Equation for the Sample Regression Line: Example

$$\hat{Y}_i = b_0 + b_1 X_i$$

$$= 1636.415 + 1.487 X_i$$

From Excel Printout:

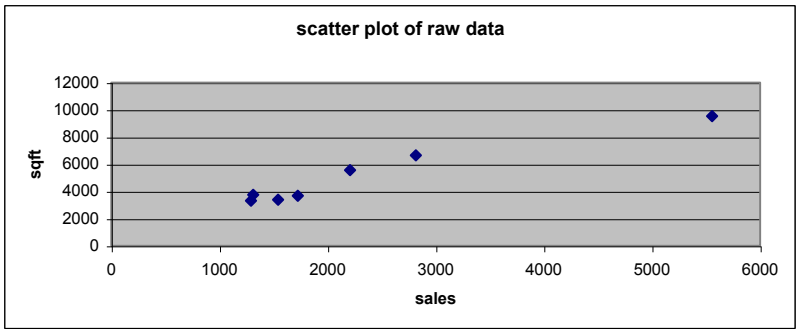
	Coefficients
Intercept	1636.414726
X Variable	1.486633657

These coefficients are generated by Excel.

[Trend line in Excel is a quick why to get a least squares estimate. Select option for equation and R2 value.]

The intercept CI is of little value, implies x=0 point. We can select a particular CI level in

1	1726	3681
2	1542	3395
3	2816	6653
4	5555	9543
5	1292	3318
6	2208	5563
7	1313	3760



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.970557204
R Square	0.941981286
Adjusted R Square	0.930377543
Standard Error	611.7515173
Observations	7

this is r, correlation coeff

R^2 = .94 tells us that 94% of the variation in price Y is explained by the square footage X. 94% is very good.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	30380456.12	30380456.12	81.17909015	0.000281201
Residual	5	1871199.595	374239.919		
Total	6	32251655.71			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1636.414726	451.4953308	3.624433331	0.015148819	475.8090301	2797.020422	475.8090301	2797.020422
sqft	1.486633657	0.164999212	9.009943959	0.000281201	1.062489679	1.910777635	1.062489679	1.910777635
							confidence intervals	

These are the b estimates of betas.

The intercept is of little importance. The slope term tells us Y will increase about 1.48 for every 1 increase in x.

sales up 1.48 units per sqft

the intercept ci is of little value, implies x=0 point.

The data analysis|regression dialog box. What would negative value for ci mean? Such as -.5 and 1.5? Well it crosses 0 which means beta 1 may actually be 0 meaning there is no relationship (flat line). This would say there is no relationship between the two variables. MUST LOOK FOR THIS IN THE CI's. FOR BETA 1 THIS IS IMPORTANT. Would mean x has no effect on y, model may not be valid. for beta 0 this is not really important. We find: Yest. = 1636.415+1.487Xi

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.829626548
R Square	0.688280209
Adjusted R Square	0.688280209
Standard Error	0.688280209
Observations	0.688280209

R^2 says 68.82% of variation is explained by the model. This is a measure of how close the points lie to the line.

This is similar to a P-value.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	21386.83972	21386.83972	117.0244949	5.01324E-15
Residual	53	9686.027756	182.7552407		
Total	54	31072.86748			

Used in Hypothesis Test.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	109.0013634	6.35102954	17.1627864	2.68072E-23	96.26281162	121.7399152	96.26281162	121.7399152
Average Temperature	-1.235404432	0.114201226	-10.81778604	5.01324E-15	-1.464463079	-1.006345785	-1.464463079	-1.006345785

b_1 coefficient Confidence Intervals cannot cross 0. If it does there is a chance the X & Y data has no relationship.

Perform all these checks each time a regression analysis is preformed in order to check the validity of the model.

If the b_1 confidence interval crosses 0 then β_1 may very well be 0 and there is no correlation between the X and Y data. Weather or not the b0 confidence interval crosses 0 is of no importance because the b0 intercept is of very little significance.

P-Value of Coefficients

$H_0: \beta = 0$

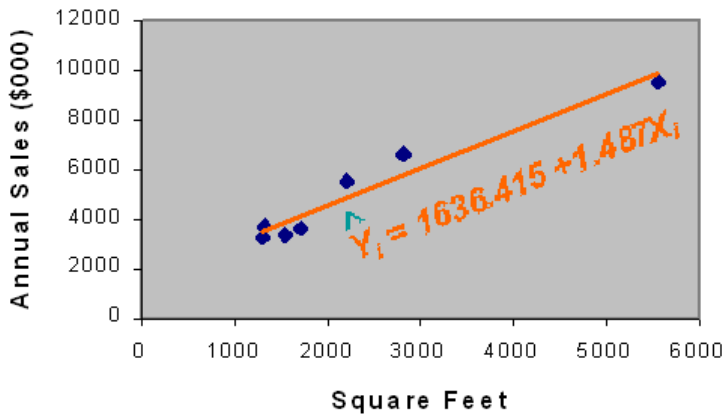
$H_1: \beta \neq 0$

In this case we always use a 2-sided test.

We always test equal / not-equal against 0.

The P-value is used to validate the coefficients. What is important here is the P-Value. It is used to evaluate the hypothesis which is that the intercept or slope (β_0 or β_1) is equal to 0. Compare it to the α which has been established. WE ARE EVALUATING THE POPULATION COEFFICIENTS β_0 and β_1, not the estimates! When the P-Value is low we reject the null hypothesis in favor of the alternative hypothesis. When we accept the null we are saying that the statistic is significant and that the beta value is highly unlikely to be 0. Keep in mind that β_1 is the more important term here, the slope. In many cases the intercept has no meaning. Slope of 0 indicates horizontal line which means no correlation! The P-value corresponds to the event of 0 being in the confidence interval.

Graph of the Sample Regression Line: Example



Easy way to plot a regression line is to create a scatter plot in Excel and then add a trend line. Use the linear and ask for the equation and the R^2 value.

Interpretation of Results: Example

$$\hat{Y}_i = 1636.415 + 1.487X_i$$

The slope of 1.487 means that for each increase of one unit in X, we predict the average of Y to increase by an estimated 1.487 units.

The model estimates that for each increase of one square foot in the size of the store, the expected annual sales are predicted to increase by $1487.

How Good is the regression?

- R^2
- Confidence Intervals
- Residual Plots
- Analysis of Variance
- Hypothesis (*t*) tests

How good is this model we have built? The 5 items in the list to the left will answer this question. Is there still a lot of random error? Is there still a lot of variation? If these conditions exist the model may not be very good and there may be relationships in the data we have not seen.

Measure of Variation: The Sum of Squares

$$SST = SSR + SSE$$

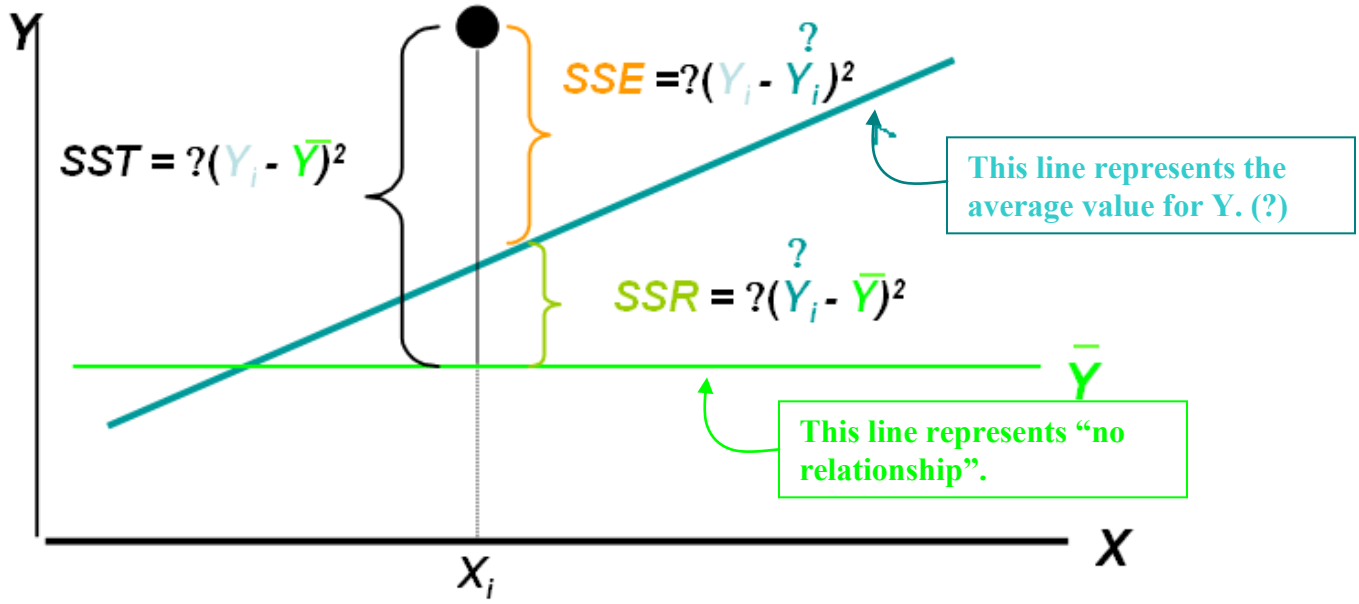
Total Sample Variability = **Explained Variability** + **Unexplained Variability**

(Refer to graph on next page)

Measure of Variation: The Sum of Squares

- SST = total sum of squares
  - Measures the variation of the Y_i values around their mean Y
- SSR = regression sum of squares
  - Explained variation attributable to the relationship between X and Y
- SSE = error sum of squares
  - Variation attributable to factors other than the relationship between X and Y

Measure of Variation: The Sum of Squares



How good is our Y line? Well the line is a good line if the points (observations) are very close to it. SSE, sum of the squared errors, is a measure of this. Excel is solving to minimize the SSE. The best line would have all zero error values and SSE would be 0. The maximum error would be the case where there is no correlation and all the error points are some random distance from the horizontal line.

SSE = Sum of Square Error.

The Coefficient of Determination

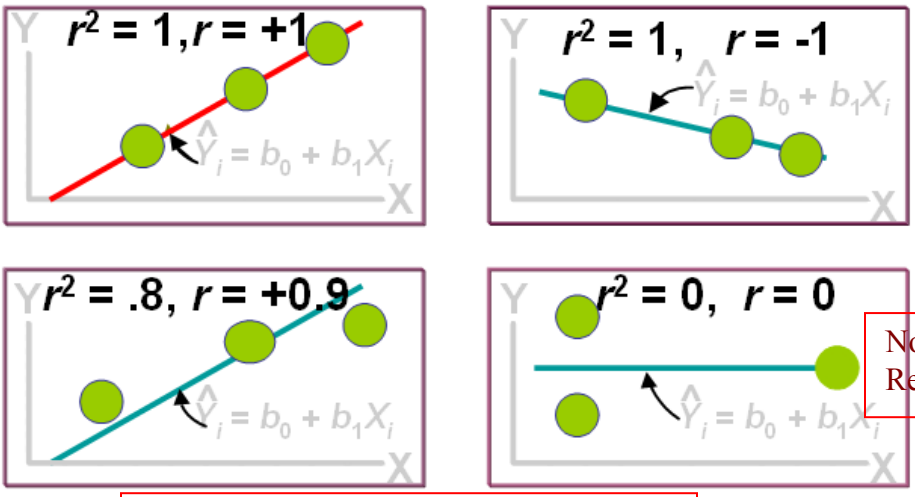
- $$r^2 = \frac{SSR}{SST} = \frac{\text{Regression Sum of Squares}}{\text{Total Sum of Squares}}$$

- Measures the proportion of variation in Y that is explained by the independent variable X in the regression model

r^2 = the percent of variation explained by the model.

Coefficients of Determination (r^2) and Correlation (r)

IDEAL PERFECT



No Relationship

R is the correlation coefficient.

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

r^2 tells you how good model is or how much randomness is left in the data.

r = 0 means all of our residual errors (SSE's) would be equal to the SST's. In the r=0 case the model is complete junk, it predicts nothing, no relationship between x and y. The minimum value for r is 0 (no relationship) and the maximum value is 1 (ideal, perfect relationship). The closer to 1 the better because we have lower residuals. In business r^2 around .7 is considered good.

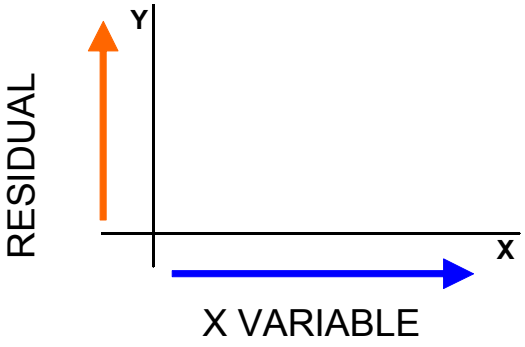
Linear Regression Assumptions

1. Linearity
2. Normality — Residuals are distributed about the line in a Normal distribution
   - Y values are normally distributed for each X
   - Probability distribution of error is normal
3. Homoscedasticity (Constant Variance)
4. Independence of Errors

We can examine these properties through Residual Analysis (check the residual box in Regression Analysis). To analyze we must plot the residuals.

Residual Analysis

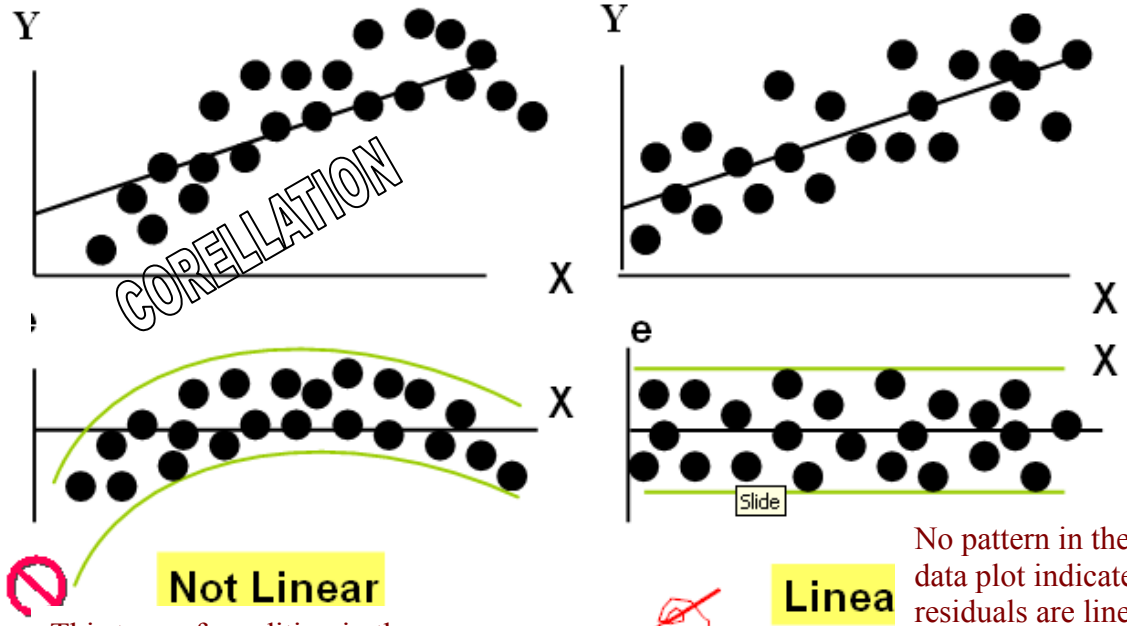
- Purposes
  - Examine linearity
  - Evaluate violations of assumptions
- Graphical Analysis of Residuals
  - Plot residuals vs. X_i, Y_i and time



The XY plot of the residuals gives us a graphical picture of the distance of each data point from the estimate line.

MOST IMPORTANT IS TO PLOT THE RESIDUALS AND CHECK FOR PATTERN.

Residual Analysis for Linearity



The horizontal line represents the r2=0 condition. (no correlation).

Not Linear

This type of condition in the residual analysis will lower the R^2 value.

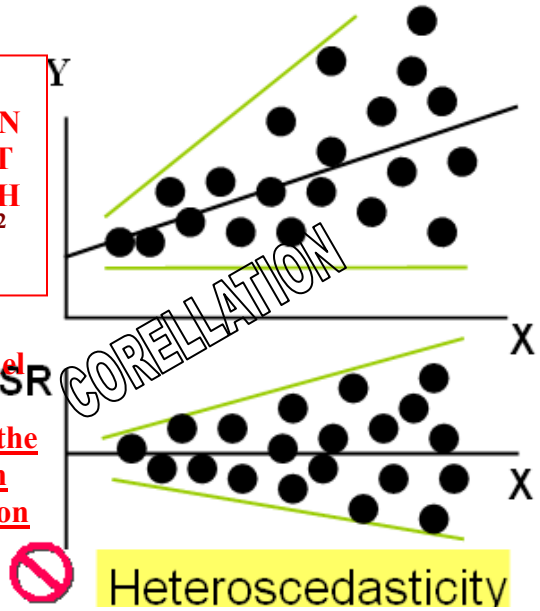
Linea

No pattern in the residual data plot indicates that the residuals are linear. **This is the condition we want.**

means **constant variance**

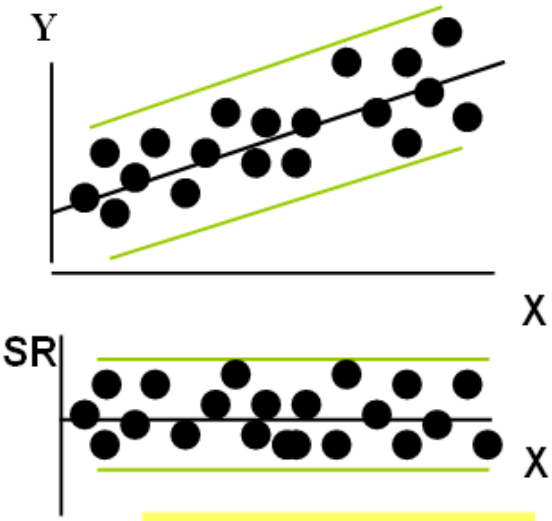
Residual Analysis for Homoscedasticity

THIS CONDITION CAN EXIST EVEN WITH A HIGH R^2 VALUE



Means model is not valid. Do not use the model when this condition exist.

Heteroscedasticity



Homoscedasticity

Horizontal line has R^2 = 0

Heteroscedasticity means non-constant variance or the absence of homoscedasticity.

NOT GOOD. The magnitude of the difference between the estimate line and the data point fans out or in. This means the error is varying with X. Error gets larger or smaller as X gets larger or smaller. This shows a correlation between X and the error, which would be very wrong!

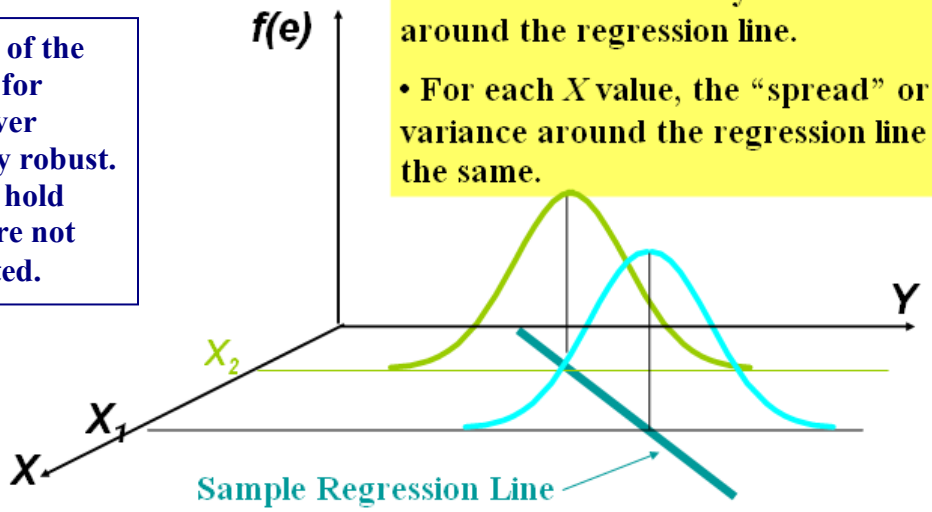
GOOD. Defined as Constant Variance in the difference between the residuals and the estimate line.

me

Variation of Errors around the Regression Line

Make a histogram of the residuals to check for Normality. However regression is pretty robust. Most assumptions hold even if residuals are not normally distributed.

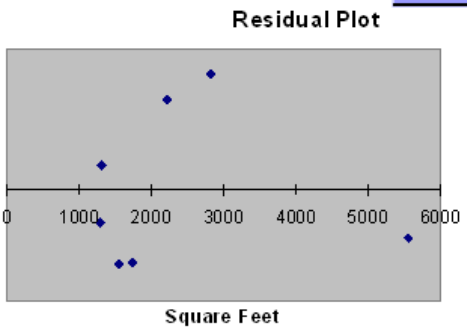
- Y values are normally distributed around the regression line.
- For each X value, the "spread" or variance around the regression line is the same.



Residual Analysis:Excel Output for Produce Stores Example

Excel Output

Observation	Predicted Y	Residuals
1	4202.344417	-521.3444173
2	3928.803824	-533.8038245
3	5822.775103	830.2248971
4	9894.664688	-351.6646882
5	3557.14541	-239.1454103
6	4918.90184	644.0981603
7	3588.364717	171.6352829



A Histogram of these residuals would/should look normal.

It is always bad to see any kind of pattern in a residual plot. We always want to see no pattern.

Residual Analysis for Independence

Graphical Approach

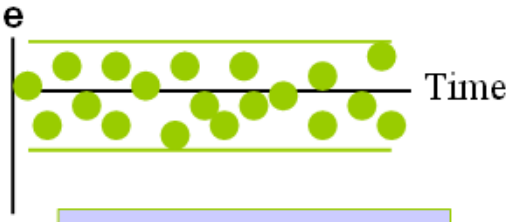
It is only necessary to perform this check when the observations are taken over time such as GDP values per quarter.

Not Independent

Cyclical Pattern

The BAD condition. Do NOT use.

Independent



No Particular Pattern

The GOOD condition.

Residual is plotted against time to detect any autocorrelation

Note we are checking the observations with respect to **TIME!**

HORIZONTAL LINE HAS R^2 VALUE OF 0.

Check for Independence is the last use of the residual data.

The ANOVA Table in Excel

ANOVA					
	df	SS	MS	F	Significance F
Regression	p	SSR	MSR =SSR/p	MSR/MSE	P-value of the F Test
Residuals	n-p-1	SSE	MSE =SSE/(n-p-1)		
Total	n-1	SST			

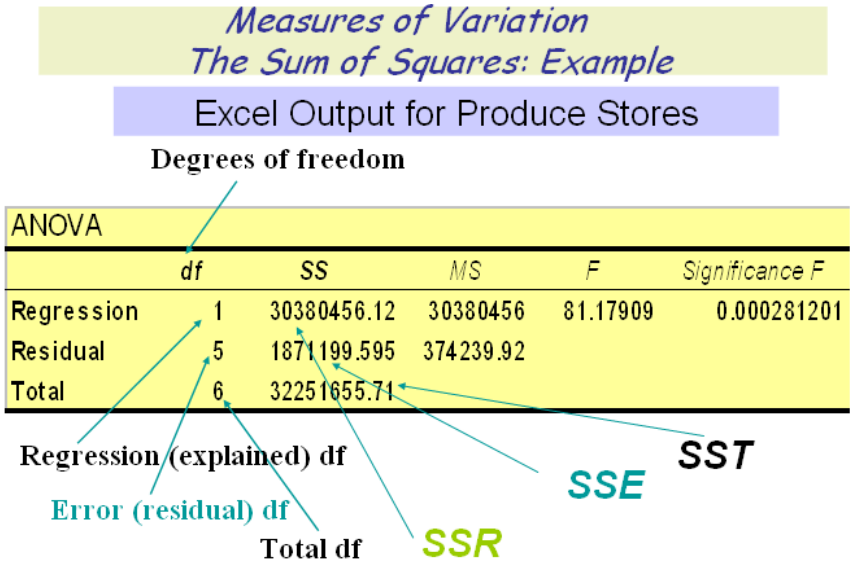
Another measure of the validity or significance of the model is the **ANOVA Significance F** value which is also a P-value.

H_0: model is junk (large Significance F value)

H_1: model is not junk(small Significance F value)

A small Significance F says reject the null and accept the alternative hypothesis. Compare the Significance F to α which is usually set at .05.

Significance F is also a P-value and we are also using it to evaluate a hypothesis. The null hypothesis in this case is, in non-technical jargon, is weather the model is "H_0: model is junk" or "H_1: model not junk". If the value is high, say 0.1, we will (in most cases) not reject the null hypothesis meaning we do not have evidence that the model we've constructed in invalid. If we have a small P-value, very small, we conclude that the model is not junk, there is a relationship between the data. For instance, square footage does explain sales.



We have seen that we can have one dependent variable. In a similar fashion we will see how we can have multiple dependent variables in Multiple Regression. In the last case we will demonstrate how to predict the sales price of a house based on square footage, architectural style, number of bedrooms, fireplaces… Many factors may come into play to predict the sales price of a house. In this case we will have multiple explaining factors, not just one.

ONE X VARIABLE IS A SIMPLE EXPLAINING FACTOR AND IS THEREFORE CALLED **SIMPLE LINEAR REGRESSION**.

The point here is that once we have run a regression we should run all these little checks to see if our results make sense.

Regression Analysis Items to Check

- Go over R2 to see that it is high enough.
- Check that the coefficients make sense to you (for instance, it would be counter intuitive to come to the result that bigger stores have lower sales).
- Look for 0 in the confidence intervals.
- Look at the P-values to see if they are significantly low (which would mean that these parameters are significantly different from 0).
- Look at the residual plots and be happy if we see that there are no patterns in the plots.
  - No patterns indicates linear model
  - Constant variance if not fanning in or out
  - If the observations are taken over time we plot residuals over time and see no patterns.

The power of regression is the ability to do predictions with values for which we did not have any observation in the past.