Inspire…Educate…Transform.

Statistics and Probability in Decision Modeling

Linear Regression

Dr. Sridhar Pappu
Executive VP – Academics, INSOFE

January 06, 2018

Multiple Linear Regression

THE OUTPUT

Multiple Linear Regression

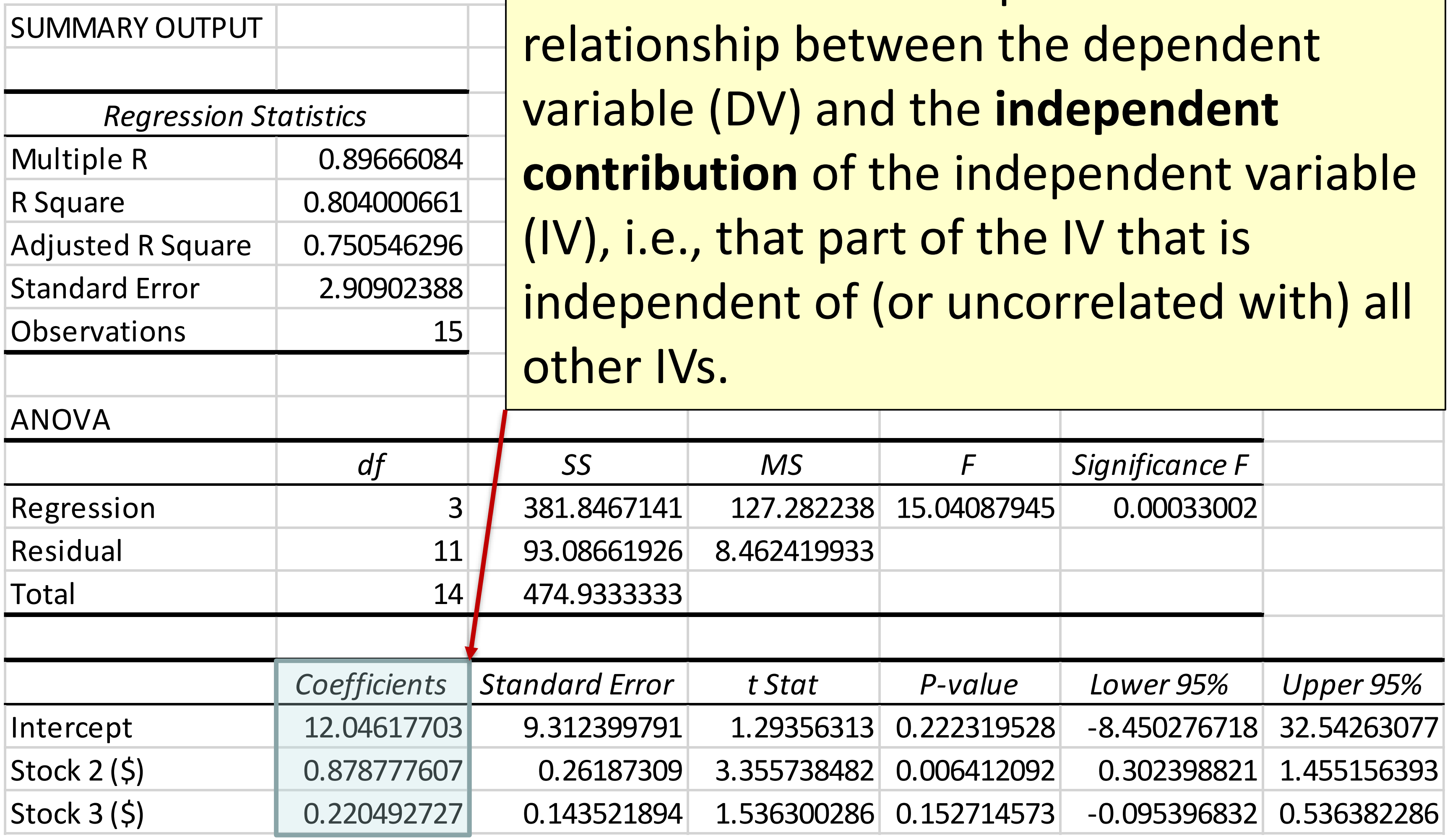
- Simple Linear Regression models the effect of one independent variable, x, on one dependent variable, y
- Multiple Regression models the effect of several independent variables, x_1, x_2 etc., on one dependent variable, y
- The different x variables are combined in a linear way and each has its own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

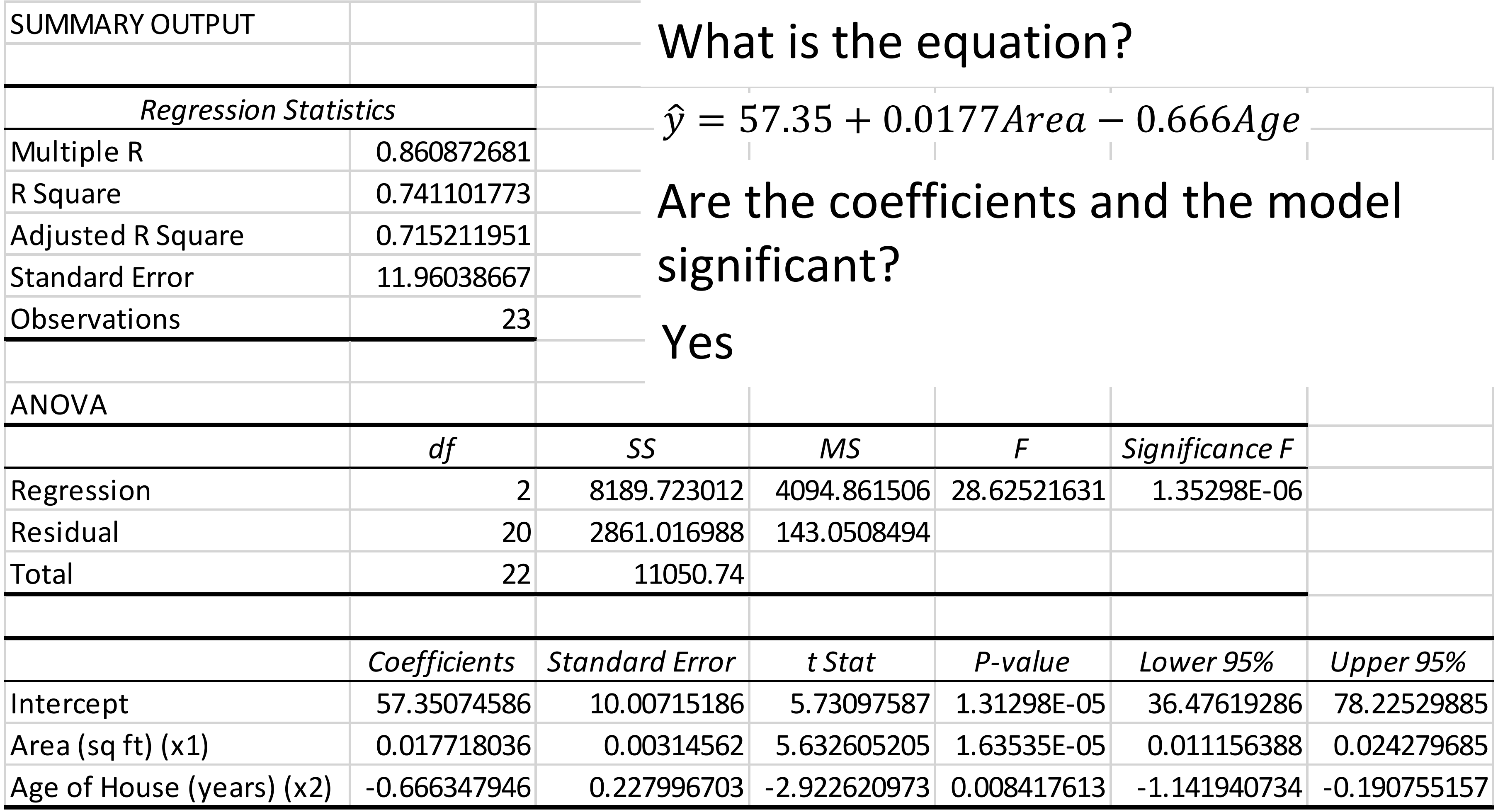
Determining the Multiple Regression Equation - Excel

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



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

What is the equation?

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

Are the coefficients and the model significant?

Yes

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, $y - \hat{y}$.

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.

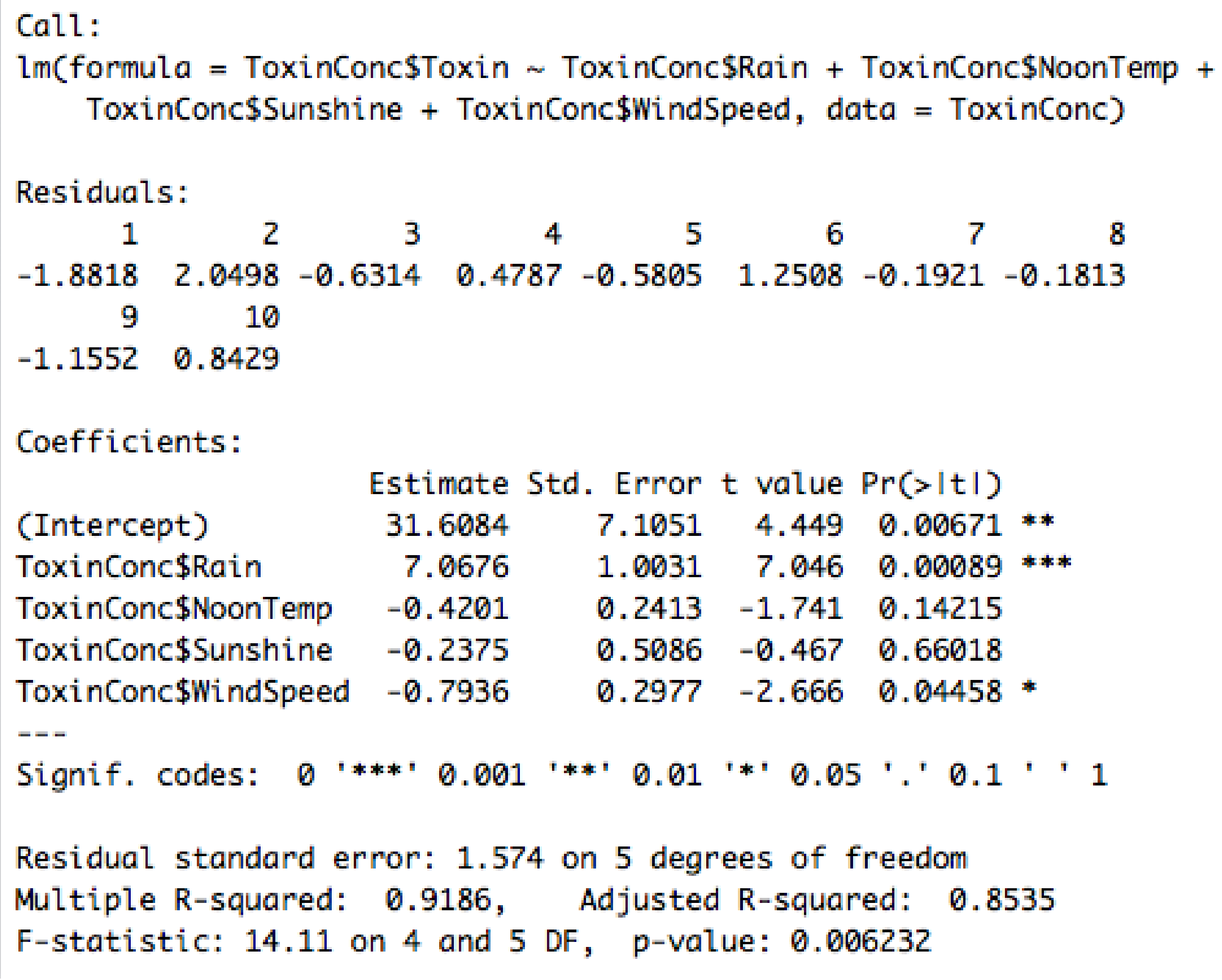
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

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
```

Multiple Linear Regression

HANDLING SPECIAL SITUATIONS

Nonlinear Models – Polynomial Regression

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

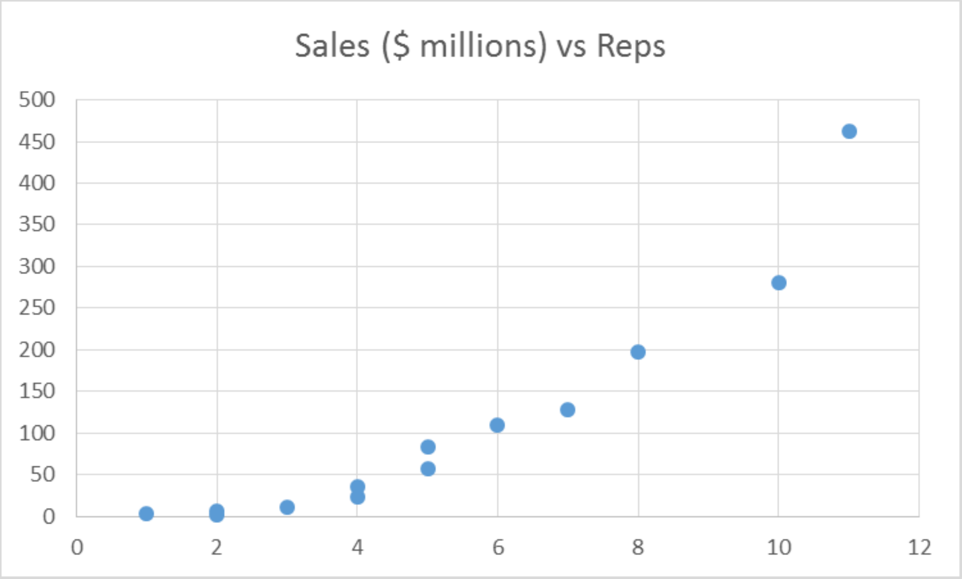
How is this a special case of the general linear model?

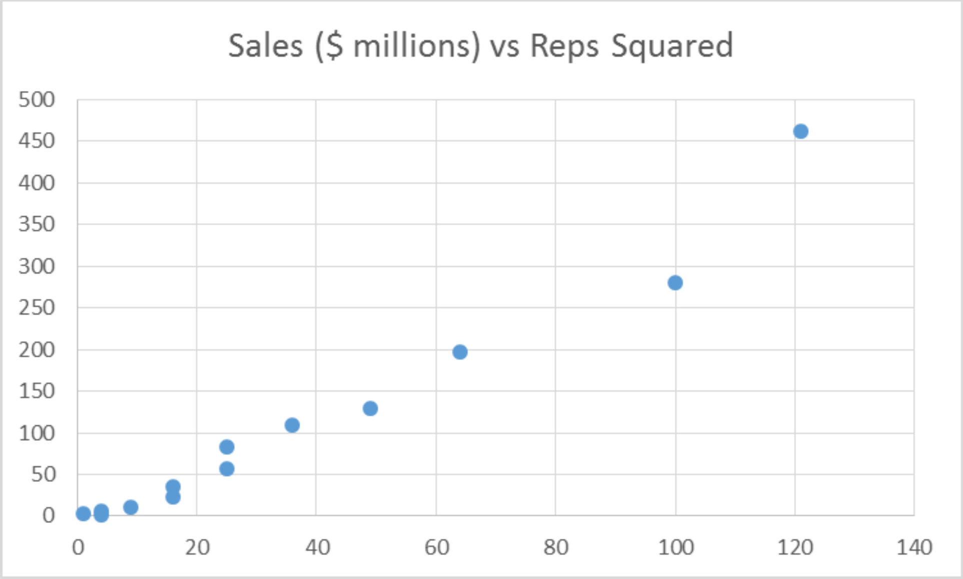
Replace ${x_1}^2$ with x_2, so that $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

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

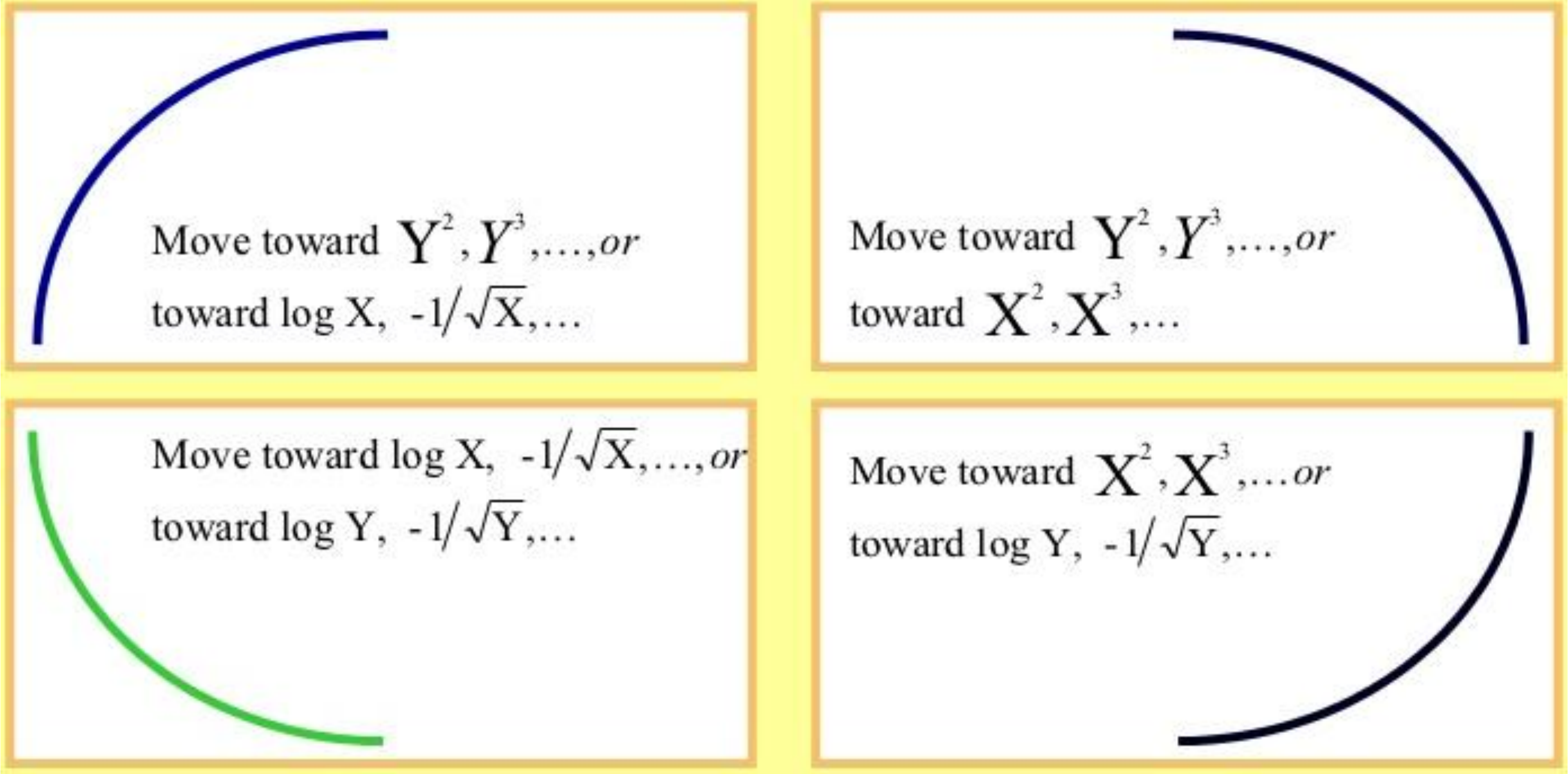




Tukey’s Ladder of Transformations

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



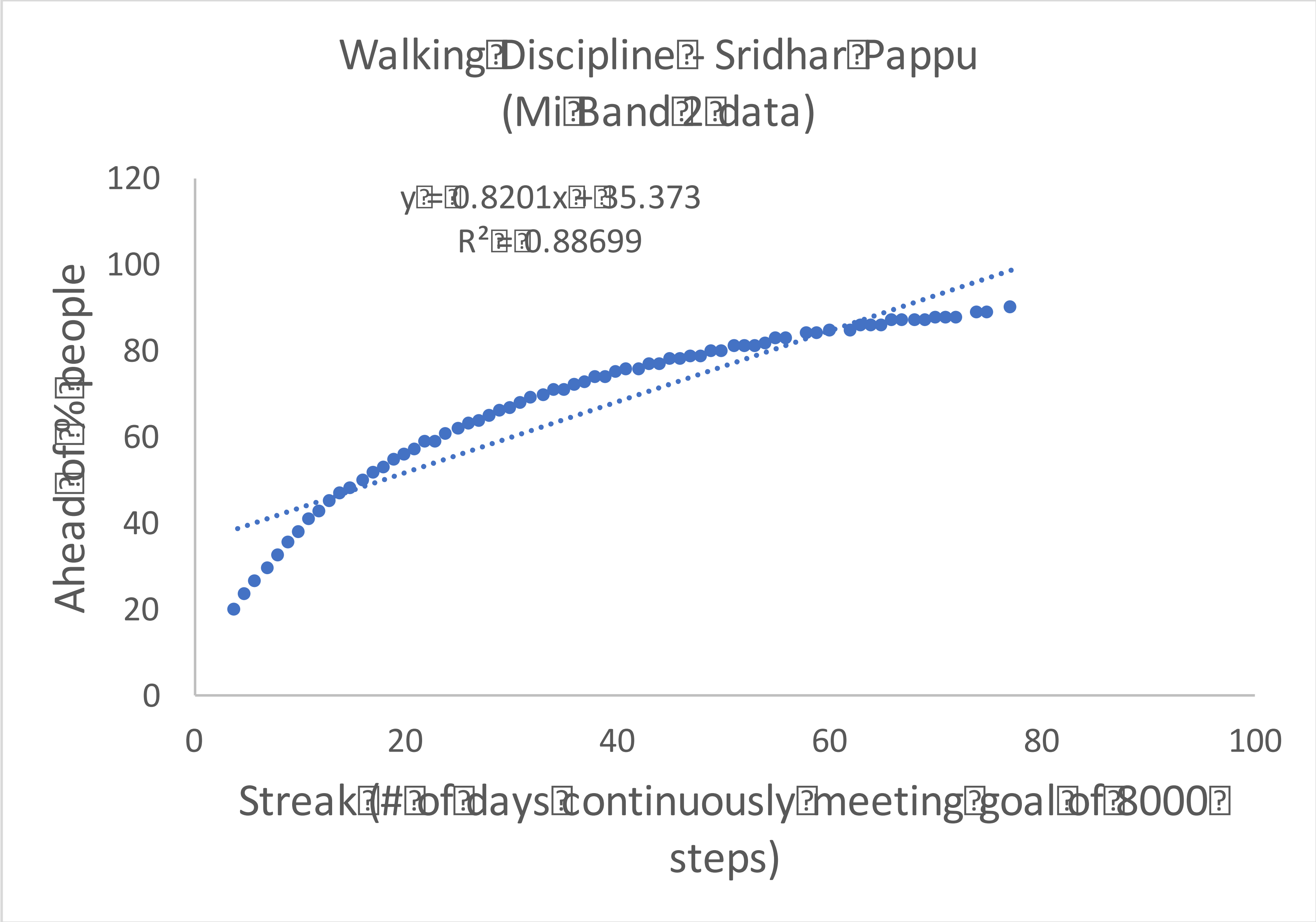
Move toward $Y^2, Y^3, \ldots, or$ toward log X, $-1/\sqrt{X}, \ldots$

Move toward $Y^2, Y^3, \ldots, or$ toward $X^2, X^3, \ldots$

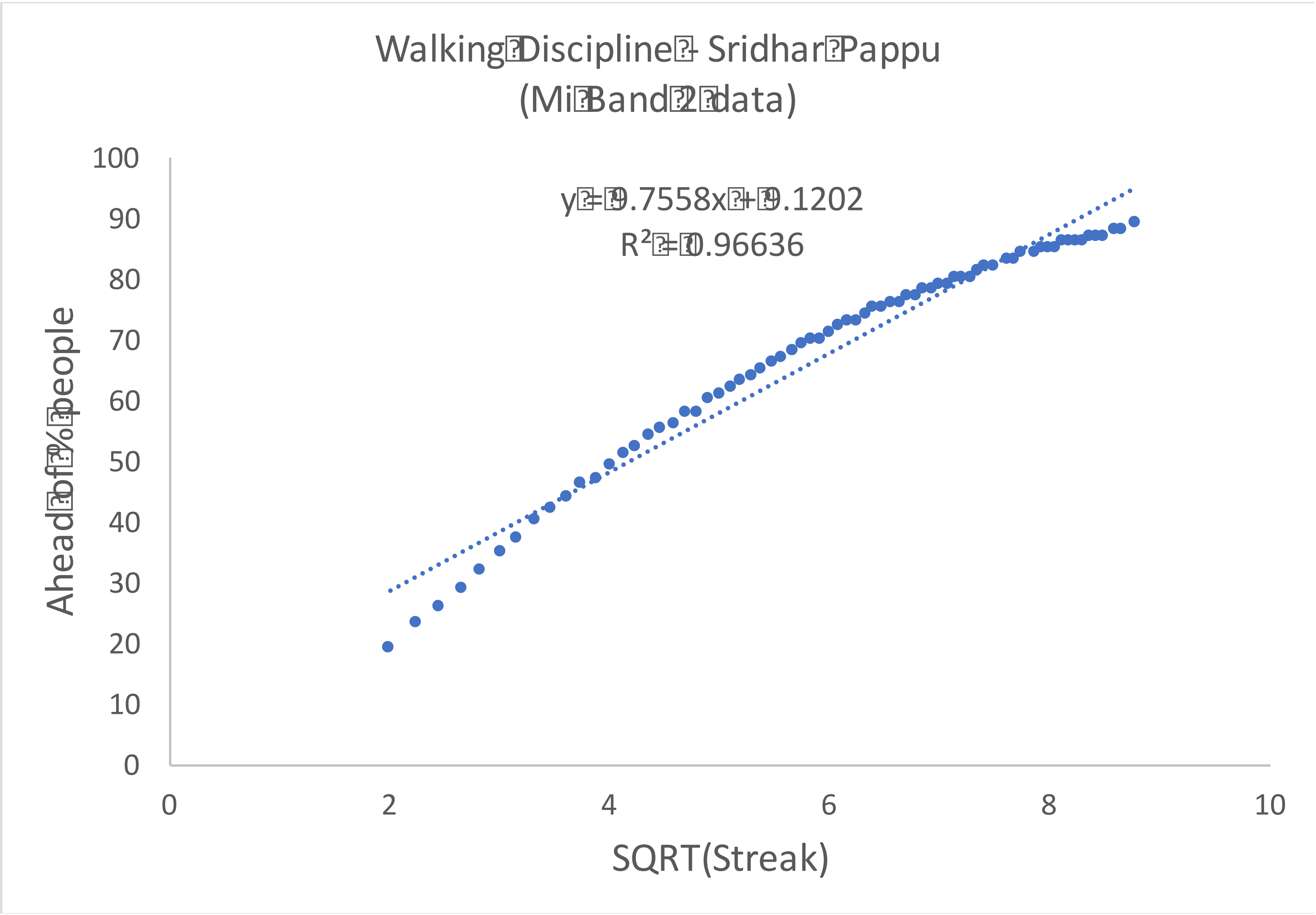
Move toward log X, $-1/\sqrt{X}, \ldots, or$ toward log Y, $-1/\sqrt{Y}, \ldots$

Move toward $X^2, X^3, \ldots or$ toward log Y, $-1/\sqrt{Y}, \ldots$

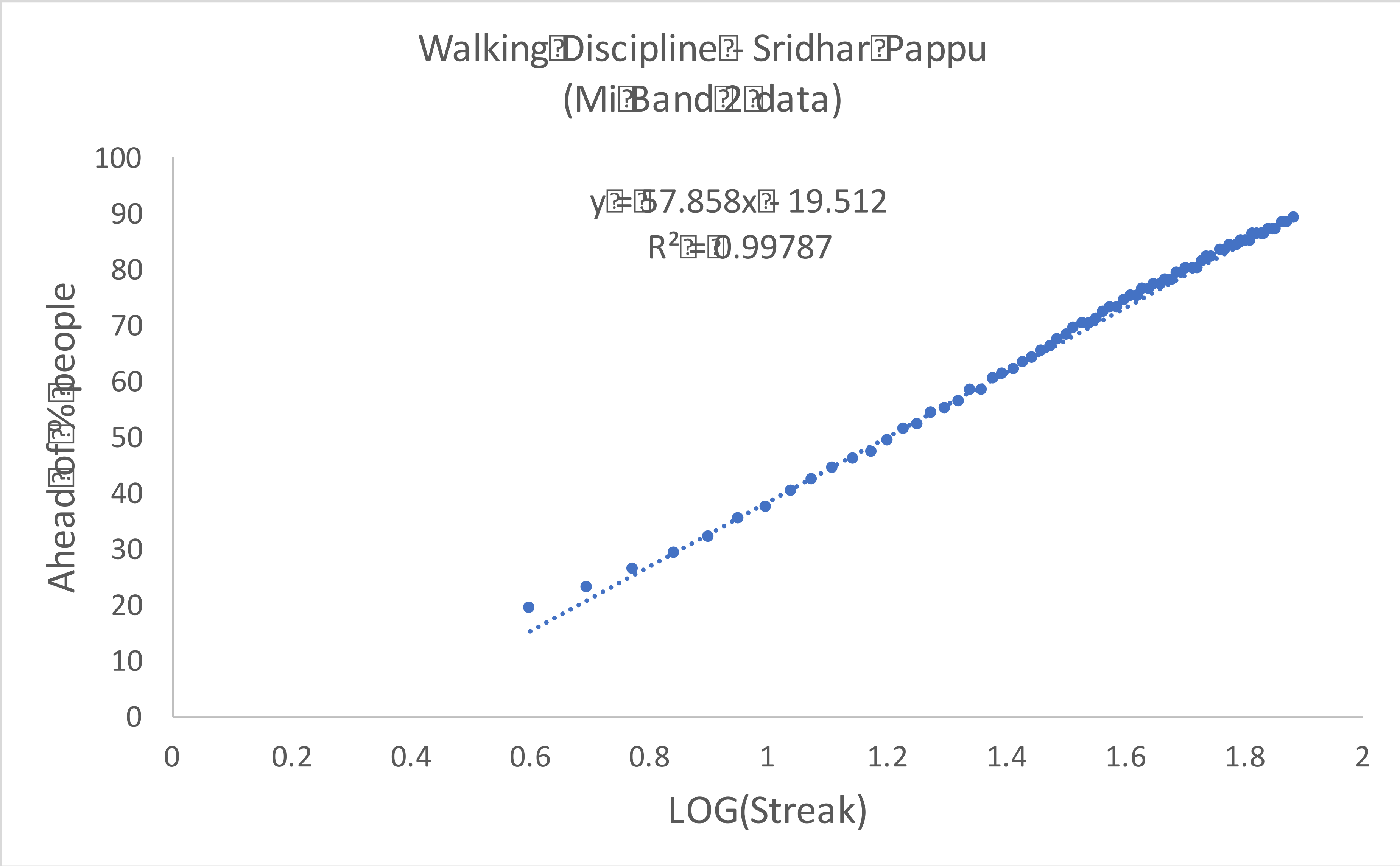
Based on Tukey’s 4-Quadrant Approach, what transformation do you recommend?



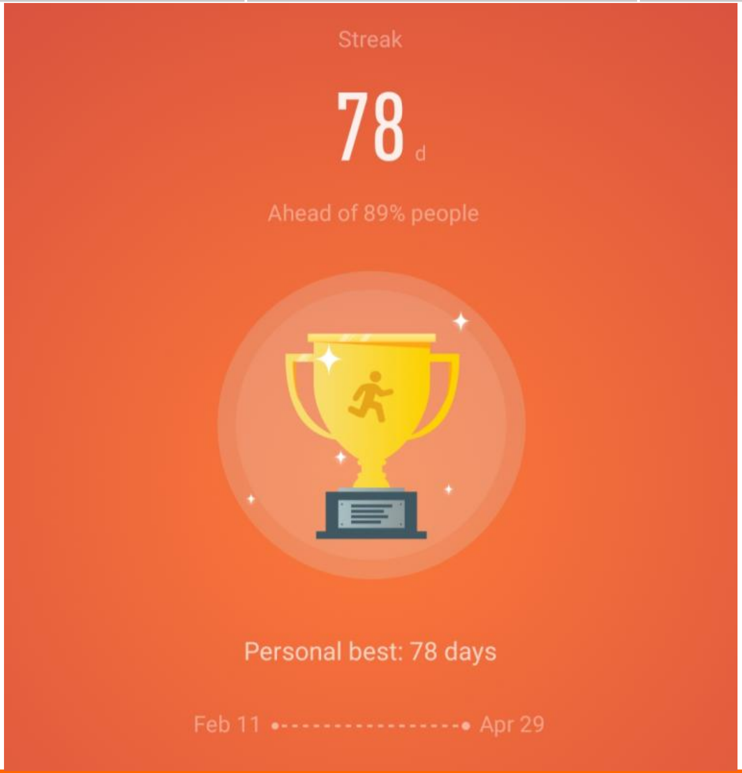
SQRT Transformation on X



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



Streak

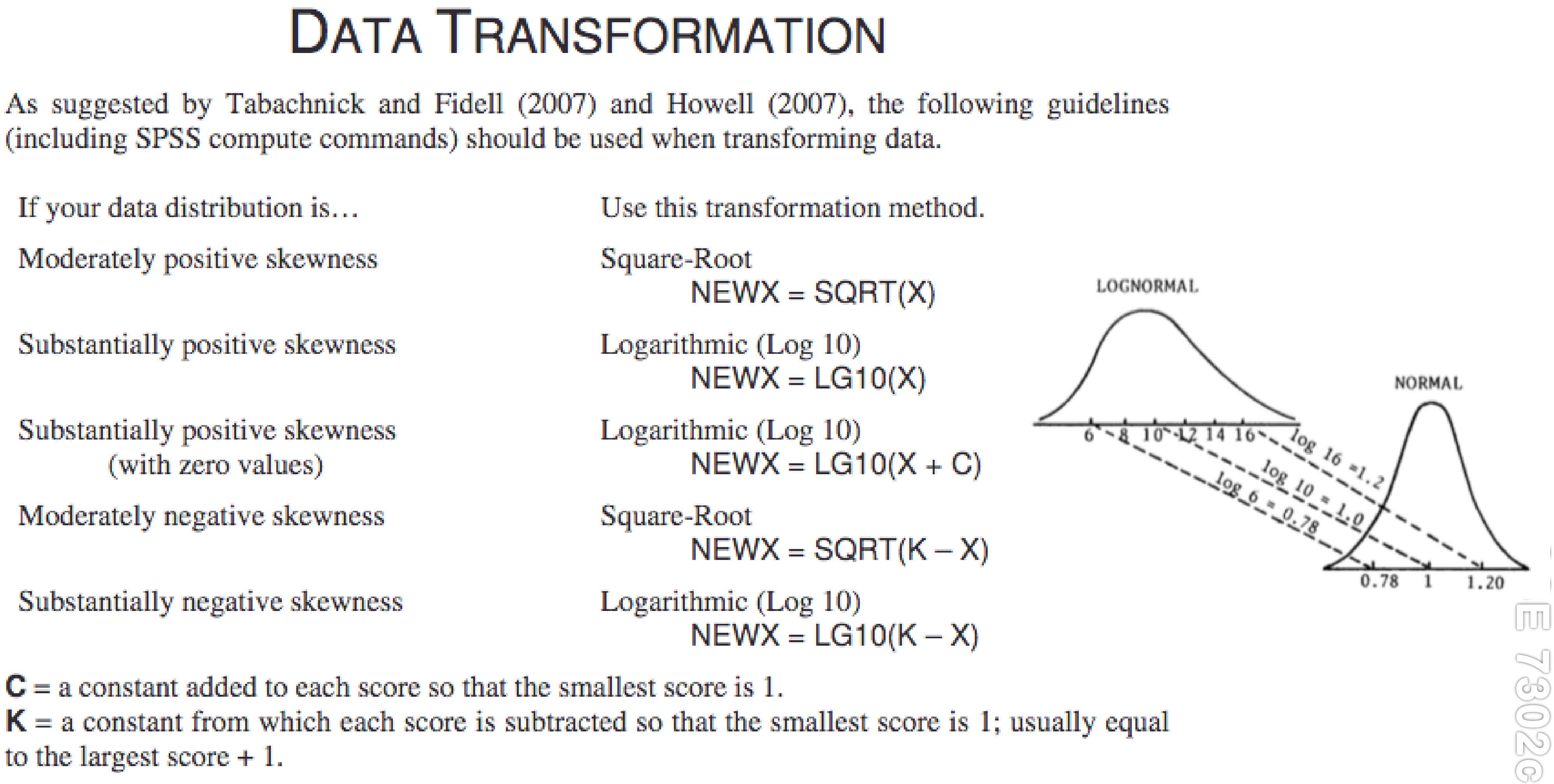
78 d

Ahead of 89% people

Personal best: 78 days

Feb 11 • - - - - - - - - - - - - - - - - • Apr 29

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.

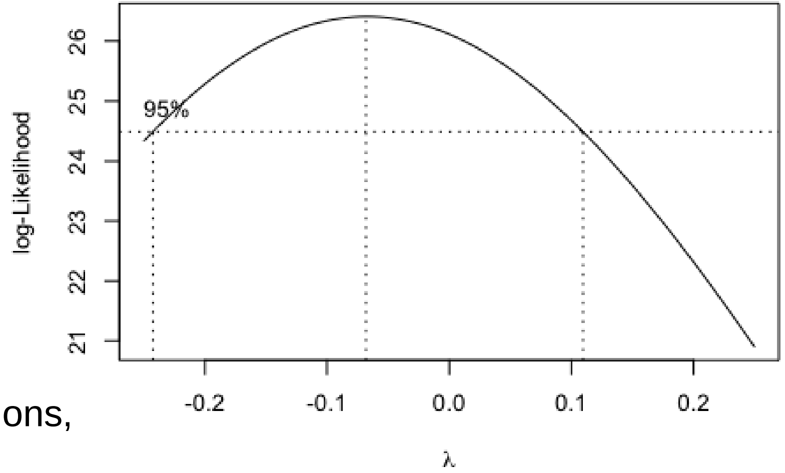
If your data distribution is…	Use this transformation method.
Moderately positive skewness	Square-Root NEWX = SQRT(X)
Substantially positive skewness	Logarithmic (Log 10) NEWX = LG10(X)
Substantially positive skewness (with zero values)	Logarithmic (Log 10) NEWX = LG10(X + C)
Moderately negative skewness	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: $X \rightarrow \sqrt{X}$
  - 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: $X \rightarrow \frac{1}{X}$
  - 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.
    2. Then transform X to remove non-linearity.

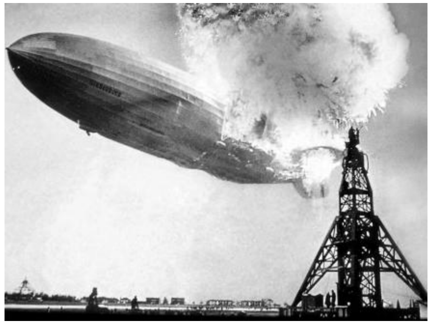
Nonlinear Models – With Interaction

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

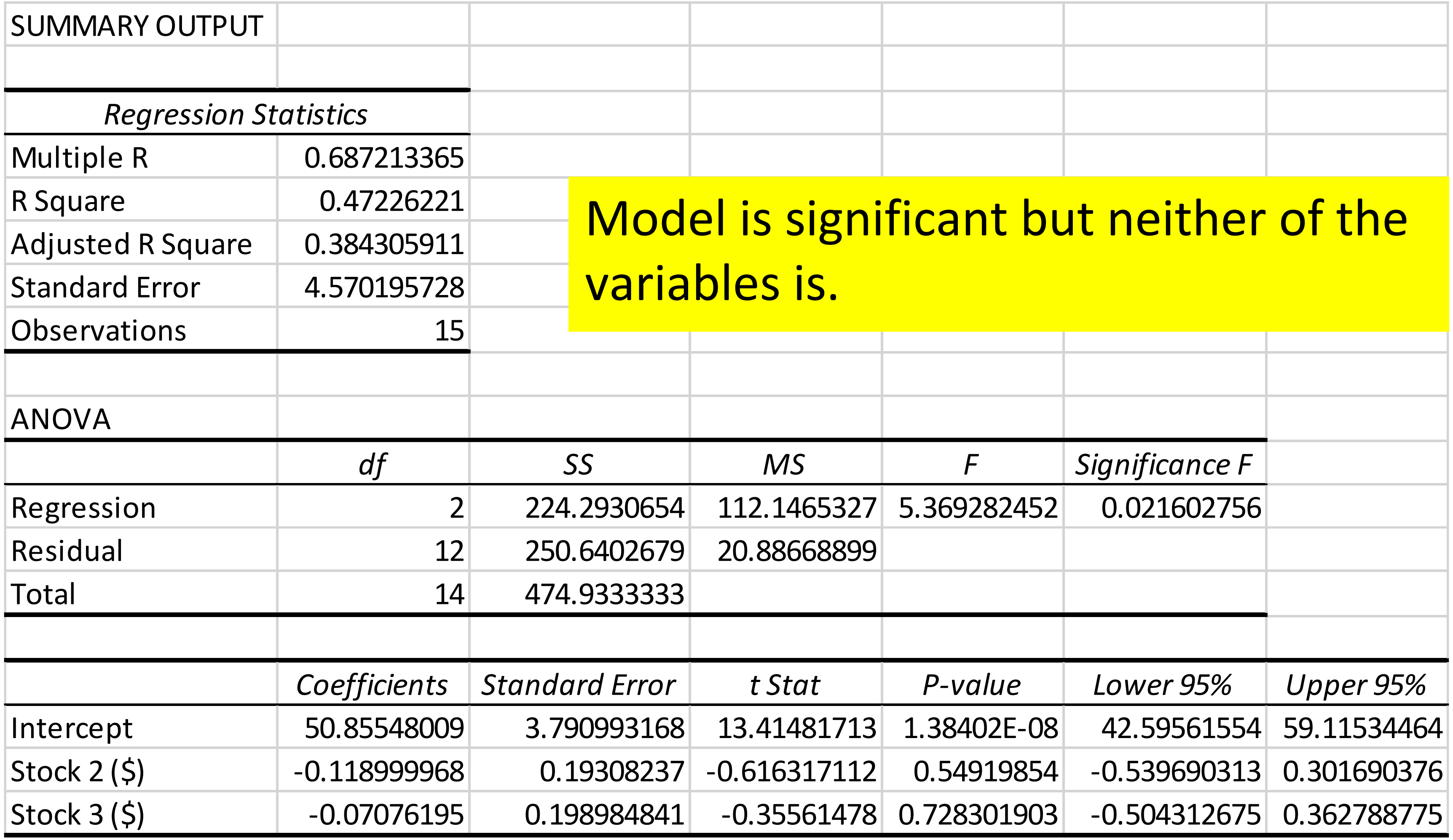
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

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.



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

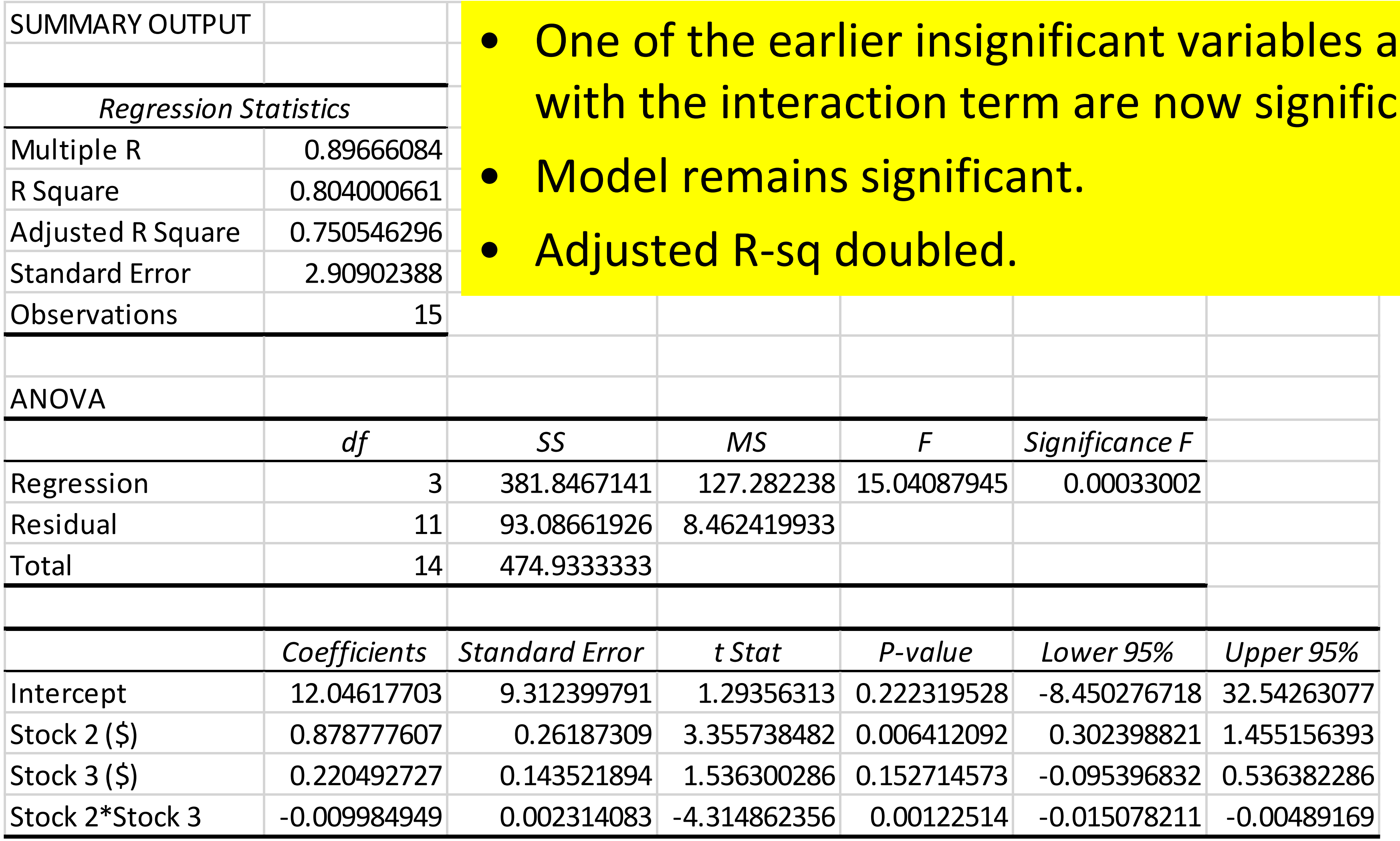
ANOVA

	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.

Nonlinear Models – With Interaction - Excel



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

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

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

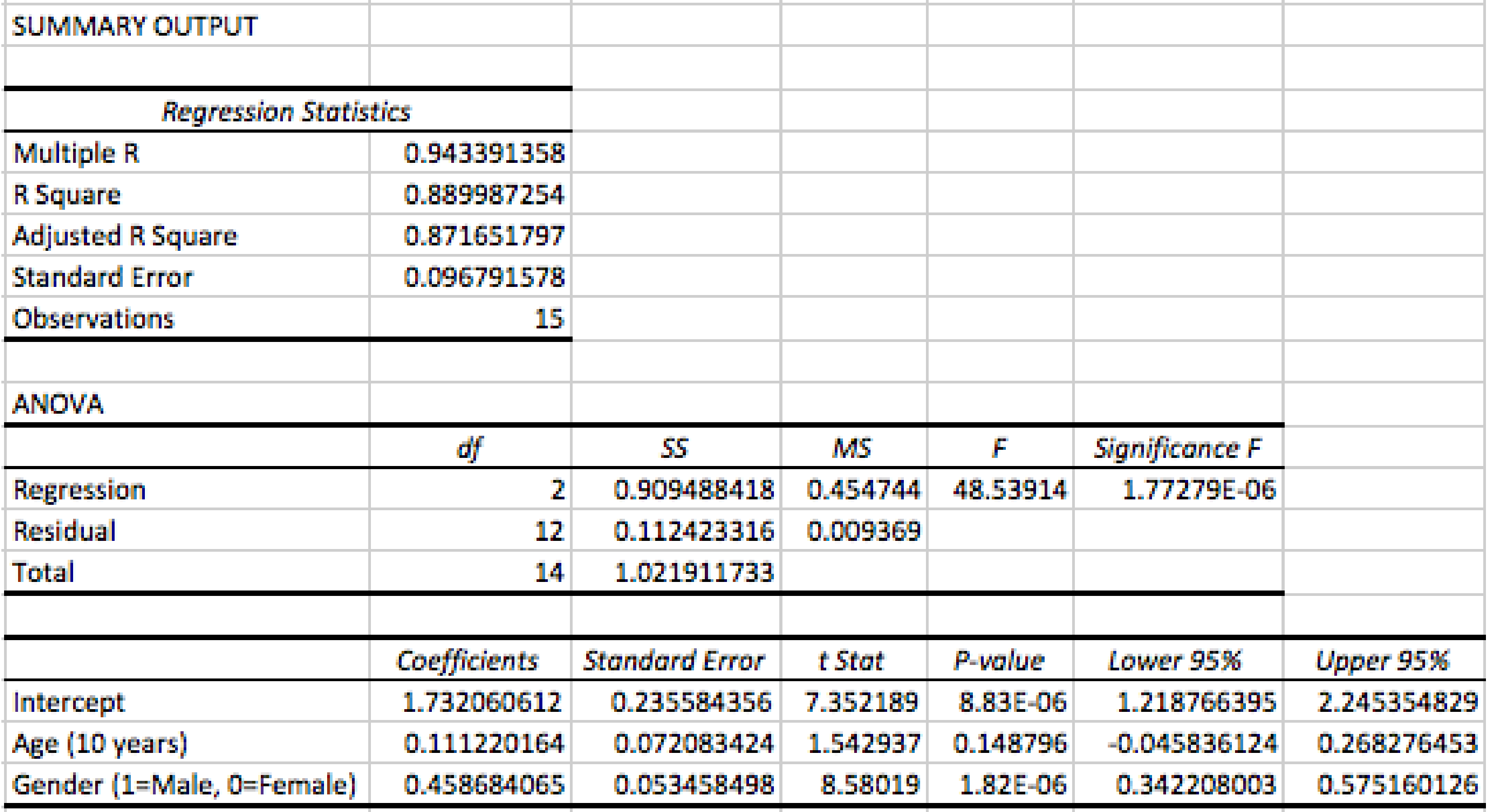
Indicator (Dummy) Variables - Excel

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

FOR SAME JOB, INDIAN MEN EARN 27% MORE THAN WOMEN

TOP STORIES MAMATA WINS WEST BENGAL

Indicator (Dummy) Variables - Excel



SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.943391358					
R Square	0.889987254					
Adjusted R Square	0.871651797					
Standard Error	0.096791578					
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	0.909488418	0.454744	48.53914	1.77279E-06	
Residual	12	0.112423316	0.009369			
Total	14	1.021911733				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.732060612	0.235584356	7.352189	8.83E-06	1.218766395	2.245354829
Age (10 years)	0.111220164	0.072083424	1.542937	0.148796	-0.045836124	0.268276453
Gender (1=Male, 0=Female)	0.458684065	0.053458498	8.58019	1.82E-06	0.342208003	0.575160126

Separate equation for each gender

Multiple Linear Regression

MODEL BUILDING METHODS

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?

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.

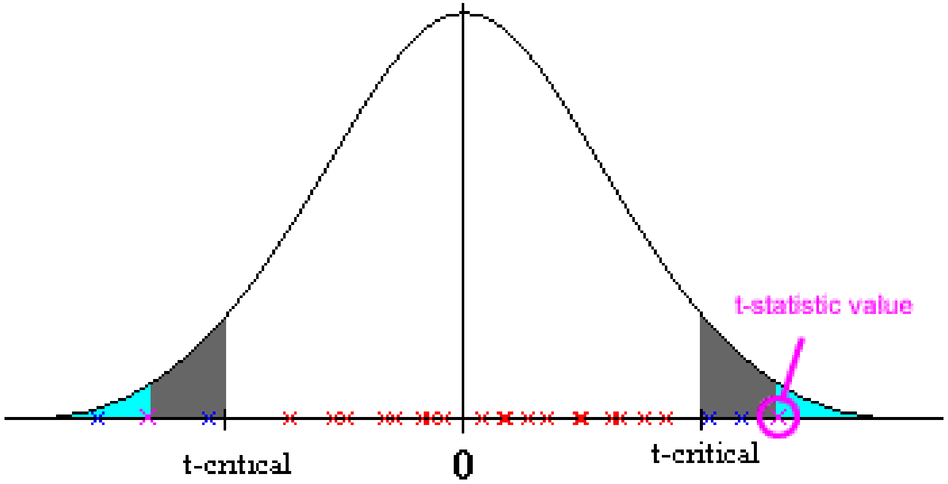
Search Procedures: Stepwise Regression

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).

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.

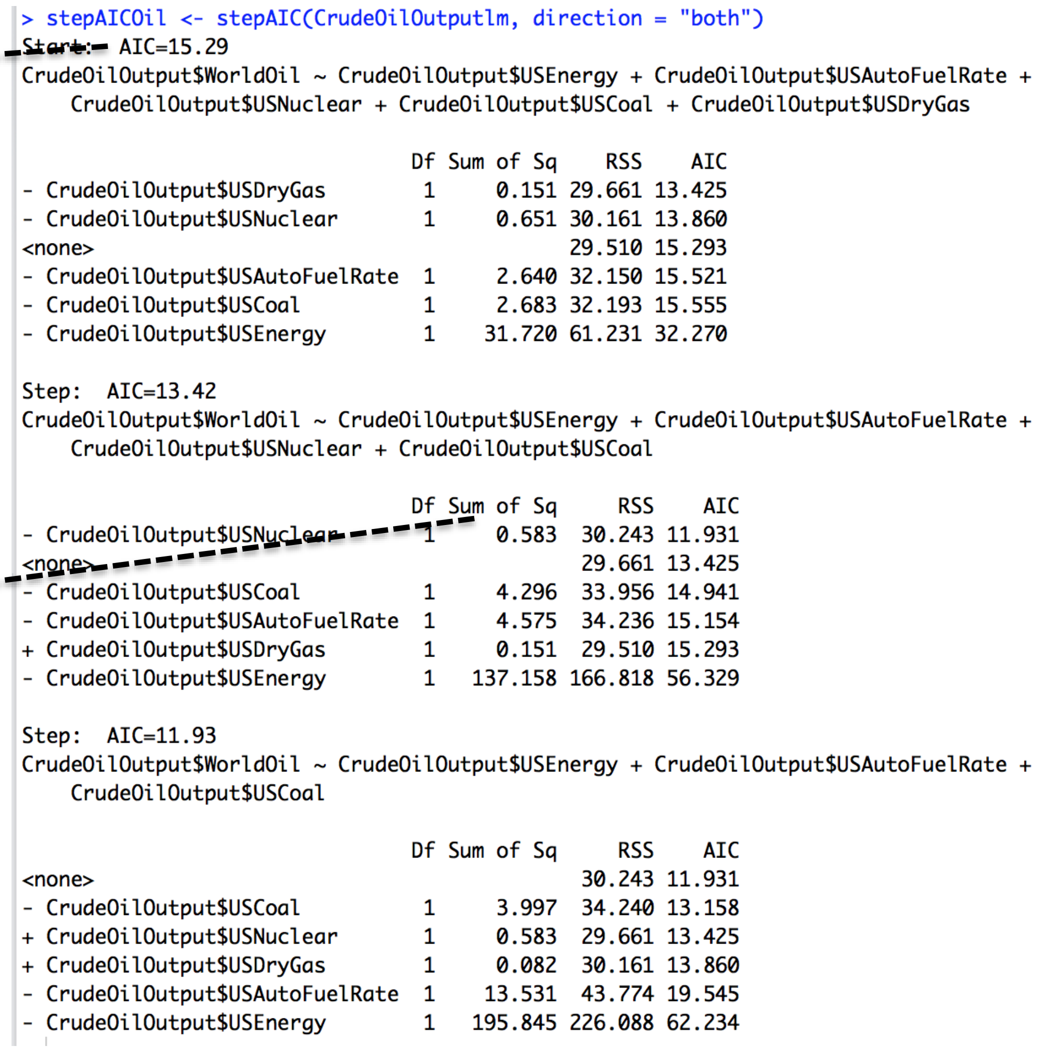
Search Procedures: Stepwise Regression - R

AIC (Akaike's Information Criterion)

AIC = $2k + n\ln(RSS/n)$ 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
```

CSE 7302c

Multiple Linear Regression

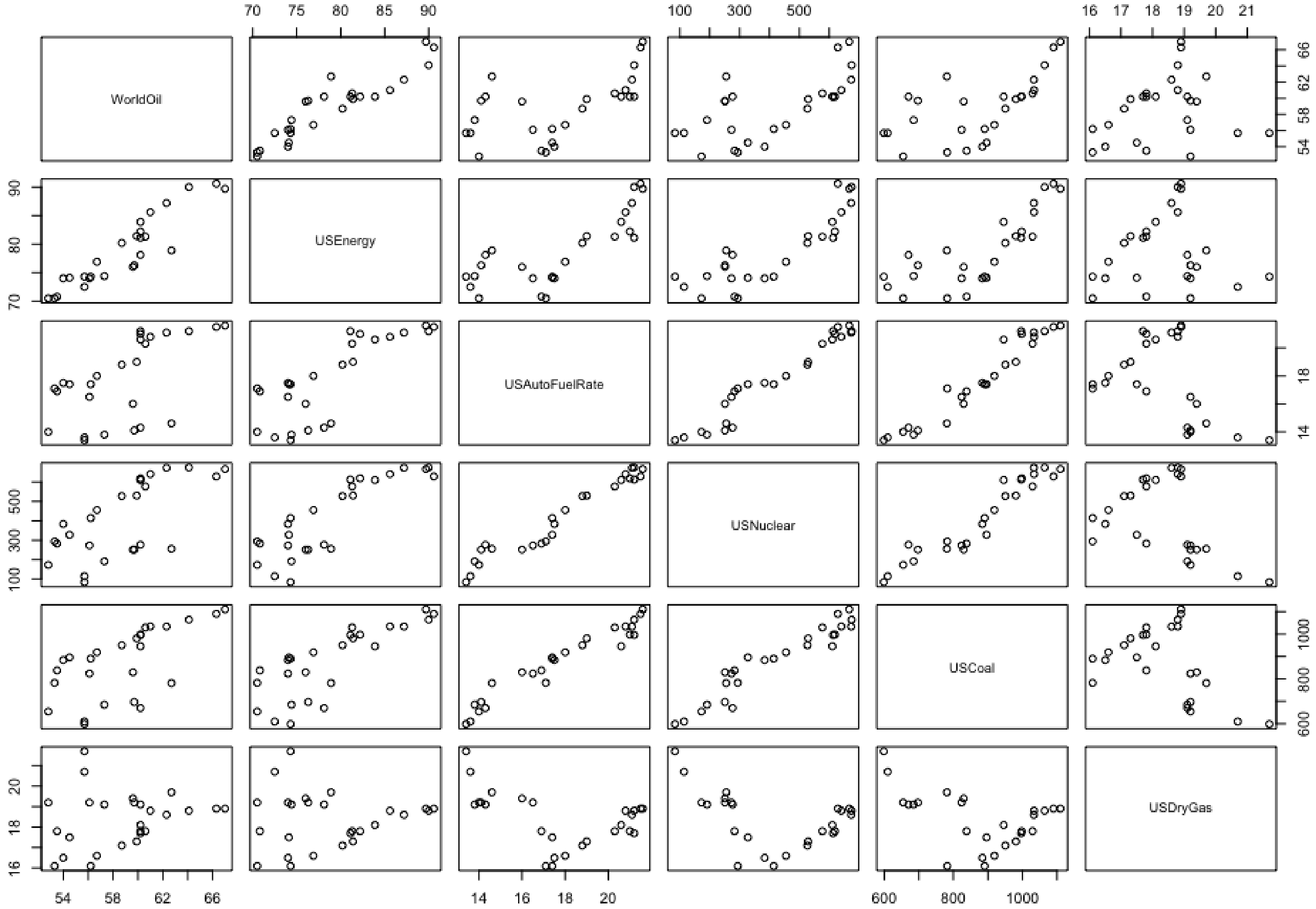
HANDLING MULTICOLLINEARITY

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)

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

Correlation between stock 2 and stock 3 is 0.96

Multicollinearity

- Stepwise regression prevents this problem to a great extent.

- 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 > 10 or ${R_i}^2$>0.90 for the largest VIFs indicates a severe multicollinearity.

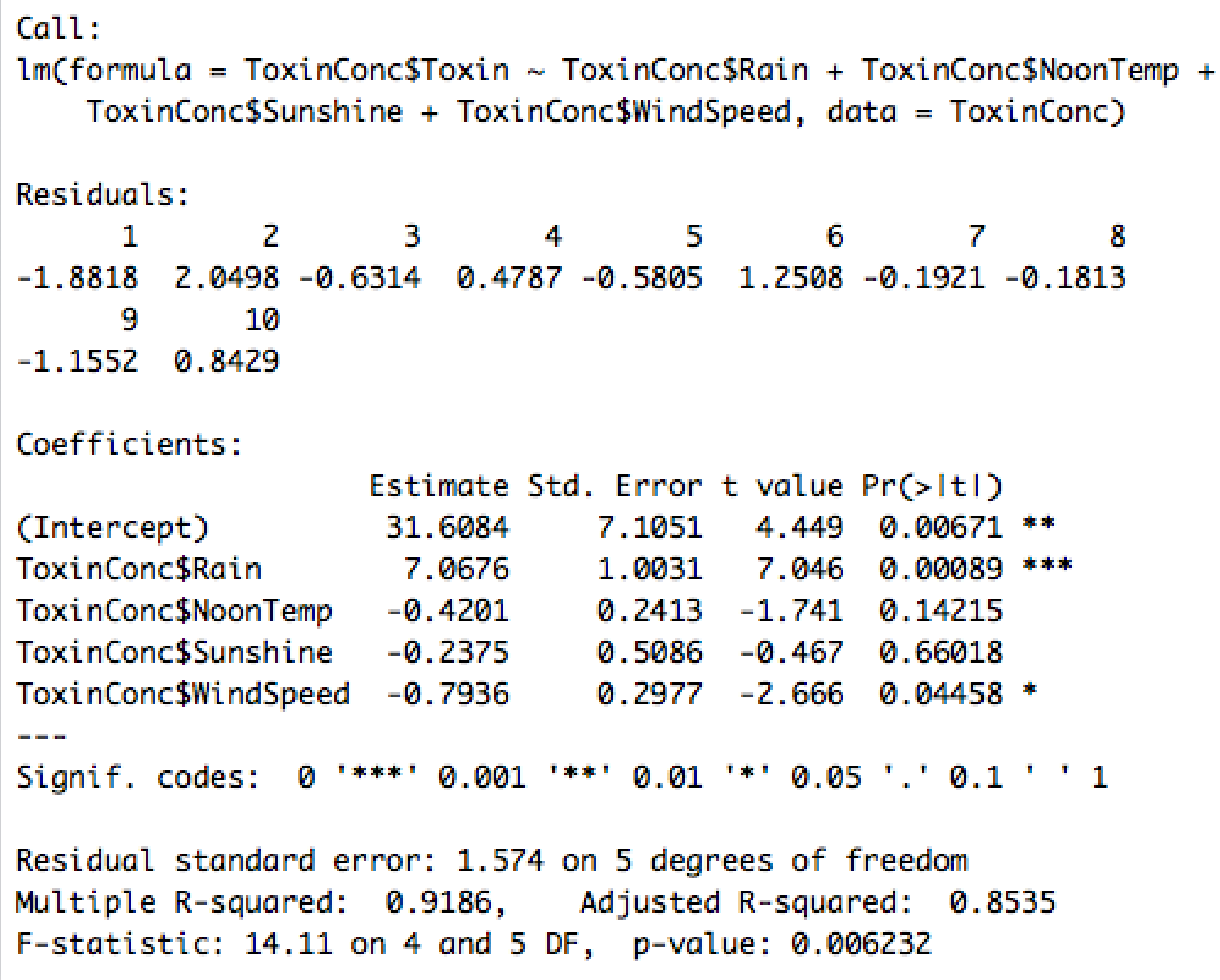
Model Building – R

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 (°C)
- Sunshine (h/day)
- Wind speed (km/h)

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

Model Building – R

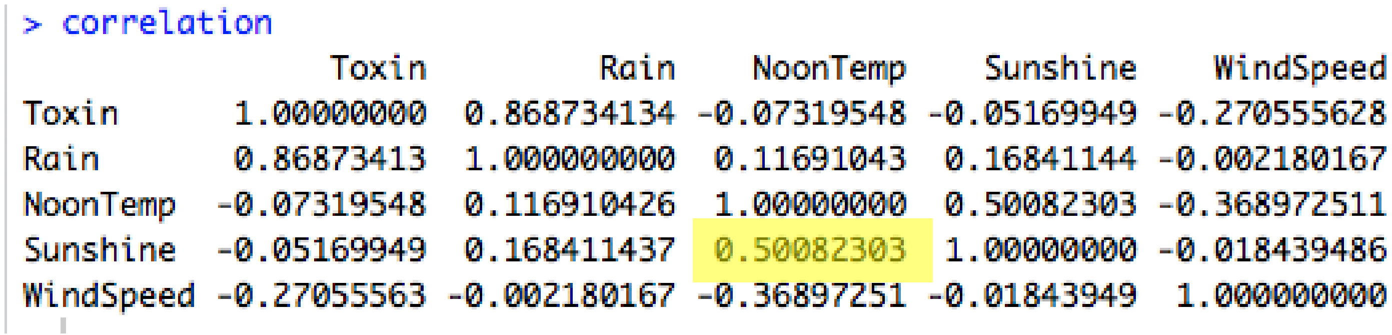


```
Call:
lm(formula = ToxinConc$Toxin ~ ToxinConc$Rain + ToxinConc$NoonTemp +
    ToxinConc$Sunshine + ToxinConc$WindSpeed, data = ToxinConc)

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.

Model Building – R

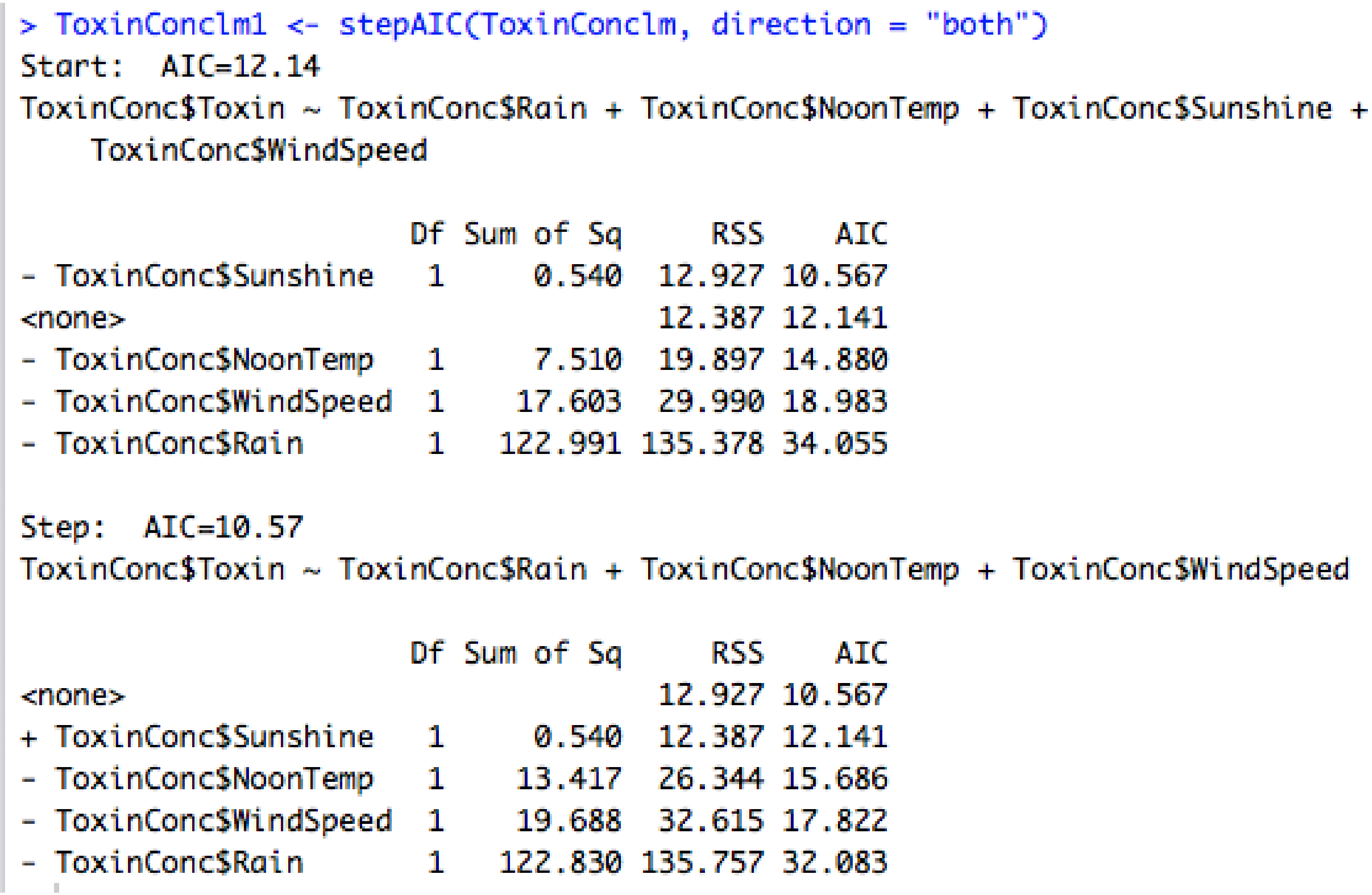
Multiple regression tends to remove correlated pairs of IVs, as in the case of Noon Temperature and Sunshine here.



```
> 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
```

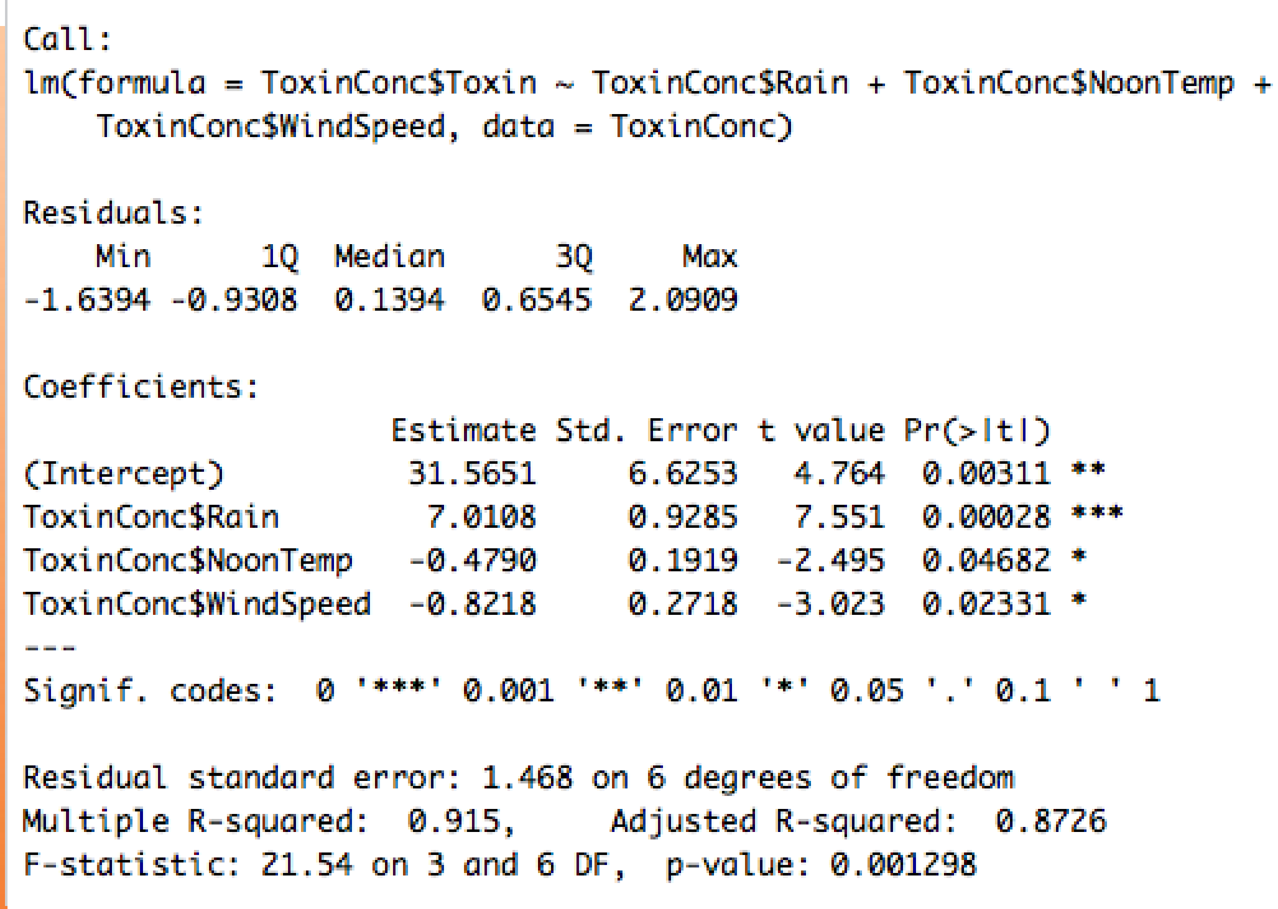
It may be worthwhile to build another model keeping one of the correlated variables in the model. The more significant can be preferred but business intuition may be cautiously used to include other statistically insignificant variable(s).

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
```

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
```

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

4. 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

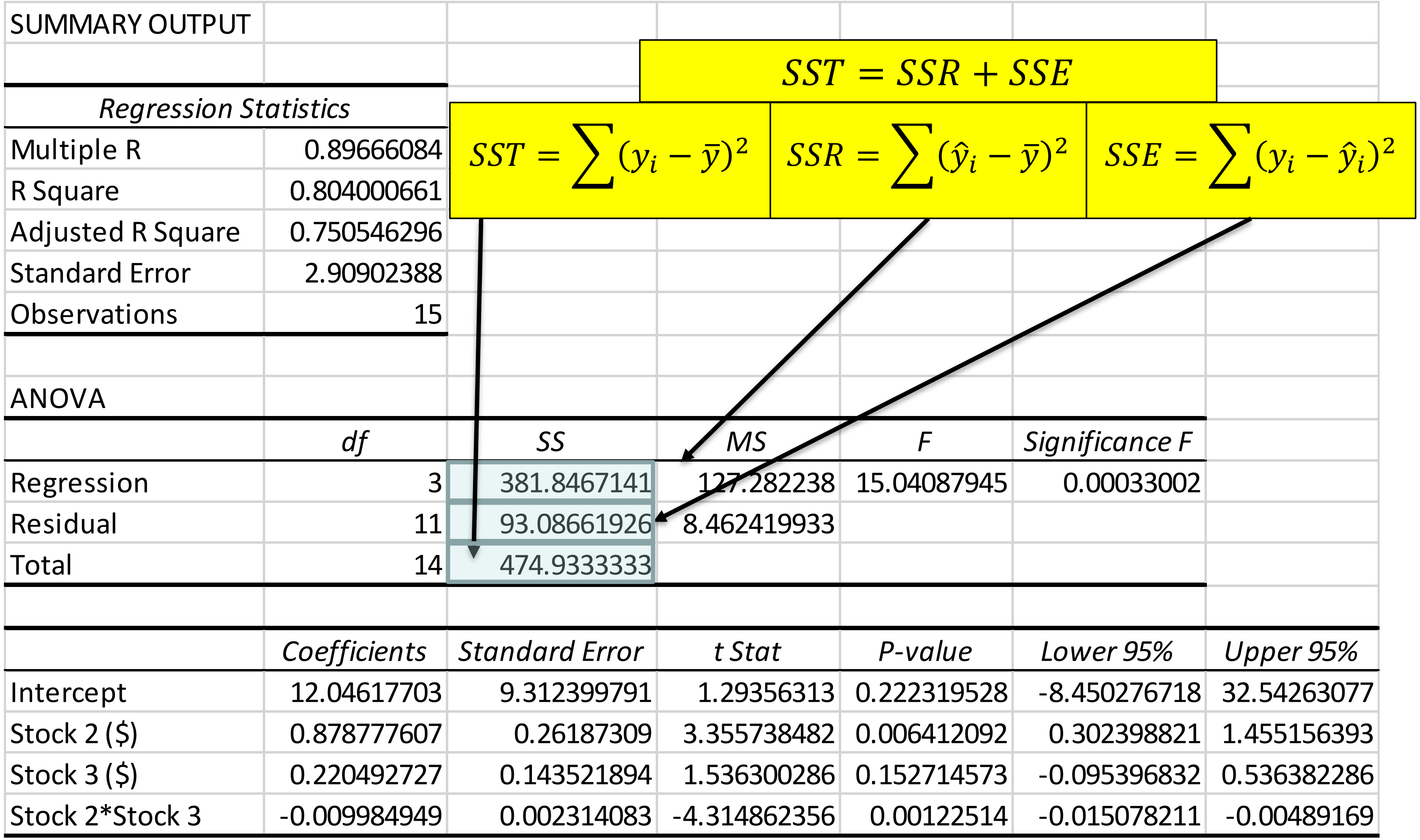
Last accessed: January 05, 2018

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

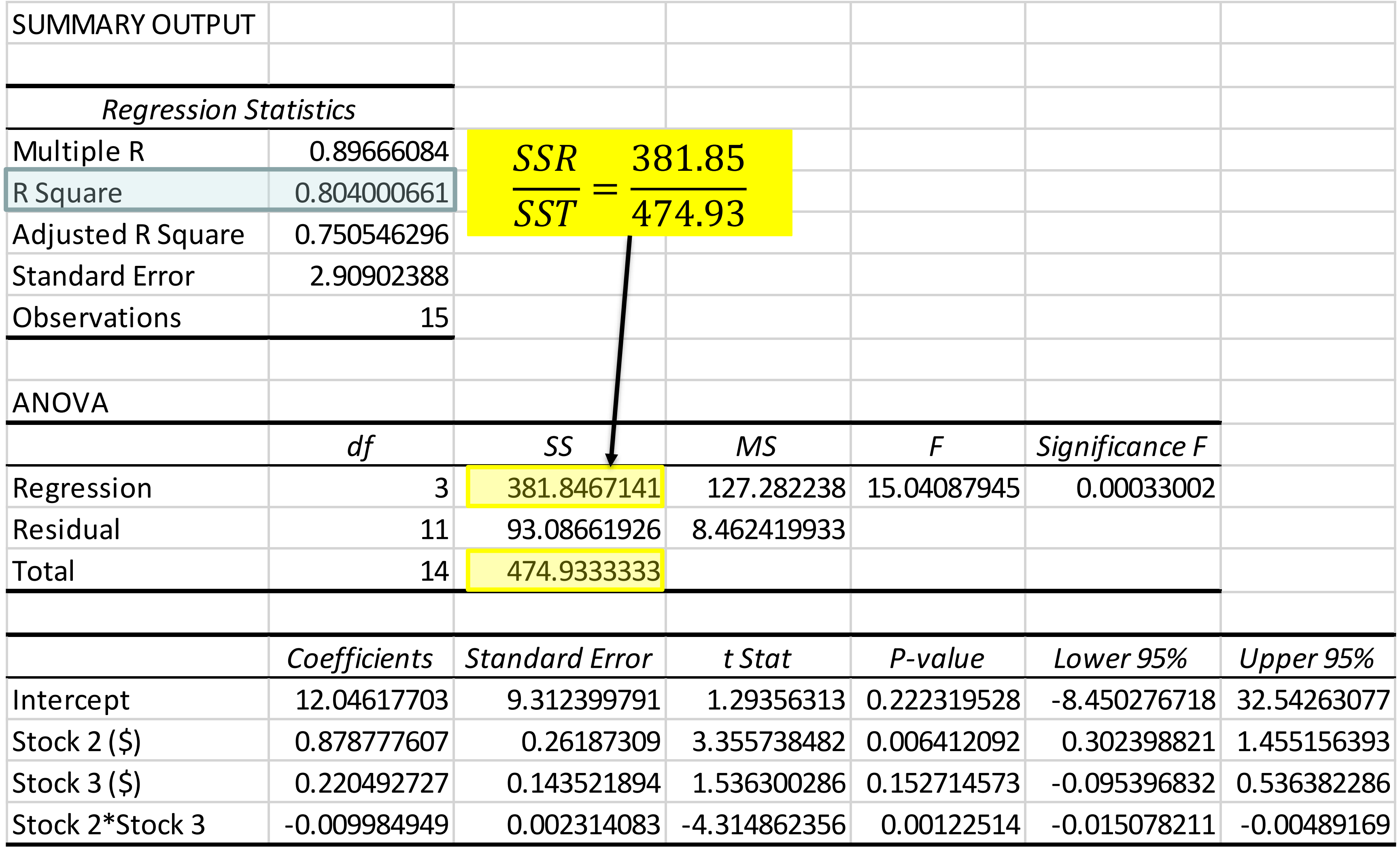
ANOVA

	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?

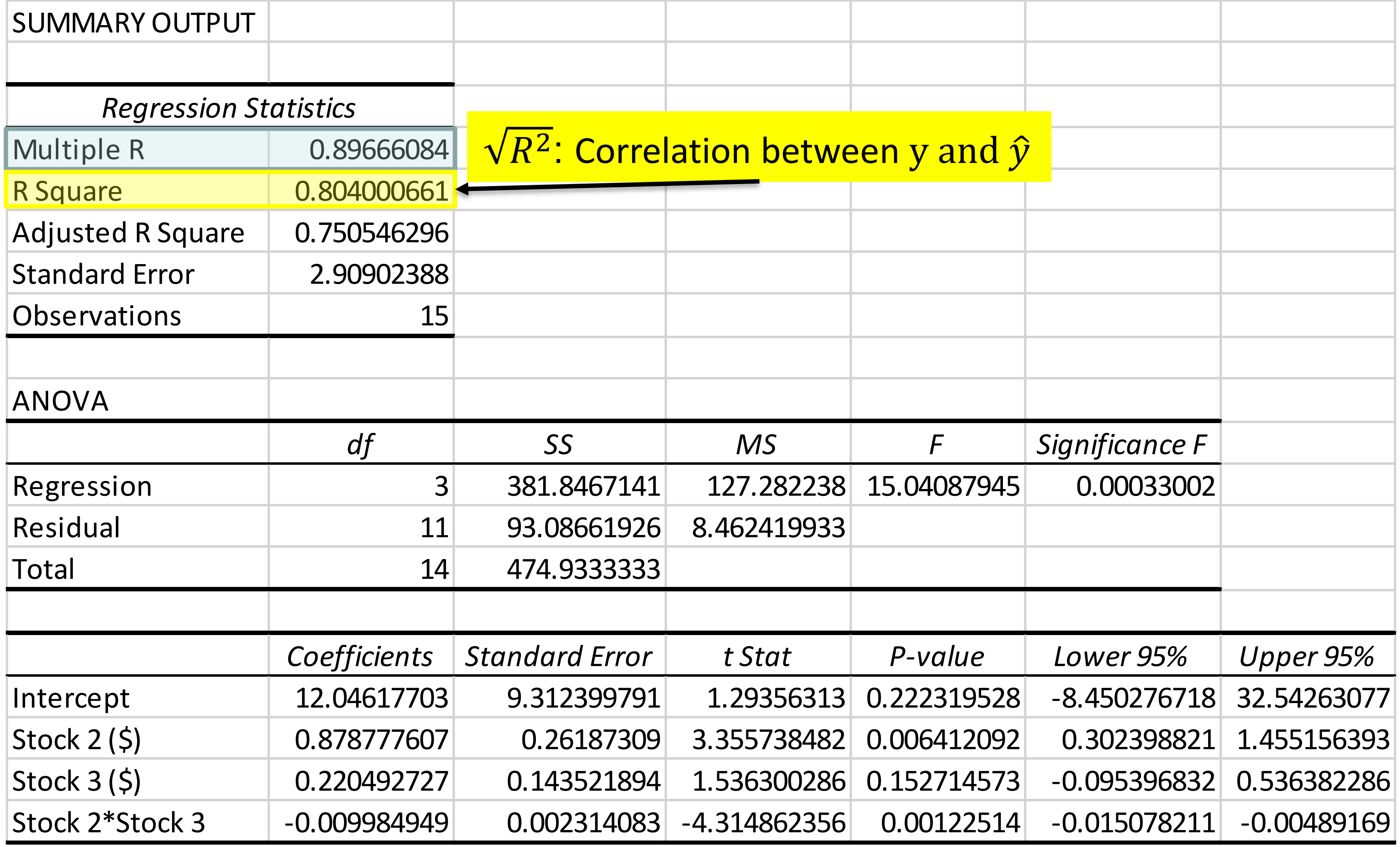


SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.89666084					
R Square	0.804000661					
Adjusted R Square	0.750546296					
Standard Error	2.90902388					
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	381.8467141	127.282238	15.04087945	0.00033002	
Residual	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

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

Output Analysis - Recap

What is the correlation between actual and expected values?

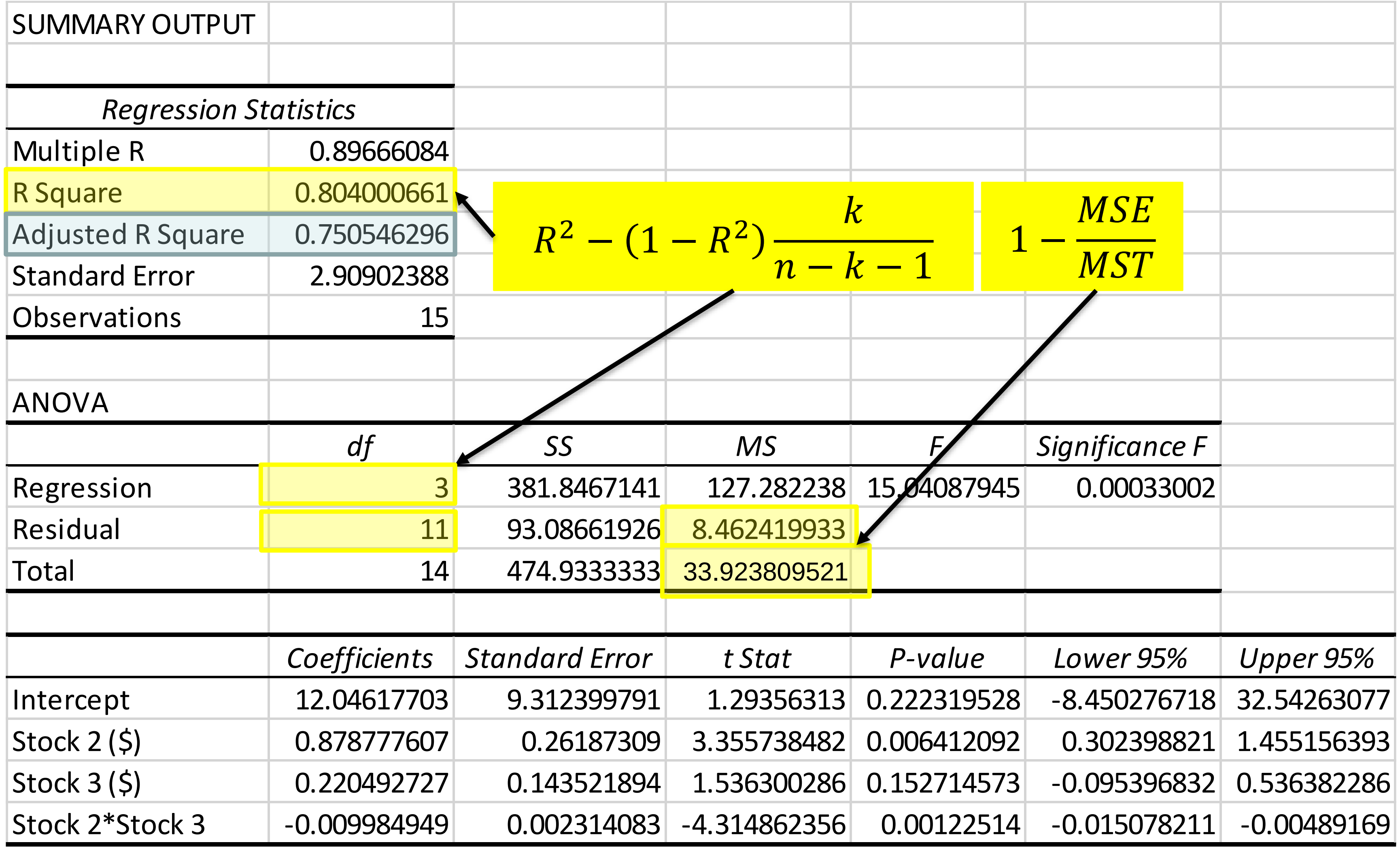


SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.89666084					
R Square	0.804000661					
Adjusted R Square	0.750546296					
Standard Error	2.90902388					
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	381.8467141	127.282238	15.04087945	0.00033002	
Residual	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

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

Output Analysis - Recap

How much of total variation can be explained by variation in independent variables (IVs) that *actually affect* the DV?



SUMMARY OUTPUT

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

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

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

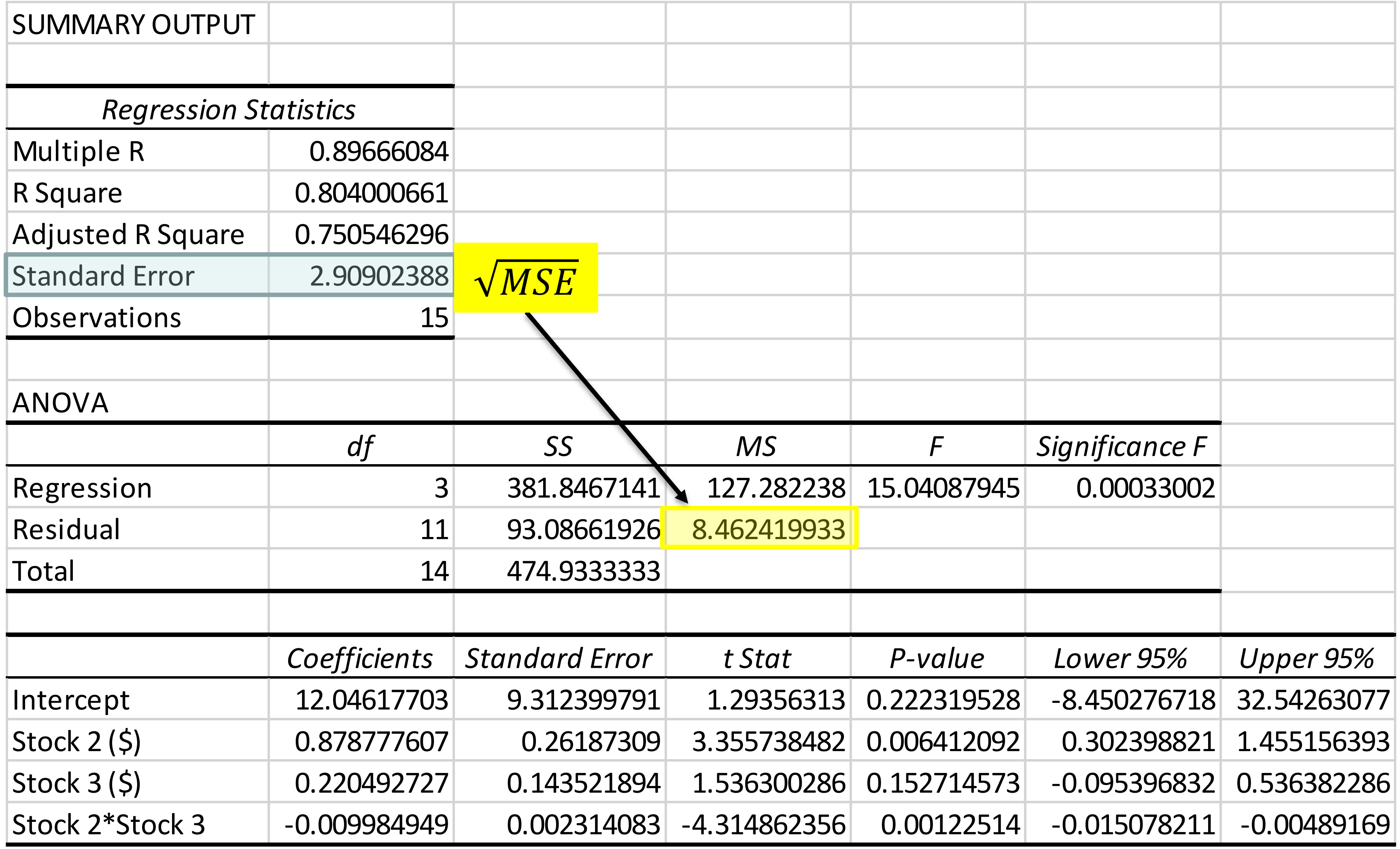
ANOVA

	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

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
Observations	15

$\sqrt{MSE}$

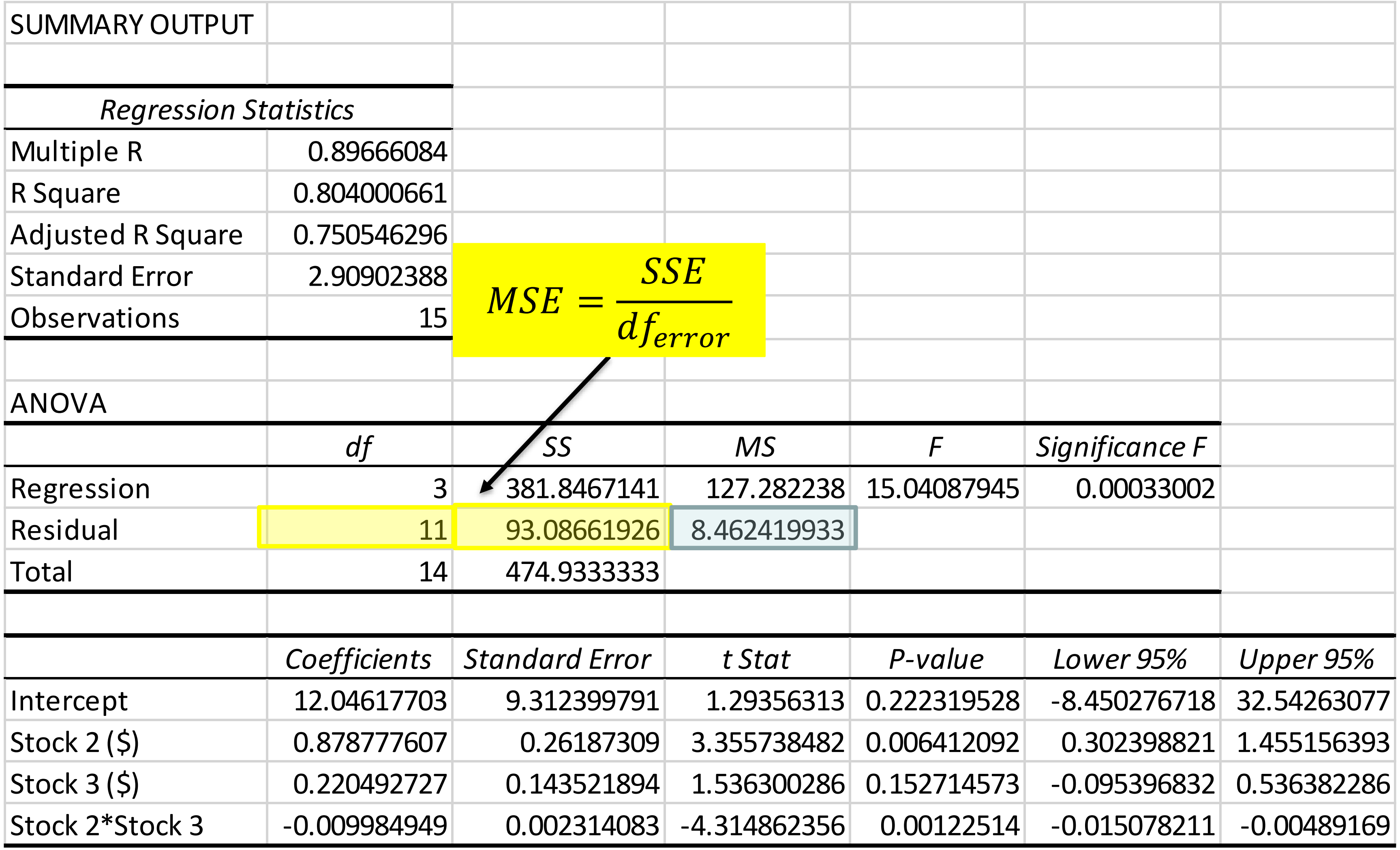
ANOVA

	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

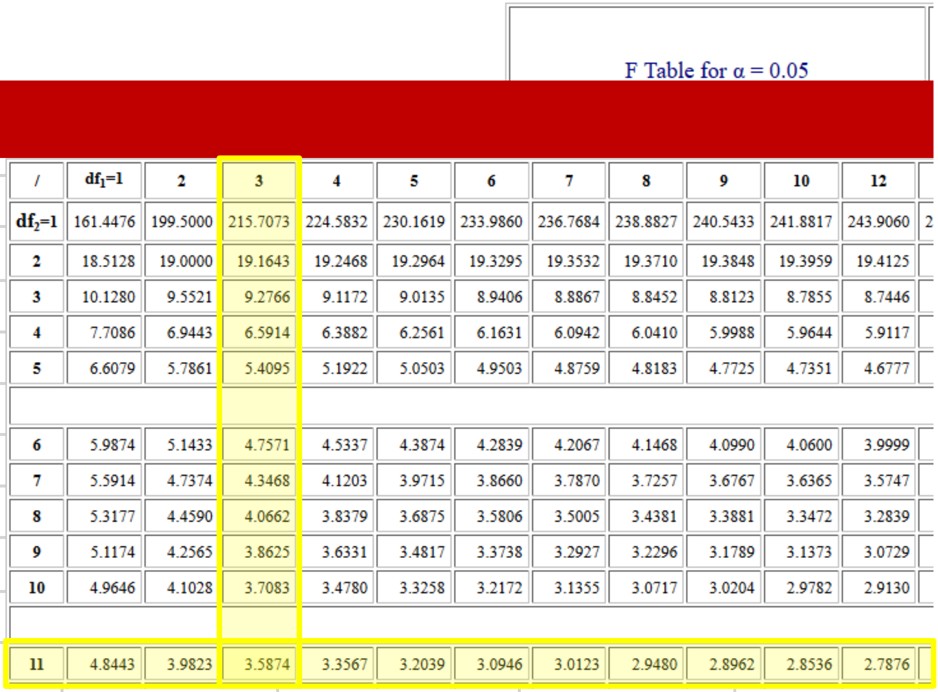
Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

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

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



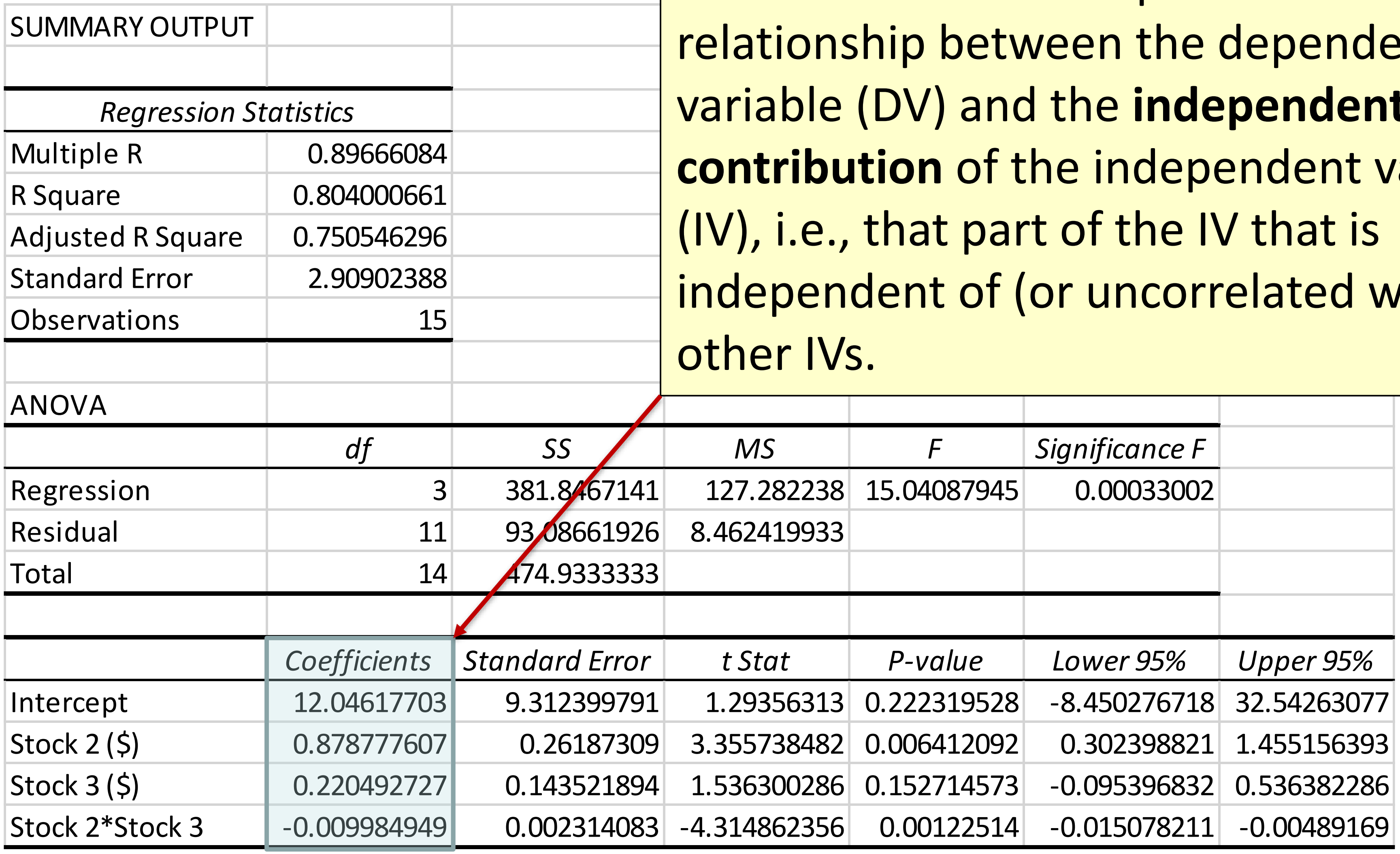
F Table for α = 0.05

/	df_1=1	2	3	4	5	6	7	8	9	10	12	
df_2=1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817	243.9060	2
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959	19.4125	
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855	8.7446	
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644	5.9117	
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.6777	
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600	3.9999	
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747	
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2839	
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729	
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130	
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876	

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 parameter?

SUMMARY OUTPUT

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 *X*s 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

Output Analysis - Recap

Are the coefficients 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*
Regression	3	381.8467141	127.282238	15.0408794
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

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

Table entry for p and C is the point t^* with probability p lying above it and probability C lying between $-t^*$ and t^*.

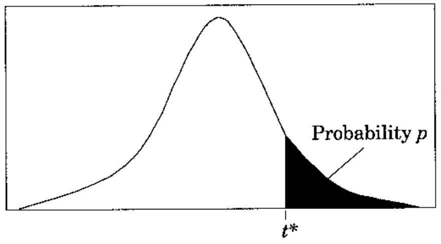


Table B t distribution critical values

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Output Analysis - Recap

What are the confidence intervals for the coefficients?

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

SUMMARY OUTPUT

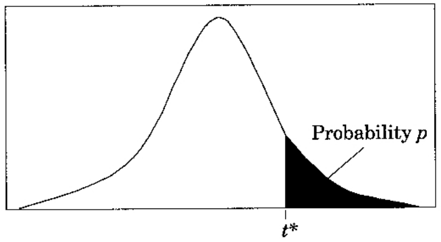
Regression Statistics	
Multiple R	0.89666084
R Square	0.804000661
Adjusted R Square	0.750546296
Standard Error	2.90902388
Observations	15

ANOVA

	df	*SS*	*MS*	*F*
Regression	3	381.8467141	127.282238	15.0408794
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

Table entry for p and C is the point t^* with probability p lying above it and probability C lying between $-t^*$ and t^*.



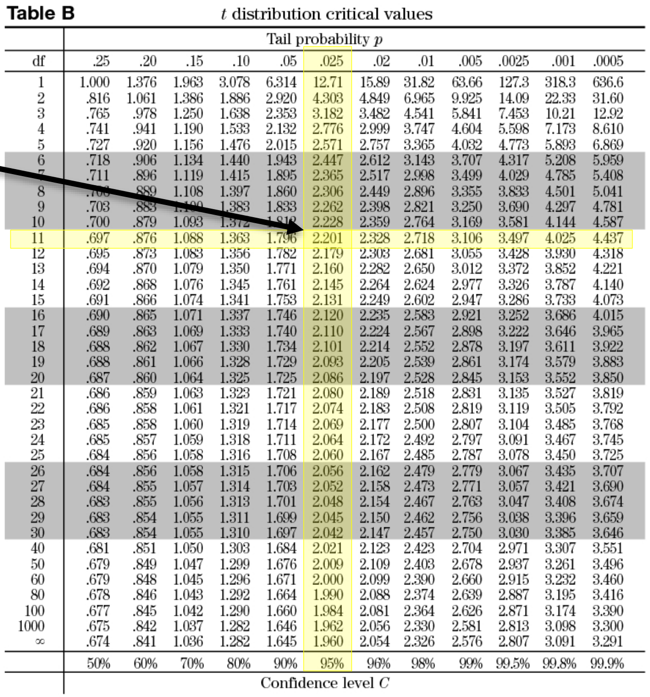


Table B t distribution critical values

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Multiple Linear Regression

CASE - MONEYBALL

Case – Oakland A’s 2002 Success (Moneyball)

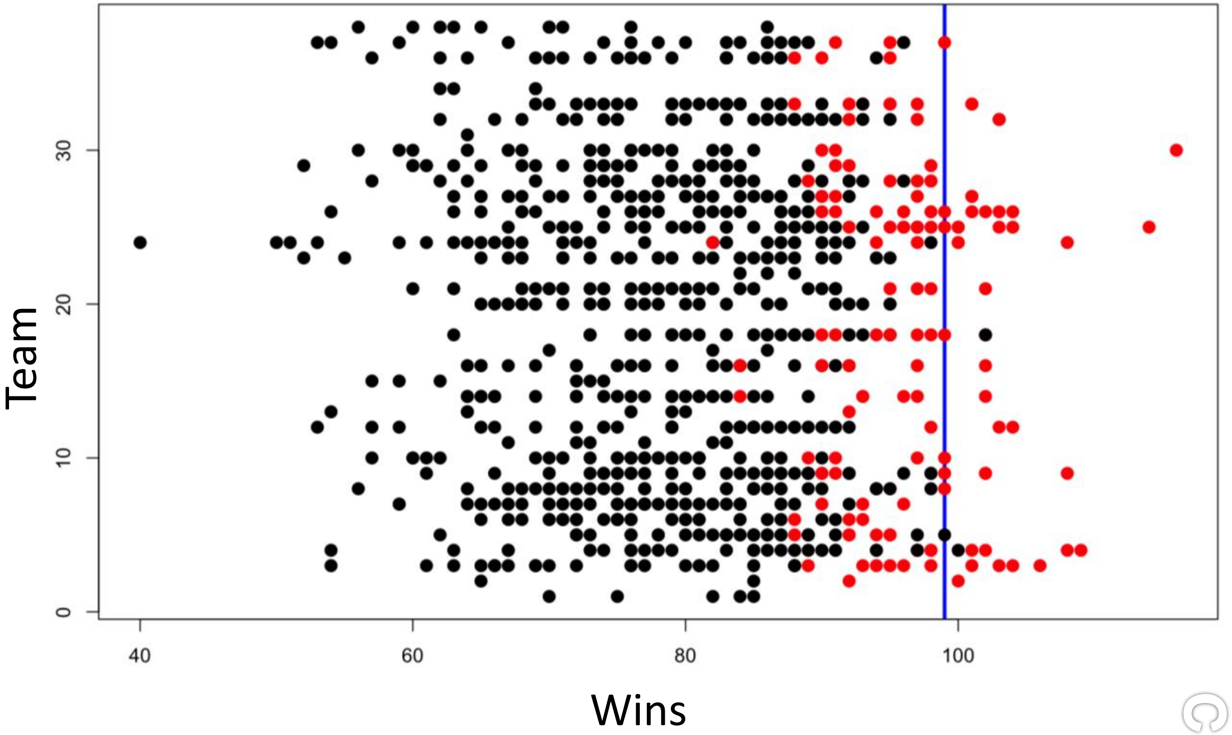
Case Study – Data (baseball-reference.com and MITx)

- 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

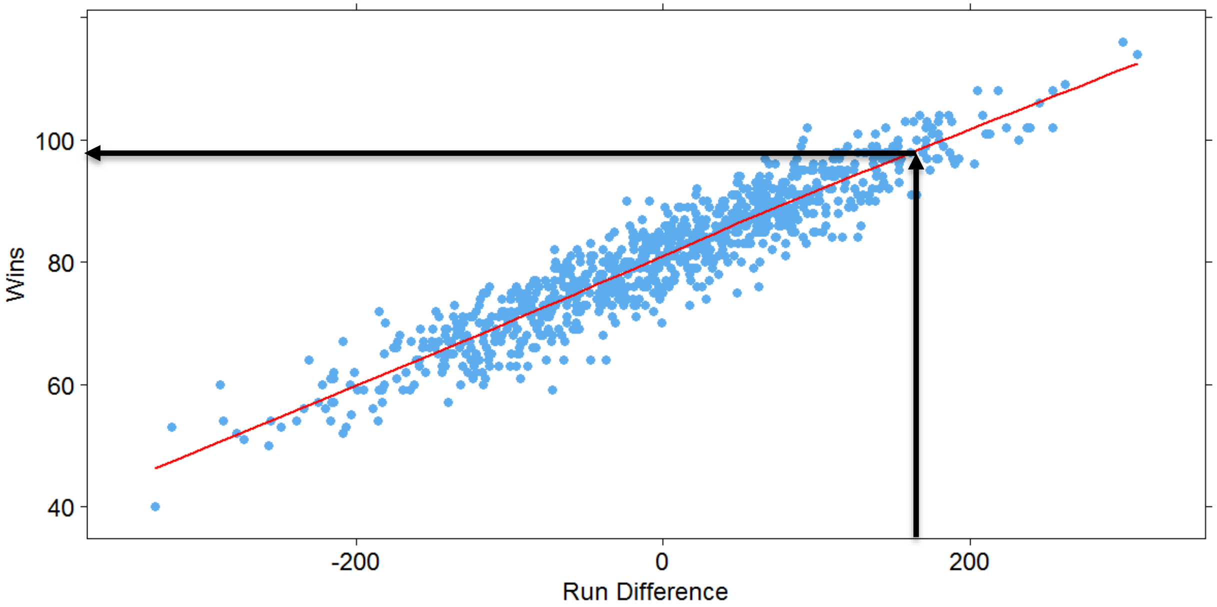
Case Study – Scatter plot

- 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

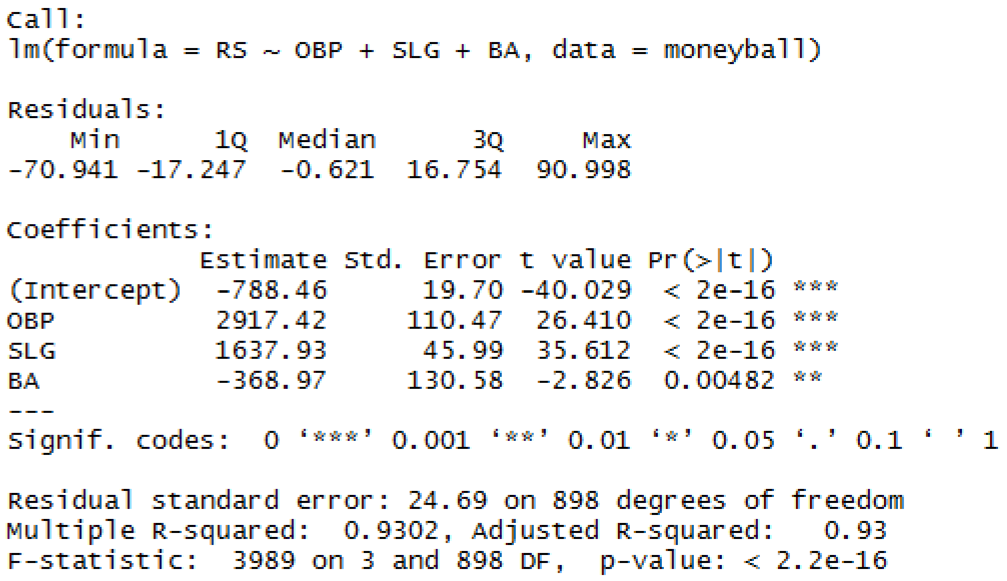


$$W = 80.881375 + 0.105766 * RD$$
$$W = 80.881375 + 0.105766 * 169 = 98.8$$

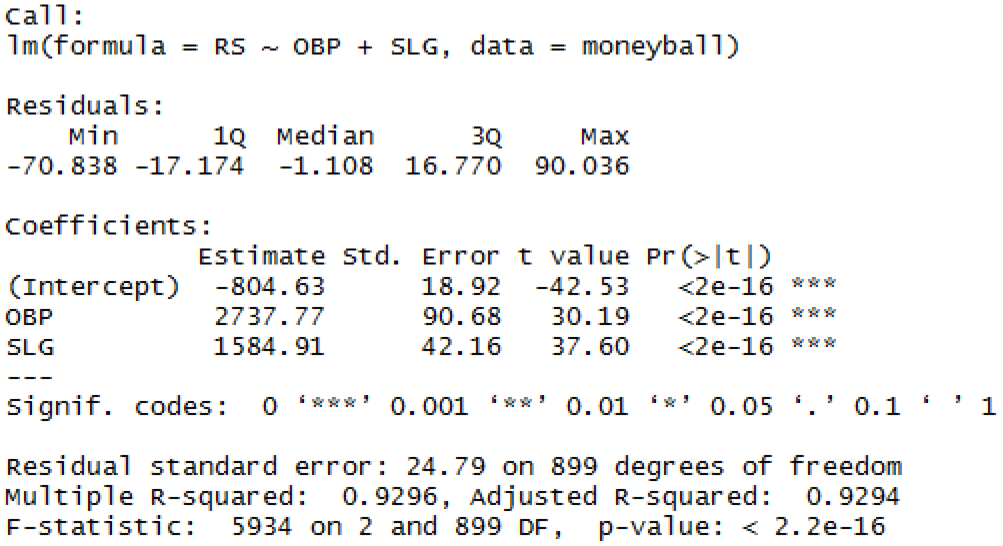
Case Study – Regression for RS

- 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

- 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



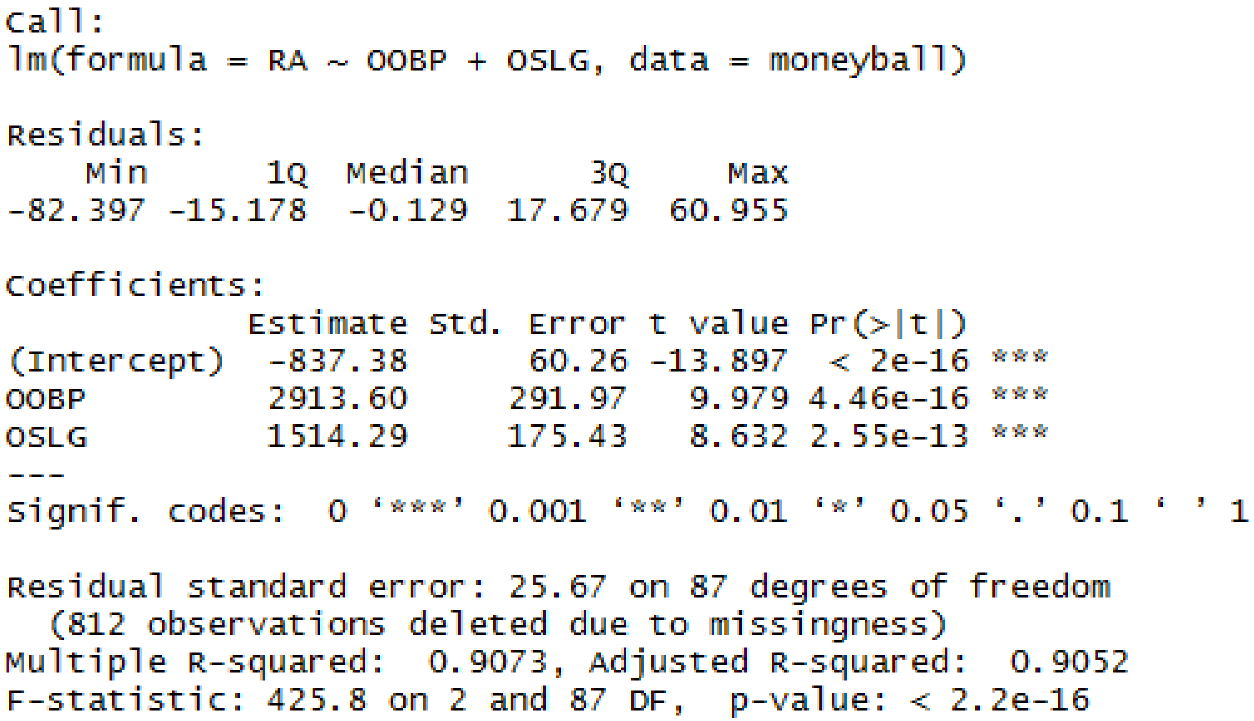
```
Call:
lm(formula = RS ~ OBP + SLG + BA, data = moneyball)

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
```



```
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

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

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

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

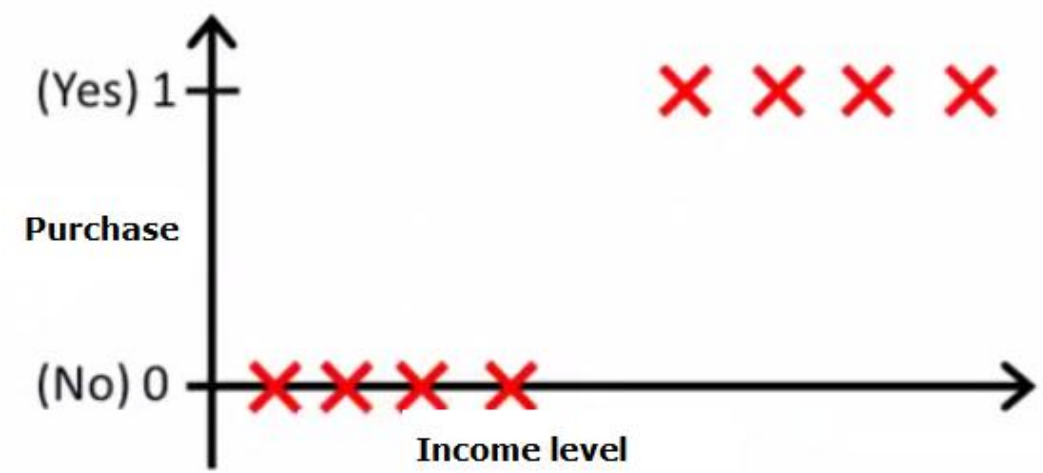
- Results

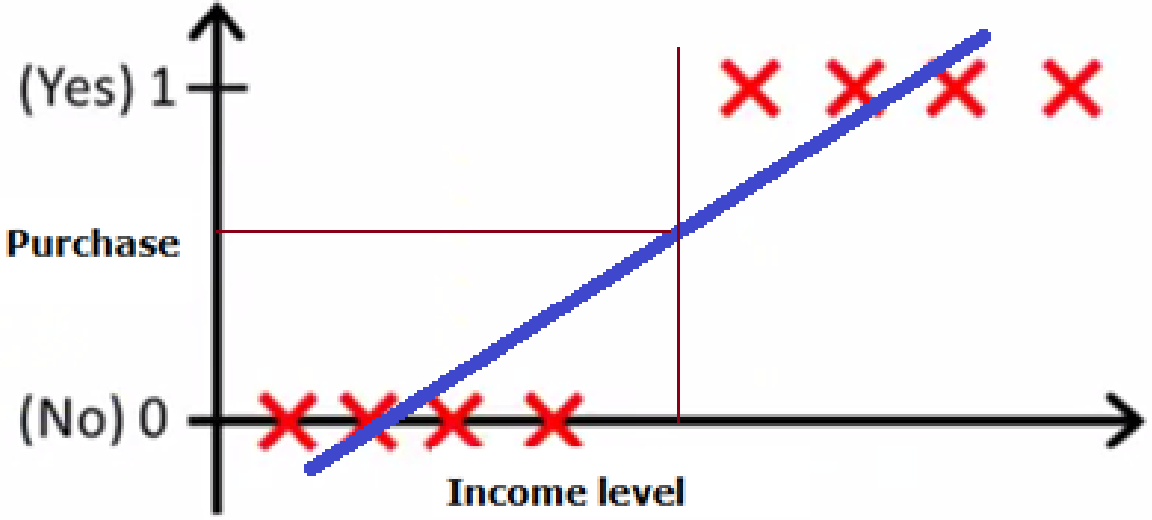
Metric	Model Prediction	DePodesta’s Estimate	Actual
RS	805	800-820	800
RA	622	650-670	654
Wins	100	93-97	103

Classification

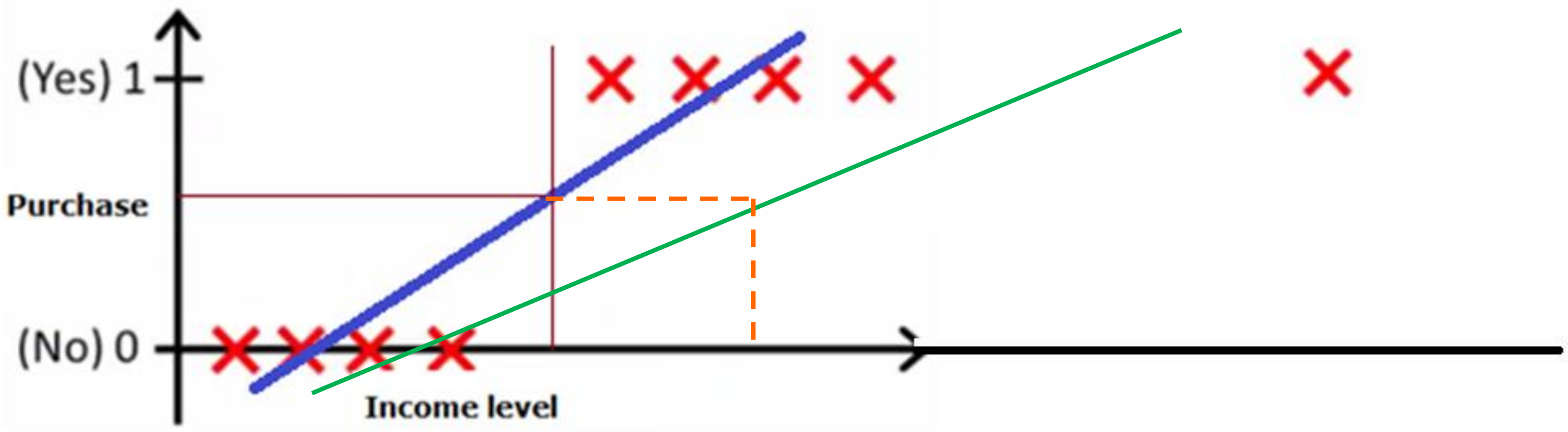
LOGISTIC REGRESSION

Classification Tasks: Regression





It could fail



- In addition, linear regression hypothesis can be much larger than 1 or much smaller than zero and hence thresholding becomes difficult.

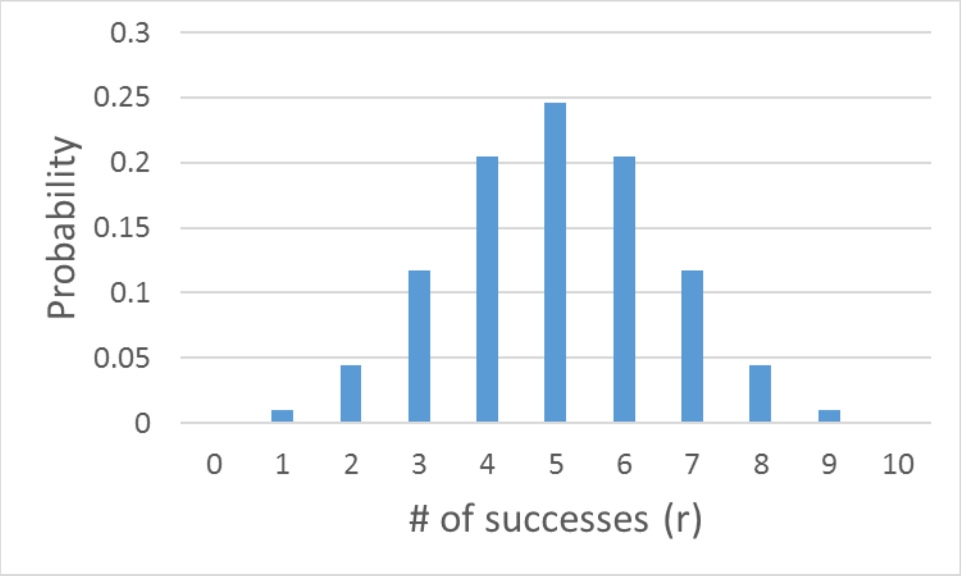
Avoids Assumptions of OLS

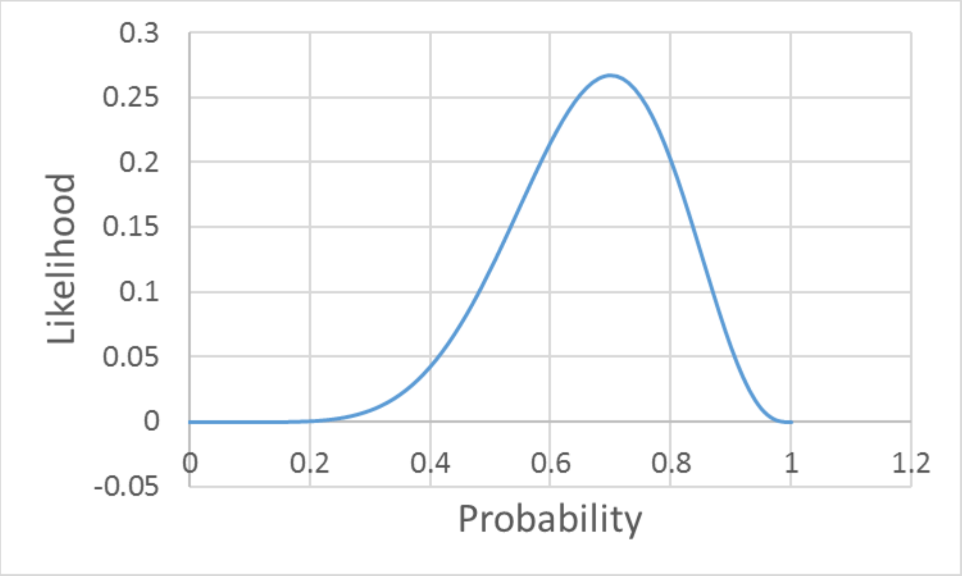
Ordinary Least Squares (OLS) is inappropriate. Maximum Likelihood Estimation (MLE) is used instead.

Hence avoids assumptions regarding normality and homoscedasticity of errors, and linearity between dependent and independent variables. Errors need to be independent, though.

Probability vs Likelihood - Excel

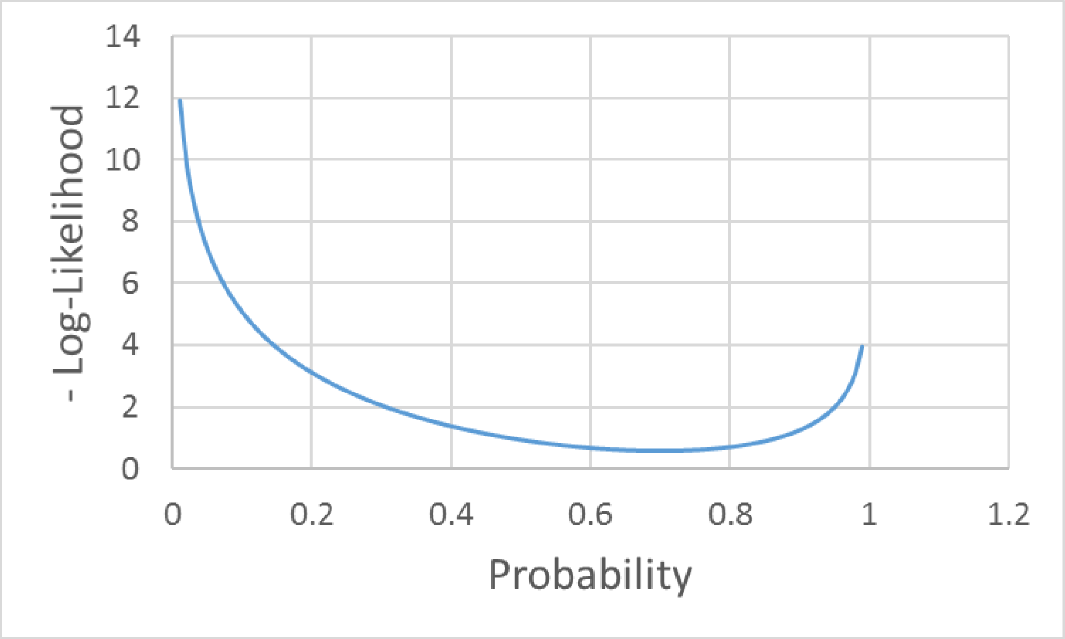
- Likelihood is also known as reverse probability.
- In Probability, we **predict data** based on **known parameters**.
  (Recall B(n,p), Geo(p), Po(λ), N(μ, σ^2), etc.)
- In Likelihood, we **predict parameters** based on **known data**.

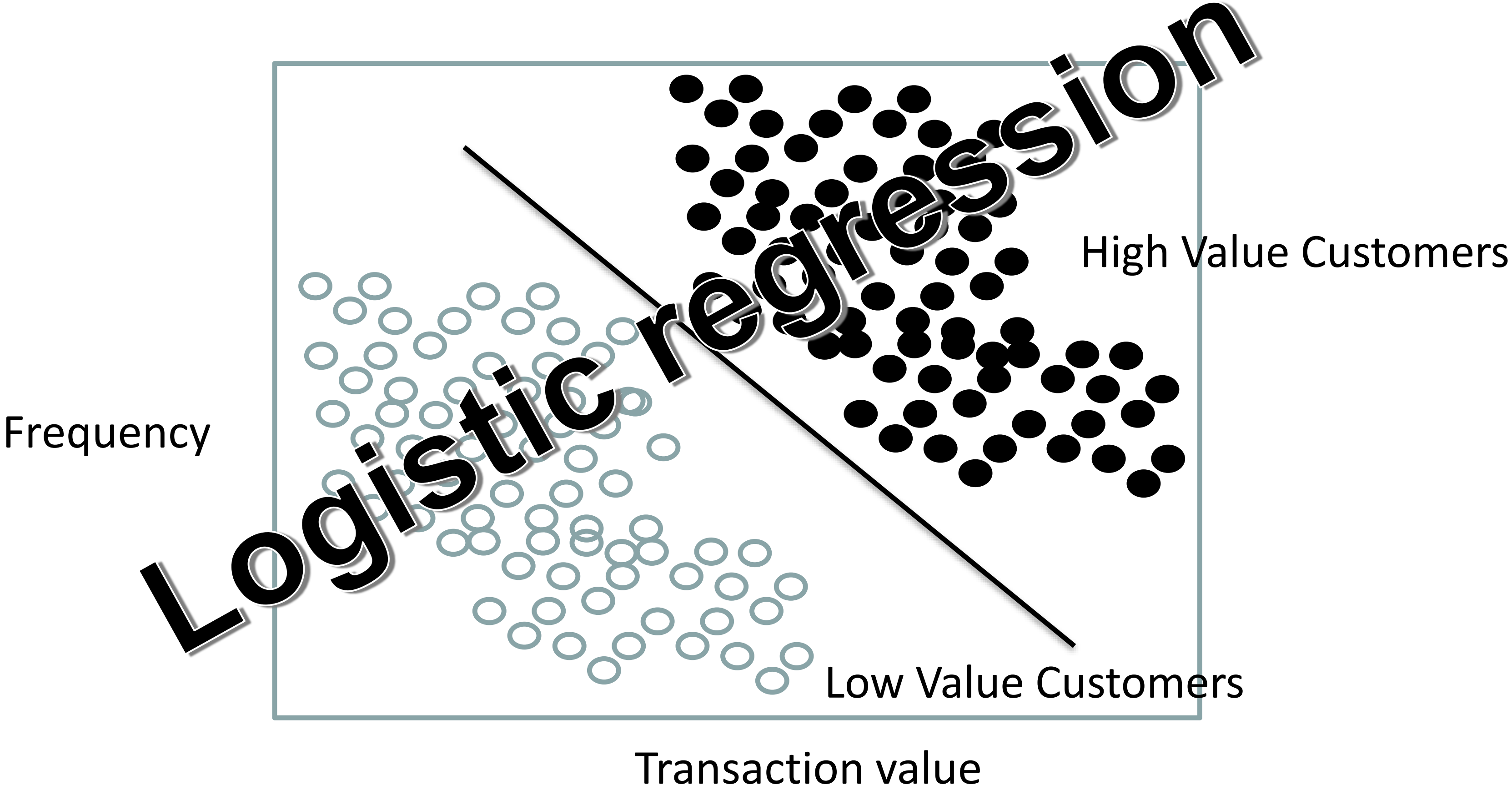




MLE

- Goal is to maximize likelihood.
- In most Data Science optimizations, the goal is to find minima using calculus (minimize sum of squared errors in linear regression, and so on) or numerical techniques like Gradient Descent (minimize deviance in logistic regression, and so on).
- Maximum Likelihood => Minimum of Negative Log-Likelihood.





Logistic regression

High Value Customers

Frequency

Low Value Customers

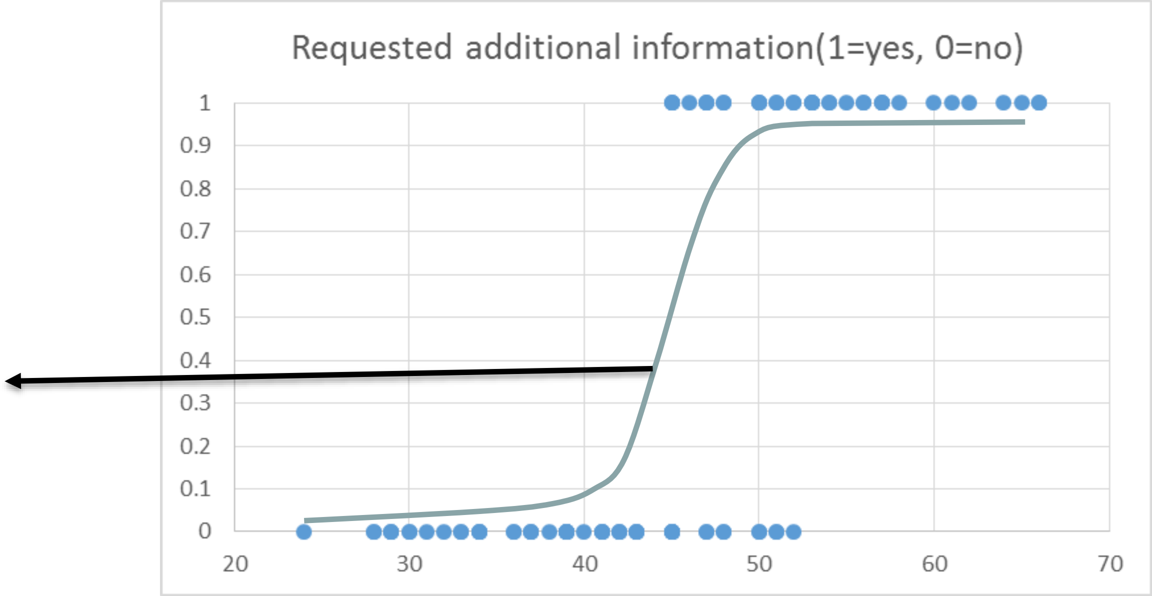
Transaction value

Example

An auto club mails a flier to its members offering to send more information regarding a supplemental health insurance plan if the member returns a brief enclosed form.

Can a model be built to predict if a member will return the form or not?

Example



$$f(x) = p = \frac{1}{1 + e^{-\mu}} = \frac{e^{\mu}}{1 + e^{\mu}}$$

where $\mu = \beta_0 + \beta_1 x_1$ (also known as the systematic or the structural component or linear predictor).

This is a logistic model. The function is also known as the inverse link function, which links the response with the systematic component.

p is the probability that a club member fits into group 1 (returns the form; success; P(Y=1|X)).

Logistic model

$$f(x) = p = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}$$

Odds Ratio is obtained by the probability of an event occurring divided by the probability that it will not occur.

Logistic model can be transformed into an odds ratio:

$$S = Odds\ ratio = \frac{p}{1 - p}$$

Attention Check – Probability and Odds

If the probability of winning is 6/12, what are the odds of winning?	1:1 (Note, the probability of losing also is 6/12)
If the odds of winning are 13:2, what is the probability of winning?	13/15
If the odds of winning are 3:8, what is the probability of losing?	8/11
If the probability of losing is 6/8, what are the odds of winning?	2:6 or 1:3

Logistic model

$$S = Odds\,ratio = \frac{p}{1-p}$$

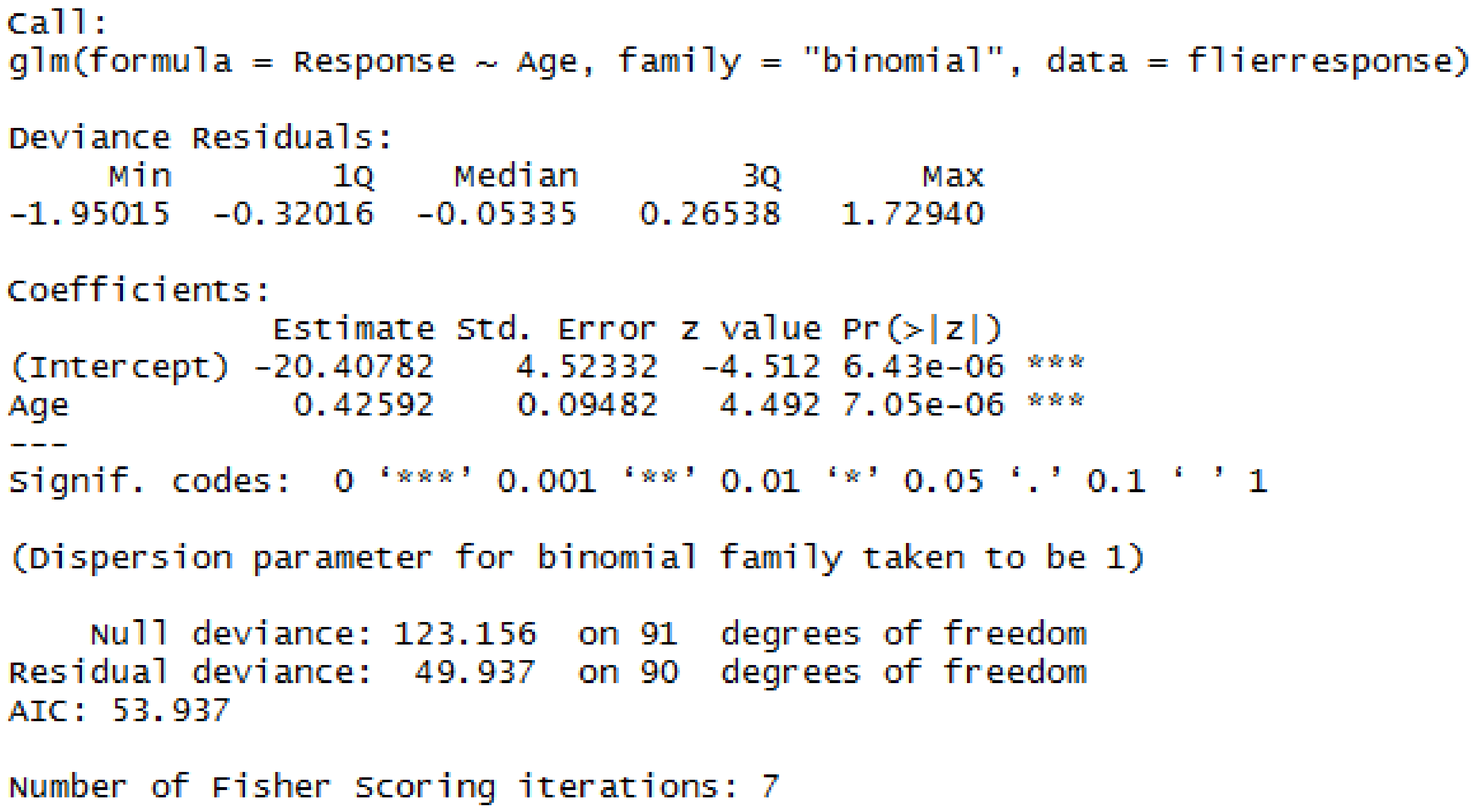
$$S = \frac{\dfrac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}}{1 - \dfrac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}} \qquad \therefore, S = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}$$

$$\ln(S) = \ln\left(e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Logistic model

The log of the odds ratio is called logit, and the transformed model is linear in βs.

R and Interpreting the output



```
Call:
glm(formula = Response ~ Age, family = "binomial", data = flierresponse)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.95015  -0.32016  -0.05335   0.26538   1.72940

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -20.40782    4.52332  -4.512 6.43e-06 ***
Age           0.42592    0.09482   4.492 7.05e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.156  on 91  degrees of freedom
Residual deviance:  49.937  on 90  degrees of freedom
AIC: 53.937

Number of Fisher Scoring iterations: 7
```

What is the logit equation?

$$\ln(S) = -20.40782 + 0.42592Age$$

Determining Logistic Regression Model

Suppose we want a probability that a 50-year old club member will return the form.

$$\ln(S) = -20.40782 + 0.42592 * 50 = 0.89$$
$$S = e^{0.89} = 2.435$$

The odds that a 50-year old returns the form are 2.435 to 1.

Determining Logistic Regression Model

$$\hat{p} = \frac{S}{S+1} = \frac{2.435}{2.435+1} = 0.709$$

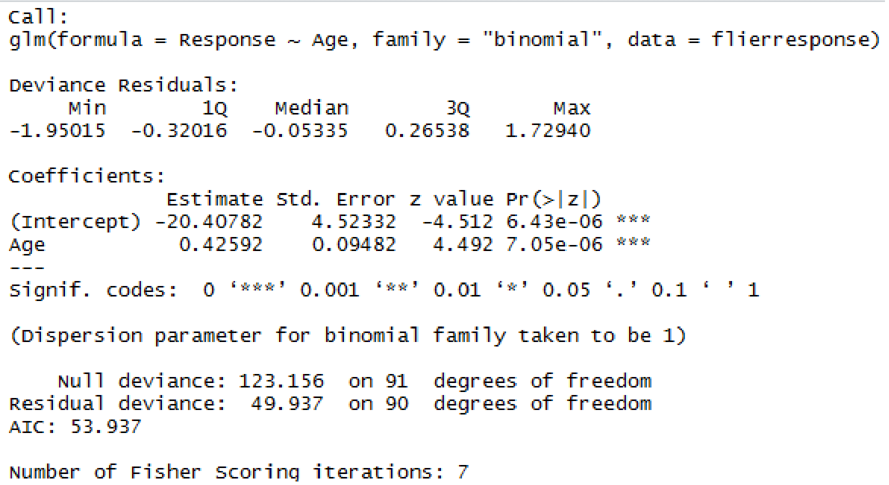
Using a probability of 0.50 as a cutoff between predicting a 0 or a 1, this member would be classified as a 1.

Interpreting Output - Deviances

Deviance or **Residual Deviance** is *similar to SSE* in the sense it measures how much remains unexplained by the model built with predictors included.

$$D = -2LL,$$

where LL is the log-likelihood.



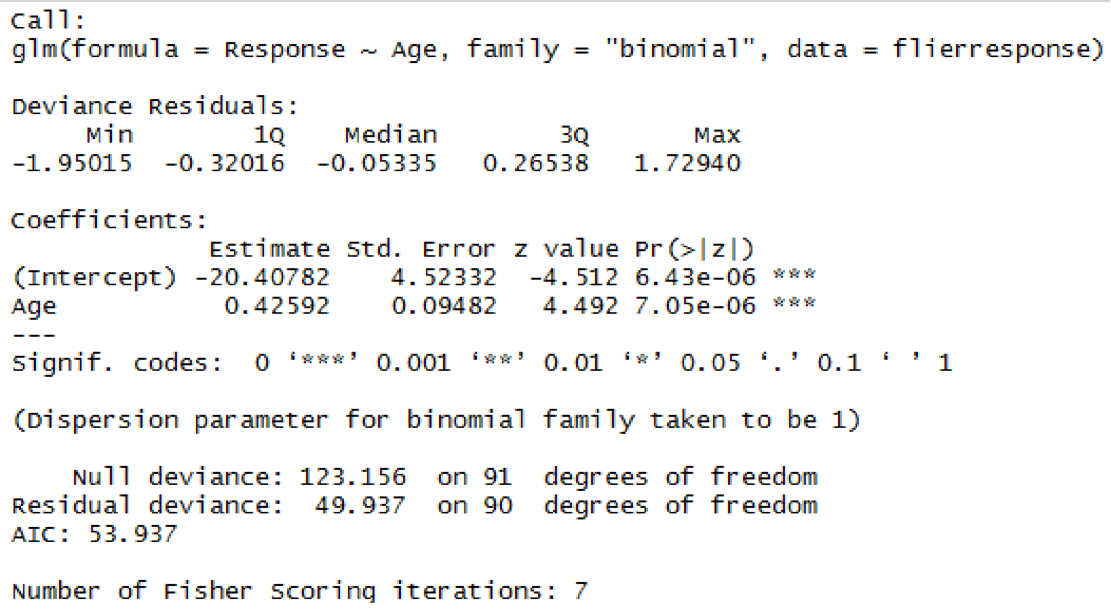
```
Call:
glm(formula = Response ~ Age, family = "binomial", data = flierresponse)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-1.95015  -0.32016  -0.05335   0.26538   1.72940  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -20.40782    4.52332  -4.512 6.43e-06 ***
Age           0.42592    0.09482   4.492 7.05e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.156  on 91  degrees of freedom
Residual deviance:  49.937  on 90  degrees of freedom
AIC: 53.937

Number of Fisher Scoring iterations: 7
```

Null Deviance shows how well the model predicts the response with only the intercept as a parameter. The intercept is the logarithm of the ratio of cases with *y=1* to the number of cases with *y=0*. This is *similar to SST*, which gives total variation when all coefficients are zero (null hypothesis).

Interpreting Output – Testing the Overall Model

The *z*-values and the associated *p*-values provide significance of individual predictor variables.

R outputs AIC (Akaike’s Information Criterion) and you need to pick the model with the lowest AIC.



```
Call:
glm(formula = Response ~ Age, family = "binomial", data = flierresponse)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-1.95015  -0.32016  -0.05335   0.26538   1.72940  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -20.40782    4.52332  -4.512 6.43e-06 ***
Age           0.42592    0.09482   4.492 7.05e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.156  on 91  degrees of freedom
Residual deviance:  49.937  on 90  degrees of freedom
AIC: 53.937

Number of Fisher Scoring iterations: 7
```

Interpreting Output – Testing the Overall Model

- AIC provides a means for model selection.
- ***AIC = D + 2k***, where k is the # of parameters in the model including the intercept. Recall in Linear Regression, it is calculated as ***AIC = nln(RSS/n) + 2k***.
- AIC is *similar to Adjusted R*^{2} in the sense it penalizes for adding more parameters to the model.
- It offers a relative estimate of the information lost when a model is used to represent the process that generated the data.
- It does not test a model in the sense of null hypothesis and hence doesn’t tell anything about the quality of the model. It is only a relative measure between multiple models.

Applications

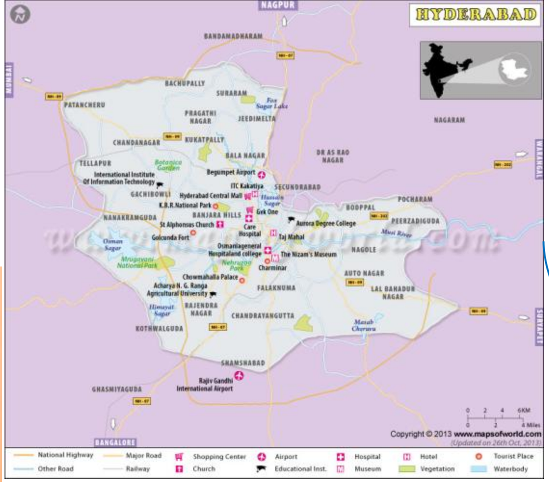
- Predicting stock price movement (up/down)
- Predict whether a patient has diabetes or not
- Predict whether a customer will buy or not
- Predict the likelihood of loan default

Diagnostic Hints

- Coefficients that tend to infinity could be a sign that an input is perfectly correlated with a subset of your responses. Or put another way, it could be a sign that this input is only really useful on a subset of your data, so perhaps it is time to segment the data.

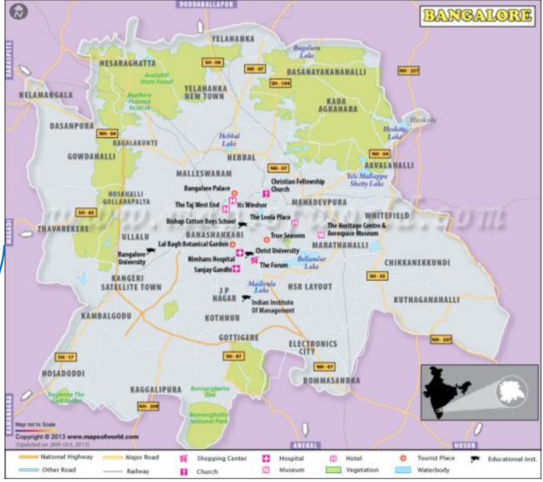
Diagnostic Hints

- Overly large coefficient magnitudes, overly large error bars on the coefficient estimates, and the wrong sign on a coefficient could be indications of correlated inputs.
- VIF can be used to check for multicollinearity. "car" package in R outputs a Generalized Variance Inflation Factor, which is obtained by correcting VIF to the degrees of freedom for categorical predictors. $GVIF = VIF^{\left(\frac{1}{2*df}\right)}$


HYDERABAD
NAGPUR
SECUNDERABAD
GACHIBOWLI
Charminar
Rajiv Gandhi International Airport
BANGALORE
Copyright © 2013 www.mapsofworld.com

INSOFE

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

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.