\sum x$, $\sum y$, $\sum xy$, and $\sum x^2$.
- For multiple regression model with 2 independent variables, we need $\sum x_1$, $\sum x_2$, $\sum y$, $\sum {x_1}^2$, $\sum {x_2}^2$, $\sum x_1 x_2$, $\sum x_1 y$, and $\sum x_2 y$.

Determining the Multiple Regression Equation - Excel

In a real estate study, multiple variables were explored to determine the price of a house.

- # of bedrooms
- # of bathrooms
- Age of the house
- # of square feet of living space
- Total # of square feet of space
- # of garages

Find the equation if you want to predict the price of the house by total square feet and age of the house.

Determining the multiple regression equation – Interpreting the output

What is the equation?

$$\hat{y} = 57.35 + 0.0177Area - 0.666Age$$

Are the coefficients and the model significant?

Yes

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.860872681
R Square	0.741101773
Adjusted R Square	0.715211951
Standard Error	11.96038667
Observations	23

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	8189.723012	4094.861506	28.62521631	1.35298E-06
Residual	20	2861.016988	143.0508494		
Total	22	11050.74			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	57.35074586	10.00715186	5.73097587	1.31298E-05	36.47619286	78.22529885
Area (sq ft) (x1)	0.017718036	0.00314562	5.632605205	1.63535E-05	0.011156388	0.024279685
Age of House (years) (x2)	-0.666347946	0.227996703	-2.922620973	0.008417613	-1.141940734	-0.190755157

CSE 7302c

Residuals – Practice Assignment

Residuals are determined the same way as in simple linear regression. The predicted value is calculated by substituting the predictor values of interest. The residual is again the difference between the observed and the predicted values, $y - \hat{y}$.

CSE 7302c

SSE and Standard Error of the Estimate, SE – Practice Assignment

$$SSE = \sum (y - \hat{y})^2$$

$$SE = \sqrt{\frac{SSE}{n - k - 1}}$$

Coefficient of Multiple Determination, R^2 – Practice Assignment

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Adjusted R^2 - Excel

As additional independent variables are added to the regression model, the value of R^2 increases.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

However, sometimes these variables are insignificant and add no real value, yet inflating the R^2 value.

Adjusted R^2 takes into consideration both the additional information and the changed degrees of freedom.

$$Adjusted\ R^2 = 1 - \frac{\frac{SSE}{(n-k-1)}}{\frac{SST}{n-1}} = R^2 - (1 - R^2)\frac{k}{n-k-1} = 1 - \frac{MSE}{MST}$$

Sample R Output

```
Call:
lm(formula = ToxinConc$Toxin ~ ToxinConc$Rain + ToxinConc$NoonTemp +
    ToxinConc$Sunshine + ToxinConc$WindSpeed, data = ToxinConc)

Residuals:
      1       2       3       4       5       6       7       8
-1.8818  2.0498 -0.6314  0.4787 -0.5805  1.2508 -0.1921 -0.1813
      9      10
-1.1552  0.8429

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          31.6084     7.1051   4.449  0.00671 **
ToxinConc$Rain        7.0676     1.0031   7.046  0.00089 ***
ToxinConc$NoonTemp   -0.4201     0.2413  -1.741  0.14215
ToxinConc$Sunshine   -0.2375     0.5086  -0.467  0.66018
ToxinConc$WindSpeed  -0.7936     0.2977  -2.666  0.04458 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.574 on 5 degrees of freedom
Multiple R-squared:  0.9186,	Adjusted R-squared:  0.8535
F-statistic: 14.11 on 4 and 5 DF,  p-value: 0.006232
```

CSE 7302c

Multiple Linear Regression

HANDLING SPECIAL SITUATIONS

Nonlinear Models – Polynomial Regression

For example, $y = \beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2 + \epsilon$

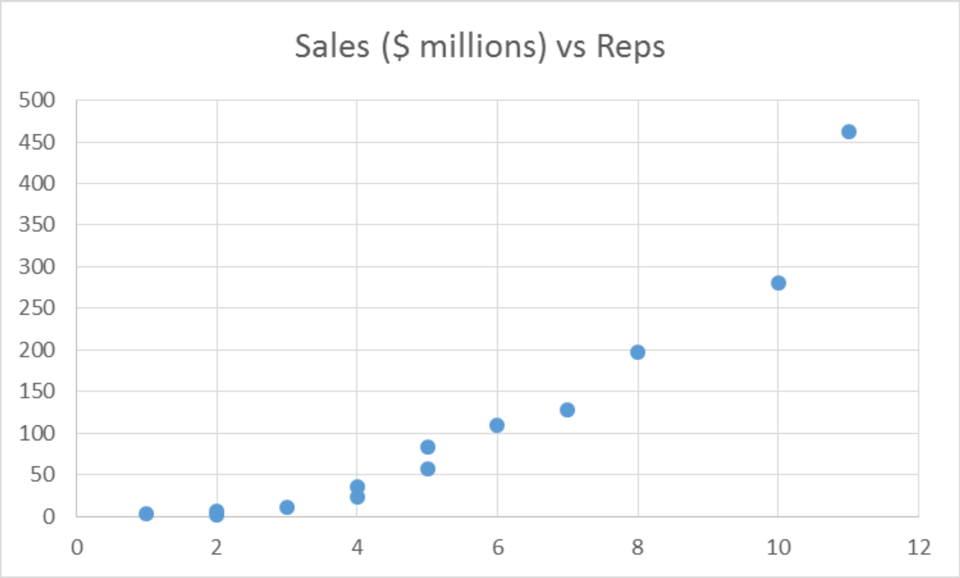
How is this a special case of the general linear model?

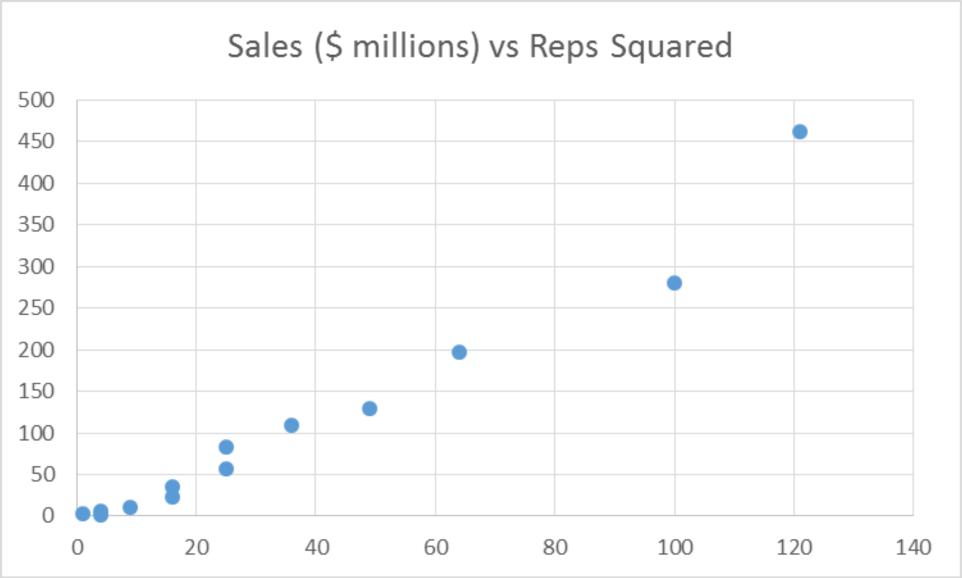
Replace ${x_1}^2$ with x_2, so that $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

Multiple linear regression assumes a linear fit of the regression coefficients and regression constant, but not necessarily a linear relationship of the independent variable values.

Nonlinear Models – Polynomial Regression - Excel

Sales volume versus # of sales reps and # of sales reps squared



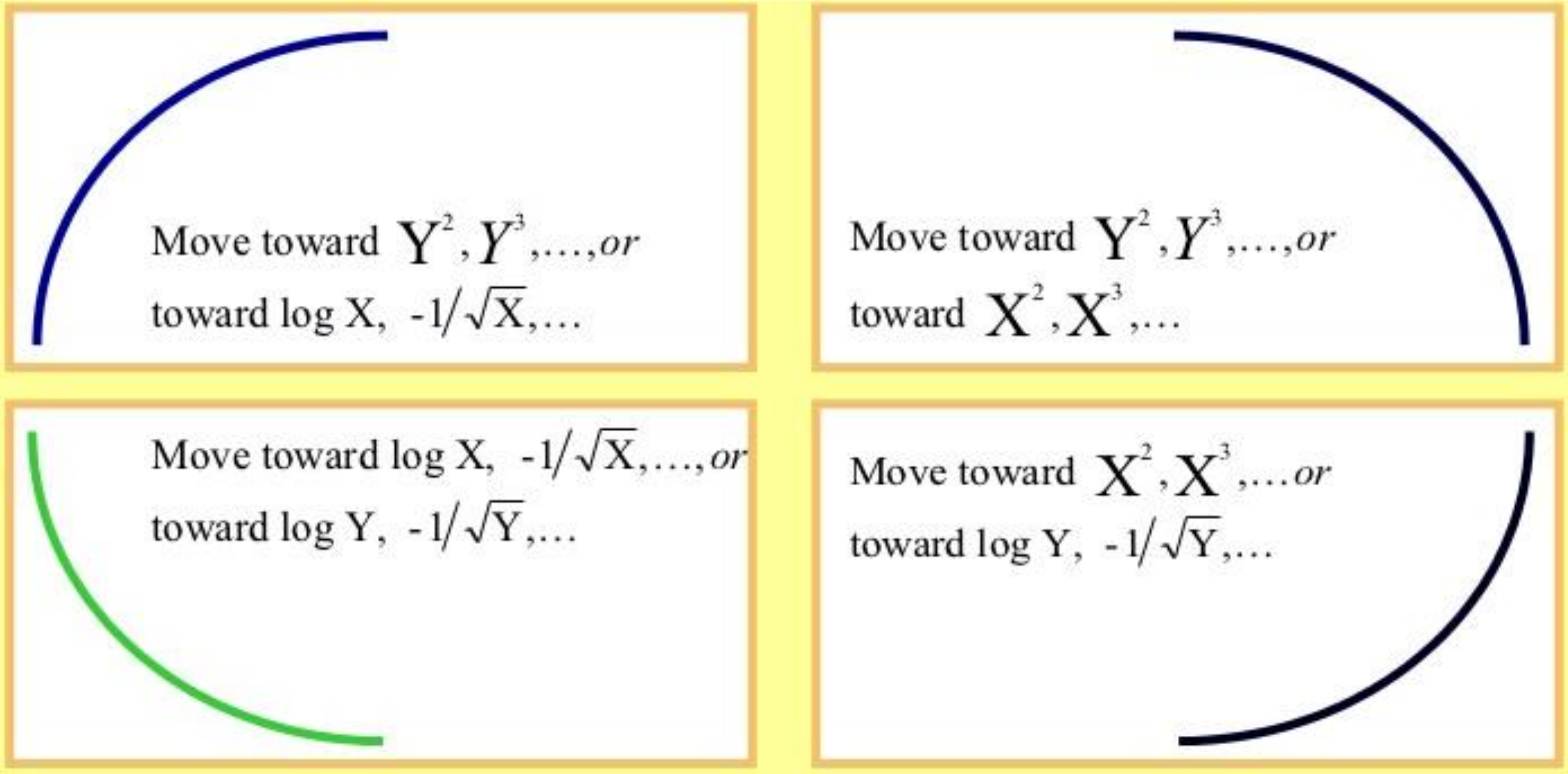


CSE 7302c

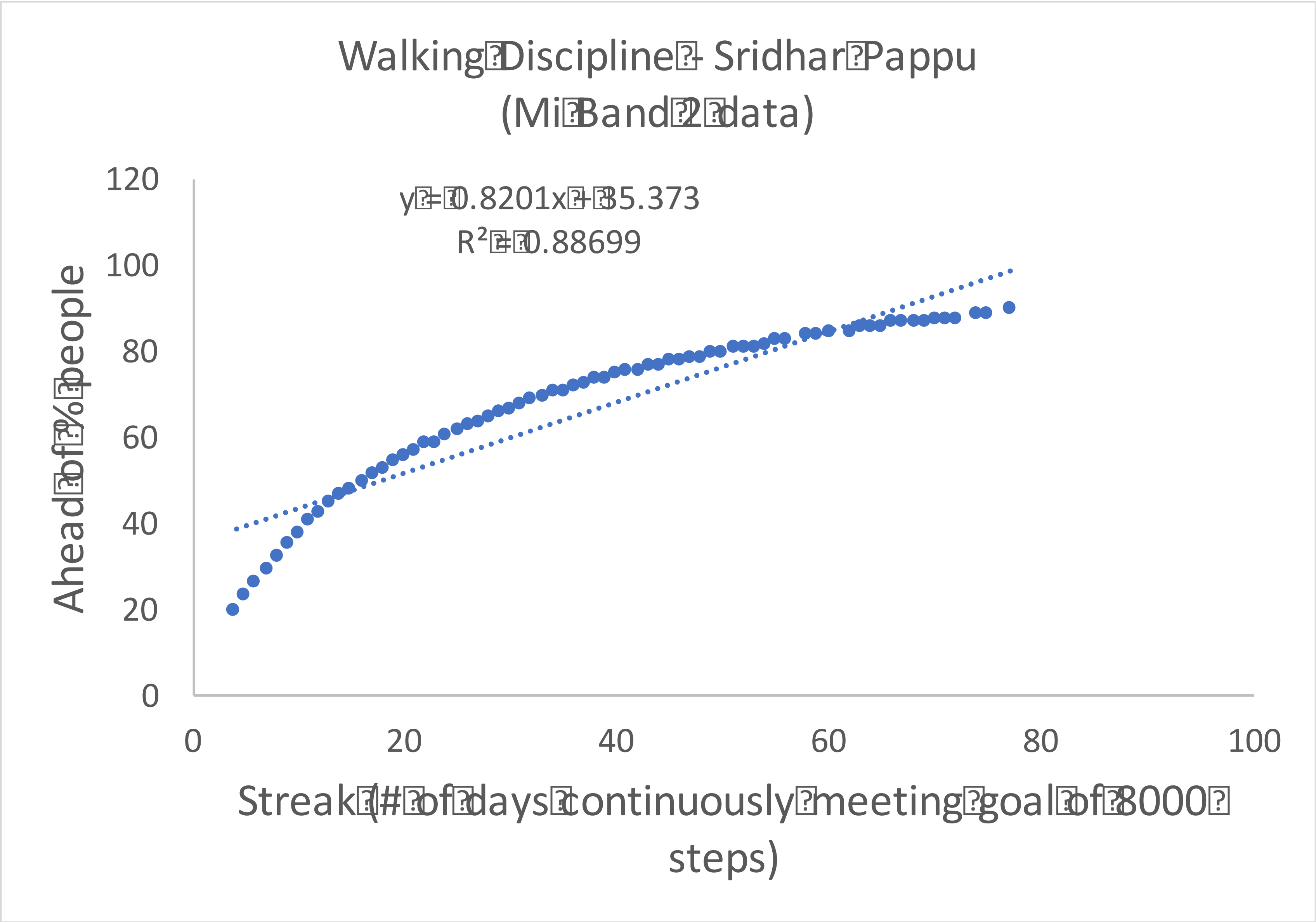
Tukey's Ladder of Transformations

Ladder for x		
Up ladder	Neutral	Down ladder
$\ldots, x^4, x^3, x^2, x$	$\sqrt{x}, x, logx$	$-\frac{1}{\sqrt{x}}, -\frac{1}{x}, -\frac{1}{x^2}, -\frac{1}{x^3}, \ldots$
Ladder for y		
Up ladder	Neutral	Down ladder
$\ldots, y^4, y^3, y^2, y$	$\sqrt{y}, y, logy$	$-\frac{1}{\sqrt{y}}, -\frac{1}{y}, -\frac{1}{y^2}, -\frac{1}{y^3}, \ldots$

Tukey’s Four-Quadrant Approach

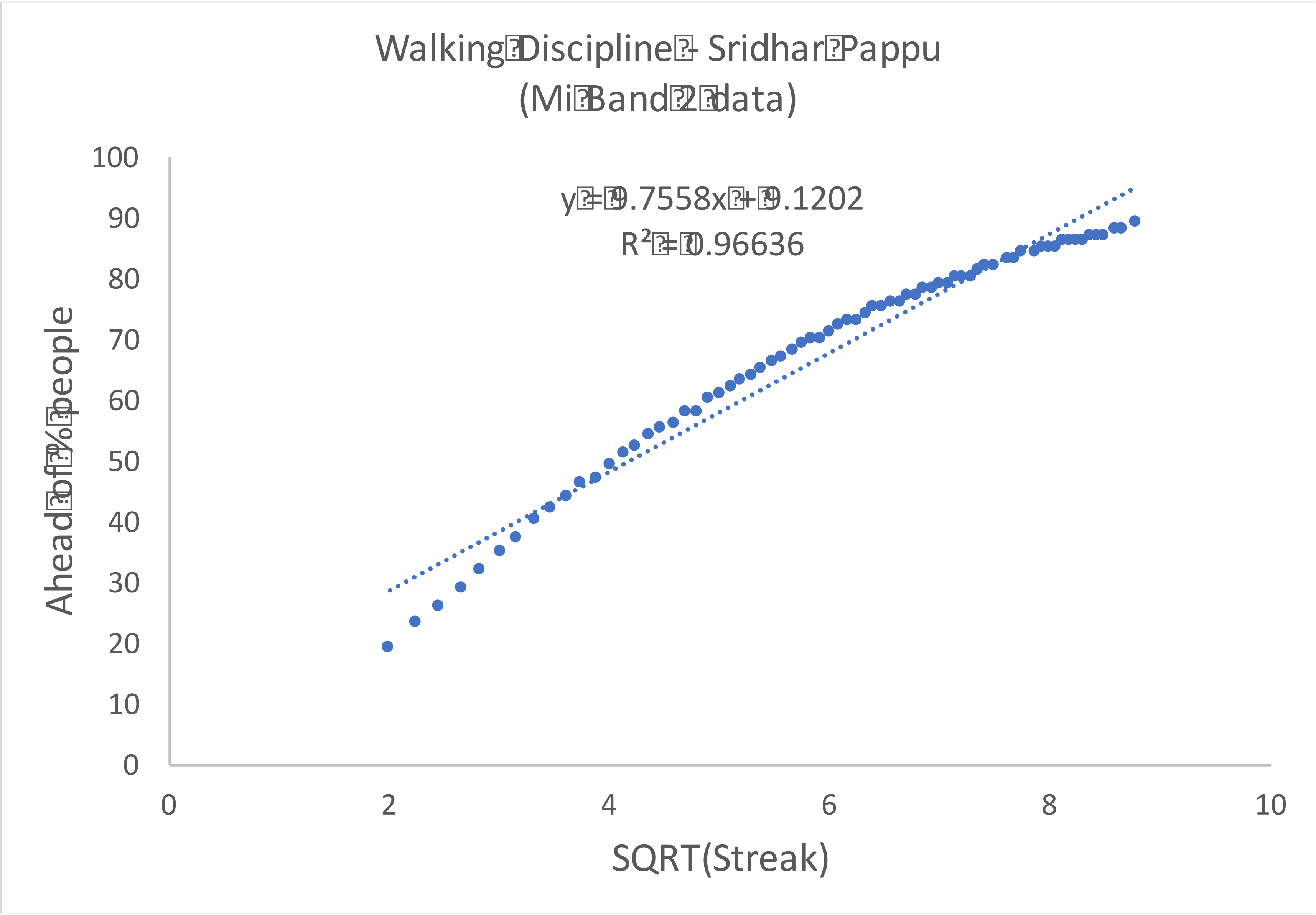


Based on Tukey's 4-Quadrant Approach, what transformation do you recommend?

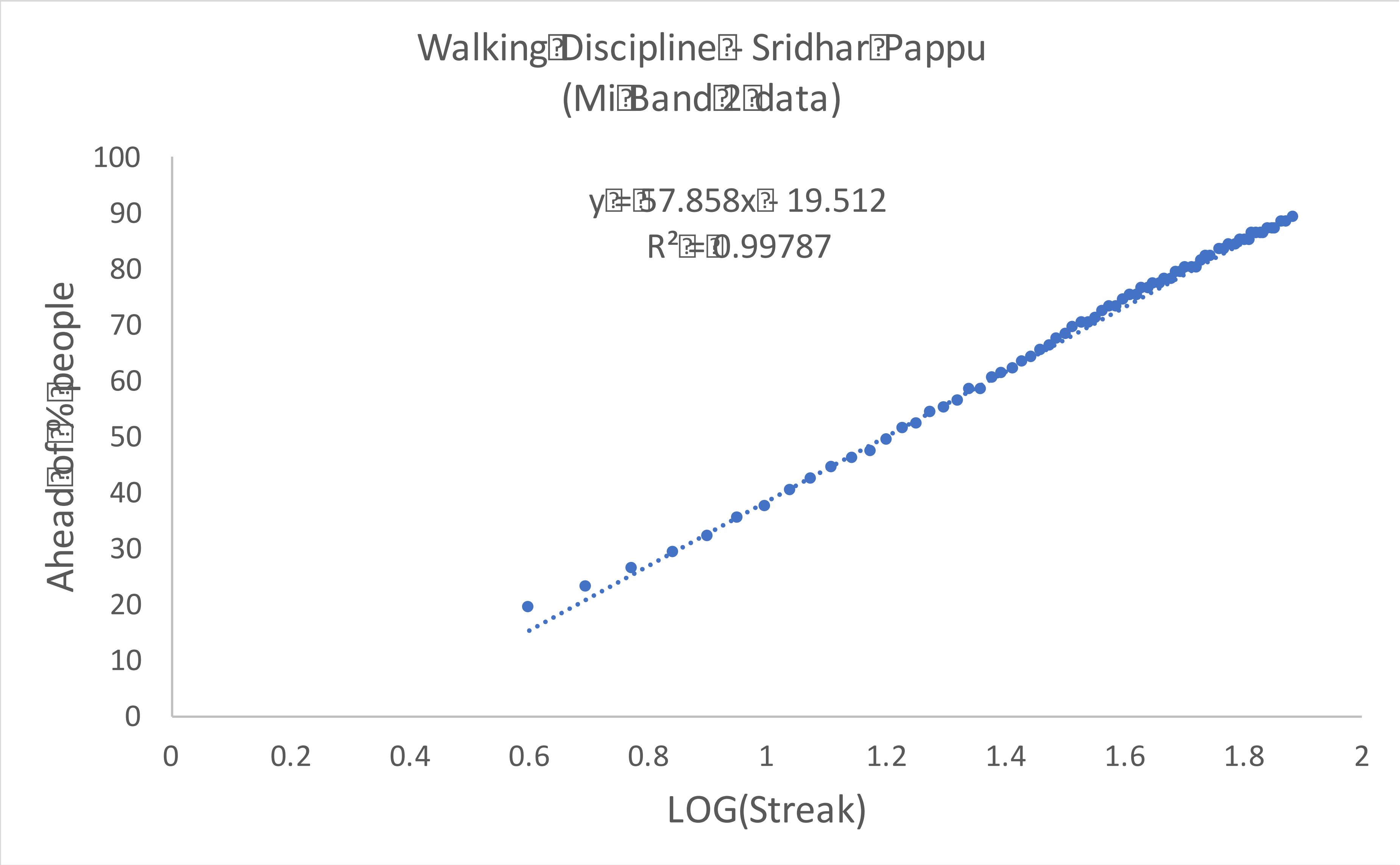


CSE 7302c

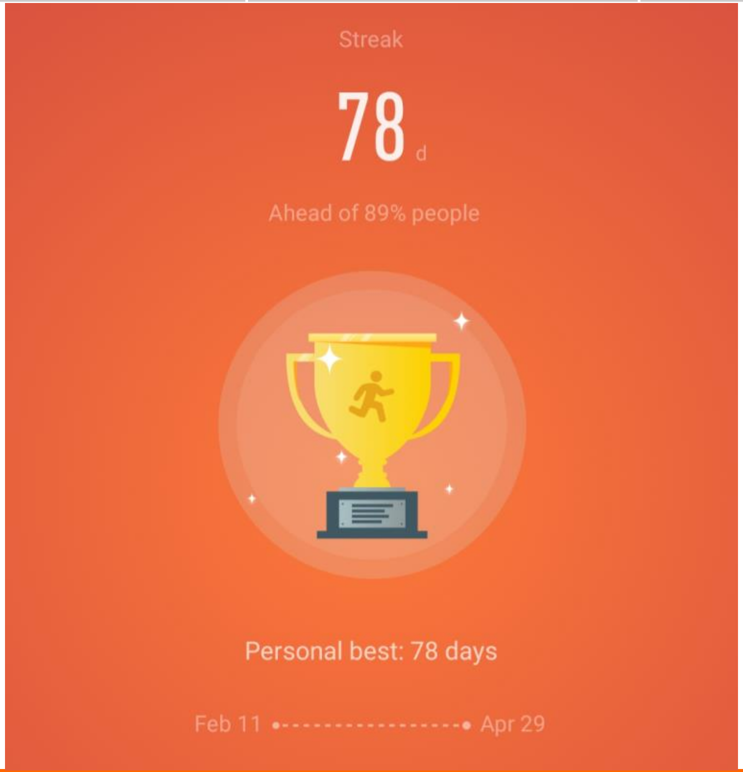
SQRT Transformation on X



LOG Transformation on X



Data	Equation	R-Squared	Ahead of % People (Prediction for Day 78)
Original	0.8201x + 35.373	88.7%	99.34
Square Root on X	9.7558x + 9.1202	96.6%	95.28
Log on X	57.858x – 19.512	99.8%	89.96

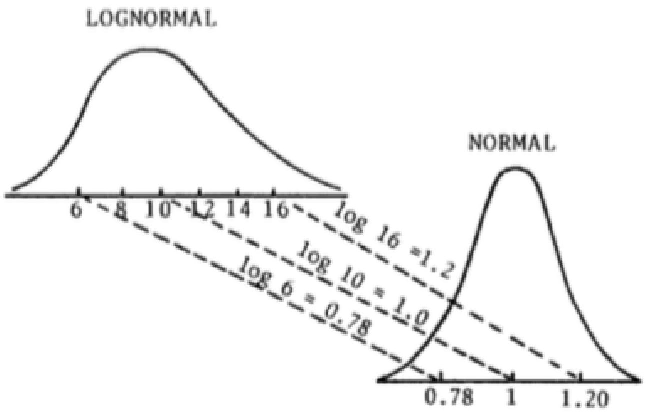


More thoughts on Transformations

DATA TRANSFORMATION

As suggested by Tabachnick and Fidell (2007) and Howell (2007), the following guidelines (including SPSS compute commands) should be used when transforming data.

If your data distribution is…	Use this transformation method.
Moderately positive skewness	Square-Root NEWX = SQRT(X)
Substantially positive skewness	Logarithmic (Log 10) NEWX = LG10(X)
Substantially positive skewness (with zero values)	Logarithmic (Log 10) NEWX = LG10(X + C)
Moderately negative skewness	Square-Root NEWX = SQRT(K – X)
Substantially negative skewness	Logarithmic (Log 10) NEWX = LG10(K – X)

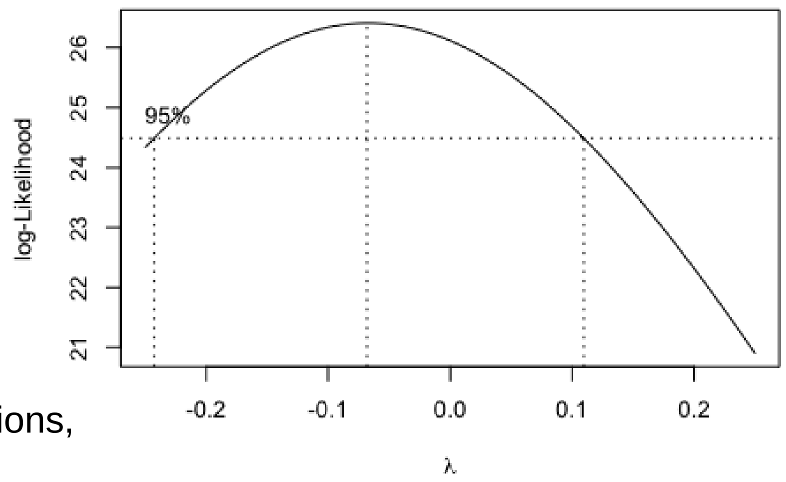


C = a constant added to each score so that the smallest score is 1.
K = a constant from which each score is subtracted so that the smallest score is 1; usually equal to the largest score + 1.

Source: http://oak.ucc.nau.edu/rh232/courses/eps625/handouts/data%20transformation%20handout.pdf
Last accessed: May 12, 2016

More thoughts on Transformations

- Square-root transformation: $X \to \sqrt{X}$
  - Use where variance is proportional to mean $(\sigma^2 \propto \mu)$. Occurs when data consists of counts, such as in urine or blood analyses or microbiological data.
  - If some values are zero or very small, use instead $\sqrt{X} + \sqrt{X+1}$.
  - Poisson variables, where mean = variance, square-root transformation will lead to homoscedasticity.
- Reciprocal transformation: $X \to \frac{1}{X}$
  - Use where standard deviation is proportional to the square of the mean $(\sigma \propto \mu^2)$.
- boxcox() in MASS package of R
- PROC TRANSREG in SAS



Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society*, Series B, *26*, 211-252.

Approach to determine whether to transform X or Y to achieve **linearity**, **homoscedasticity** and **normality**:

1. Often, a transformation that fixes one, fixes all.
2. In general, transforming both is not required, although sometimes it is.
3. A general rule of thumb:
   1. Transform Y first to remove heteroscedasticity and non-normality.
   2. Then transform X to remove non-linearity.

CSE 7302c

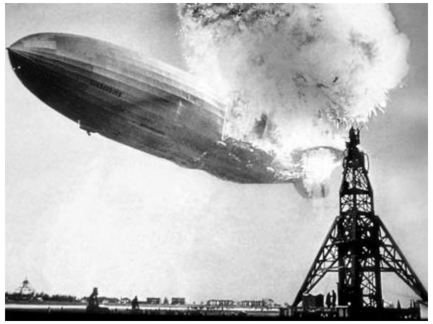
Nonlinear Models – With Interaction

Interaction can be examined as a separate independent variable in regression.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

For example,

- Individually each of two drugs might improve symptoms, but when taken together, they may interact and cause a decline in health.
- Fire increases a balloon's levity (hot air balloon). Hydrogen also increases levity as in the Zeppelins. But fire and hydrogen dramatically reduce the levity.



CSE 7302c

Nonlinear Models – Without Interaction - Excel

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.687213365					
R Square	0.47226221					
Adjusted R Square	0.384305911					
Standard Error	4.570195728					
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	224.2930654	112.1465327	5.369282452	0.021602756	
Residual	12	250.6402679	20.88668899			
Total	14	474.9333333				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	50.85548009	3.790993168	13.41481713	1.38402E-08	42.59561554	59.11534464
Stock 2 ($)	-0.118999968	0.19308237	-0.616317112	0.54919854	-0.539690313	0.301690376
Stock 3 ($)	-0.07076195	0.198984841	-0.35561478	0.728301903	-0.504312675	0.362788775

Model is significant but neither of the variables is.

CSE 7302c

Nonlinear Models – With Interaction - Excel

- One of the earlier insignificant variables along with the interaction term are now significant.
- Model remains significant.
- Adjusted R-sq doubled.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	381.8467141	127.282238	15.04087945	0.00033002
Residual	11	93.08661926	8.462419933		
Total	14	474.9333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

CSE 7302c

Indicator (Dummy) Variables

Categorical variables such as gender, geographic region, occupation, marital status, level of education, economic class, religion, buying/renting a home, etc. can also be used in multiple regression analysis.

If there are *n* levels in a category, *n-1* dummy variables need to be inserted into the regression analysis replacing that category.

Indicator (Dummy) Variables

If a survey question asks about the region of country your office is located in, with North, South, East and West as the options, the **recoding** can be done as follows:

Region
North
East
North
South
West
West
East

North	West	South
1	0	0
0	0	0
1	0	0
0	0	1
0	1	0
0	1	0
0	0	0

Indicator (Dummy) Variables - Excel

Consider the issue of gender discrimination in the salary earnings of workers in some industries. If there is discrimination, how much is one gender earning more than the other?

Indicator (Dummy) Variables - Excel

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.933293402							
R Square	0.871036574							
Adjusted R Square	0.869727301							
Standard Error	0.095635901							
Observations	200							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	12.16964651	6.084823253	665.2824405	2.40412E-88			
Residual	197	1.801806432	0.009146226					
Total	199	13.97145294						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Jpper 95.0%*
Intercept	1.821901302	0.059565421	30.58655988	9.46086E-77	1.704433585	1.939369019	1.704434	1.939369
Age (10 years)	0.083754451	0.018135789	4.618186202	6.97762E-06	0.047989241	0.11951966	0.047989	0.11952
Gender (1=Male, 0=Female)	0.467628629	0.014321506	32.65219766	2.00282E-81	0.439385488	0.49587177	0.439385	0.495872

Separate equation for each gender

Indicator (Dummy) Variables – Interpreting Coefficients and Relationship to ANOVA

ANOVA

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Exp-Fresher	55	119.7279	2.176871	0.096379		
Exp-Low	70	168.6399	2.409142	0.045699		
Exp-Med	75	179.2534	2.390045	0.049032		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.985371	2	0.992685	16.31551	2.78E-07	3.04175303
Within Groups	11.98608	197	0.060843			
Total	13.97145	199				

OLS

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.376964139					
R Square	0.142101962					
Adjusted R Square	0.133392337					
Standard Error	0.246663853					
Observations	200					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	1.985370871	0.992685	16.31551	2.77596E-07	
Residual	197	11.98608207	0.060843			
Total	199	13.97145294				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.176871087	0.033260147	65.44983	1.3E-135	2.111279448	2.242463
Exp-Low	0.232270784	0.044445741	5.22594	4.4E-07	0.14462027	0.319921
Exp-Med	0.213174092	0.043789018	4.868209	2.3E-06	0.126818687	0.299529

- Mean of the reference group in ANOVA is the intercept in OLS.
- Differences between means of groups are the coefficients in OLS.

Indicator (Dummy) Variables – Interpreting Coefficients and Relationship to ANOVA

Choice of reference group is not important; end results remain the same.

What will be the salary of a fresher in the two cases below where *Fresher* is the reference group in the 1st case and *Low experience* is the reference group in the 2nd?

Multiple R	0.376964139					
R Square	0.142101962					
Adjusted R Square	0.133392337					
Standard Error	0.246663853					
Observations	200					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	1.985370871	0.992685	16.31551	2.77596E-07	
Residual	197	11.98608207	0.060843			
Total	199	13.97145294				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.176871087	0.033260147	65.44983	1.3E-135	2.111279448	2.242463
Exp-Low	0.232270784	0.044445741	5.22594	4.4E-07	0.14462027	0.319921
Exp-Med	0.213174092	0.043789018	4.868209	2.3E-06	0.126818687	0.299529

Regression Statistics						
Multiple R	0.376964139					
R Square	0.142101962					
Adjusted R Square	0.133392337					
Standard Error	0.246663853					
Observations	200					
ANOVA						
	df	*SS*	*MS*	*F*	*ignificance F*	
Regression	2	1.985370871	0.992685	16.31551	2.78E-07	
Residual	197	11.98608207	0.060843			
Total	199	13.97145294				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.409141871	0.029481969	81.71577	6.3E-154	2.351001	2.467283
Exp-Fresher	-0.232270784	0.044445741	-5.22594	4.4E-07	-0.31992	-0.14462
Exp-Med	-0.019096692	0.040993015	-0.46585	0.641836	-0.09994	0.061745

p-values here indicate if the level (or group) is significantly different from the reference level (or group).

What might you do if there is no significant difference as is the case between low and medium experience? Also, check the averages in ANOVA output.

A possible action could be to combine Low and Medium groups into a single group

Indicator (Dummy) Variables – Interpreting Coefficients and Relationship to ANOVA

Interpret the coefficients of the numeric and categorical variables below.

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.948085877					
R Square	0.898866831					
Adjusted R Square	0.896792304					
Standard Error	0.08512366					
Observations	200					
ANOVA						
	df	*SS*	*MS*	*F*	*ignificance F*	
Regression	4	12.55848	3.139619	433.2877	8.41E-96	
Residual	195	1.412977	0.007246			
Total	199	13.97145				
	Coefficients	*andard Errc*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.631967642	0.059023	27.64988	2.15E-69	1.515563	1.748372
Age (10 years)	0.122503981	0.016996	7.20789	1.22E-11	0.088985	0.156023
Gender (1=Male, 0=Female)	0.430437318	0.013721	31.37032	3.96E-78	0.403376	0.457498
Exp-Low	0.114744786	0.016665	6.885566	7.7E-11	0.081879	0.147611
Exp-Med	0.100583631	0.016081	6.254777	2.47E-09	0.068868	0.132299

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.948085877					
R Square	0.898866831					
Adjusted R Square	0.896792304					
Standard Error	0.08512366					
Observations	200					
ANOVA						
	df	*SS*	*MS*	*F*	*ignificance F*	
Regression	4	12.55848	3.139619	433.2877	8.41E-96	
Residual	195	1.412977	0.007246			
Total	199	13.97145				
	Coefficients	*andard Errc*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.746712428	0.054233	32.20771	5.32E-80	1.639754	1.85367
Age (10 years)	0.122503981	0.016996	7.20789	1.22E-11	0.088985	0.156023
Gender (1=Male, 0=Female)	0.430437318	0.013721	31.37032	3.96E-78	0.403376	0.457498
Exp-Fresher	-0.114744786	0.016665	-6.88557	7.7E-11	-0.14761	-0.08188
Exp-Med	-0.014161156	0.01419	-0.99797	0.319532	-0.04215	0.013824

- **Numeric:** For unit change in Age (numeric), Salary **increases** by 0.1225 (x 1000 $).
- **Categorical (Dummy):** If a person is a fresher, (s)he makes 0.1147 (x 1000$) **less** than a person with low experience.

Multiple Linear Regression

MODEL BUILDING METHODS

CSE 7302c

Model Building: Search Procedures

Suppose a model to predict the world crude oil production (barrels per day) is to be developed and the predictors used are:

- US energy consumption (BTUs)
- Gross US nuclear electricity generation (kWh)
- US coal production (short-tons)
- Total US dry gas (natural gas) production (cubic feet)
- Fuel rate of US-owned automobiles (miles per gallon)

What does your intuition say about how each of these variables would affect the oil production?

CrudeOilOutput

WorldOil	USEnergy	USAutoFuelRate	USNuclear	USCoal	USDryGas
55.7	74.3	13.4	83.5	598.6	21.7
55.7	72.5	13.6	114	610	20.7
52.8	70.5	14	172.5	654.6	19.2
57.3	74.4	13.8	191.1	684.9	19.1
59.7	76.3	14.1	250.9	697.2	19.2
60.2	78.1	14.3	276.4	670.2	19.1
62.7	78.9	14.6	255.2	781.1	19.7
59.6	76	16	251.1	829.7	19.4
56.1	74	16.5	272.7	823.8	19.2
53.5	70.8	16.9	282.8	838.1	17.8
53.3	70.5	17.1	293.7	782.1	16.1
54.5	74.1	17.4	327.6	895.9	17.5
54	74	17.5	383.7	883.6	16.5
56.2	74.3	17.4	414	890.3	16.1
56.7	76.9	18	455.3	918.8	16.6
58.7	80.2	18.8	527	950.3	17.1
59.9	81.4	19	529.4	980.7	17.3
60.6	81.3	20.3	576.9	1029.1	17.8
60.2	81.1	21.2	612.6	996	17.7
60.2	82.2	21	618.8	997.5	17.8
60.2	83.9	20.6	610.3	945.4	18.1
61	85.6	20.8	640.4	1033.5	18.8
62.3	87.2	21.1	673.4	1033	18.6
64.1	90	21.2	674.7	1063.9	18.8
66.3	90.6	21.5	628.6	1089.9	18.9
67	89.7	21.6	666.8	1109.8	18.9

CSE 7302c

Model Building: Search Procedures

Two considerations in model building:

- Explaining most variation in dependent variable
- Keeping the model simple AND economical

Quite often, the above two considerations are in conflict of each other.

If 3 variables can explain the variation nearly as well as 5 variables, the simpler model is better. Search procedures help choose the more attractive model.

Search Procedures: All Possible Regressions

All variables used in all combinations. For a dataset containing *k* independent variables, 2^k-1 models are examined. In the example of the oil production, 31 models are examined.

Tedious, Time-Consuming, Inefficient, Overwhelming.

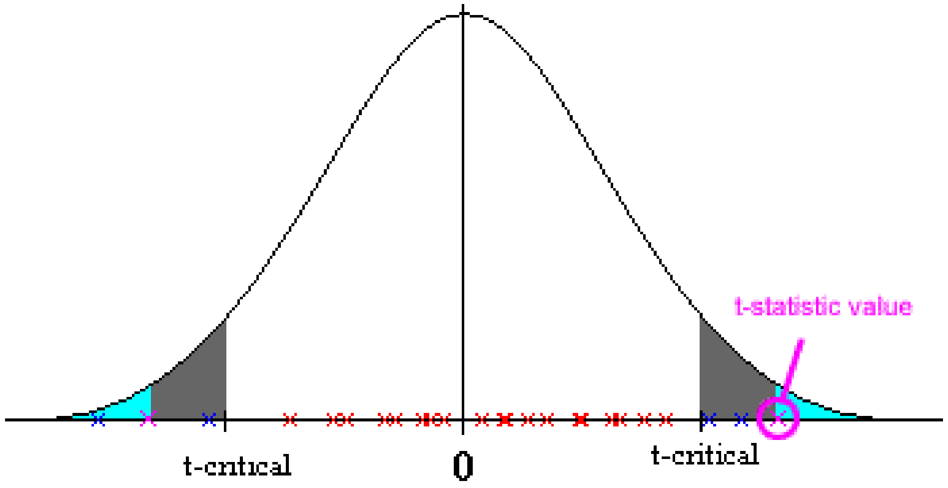
Search Procedures: Stepwise Regression

Starts a model with a single predictor and then adds or deletes predictors one step at a time.

- Step 1
  - Simple regression model for each of the independent variables one at a time.
  - Model with largest absolute value of *t* selected and the corresponding independent variable considered the best single predictor, denoted x_1.
  - If no variable produces a significant *t*, the search stops with no model.

Why LARGEST absolute *t* value and not the SMALLEST?

*Visualize the normal (or **t**) distribution, recall hypothesis testing, think of what the null hypothesis is and then understand what the largest and smallest absolute **t** values mean in terms of the distance from the null value.*



Search Procedures: Stepwise Regression

- Step 2
  - All possible two-predictor regression models with x_1 as one variable.
  - Model with largest absolute *t* value in conjunction with x_1 and one of the other k-1 variables denoted x_2.
  - Occasionally, if x_1 becomes insignificant, it is dropped and search continued with x_2.
  - If no other variables are significant, procedure stops.
- The above process continues with the 3rd variable added to the above 2 selected and so on.

Search Procedures: Stepwise Regression - Excel

Step 1

Dependent Variable	Independent Variable	*t* Ratio	*p*-value	R^2
Oil production	Energy consumption	11.77	1.86e-11	85.2%
Oil production	Nuclear	4.43	0.000176	45.0
Oil production	Coal	3.91	0.000662	38.9
Oil production	Dry gas	1.08	0.292870	4.6
Oil production	Fuel rate	3.54	0.00169	34.2

$$y = 13.075 + 0.580x_1$$

Search Procedures: Stepwise Regression - Excel

Step 2

Dependent Variable, *y*	Independent Variable, x_1	Independent Variable, x_2	*t* Ratio of x_2	*p*-value	R^2
Oil production	Energy consumption	Nuclear	-3.60	0.00152	90.6%
Oil production	Energy consumption	Coal	-2.44	0.0227	88.3
Oil production	Energy consumption	Dry gas	2.23	0.0357	87.9
Oil production	Energy consumption	Fuel rate	-3.75	0.00106	90.8

$$y = 7.14 + 0.772x_1 - 0.517x_2$$

t value for Energy Consumption is now at 11.91 and still significant (2.55e-11).

CSE 7302c

Search Procedures: Stepwise Regression - R

Step 3

Dependent Variable, *y*	Independent Variable, x_1	Independent Variable, x_2	Independent Variable, x_3	*t* Ratio of x_3	*p*-value
Oil production	Energy consumption	Fuel rate	Nuclear	-0.43	0.672
Oil production	Energy consumption	Fuel rate	Coal	1.71	0.102
Oil production	Energy consumption	Fuel rate	Dry gas	-0.46	0.650

No t ratio is significant at $\alpha = 0.05$. No new variables are added to the model.

Search Procedures: Stepwise Regression - R

AIC (Akaike's Information Criterion)

AIC = 2*k* + *n*ln(RSS/*n*) where RSS is Residual Sum of Squares or SSE.

k is the number of parameters including intercept.

Sum of Sq is the additional reduction in SSE due to the addition of a variable or additional increase in SSE due to the removal of a variable.

```
> stepAICOil <- stepAIC(CrudeOilOutputlm, direction = "both")
Start:  AIC=15.29
CrudeOilOutput$WorldOil ~ CrudeOilOutput$USEnergy + CrudeOilOutput$USAutoFuelRate +
    CrudeOilOutput$USNuclear + CrudeOilOutput$USCoal + CrudeOilOutput$USDryGas

                               Df Sum of Sq    RSS    AIC
- CrudeOilOutput$USDryGas       1     0.151 29.661 13.425
- CrudeOilOutput$USNuclear      1     0.651 30.161 13.860
<none>                                      29.510 15.293
- CrudeOilOutput$USAutoFuelRate 1     2.640 32.150 15.521
- CrudeOilOutput$USCoal         1     2.683 32.193 15.555
- CrudeOilOutput$USEnergy       1    31.720 61.231 32.270

Step:  AIC=13.42
CrudeOilOutput$WorldOil ~ CrudeOilOutput$USEnergy + CrudeOilOutput$USAutoFuelRate +
    CrudeOilOutput$USNuclear + CrudeOilOutput$USCoal

                               Df Sum of Sq     RSS    AIC
- CrudeOilOutput$USNuclear      1     0.583  30.243 11.931
<none>                                       29.661 13.425
- CrudeOilOutput$USCoal         1     4.296  33.956 14.941
- CrudeOilOutput$USAutoFuelRate 1     4.575  34.236 15.154
+ CrudeOilOutput$USDryGas       1     0.151  29.510 15.293
- CrudeOilOutput$USEnergy       1   137.158 166.818 56.329

Step:  AIC=11.93
CrudeOilOutput$WorldOil ~ CrudeOilOutput$USEnergy + CrudeOilOutput$USAutoFuelRate +
    CrudeOilOutput$USCoal

                               Df Sum of Sq     RSS    AIC
<none>                                       30.243 11.931
- CrudeOilOutput$USCoal         1     3.997  34.240 13.158
+ CrudeOilOutput$USNuclear      1     0.583  29.661 13.425
+ CrudeOilOutput$USDryGas       1     0.082  30.161 13.860
- CrudeOilOutput$USAutoFuelRate 1    13.531  43.774 19.545
- CrudeOilOutput$USEnergy       1   195.845 226.088 62.234
```

Multiple Linear Regression

HANDLING MULTICOLLINEARITY

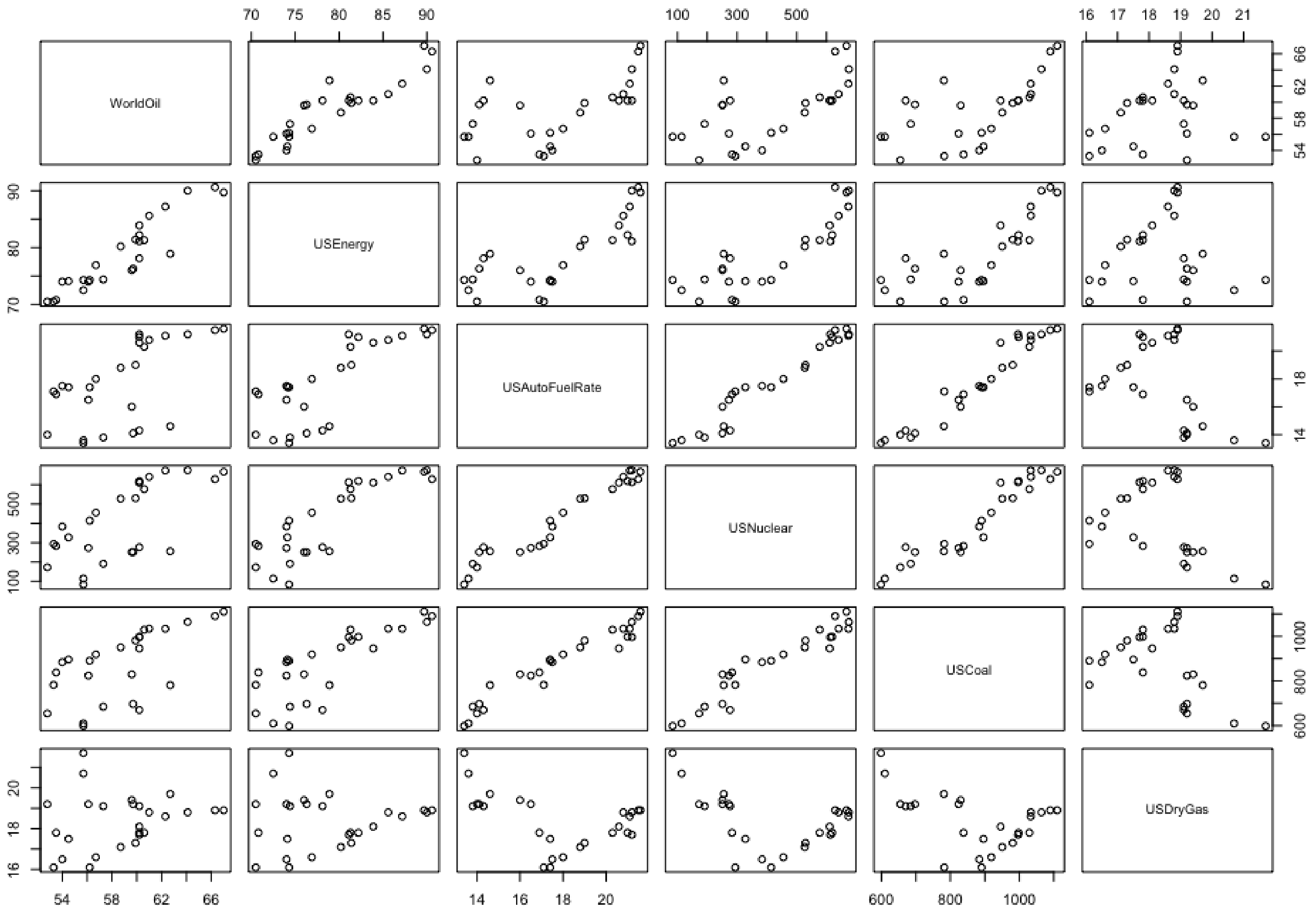
CSE 7302c

Multicollinearity - R

Two or more **independent variables** are highly correlated.

	Energy consumption	Nuclear	Coal	Dry gas	Fuel rate
Energy consumption	1				
Nuclear	0.856	1			
Coal	0.791	0.952	1		
Dry gas	0.057	-0.404	-0.448	1	
Fuel rate	0.791	0.972	0.968	-0.423	1

Multicollinearity - R



Multicollinearity

Sign of estimated regression coefficient when interacting may be opposite of the signs when used as individual predictors.

For example, fuel rate and coal production are highly correlated (0.968).

$$\hat{y} = 44.869 + 0.7838(fuel\ rate)$$

$$\hat{y} = 45.072 + 0.0157(coal)$$

$$\hat{y} = 45.806 + 0.0277(coal) - 0.3934(fuel\ rate)$$

Multicollinearity

Multicollinearity can lead to a model where the model (*F* value) is significant but all individual predictors (*t* values) are insignificant.

(Recall the with- and without-interaction example)

Correlation between stock 2 and stock 3 is 0.96

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.687213365
R Square	0.47226221
Adjusted R Square	0.384305911
Standard Error	4.570195728
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	224.2930654	112.1465327	5.369282452	0.021602756
Residual	12	250.6402679	20.88668899		
Total	14	474.9333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	50.85548009	3.790993168	13.41481713	1.38402E-08	42.59561554	59.11534464
Stock 2 ($)	-0.118999968	0.19308237	-0.616317112	0.54919854	-0.539690313	0.301690376
Stock 3 ($)	-0.07076195	0.198984841	-0.35561478	0.728301903	-0.504312675	0.362788775

Multicollinearity

- Variance Inflation Factor (VIF): A regression analysis is conducted to predict an independent variable by the other independent variables. The independent variable being predicted becomes the dependent variable in this analysis.

$$VIF = \frac{1}{1 - {R_i}^2}$$

- VIF quantifies how much the variance of an estimated coefficient gets inflated in the presence of correlated predictors, compared to the baseline variance when only that one variable is present.

Recall the $Standard\ Error\ of\ the\ Slope = \frac{SE}{\sqrt{SS_{xx}}}$ where $SS_{xx} = \sum(x - \bar{x})^2$ and hence the baseline **variance** of the slope (coefficient) is $\frac{\sigma^2}{\sum(x-\bar{x})^2}$

Multicollinearity - VIF

- VIF > 4 (R_i^2>0.75), 5 (R_i^2>0.80) and 10 (R_i^2>0.90) are commonly used as rules of thumb to indicate severe multicollinearity.
- In practical situations, sometimes even 1.5 is considered as large VIF.
- Remove such variables, rebuild models and compare with earlier model. Make decision based on whether **accuracy of prediction** is more important to the business or **interpretation of the model and the coefficients**.
- Let us look at 2 cases to understand why blindly using the rules of thumb for VIF may be impractical. Stepwise regression prevents multicollinearity to a great extent.

Case 1: Motor Trend Car Road Tests – mtcars dataset in R

Data was extracted from the *Motor Trend* US magazine with a goal to predicting the fuel consumption (mpg) using 10 variables dealing with automobile design and performance.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1

mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

Case 1: mtcars – Model Building

```
Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4506 -1.6044 -0.1196  1.2193  4.6271 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 12.30337   18.71788   0.657   0.5181  
cyl         -0.11144    1.04502  -0.107   0.9161  
disp         0.01334    0.01786   0.747   0.4635  
hp          -0.02148    0.02177  -0.987   0.3350  
drat         0.78711    1.63537   0.481   0.6353  
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739  
vs           0.31776    2.10451   0.151   0.8814  
am           2.52023    2.05665   1.225   0.2340  
gear         0.65541    1.49326   0.439   0.6652  
carb        -0.19942    0.82875  -0.241   0.8122  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,	Adjusted R-squared:  0.8066 
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

- Very good Adjusted R^2
- No significant variable at 5% significance level
- Model is significant
- Indicates multicollinearity

```
> vif(mtcarslm)
      cyl      disp        hp      drat        wt      qsec 
15.373833 21.620241  9.832037  3.374620 15.164887  7.527958 
       vs        am      gear      carb 
 4.965873  4.648487  5.357452  7.908747 
```

- Rules of thumb indicate almost everything is highly collinear
- Let’s run StepAIC

Case 1: mtcars – Model Building

```
> mtcarsStepAIC <- stepAIC(mtcarslm)
Start:  AIC=70.9
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq    RSS    AIC
- cyl   1    0.0799 147.57 68.915
- vs    1    0.1601 147.66 68.932
- carb  1    0.4067 147.90 68.986
- gear  1    1.3531 148.85 69.190
- drat  1    1.6270 149.12 69.249
- disp  1    3.9167 151.41 69.736
- hp    1    6.8399 154.33 70.348
- qsec  1    8.8641 156.36 70.765
<none>              147.49 70.898
- am    1   10.5467 158.04 71.108
- wt    1   27.0144 174.51 74.280

Step:  AIC=68.92
mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq    RSS    AIC
- vs    1    0.2685 147.84 66.973
- carb  1    0.5201 148.09 67.028
- gear  1    1.8211 149.40 67.308
- drat  1    1.9826 149.56 67.342
- disp  1    3.9009 151.47 67.750
- hp    1    7.3632 154.94 68.473
<none>              147.57 68.915
- qsec  1   10.0933 157.67 69.032
- am    1   11.8359 159.41 69.384
- wt    1   27.0280 174.60 72.297

Step:  AIC=66.97
mpg ~ disp + hp + drat + wt + qsec + am + gear + carb

       Df Sum of Sq    RSS    AIC
- carb  1    0.6855 148.53 65.121
- gear  1    2.1437 149.99 65.434
```

```
> mtcarsStepAIC

Call:
lm(formula = mpg ~ wt + qsec + am, data = mtcars)

Coefficients:
(Intercept)           wt         qsec           am
      9.618       -3.917        1.226        2.936
```

- StepAIC identified 3 variables as significant
- Let us build the model with these 3

Case 1: mtcars – Model Building

```
Call:
lm(formula = mpg ~ am + qsec + wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4811 -1.5555 -0.7257  1.4110  4.6610 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.6178     6.9596   1.382 0.177915    
am            2.9358     1.4109   2.081 0.046716 *  
qsec          1.2259     0.2887   4.247 0.000216 ***
wt           -3.9165     0.7112  -5.507 6.95e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared:  0.8497,	Adjusted R-squared:  0.8336 
F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```
> vif(mtcarslm2)
      am     qsec       wt 
2.541437 1.364339 2.482952 
```

- Adjusted R^2 improved
- All variables are significant
- Model is significant
- VIF values are around 2.5 or less

Case 2: Predicting Fungal Toxin Contamination

A drug precursor molecule is extracted from a type of nut, which is commonly contaminated by a fungal toxin that is difficult to remove during the purification process. The suspected predictors of the amount of fungus are:

- Rainfall (cm/week)
- Noon temperature (^{o}C)
- Sunshine (h/day)
- Wind speed (km/h)

The fungal toxin concentration is measured in µg/100 g.

FungalToxinContamination

Toxin	Rain	NoonTemp	Sunshine	WindSpeed
18.1	1.3	20.9	6.23	13.3
28.6	2.28	25.4	8.13	10.8
15.9	1.11	28.2	10.21	10.9
19.2	0.74	23.7	6.96	8.2
19.3	1.32	26.5	9.04	9.8
14.8	0.51	23.9	7.84	12.3
21.7	1.56	26.7	6.69	10
16.5	1.32	30	8.3	12.2
23.8	2.05	24.9	9.22	10.7
19	1.37	22	8.37	15

Case 2: Model Building

```
Call:
lm(formula = ToxinConc$Toxin ~ ToxinConc$Rain + ToxinConc$NoonTemp +
    ToxinConc$Sunshine + ToxinConc$WindSpeed, data = ToxinConc)

Residuals:
      1       2       3       4       5       6       7       8
-1.8818  2.0498 -0.6314  0.4787 -0.5805  1.2508 -0.1921 -0.1813
      9      10
-1.1552  0.8429

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          31.6084     7.1051   4.449  0.00671 **
ToxinConc$Rain        7.0676     1.0031   7.046  0.00089 ***
ToxinConc$NoonTemp   -0.4201     0.2413  -1.741  0.14215
ToxinConc$Sunshine   -0.2375     0.5086  -0.467  0.66018
ToxinConc$WindSpeed  -0.7936     0.2977  -2.666  0.04458 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.574 on 5 degrees of freedom
Multiple R-squared:  0.9186,    Adjusted R-squared:  0.8535
F-statistic: 14.11 on 4 and 5 DF,  p-value: 0.006232
```

Multiple regression tends to remove correlated pairs of IVs, as in the case of Noon Temperature and Sunshine here.

Case 2: Model Building – R

```
> vif(ToxinConclm)
    ToxinConc$Rain  ToxinConc$NoonTemp  ToxinConc$Sunshine ToxinConc$WindSpeed 
          1.031045            1.616535            1.415269            1.209717 
```

```
> correlation
                Toxin         Rain    NoonTemp    Sunshine    WindSpeed
Toxin      1.00000000  0.868734134 -0.07319548 -0.05169949 -0.270555628
Rain       0.86873413  1.000000000  0.11691043  0.16841144 -0.002180167
NoonTemp  -0.07319548  0.116910426  1.00000000  0.50082303 -0.368972511
Sunshine  -0.05169949  0.168411437  0.50082303  1.00000000 -0.018439486
WindSpeed -0.27055563 -0.002180167 -0.36897251 -0.01843949  1.000000000
```

There doesn't appear to be any strongly correlated variables either using correlation values or the VIF, although in some situations, a VIF of 1.5 is considered high.

It may be worthwhile to build another model keeping one of the correlated variables in the model. The more significant can be preferred but business intuition may be cautiously used to include other statistically insignificant variable(s).

Let us do StepAIC first.

Case 2: Model Building – R

```
> ToxinConclm1 <- stepAIC(ToxinConclm, direction = "both")
Start:  AIC=12.14
ToxinConc$Toxin ~ ToxinConc$Rain + ToxinConc$NoonTemp + ToxinConc$Sunshine +
    ToxinConc$WindSpeed

                      Df Sum of Sq     RSS    AIC
- ToxinConc$Sunshine   1     0.540  12.927 10.567
<none>                              12.387 12.141
- ToxinConc$NoonTemp   1     7.510  19.897 14.880
- ToxinConc$WindSpeed  1    17.603  29.990 18.983
- ToxinConc$Rain       1   122.991 135.378 34.055

Step:  AIC=10.57
ToxinConc$Toxin ~ ToxinConc$Rain + ToxinConc$NoonTemp + ToxinConc$WindSpeed

                      Df Sum of Sq     RSS    AIC
<none>                              12.927 10.567
+ ToxinConc$Sunshine   1     0.540  12.387 12.141
- ToxinConc$NoonTemp   1    13.417  26.344 15.686
- ToxinConc$WindSpeed  1    19.688  32.615 17.822
- ToxinConc$Rain       1   122.830 135.757 32.083
```

Case 2: Model Building – R

```
Call:
lm(formula = ToxinConc$Toxin ~ ToxinConc$Rain + ToxinConc$NoonTemp +
    ToxinConc$WindSpeed, data = ToxinConc)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6394 -0.9308  0.1394  0.6545  2.0909

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          31.5651     6.6253   4.764  0.00311 **
ToxinConc$Rain        7.0108     0.9285   7.551  0.00028 ***
ToxinConc$NoonTemp   -0.4790     0.1919  -2.495  0.04682 *
ToxinConc$WindSpeed  -0.8218     0.2718  -3.023  0.02331 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.468 on 6 degrees of freedom
Multiple R-squared:  0.915,     Adjusted R-squared:  0.8726
F-statistic: 21.54 on 3 and 6 DF,  p-value: 0.001298
```

```
> vif(ToxinConclm1)
     ToxinConc$Rain  ToxinConc$NoonTemp ToxinConc$WindSpeed
           1.015857            1.175947            1.159879
```

Toxin concentrations increase with increasing rainfall and decrease in drier climates characterized by higher temperatures and wind speeds.

The business can take a decision to rent farms in drier climates if the cost benefits of saved nuts versus higher rents are high.

Multicollinearity and Standardization - Excel

1. If **interaction** terms are used in regression, standardizing the variables first reduces collinearity.

2. If **power** terms (polynomial regression) are included, standardization again reduces collinearity.

3. Except in above cases, standardization has no bearing on multicollinearity among main effects.

4. Standardization does not improve model performance or R-squared, etc.

5. If interpreting the magnitude of coefficients in terms of the **weightage of the corresponding variable** is desired, then standardizing is required. The raw coefficients do not carry any such interpretation.

Also read: http://www.listendata.com/2017/04/how-to-standardize-variable-in-regression.html

Last accessed: January 05, 2018

Multiple Linear Regression

RECAP - OUTPUT ANALYSIS

Output Analysis - Recap

What is the total variation and its explainable and unexplainable components?

$$SST = SSR + SSE$$

$$SST = \sum (y_i - \bar{y})^2 \qquad SSR = \sum (\hat{y}_i - \bar{y})^2 \qquad SSE = \sum (y_i - \hat{y}_i)^2$$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	381.8467141	127.282238	15.04087945	0.00033002
Residual	11	93.08661926	8.462419933		
Total	14	474.9333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Output Analysis - Recap

How much of total variation can be explained by variation in independent variables?

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.89666084					
R Square	0.804000661					
Adjusted R Square	0.750546296					
Standard Error	2.90902388					
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	381.8467141	127.282238	15.04087945	0.00033002	
Residual	11	93.08661926	8.462419933			
Total	14	474.9333333				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

$$\frac{SSR}{SST} = \frac{381.85}{474.93}$$

CSE 7302c

Output Analysis - Recap

What is the correlation between actual and expected values?

$\sqrt{R^2}$: Correlation between y and $\hat{y}$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.89666084					
R Square	0.804000661					
Adjusted R Square	0.750546296					
Standard Error	2.90902388					
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	381.8467141	127.282238	15.04087945	0.00033002	
Residual	11	93.08661926	8.462419933			
Total	14	474.9333333				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Output Analysis - Recap

How much of total variation can be explained by variation in independent variables (IVs) that *actually affect* the DV? Don't forget that this does not mean those are not important or that they don't have

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

$$R^2 - (1 - R^2)\frac{k}{n-k-1}$$

$$1 - \frac{MSE}{MST}$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	381.8467141	127.282238	15.04087945	0.00033002
Residual	11	93.08661926	8.462419933		
Total	14	474.9333333	33.923809521		

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

CSE 7302c

Output Analysis - Recap

What is the "average" deviation of the actual values from the expected values?

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.89666084					
R Square	0.804000661					
Adjusted R Square	0.750546296					
Standard Error	2.90902388	$\sqrt{MSE}$				
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	381.8467141	127.282238	15.04087945	0.00033002	
Residual	11	93.08661926	8.462419933			
Total	14	474.9333333				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Output Analysis - Recap

What is the average of the squared errors?

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.89666084					
R Square	0.804000661					
Adjusted R Square	0.750546296					
Standard Error	2.90902388					
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	381.8467141	127.282238	15.04087945	0.00033002	
Residual	11	93.08661926	8.462419933			
Total	14	474.9333333				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

$$MSE = \frac{SSE}{df_{error}}$$

Output Analysis - Recap

Is the model significant?

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.89666084					
R Square	0.804000661					
Adjusted R Square	0.750546296					
Standard Error	2.90902388					
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	381.8467141	127.282238	15.04087945	0.00033002	
Residual	11	93.08661926	8.462419933			
Total	14	474.9333333				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

$$F = \frac{MSR}{MSE}$$

R code for critical F: $qf(0.05,3,11) = 3.5874$

CSE 7302c

Output Analysis – Recap

What do regression coefficients mean?

A coefficient is the slope of the linear relationship between the dependent variable (DV) and the **independent contribution** of the independent variable (IV), i.e., that part of the IV that is independent of (or uncorrelated with) all other IVs.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	381.8467141	127.282238	15.04087945	0.00033002
Residual	11	93.08661926	8.462419933		
Total	14	474.9333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Output Analysis - Recap

How much will the variation be between the estimated coefficient and the corresponding true population

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

$$SE_{b_1} = \frac{SE}{\sqrt{\sum(x_{1_i} - \bar{x}_1)^2}} \sqrt{1 - R^2{}_{(x_1, x_2 x_3)}}$$

R^2 with x_1 as dependent and other Xs as independent

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	381.8467141	127.282238	15.04087945	0.00033002
Residual	11	93.08661926	8.462419933		
Total	14	474.9333333			

		Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	b_0	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	b_1	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	b_2	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	b_3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

CSE 7302c

Output Analysis - Recap

Are the coefficients significant?

$$t = \frac{b_i - \beta_{i_{null}}}{SE_{b_i}} \quad \beta_{i_{null}} = 0$$

R code for critical t: $qt(0.025,11) = 2.201$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	381.8467141	127.282238	15.04087945	0.00033002
Residual	11	93.08661926	8.462419933		
Total	14	474.9333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Output Analysis - Recap

What are the confidence intervals for the coefficients?

$$b_i - t_{\left(\frac{\alpha}{2},\nu\right)} * SE_{b_i} \leq \beta_i \leq b_i + t_{\left(\frac{\alpha}{2},\nu\right)} * SE_{b_i}$$

R code for critical t: $qt(0.025,11) = 2.201$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	381.8467141	127.282238	15.04087945	0.00033002
Residual	11	93.08661926	8.462419933		
Total	14	474.9333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286
Stock 2*Stock 3	-0.009984949	0.002314083	-4.314862356	0.00122514	-0.015078211	-0.00489169

Multiple Linear Regression

CASE - MONEYBALL

Case – Oakland A's 2002 Success (Moneyball)

CSE 7302c

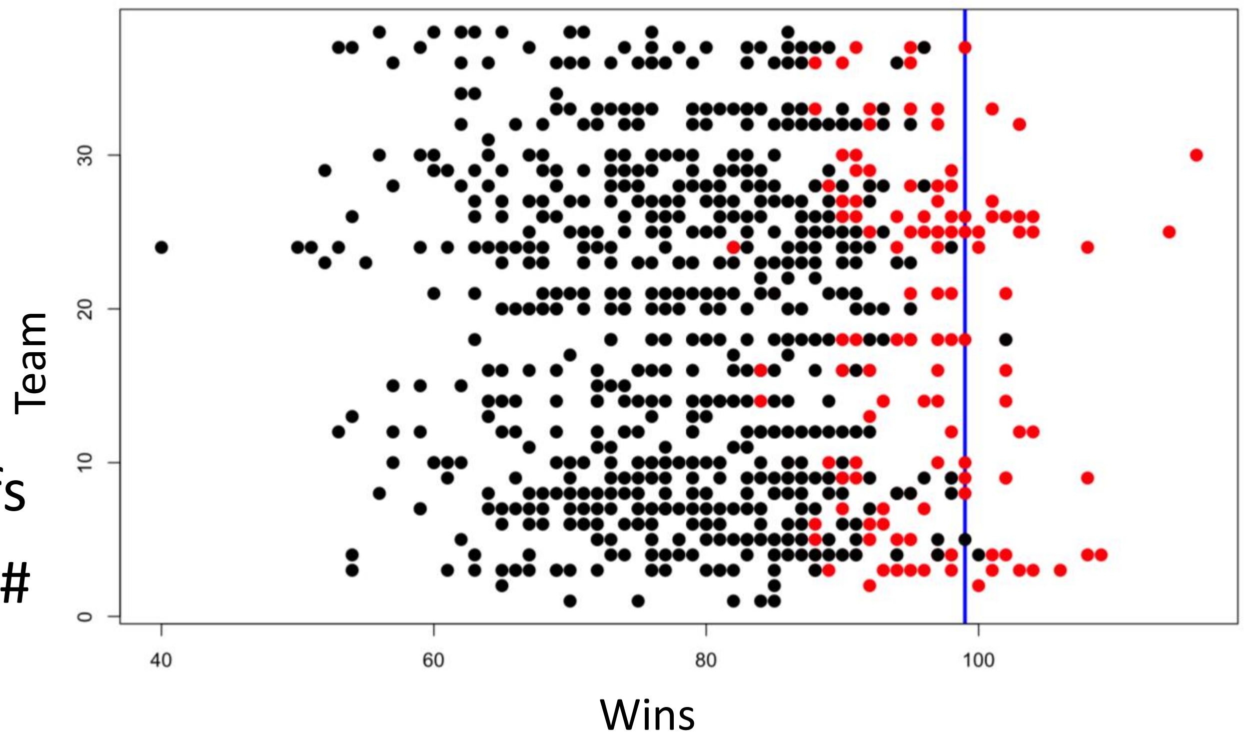
Case Study – Data (baseball-reference.com and MITx)

- 1232 rows, 15 variables
- Statistics for 40 teams from 1962 to 2012
- Oakland A was trying to make playoffs in 2002 and so, 902 rows of data from pre-2002 dates used.

Team	League	Year	RS	RA	W	OBP	SLG	BA	Playoffs	RankSeason	RankPlayoffs	G	OOBP	OSLG
ANA	AL	2001	691	730	75	0.327	0.405	0.261	0			162	0.331	0.412
ARI	NL	2001	818	677	92	0.341	0.442	0.267	1	5	1	162	0.311	0.404
ATL	NL	2001	729	643	88	0.324	0.412	0.26	1	7	3	162	0.314	0.384
BAL	AL	2001	687	829	63	0.319	0.38	0.248	0			162	0.337	0.439
BOS	AL	2001	772	745	82	0.334	0.439	0.266	0			161	0.329	0.393
CHC	NL	2001	777	701	88	0.336	0.43	0.261	0			162	0.321	0.398
CHW	AL	2001	798	795	83	0.334	0.451	0.268	0			162	0.334	0.427
CIN	NL	2001	735	850	66	0.324	0.419	0.262	0			162	0.341	0.455
CLE	AL	2001	897	821	91	0.35	0.458	0.278	1	6	4	162	0.341	0.417
COL	NL	2001	923	906	73	0.354	0.483	0.292	0			162	0.35	0.48
DET	AL	2001	724	876	66	0.32	0.409	0.26	0			162	0.357	0.461

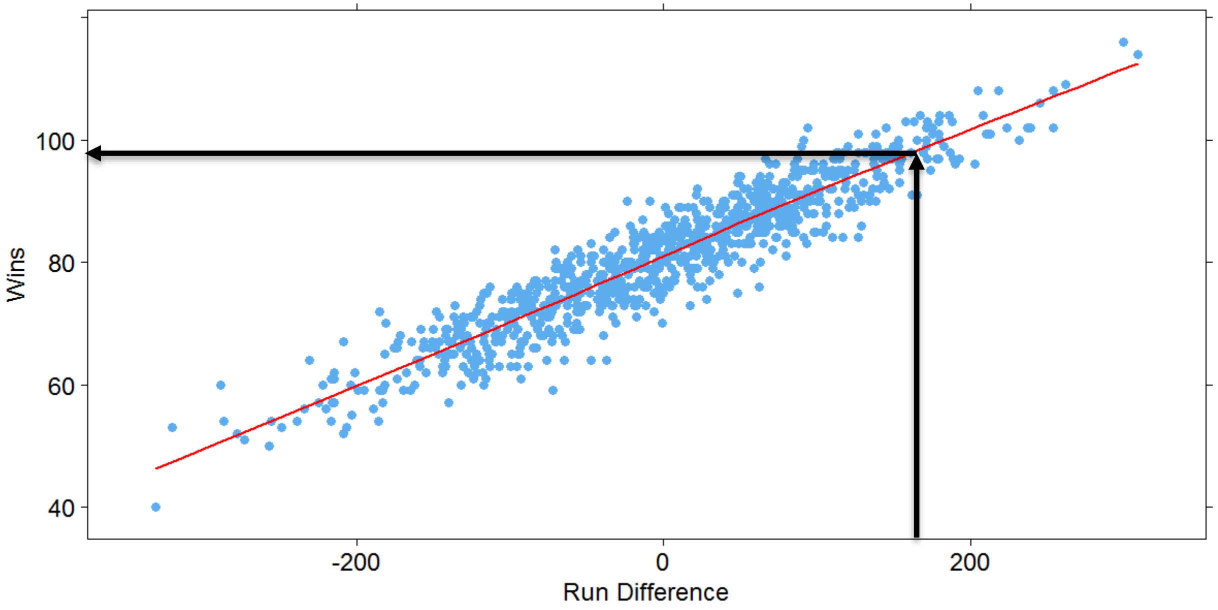
Case Study – Scatter plot

- No. of wins for each team
- Red – Case when team went to playoffs
- Black – Case when team did not go to playoffs
- Vertical blue line – DePodesta's estimate for # of wins required (99)



Case Study – Scatter plot

- DePodesta also estimated that a team on an average needed to score 169 runs more (814-645) per game than their opponent to make the 99 wins
- Strong correlation = 0.94
- Model also predicted 99 wins for a 169-run difference



$$W = 80.881375 + 0.105766 * RD$$
$$W = 80.881375 + 0.105766 * 169 = 98.8$$

CSE 7302c

Case Study – Regression for RS

- Run difference = Runs Scored (RS) – Runs Allowed (RA)
- RS is a function of OBP (On Base Percentage), SLG (Slugging Percentage) and BA (Batting Average)
- Adj. $R^2 = 0.93$

```
Call:
lm(formula = RS ~ OBP + SLG + BA, data = moneyball)

Residuals:
    Min      1Q  Median      3Q     Max 
-70.941 -17.247  -0.621  16.754  90.998 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -788.46      19.70 -40.029  < 2e-16 ***
OBP          2917.42     110.47  26.410  < 2e-16 ***
SLG          1637.93      45.99  35.612  < 2e-16 ***
BA           -368.97     130.58  -2.826  0.00482 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.69 on 898 degrees of freedom
Multiple R-squared:  0.9302, Adjusted R-squared:   0.93 
F-statistic:  3989 on 3 and 898 DF,  p-value: < 2.2e-16
```

- However, coefficient of BA is negative, which is non-intuitive (higher batting average leading to lower chance of winning!). This indicates multi-collinearity.
- Removing BA gives a model with Adj. $R^2 = 0.9294$

```
Call:
lm(formula = RS ~ OBP + SLG, data = moneyball)

Residuals:
    Min      1Q  Median      3Q     Max 
-70.838 -17.174  -1.108  16.770  90.036 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -804.63      18.92  -42.53   <2e-16 ***
OBP          2737.77      90.68   30.19   <2e-16 ***
SLG          1584.91      42.16   37.60   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.79 on 899 degrees of freedom
Multiple R-squared:  0.9296, Adjusted R-squared:  0.9294 
F-statistic:  5934 on 2 and 899 DF,  p-value: < 2.2e-16
```

Case Study – Regression for RA

- RA is a function of OOBP (Opponent On Base Percentage) and OSLG (Opponent Slugging Percentage)
- Missing values removed. 902 values got dropped to 90.
- Adj. R^2 = 0.9052

```
Call:
lm(formula = RA ~ OOBP + OSLG, data = moneyball)

Residuals:
    Min      1Q  Median      3Q     Max 
-82.397 -15.178  -0.129  17.679  60.955 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -837.38      60.26 -13.897  < 2e-16 ***
OOBP         2913.60     291.97   9.979 4.46e-16 ***
OSLG         1514.29     175.43   8.632 2.55e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.67 on 87 degrees of freedom
  (812 observations deleted due to missingness)
Multiple R-squared:  0.9073,	Adjusted R-squared:  0.9052 
F-statistic: 425.8 on 2 and 87 DF,  p-value: < 2.2e-16
```

Case Study – Prediction

- Predict how many runs A's will score and allow in 2002 indicating whether they will make the playoffs or not.
- Inputs to RS and RA models are average team OBP, SLG, OOBP and OSLG values in 2001, assuming team quality remains the same in 2002.
- Values in 2001 (data file has for the entire season including playoffs; the values below are for the regular season as predictions are for that part only)
  - OBP: 0.339
  - SLG: 0.430
  - OOBP: 0.307
  - OSLG: 0.373

Case Study – Prediction

- Equations

$$RS = -804.96 + 2737.77 * OBP + 1584.91 * SLG$$

$$RA = -837.38 + 2913.60 * OOBP + 1514.29 * OSLG$$

$$W = 80.881375 + 0.105766 * RD$$

- Calculations

$$RS = -804.96 + 2737.77 * 0.339 + 1584.91 * 0.430 = 804.66 \sim 805$$

$$RA = -837.38 + 2913.60 * 0.307 + 1514.29 * 0.373 = 621.93 \sim 622$$

$$W = 80.881375 + 0.105766 * 183 = 100.2 \sim 100$$

- Results

Metric	Model Prediction	DePodesta's Estimate	Actual
RS	805	800-820	800
RA	622	650-670	654
Wins	100	93-97	103

Theoretical World vs the Practical World - Advice

Is it true that the majority of business problems can be solved with linear and logistic regression models?

Ryan Barnes, Data Scientist at Mountain America Credit Union (2015-present)

Answered Jun 23 · Upvoted by Edward Williams, M.A. Statistics, University of Wisconsin - Madison (1968) and Martin Lukac, Ph.D. Sociology & Statistics, KU Leuven (2020) · 2 min read

Let me let you in on a secret about the difference between school (data science competitions too) and the real world. In the real world things break. Data shifts because the guy entering it into the system leaves the job and the new guy does it a little bit differently. The world changes around your model, like the NBA 3-point line gets moved back, and so your data distribution on 3 point attempts made shifts. Other things out of your control happen.

In the real world you are constantly balancing between getting the "right answer", getting an answer quickly, and getting a solution that isn't fragile, and that is easy to debug when it does break (because it will break). In my professional life, time and again I thought that a linear model wasn't powerful enough, and started with something more complicated. Then I was forced to come back to a linear model. Why?

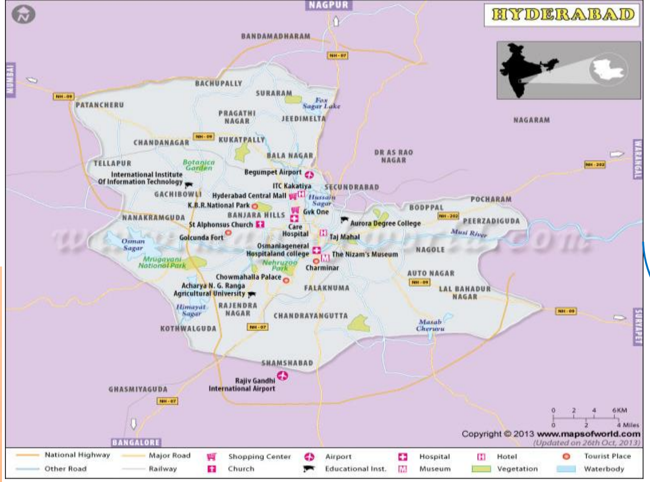
Once I had an optimization problem, I had developed a cool genetic algorithm to solve the problem for setting the optimal cutoff values for a fraud model. It would spin for a day or two up to a week depending on how complex the rule it was trying to optimize, but it would get fantastic results. It worked like a charm every time we needed to set these thresholds. Turns out nobody used it. When asked why, the humans weren't patient enough to wait for the machine to think.

So I threw together a linear regression to set the thresholds. The result wasn't nearly as optimal. But it ran in a couple of seconds. Everyone uses that system. It gets them better results than just using a gut feeling, and it is fast. Are we leaving money on the table? Maybe, depends on your perspective, if no one uses it, we are leaving way more money on the table than by doing a linear regression.

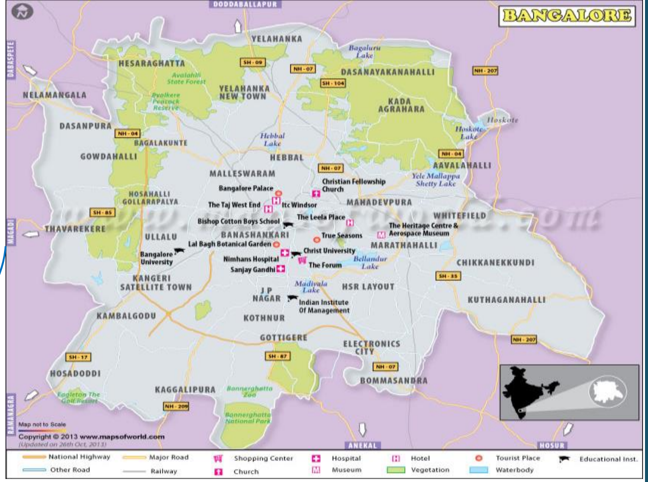
How about if that thing broke. It was nearly impossible to debug, and the results were stochastic to boot. So you never knew if you had the best possible result. With a linear model, I can write a unit test. I can figure out why it gave the answer that it did, and it is just a more solid algorithm that is nearly impossible to break.

So to answer your question, can you solve any business problem with linear regression and probability models? Probably not, I'm looking at PR or HR problems for example, but in terms of data science, they are rock solid models and should be your go to models. Only when they won't work, and you are 100% sure that they aren't working should you move onto anything else.

Source: https://www.quora.com/Is-it-true-that-the-majority-of-business-problems-can-be-solved-with-linear-and-logistic-regression-models
Last accessed: July 12, 2018



Inspire...Educate...Transform.



HYDERABAD
2nd Floor, Jyothi Imperial, Vamsiram Builders, Old Mumbai Highway, Gachibowli, Hyderabad - 500 032
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

BENGALURU
Floors 1-3, L77, 15th Cross Road, 3A Main Road, Sector 6, HSR Layout, Bengaluru – 560 102
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Social Media

Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering