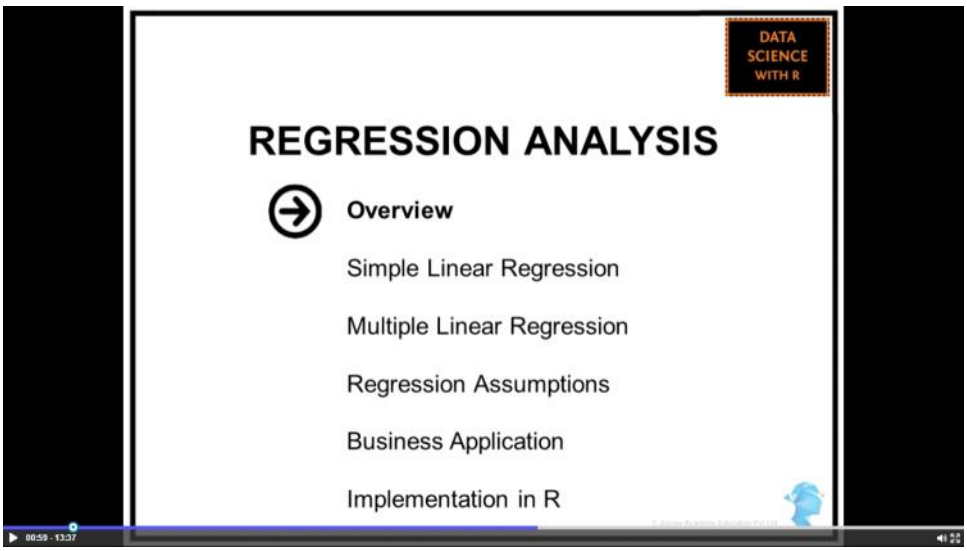
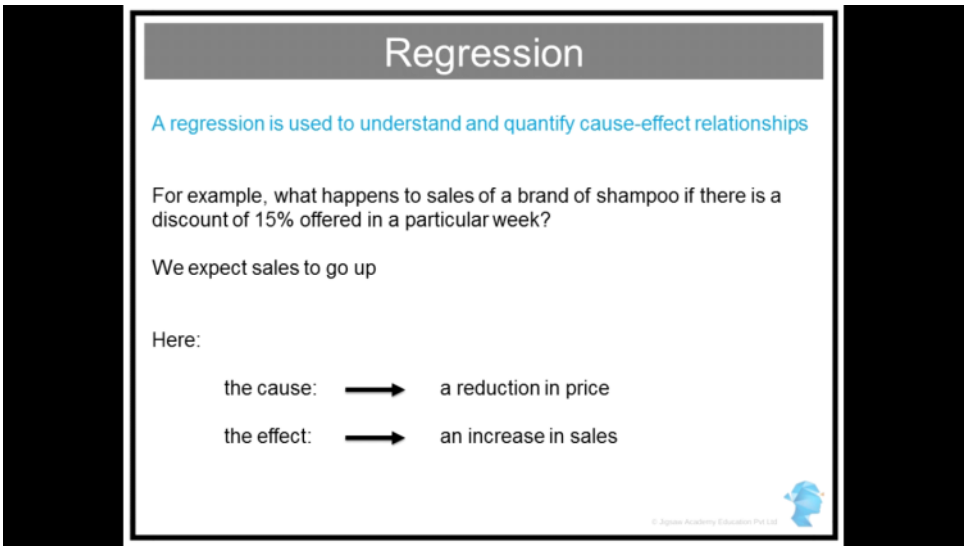
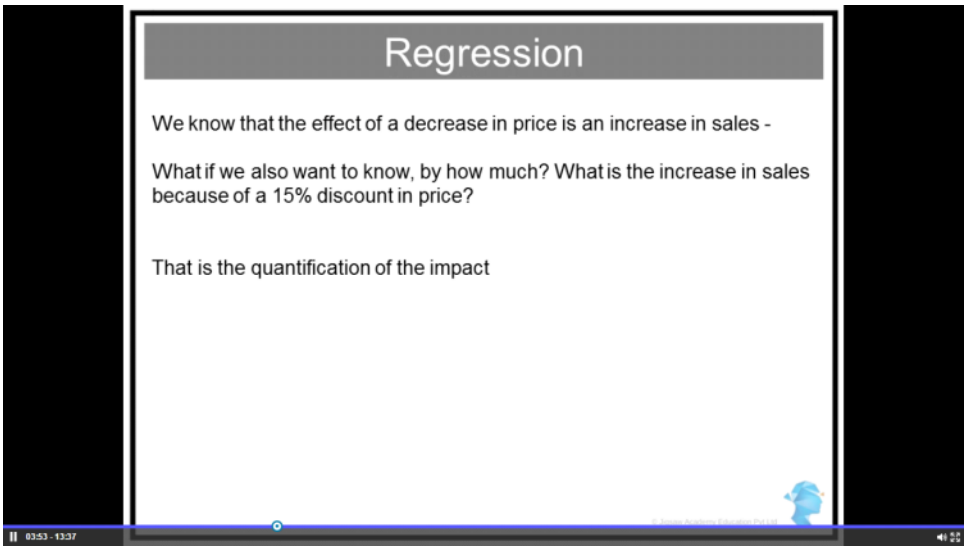
Regression Analysis

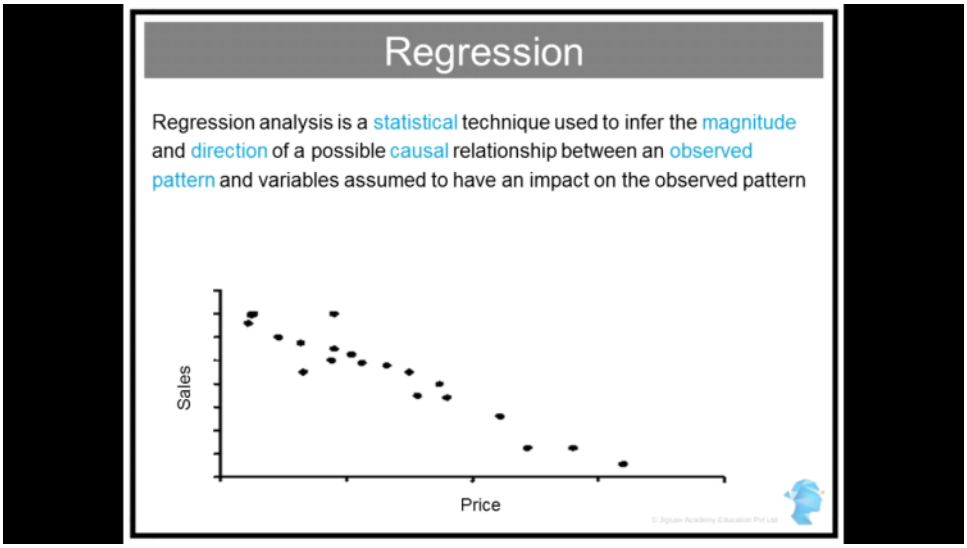
Wednesday, September 28, 2016 5:35 AM

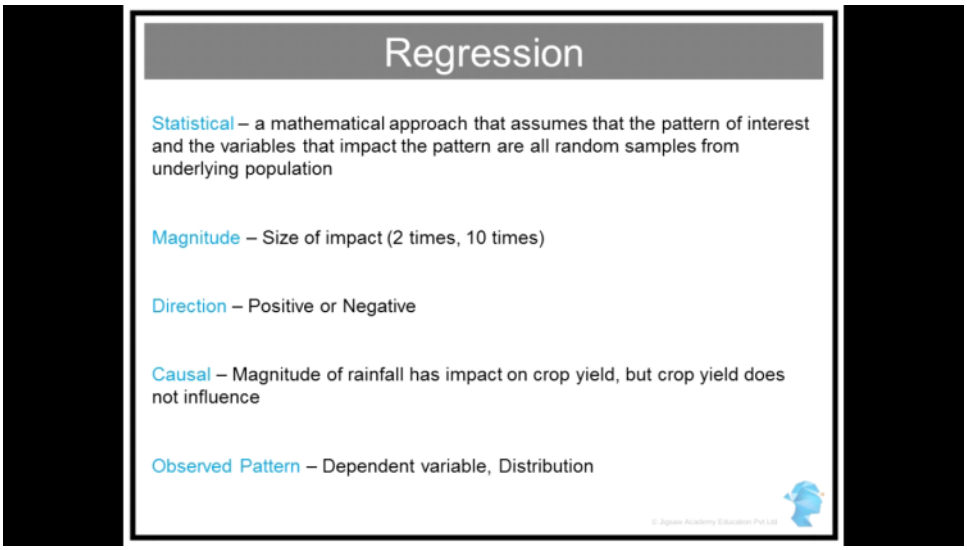






We want to know , how much of a decrease in price (discount) is neeeded for a specific increase in sales.
This is the quantification of the impact.





Regression

Example:

You work for a hospital and are looking to understand factors that may influence the birth weight of a baby

What factors would you think of?

- Maternal Diet
- Gestation period
- Maternal health issues
- Ethnicity
- Age of the mother

Use inferential statistics to evaluate sample Vs population
Data availability is variable, we might have about 70% of the data. In this situation do the best with the data that you havve.

Regression

Is it possible to analyze the population?

We can analyze a sample, and make inferences about the population based on the sample

Sample:

1115 observations from a hospital in the US

Are these the only factors that impact birth weight of a baby?

Birthweight (grams)	Weeks of gestation	Mother's education in years	Race	Smoked during pregnancy
2898	40	0	0	1
994	26	0	1	1
3977	38	2	0	0
3040	37	2	0	1
3523	38	2	0	0
3100	40	5	0	1
3670	40	6	1	0
3097	41	7	1	0
3040	39	7	1	1
3239	39	7	1	1
2955	38	8	0	0
2200	38	8	0	0
3182	40	8	1	0
3510	40	8	0	0
3381	39	8	1	1
3530	40	8	1	0
2985	38	8	1	1
3374	39	8	1	0
3765	42	8	0	0
2715	39	8	1	0
3640	39	8	1	0
3040	42	8	1	0

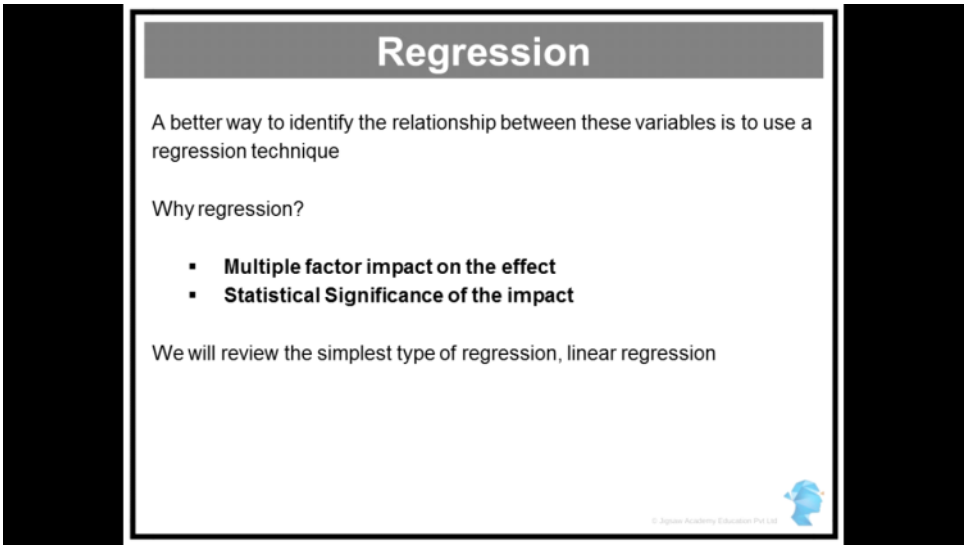
Regression

We want to analyze the relationship between the variables available & the birth weight to see causes of variation in baby birth weights:

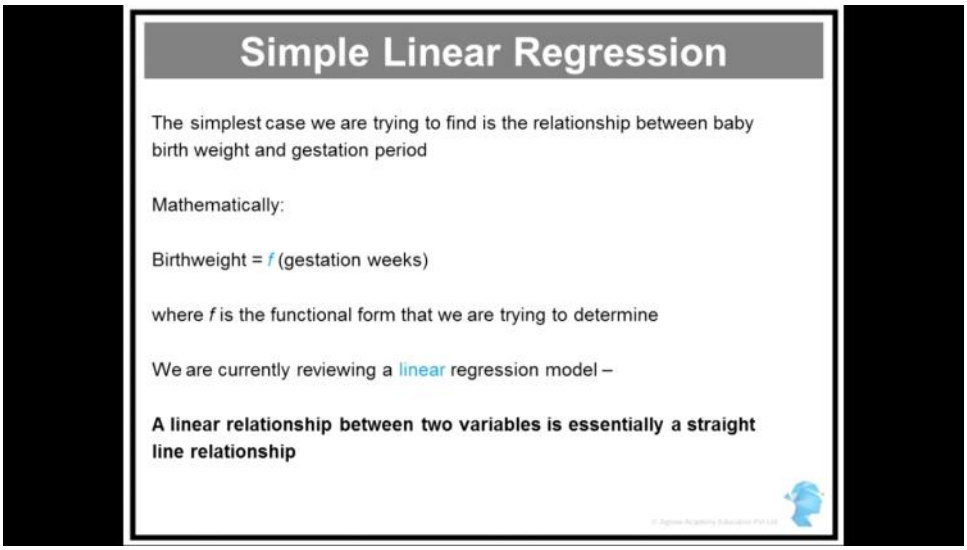
- The effect is baby birth weight
- The possible causes of baby birth weight in this dataset are gestation weeks, mother's education, race, and smoking during pregnancy

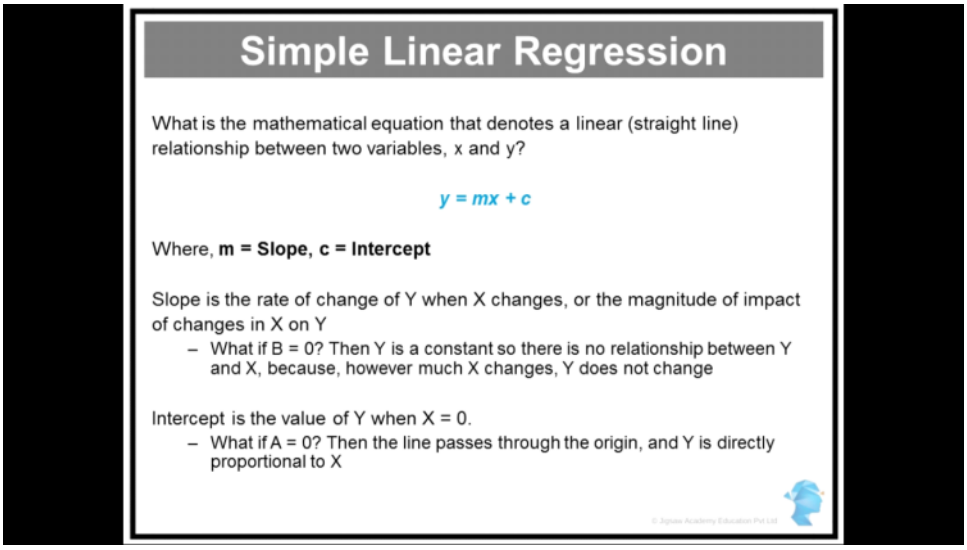
What are possible ways of assessing these relationships?

- **Graphical visualization**
- **Correlations**
- **Run regression model**



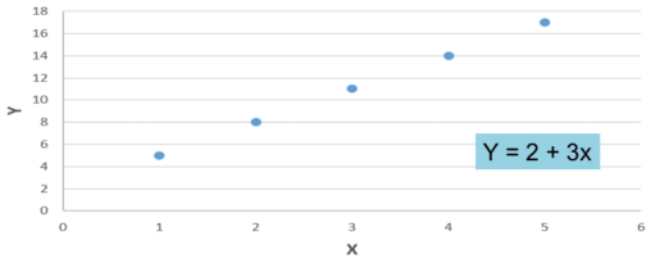
Simple Linear Regression:





Simple Linear Regression

Straight Line Relationship

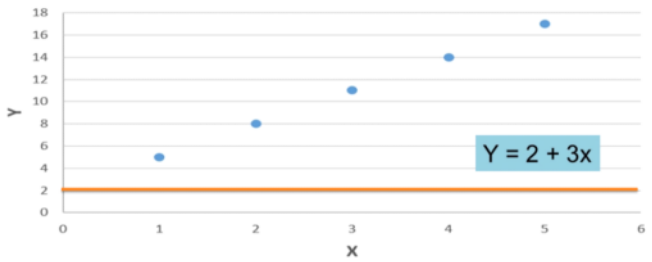


Intercept = 2
Slope = 3

For a unit change in X, Y Changes by a constant amount (3)

Simple Linear Regression

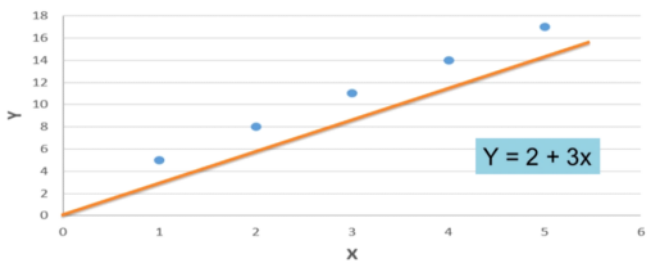
Straight Line Relationship



If m = 0?

Simple Linear Regression

Straight Line Relationship



If c = 0?

Simple Linear Regression

In a linear regression model, since we are looking at a straight line relationship, the relationship is usually shown as

$$Y = \beta_0 + \beta_1 X + e$$

where, β_0 = Intercept
β_1 = Slope
e = Error

Therefore, in order to understand the relationship between X and Y, we need to figure out what the values of the BETAs are

Simple Linear Regression

TERMINOLOGY

Dependent Variable: Y: Predicted Variables: *The variable whose behavior we hypothesize can be explained or influenced by other factors*

Independent Variable (s): X(s): Predictor(s): *The factor(s) that we hypothesize influence the dependent variable*

Beta Coefficient(s): *The estimate of magnitude of impact of changes in the predictor(s) on the predicted variable*

Error:e:u: *The impact of the unobserved variables on the dependent variable, usually calculated as the difference between the predicted value of Y given the estimated regression function and the actual value of Y*

Simple Linear Regression

In the birthweight example, we believe:

$$Birthweight = \beta_0 + \beta_1 * Gestation\ Period + e$$

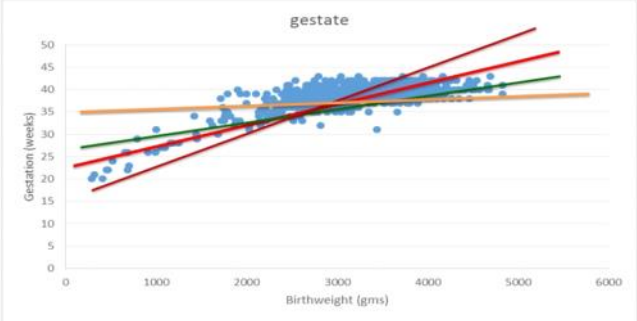
We now need to estimate what the beta coefficients values are, from the data available to us, that will best capture the relationship between Birthweight and Gestation Period

Ordinary Least Squares Regression

The Ordinary Least Squares Regression (OLS) technique estimates coefficients on the variables hypothesized to have an impact on the variable of interest by identifying the line that minimizes the sum of squared differences between points on the estimated line and the actual values of the independent variable

- **Coefficients**: Betas
- **Minimizes**: Least
- **Sum of Squared Differences:** Square of residuals
- **Estimated Line:** Regression Line
- **Actual Values:** Values in data set

OLS Regression



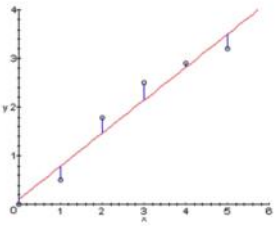
- Is there a linear relationship?
- Would it be possible to fit a straight line through these points?
- How many straight lines?

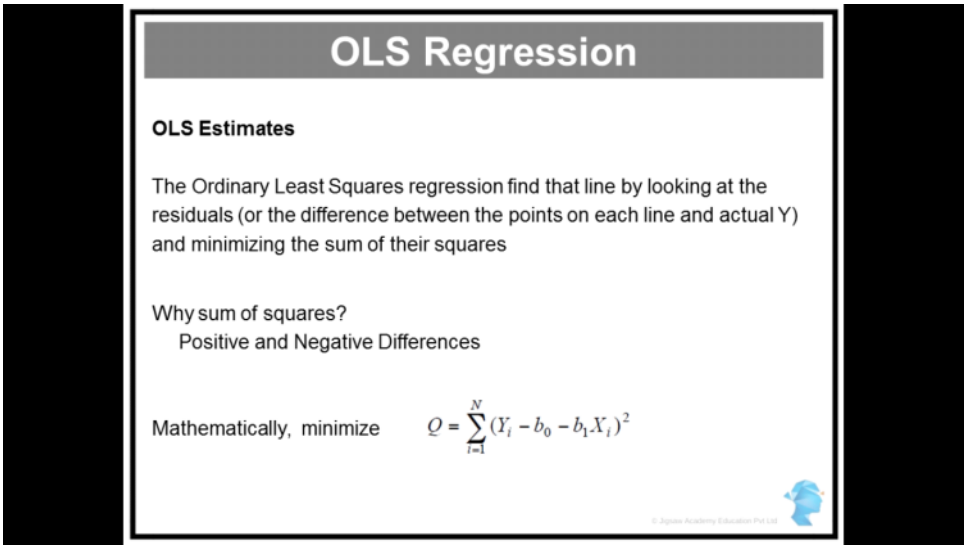
OLS Regression

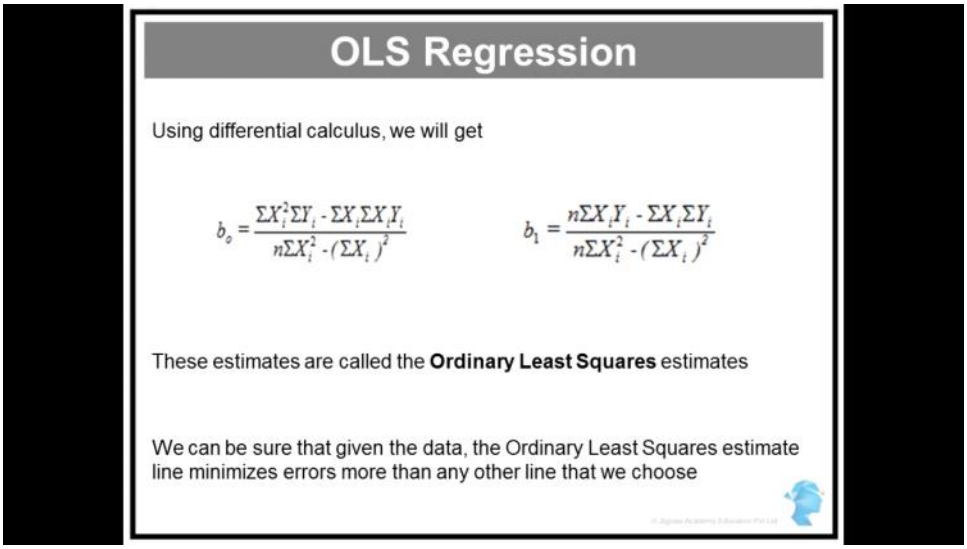
Clearly we can fit many straight lines that each will cover some of the points

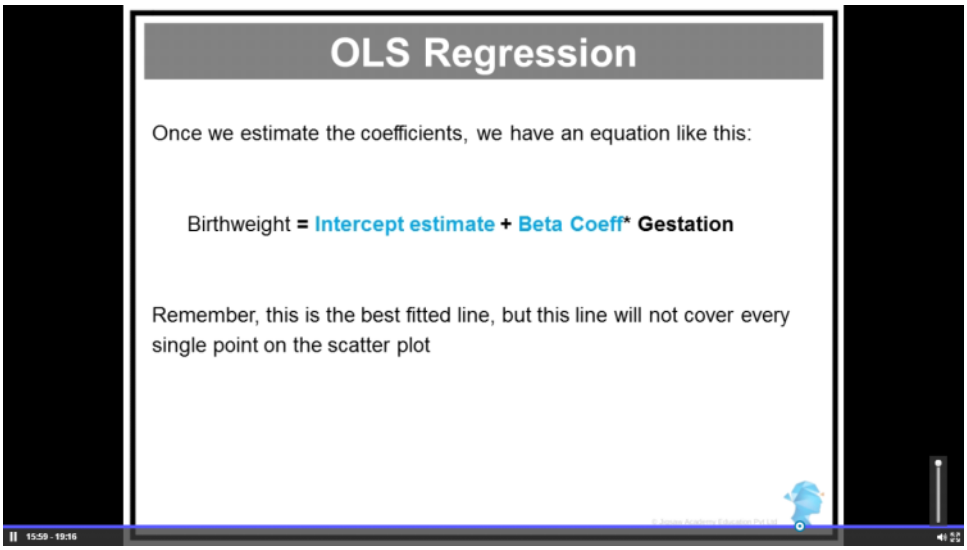
Is there a straight line that can hit all points?

One way of choosing a line among all possible lines is to identify the line that would explain most variation in Y - In other words, have least total error









Simple Linear Regression 2:

Regression

SIMPLE LINEAR REGRESSION

- ✓ Concepts - OLS
- ✓ How to Run
- ✓ **Interpret Results**

OLS Results : Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.702085646
R Square	0.492924254
Adjusted R Square	0.49246866
Standard Error	451.3259178
Observations	1115

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	220385522.7	2.2E+08	1081.938347	2.54E-166
Residual	1113	226712628.6	203695.1		
Total	1114	447098151.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
lIntercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-3632.001323	-2858.891465
ggestate	166.4462854	5.060260218	32.89283	2.54E-166	156.5175606	176.3750103

OLS Results Interpretation

Understanding the output

Starting with the bottom most table:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-
gestate	166.4462854	5.060260218	32.89283	2.54E-166	

We want to estimate a straight line that best captures the relationship between Birthweight and Gestation period –

As per this model that straight line is:

Birthweight = -3245.44 + 166 * Gestate

OLS Results Interpretation

	Coefficients	Standard Error	t Stat	P-value	
Intercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-
gestate	166.4462854	5.060260218	32.89283	2.54E-166	

Birthweight = -3245.44 + 166 * Gestate

Here, the intercept is -3245. 44
the beta coefficient on Gestate is 166

How do we interpret the Beta Coefficient?

For a unit increase in gestation period (1 week), the average increase in birthweight is 166

OLS Results Interpretation

BETA Coefficient:

- For every unit increase in Gestation Period, we expect to see an increase in Birthweight by 166 grams
- Positive sign on the coefficient on gestation implies a positive relationship between Gestation Period and Birthweight
- What does unit increase mean?
- Will every additional week of gestation automatically add 166 grams of birthweight to every baby?

OLS Results Interpretation

How do we interpret the estimated regression function?

Birthweight = -3245.44 + 166 * Gestate

INTERCEPT

With zero gestation weeks, we expect birthweight to be Negative

- It doesn't really make sense to talk about birthweight at Zero weeks
- Provides a baseline

OLS Results Interpretation

The second piece of critical information from the coefficients table is the P values

P-values denote the probability of rejecting the null hypothesis when it is in fact true

OLS Results Interpretation

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-3245.446394	197.0110519	-16.4734	9.95259E-55 -
gestate	166.4462854	5.060260218	32.89283	2.54E-166

Birthweight = -3245.44 + 166 * Gestate

Here, the intercept is -3245. 44
the beta coefficient on Gestate is 166

How do we interpret the Beta Coefficient?

For a unit increase in gestation period (1 week), the average increase in birthweight is 166

Since p value is close to zero we reject the null hypthesis(the influence of gestation on weight is zero)

OLS Results Interpretation

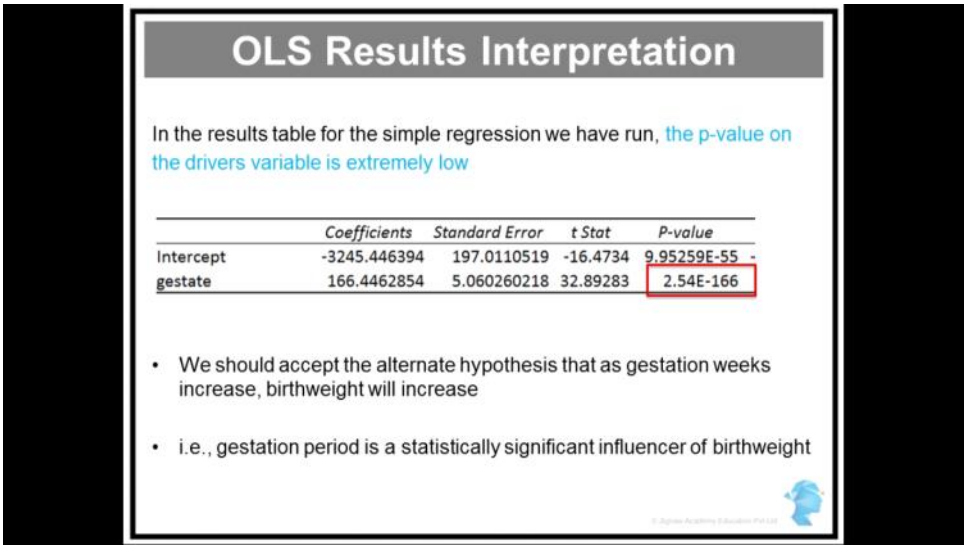
How do we actually test the hypothesis?

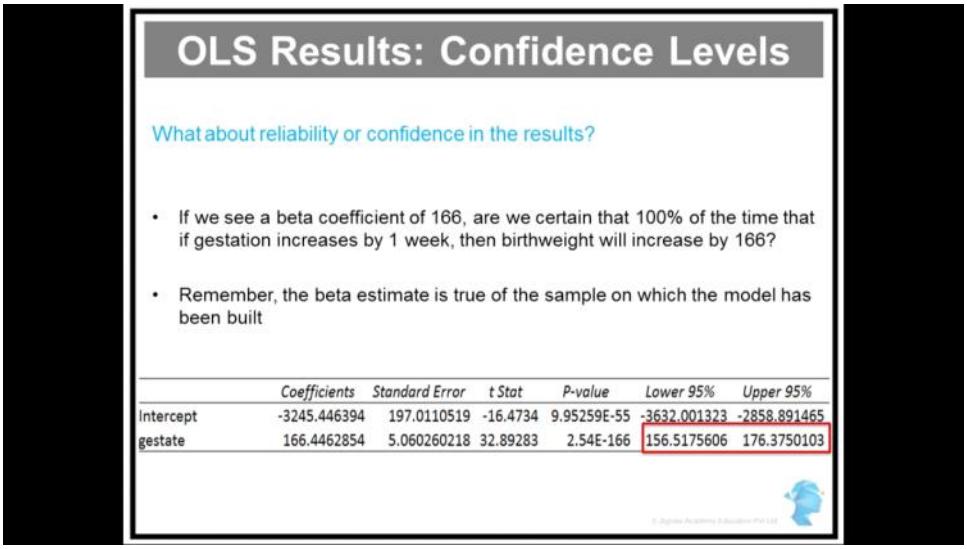
One main point to remember is that we can show the distribution of

$$(\hat{\beta}_j - \beta_j) / se(\hat{\beta}_j) \sim t_{n-k-1}$$

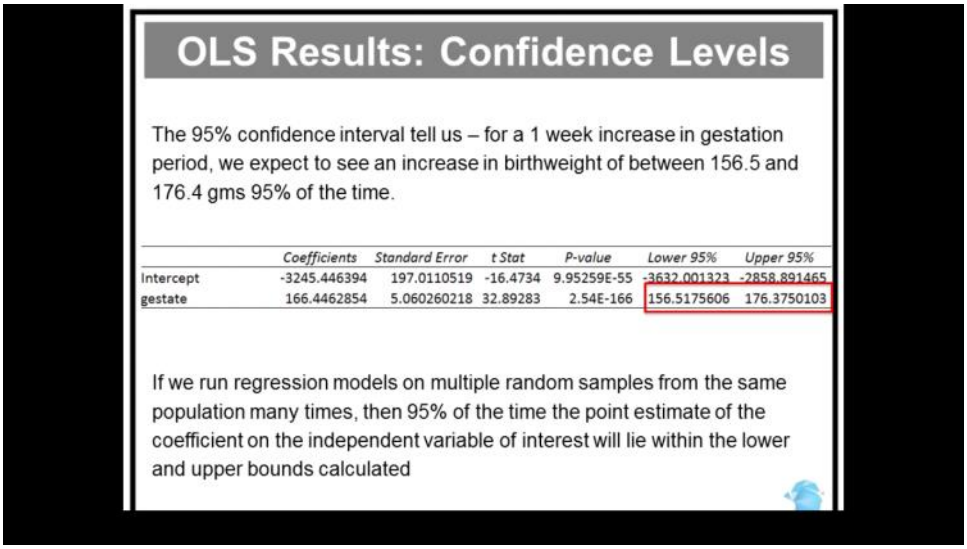
What does the above equation mean?

The difference between the estimated coefficient and the actual value in the population divided by the standard error of the estimated population is distributed as a t-distribution with n-k-1 degrees of freedom, where k+1 are the number of unknown parameters in the population model





This basically indicates that 95% of the time the values will be in the range mentioned above. The narrower the range, the more precise we are.



Measure of explainability --> R square (what percentage of the equation is explained by regression)

OLS Results Interpretation

While the regression equation is the best straight line equation possible, how do we assess the effectiveness of the overall model?

One way is to look at a measure of "explainability"; i.e., how much of the dependent variable Y is explained by X?
Or, a better way to put it is, how much of the variance in Y is explained by X?

The mathematical calculation is:

$$R^2 \equiv 1 - \frac{SS_{err}}{SS_{tot}}, \quad \text{Where,} \quad SS_{tot} = \sum_i (y_i - \bar{y})^2, \quad SS_{err} = \sum_i (y_i - f_i)^2,$$

OLS Results Interpretation

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.702085646
R Square	0.492924254
Adjusted R Square	0.49246866
Standard Error	451.3259178
Observations	1115

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	220385522.7	2.2E+08	1081.938347
Residual	1113	226712628.6	203695.1	
Total	1114	447098151.3		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-3245.446394	197.0110519	-16.4734	9.95259E-55
gestate	166.4462854	5.060260218	32.89283	2.54E-166

As per the above R square --> 49% of the Y variable behavior is being explained by the X variable

OLS Results Interpretation

The R2 estimate is 49%, which implies that 49% of the variation in birthweight is captured or explained by variation in the gestation weeks variable

- Clearly, the higher the R^2 the better the model
- However, R^2 is not the only indicator of model fit
- It is possible to have the same R^2 but different models with different fit
- R^2 also increases with addition of variables, whether relevant or not, it is better to use the adjusted R2 measure

OLS Results Interpretation

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.702085646
R Square	0.492924254
Adjusted R Square	0.49246866
Standard Error	451.3259178
Observations	1115

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	220385522.7	2.2E+08	1081.938347
Residual	1113	226712628.6	203695.1	
Total	1114	447098151.3		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-3245.446394	197.0110519	-16.4734	9.95259E-55
gestate	166.4462854	5.060260218	32.89283	2.54E-166

Then there is the ANOVA table.

OLS Results Interpretation

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	220385522.7	2.2E+08	1081.938347	2.54E-166
Residual	1113	226712628.6	203695.1		
Total	1114	447098151.3			

The ANOVA table shows us the output of the test of the hypothesis that at least one of the beta coefficients is different from zero

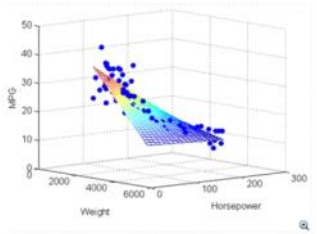
In this example, p value < 0.05, so we conclude that at least one of the beta coefficients is significant (in this case, we have only one beta)

In real life scenario we use Miltiple Linnear Regressions:

Multiple Linear Regression

In real life business situations:

- expect multiple independent variables simultaneously impacting the dependent
- We would again estimate the line across multiple dimensions that would minimize the sum of squared residuals



Multiple Linear Regression

OLS ESTIMATES

- Gestation period is not the only explanatory variable to explain differences in birthweight
- We had data on other independent variables: Mother's education level, Race, and Smoking Status
- So the actual regression equation is:

$Birthweight = \beta_0 + \beta_1 * Gestation + \beta_2 * Years\ Of\ Education + \beta_3 * Race + \beta_4 * Smoking$

Multiple Linear Regression

OLS ESTIMATES

$Birthweight = \beta_0 + \beta_1 * Gestation + \beta_2 * Years\ Of\ Education + \beta_3 * Race + \beta_4 * Smoking$

- So, we would now need to estimate 4 beta coefficients
- We would use the same OLS approach of minimizing sum of squared residuals across multiple dimensions
- It is difficult to visualize the actual process of minimizing errors in multiple dimensions (as we can do easily in 2 dimensions), but the logic of minimizing residuals is identical

Multiple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.726512578
R Square	0.527820526
Adjusted R Square	0.526118978
Standard Error	436.1074441
Observations	1115

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	235987581.2	58996895.29	310.2002602	3.9601E-179
Residual	1110	211110570.1	190189.7028		
Total	1114	447098151.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-2834.485706	215.6221999	-13.14561166	8.64385E-37	-3257.55877
YearsEduc	9.571829193	6.458197303	1.482120899	0.138591957	-3.09982207
Race (1=b)	-168.9683577	27.26026985	-6.198337677	8.03139E-10	-222.4558274
Smoke	-174.8128917	31.62426255	-5.527809271	4.04011E-08	-236.8629666
gestate	156.5115539	5.013557387	31.21766478	4.4373E-154	146.6744356

Multiple Linear Regression

	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	-2834.485706	215.6221999	-13.14561166	8.64385E-37	-3257.55877
YearsEduc	9.571829193	6.458197303	1.482120899	0.138591957	-3.09982207
Race (1=b)	-168.9683577	27.26026985	-6.198337677	8.03139E-10	-222.4558274
Smoke	-174.8128917	31.62426255	-5.527809271	4.04011E-08	-236.8629666
gestate	156.5115539	5.013557387	31.21766478	4.4373E-154	146.6744356

- What are key things to look in this output?
  Coefficient signs, and p-values
- Are all the coefficient signs as expected? What should be "expected?"
- Which of the independent variables are significant?

Multiple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.726512578
R Square	0.527820526
Adjusted R Square	0.526118978
Standard Error	436.1074441
Observations	1115

R2 is only 52%, even with the introduction of additional IVs

Multiple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.726512578
R Square	0.527820526
Adjusted R Square	0.526118978
Standard Error	436.1074441
Observations	1115

ANOVA

	df	SS	MS	F	Significance F
Regression	4	235987581.2	58996895.29	310.2002602	3.9601E-179
Residual	1110	211110570.1	190189.7028		
Total	1114	447098151.3			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	-2834.485706	215.6221999	-13.14561166	8.64385E-37	-3257.55877
YearsEduc	9.571829193	6.458197303	1.482120899	0.138591957	-3.09982207
Race (1=b)	-168.9683577	27.26026985	-6.198337677	8.03139E-10	-222.4558274
Smoke	-174.8128917	31.62426255	-5.527809271	4.04011E-08	-236.8629666
gestate	156.5115539	5.013557387	31.21766478	4.4373E-154	146.6744356

Regression Results: Model Fit

- R^2
- Fit Chart - Actual v/s Fitted Values
- MAPE – Mean Absolute Percentage Error

DATA SCIENCE WITH R

REGRESSION ANALYSIS

Overview

Simple Linear Regression

Multiple Linear Regression

Regression Assumptions

Implementation in SAS

Regression Results: Model Fit

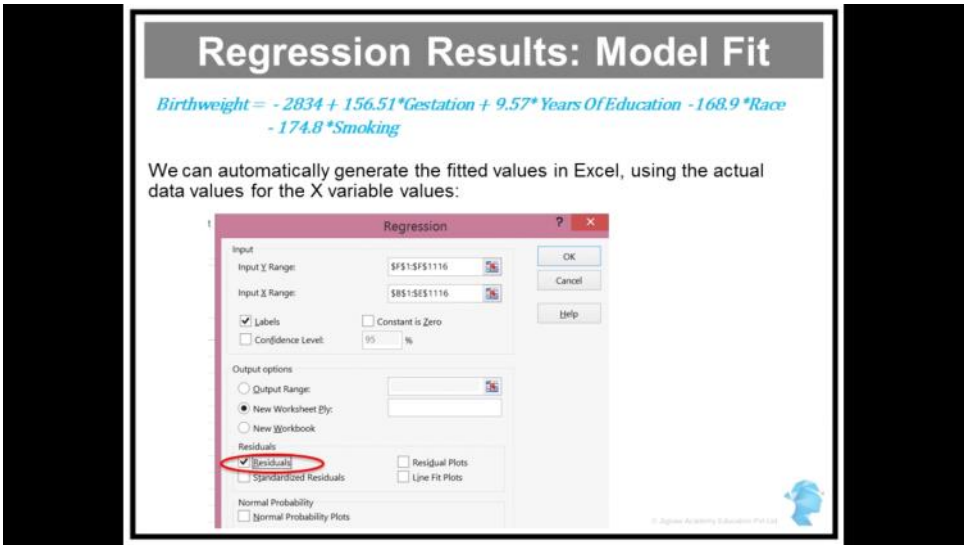
How do we validate the model?

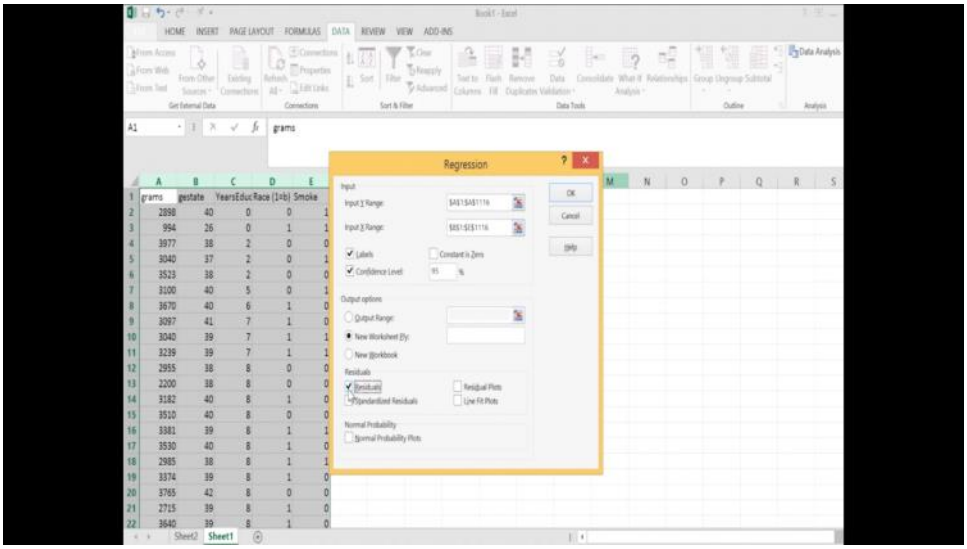
- R^2
- **Fit Chart - Actual vs Fitted Values**

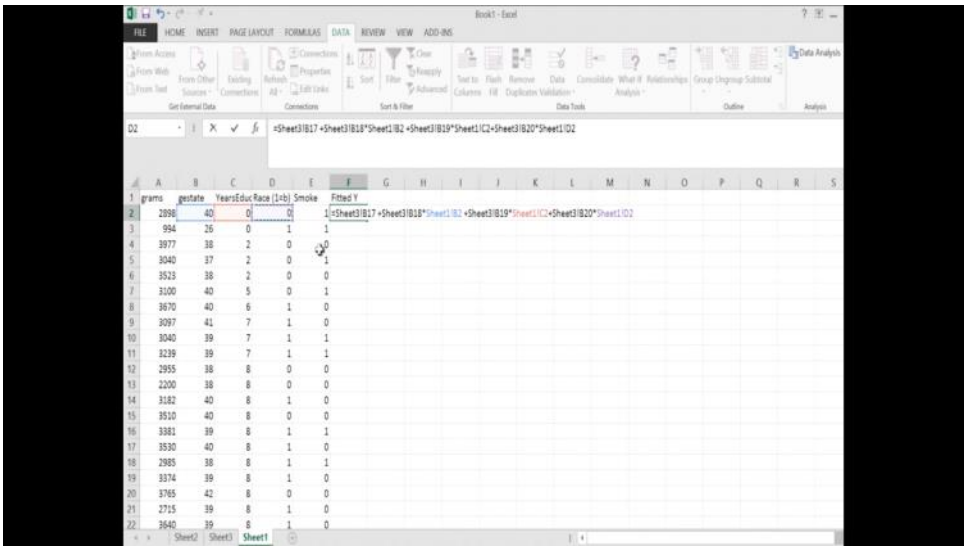
Fitted values are values of the Dependent variable (Birthweight) according to the model equation

*Birthweight = - 2834 + 156.51*Gestation + 9.57* Years Of Education -168.9 *Race - 174.8 *Smoking*

Given values of the X's (IVs), we can come up with a Fitted value for Y (DV)







But if we check the residuals check box we will get the output in the spreadsheet

Regression Results: Model Fit

This will generate predicted (fitted) values, and residuals

RESIDUAL OUTPUT

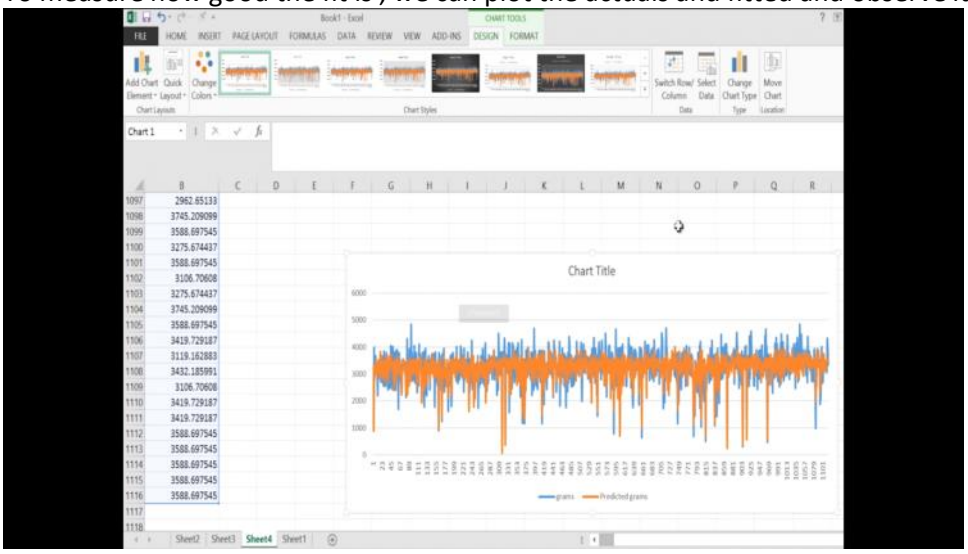
Observation	Predicted grams	Residuals
1	3251.163557	-353.1635571
2	891.0334453	102.9665547
3	3132.096999	844.9030006
4	2800.772554	239.2274461
5	3132.096999	390.9030006
6	3299.022703	-199.022703
7	3314.439066	355.5609338
8	3480.522449	-383.5224493
9	2992.68645	47.31355013
10	2992.68645	246.3135501
11	3189.527975	-234.5279746
12	3189.527975	-989.5279746
13	3333.582725	-151.5827246
14	3502.551082	7.448917694

Predicted Values are the fitted values

Residuals are the difference between Predicted Values of Y and the Actual Values of Y

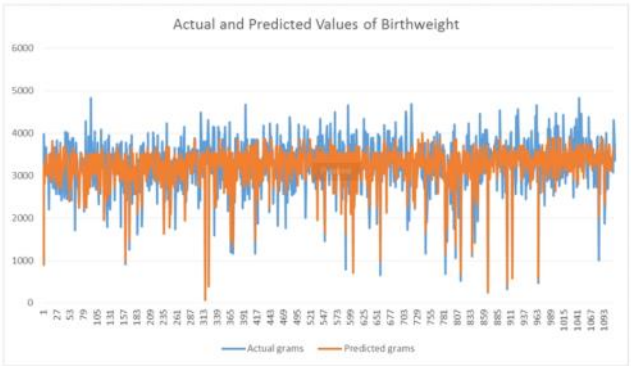
How many predicted values will be obtained?

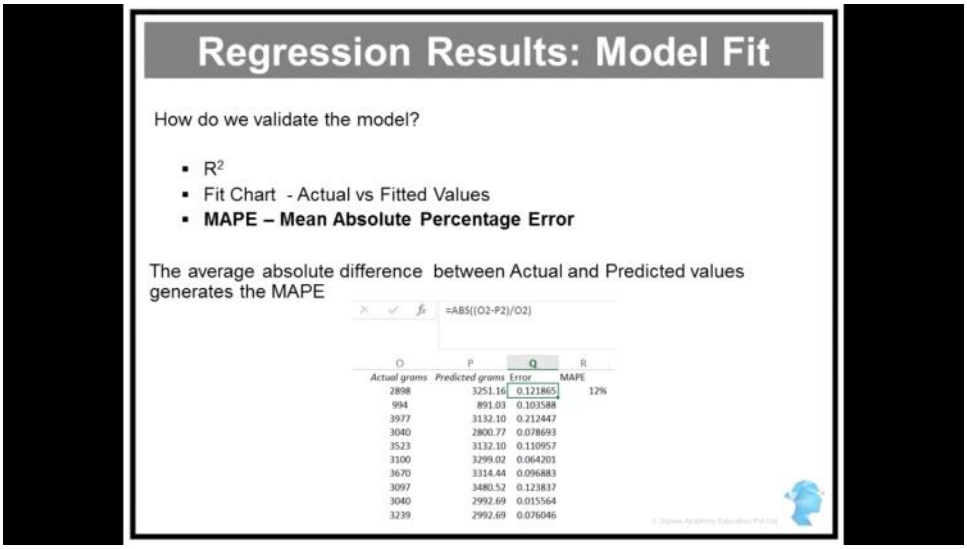
The residuals are the differnece between actual Y and predicted Y. Like an error
To measure how good the fit is , we can plot the actuals and fitted and observe it:

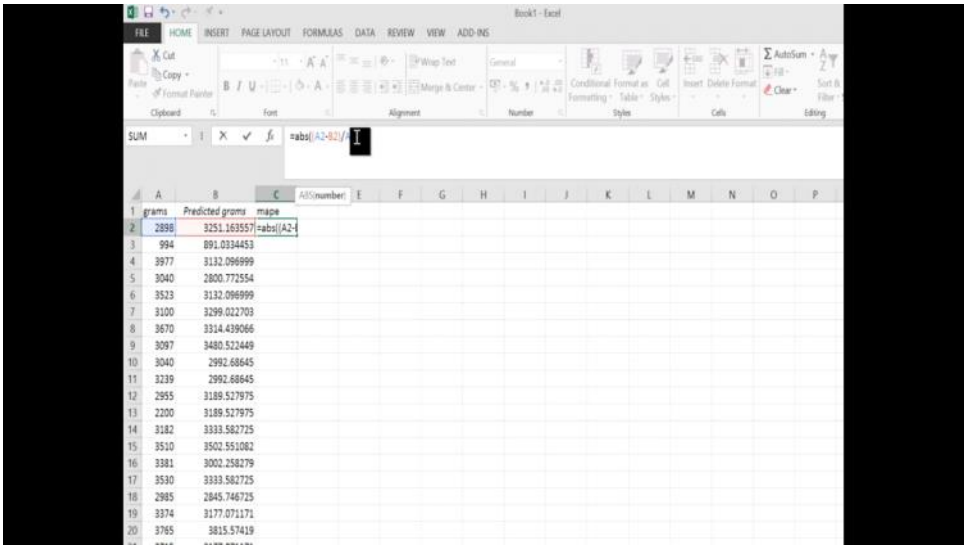


Regression Results: Model Fit

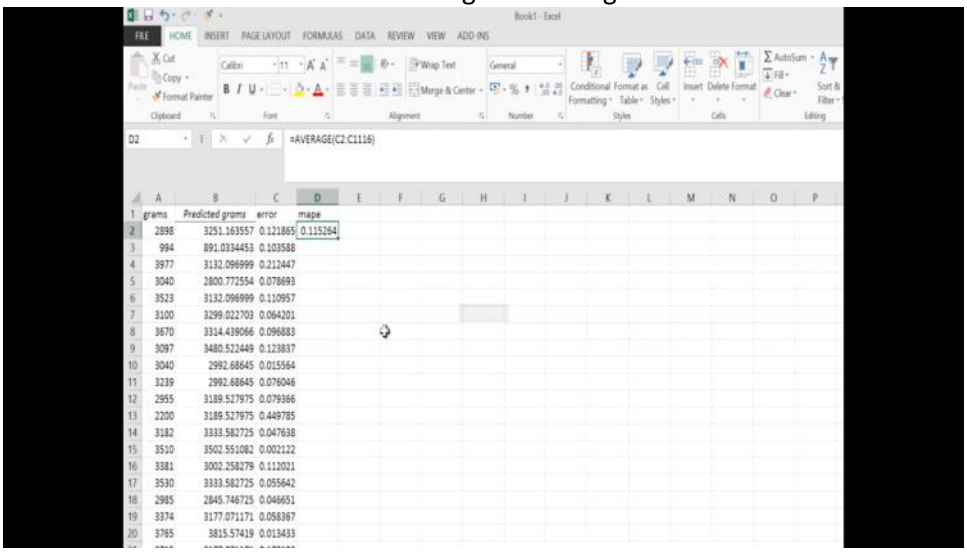
Visual Comparision of Actual and Predicted Values: FIT CHART







Use the absolute value as we want to ignore the sign on the error.



Regression Results

PREDICTIVE MODEL

How is a regression a predictive modeling technique?

- Once we have a final validated model, given the regression equation, for values of X we can predict a "Y" value
- We calculated fitted values or predicted values for the actual data values of X in our dataset
- We can use the same calculation for other (future) values of X's

Regression Results

PREDICTIVE MODEL

For example, we have a mother with 10 years of education, Race = Black (1), expected Gestation period = 40 weeks, and Smoking = No (0),

we can calculate the expected birthweight as:

Birthweight = *- 2834 + 156.51*Gestation + 9.57* Years Of Education -168.9 *Race - 174.8 *Smoking*

= -2834 + 156.51 * 40 + 9.57 *10 – 168.9 * 1 – 174.8 * 0

= 3352.76 gms

This is the predicted weight of the baby

Regression Results

PREDICTIVE MODEL

Why should we believe the predicted value?

If we have a good model (High R2, good fit, low MAPE), we can be confident about our predictions

In this example, our R2 is only 52%, and MAPE is 12%. This is not a great model

What next?

Regression Assumptions:

DATA SCIENCE WITH R

REGRESSION ANALYSIS

Overview

Simple Linear Regression

Multiple Linear Regression

Regression Assumptions

Implementation in SAS

Regression Assumptions

1. Model is *linear in parameters*
2. The data are a *random sample* of the population
   - The errors are *statistically independent* from one another
3. The expected value of the errors is always zero
4. The independent variables are not too strongly *collinear*
5. The independent variables are measured *precisely*
6. The residuals have *constant variance*

Regression Assumptions

7. The errors are normally distributed
8. The model is correctly specified

If all these conditions hold, then OLS estimators are **BLUE** –
Best Linear Unbiased Estimators

Regression Assumptions

In real life, all the conditions may not be met.

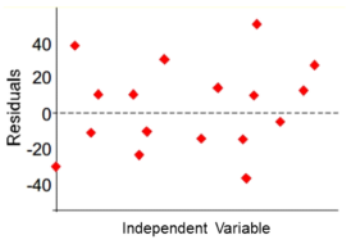
We need to:

1. Check if the assumptions are holding up
2. If not, assess how to correct for violations

Regression Results: Assumptions

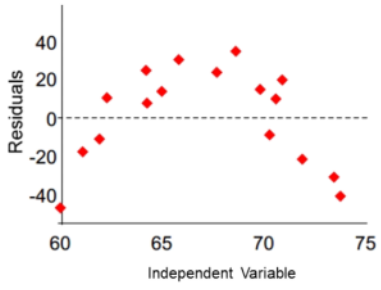
CHECKING IF ASSUMPTIONS ARE VALID

1. **Check for linearity – plot the residuals against each IV**
   - If data is linearly related, we should see no pattern in the plot



Regression Results: Assumptions

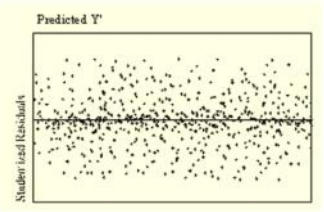
If relationship is non–linear?



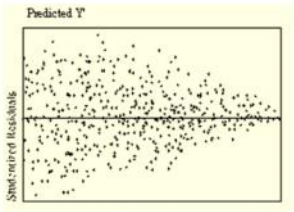
Plot the standard residuals to check for homoscedasticity:

Regression Results: Assumptions

2. **The residuals should have constant variance – homoscedasticity**
   - Plot the residuals against Predicted Y



Homoscedasticity



Heteroscedasticity

Regression Results: Assumptions

If errors are heteroscedastic?

- The presence of heteroscedasticity does not imply bias in the estimates
- Heteroscedasticty leads to bias in the standard errors, leading to issues with hypothesis testing and confidence intervals
  - Std error is a measure of variance, and therefore if standard errors are biased, then hypothesis test results will be biased leading to wrong inferences

Regression Results: Assumptions

3. **The residuals are normally distributed**
   - Histogram or probability plot for the residuals

Regression Results: Assumptions

If residuals are not normally distributed?

Hypothesis test outcomes may be invalid, though less of an issue with large samples

Regression Results: Assumptions

4. **The IVs are not too correlated - multicollinearity**
   - The IVs should not be highly correlated to one another
   - Check pairwise correlations, or generate VIF

VIF >>Variance Inflation Factor
Can be generated in R or SAS but not excel

Multiple Linear Regression

MODELING TECHNIQUES

Running a model given data is an easy task given that all the computation is done via SAS or Excel

The skill of an analyst lies in generating the right model to understand and solve for the business issue at hand

The first model, or naïve model that is generated from data is usually used as a starting point

- Depending on domain understanding and modeling techniques knowledge, typically many models are run before arriving at a final model

How can multiple models be run using the same data?

Multiple Linear Regression

MODELING TECHNIQUES

Step-wise Regression

It may be useful to run models adding (or removing) one variable at a time

Two types of step-wise regressions:

- Forward – Add one variable at a time
- Backward – Remove one variable at a time

Multiple Linear Regression

MODELING TECHNIQUES

Forward Step-wise regression

Variables are added one at a time until the model cannot be significantly improved by adding another variable

- Note that the variable order we use to add has an impact, so multiple step-wise forward regression models could be run before arriving at a best model

Backward Step-wise regression

This approach is the reverse, where we start with a model that has all explanatory variables, and variables are dropped one by one based on p-value (highest p-value dropped first).

- Re-run model without the variable dropped, and then drop next variable with highest p-value. Continue till no other variables can be dropped based on a pre-determined cut-off value (5%, 10%)

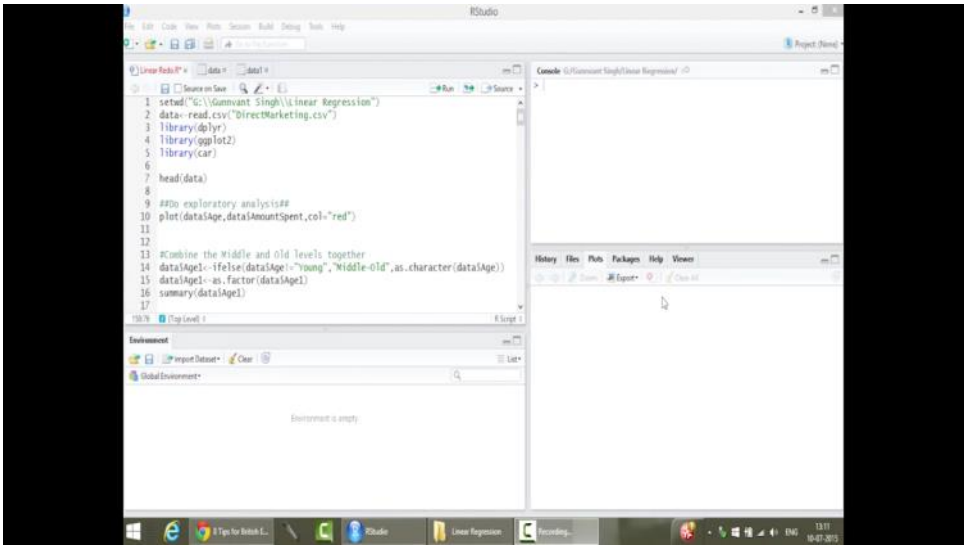
Multiple Linear Regression

MODELING TECHNIQUES

Other modeling techniques could include:

- Transforming variables – use log transformation for example
- Creating interaction variables – trying to capture impact of variable A and Variable B together
- Aggregating or disaggregating variables – Adding up marketing, or disaggregating promotions

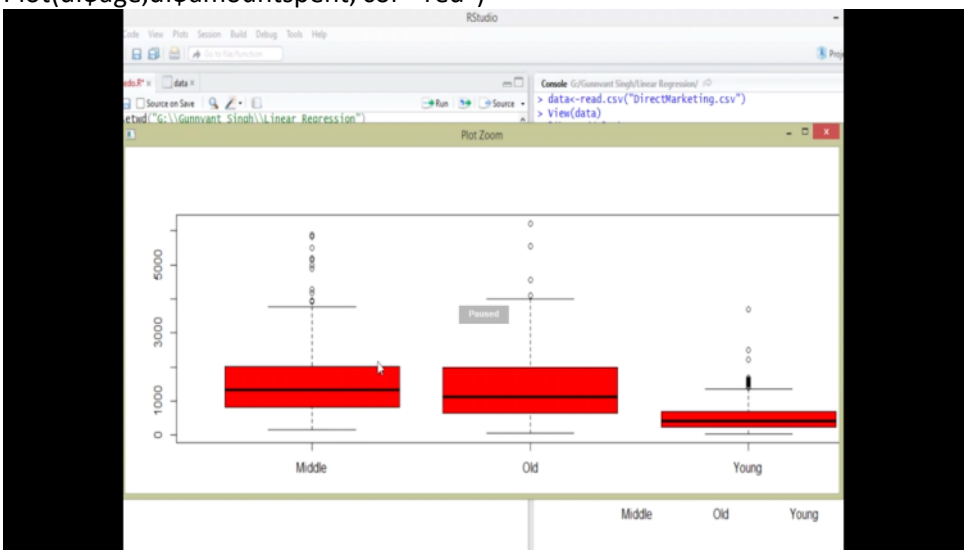
R code Demo:



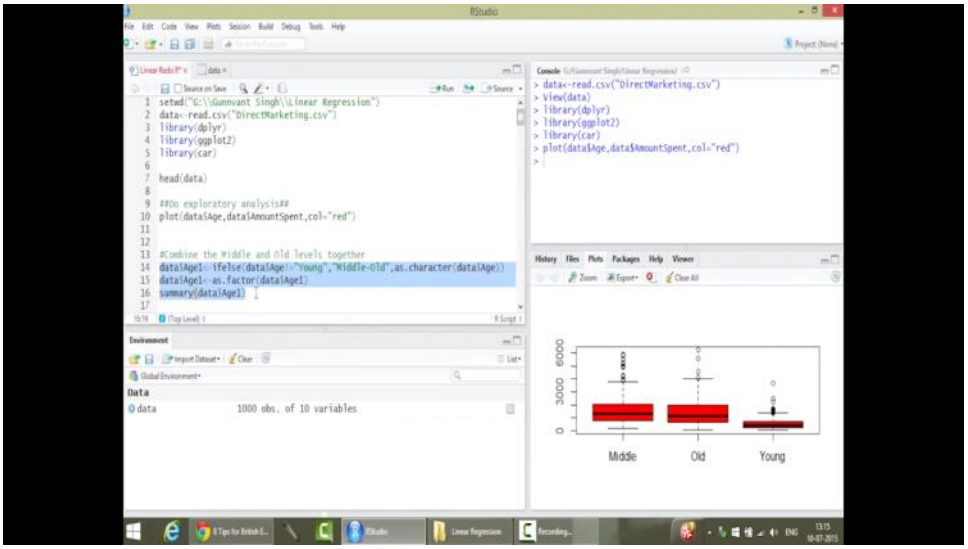
Looking at customer spending behavior:
Attributes: age(categorized as young, middle, old), Gender(M/F), OwnHome(Own/rent), Married(single/married/divorced), Location to store, not just our but others that sell similar products(Far/close), Salary(numbers), numberof childer(numbers), history (based on volume of purchase, i.e. are they high spenders , medium etc), Number of catalogs sent (total count to a customer), amount spent($)

So the linear regression will assess the amount spent based on rest of the attributes provided.

First look at the various variables, for everything that is categorical create a box plot to understand variations:
Plot(dfage,dfamountspent, col="red")



Combine the two categories (oldd,middle) as a single category as their behavior to spending is similar.



Here is how we will combine the two categorical values:

Create a new fields:

Df$age1<-ifelse(data$age!="Young","Middle-Old",as.character(data$age))

After combining, convert them to factors:

Df$age1<-as.factor(df$age1)

Run a summary to validate this:

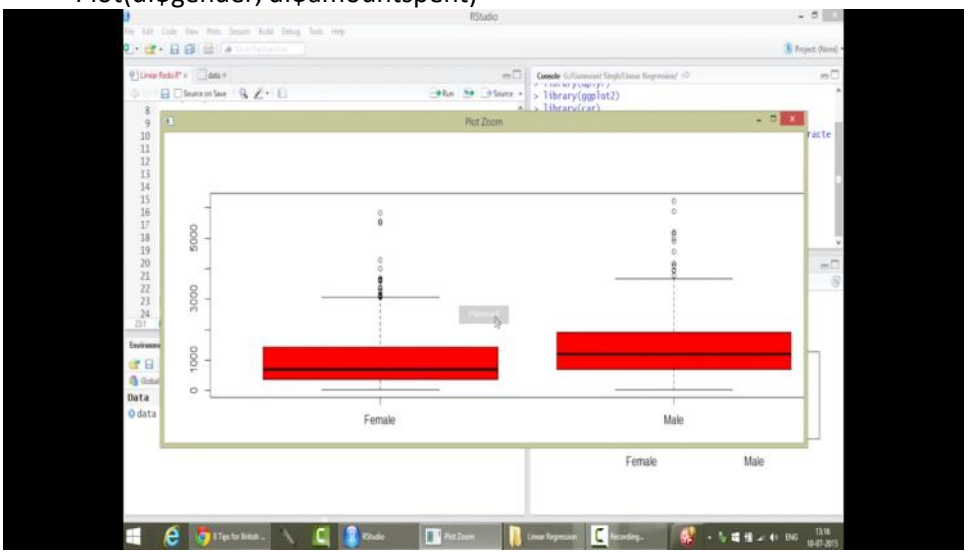
Summary(df$age1)

Run another boxplot with the new field and the amount spent:

Plot(df$age1,df$amountspent,color="red")

Let's analyze the gender

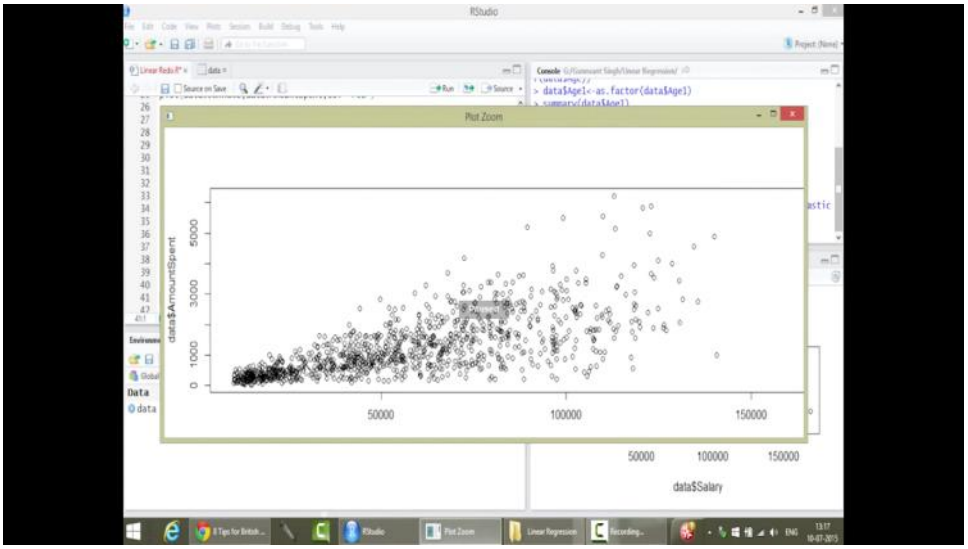
Plot(df$gender, df$amountspent)



Lets look at non-categorical variable like salary and the plot:

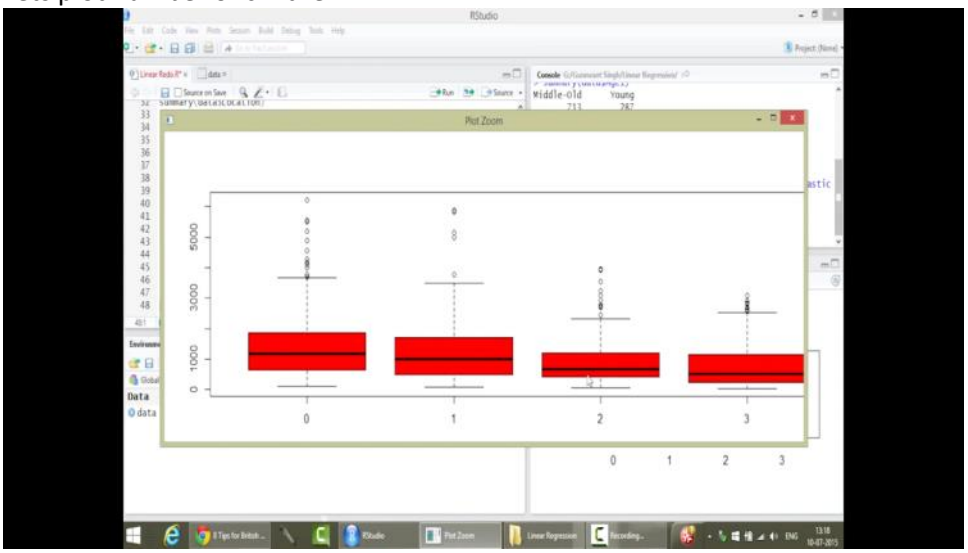
Summary(df$salary)

Plot(df$salary, df$amountspent) #might be heteroscadasticity

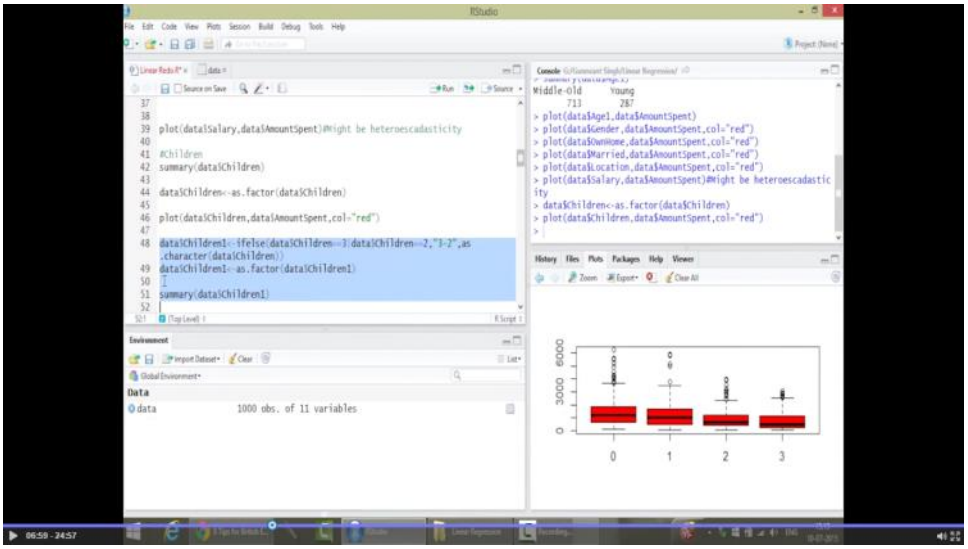


As salary increases, the amount spent also increases but the variation in amount spent also increases. This is heteroscadasticity and will impact our model.

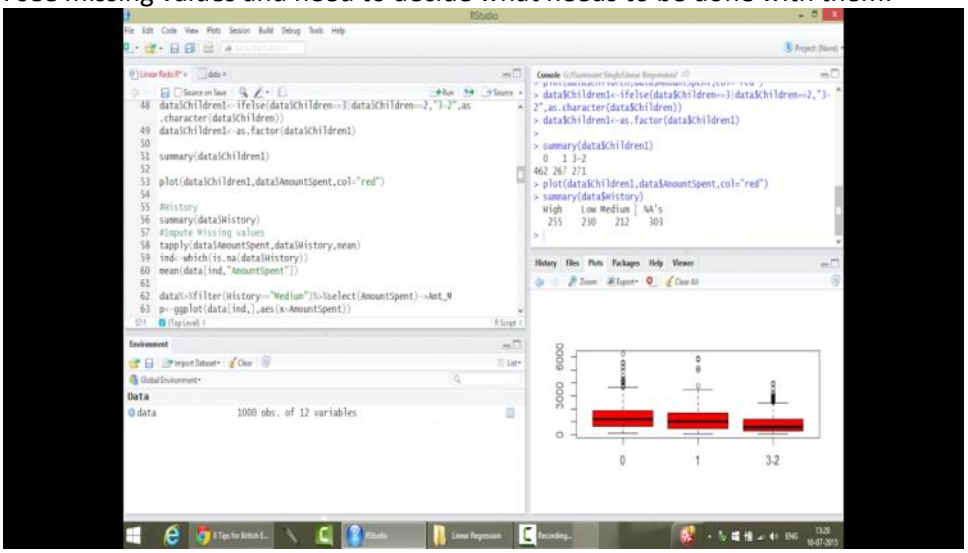
Lets plot number of children:



Since 0 and 1 behave in a simlar way while 2and 3 behave similarly we will combine to get just 2 variables here:
Df$children1<-ifelse(df$children<2,"under2",as.character(df$children))
Df$children1<-as.factor(df$children1)

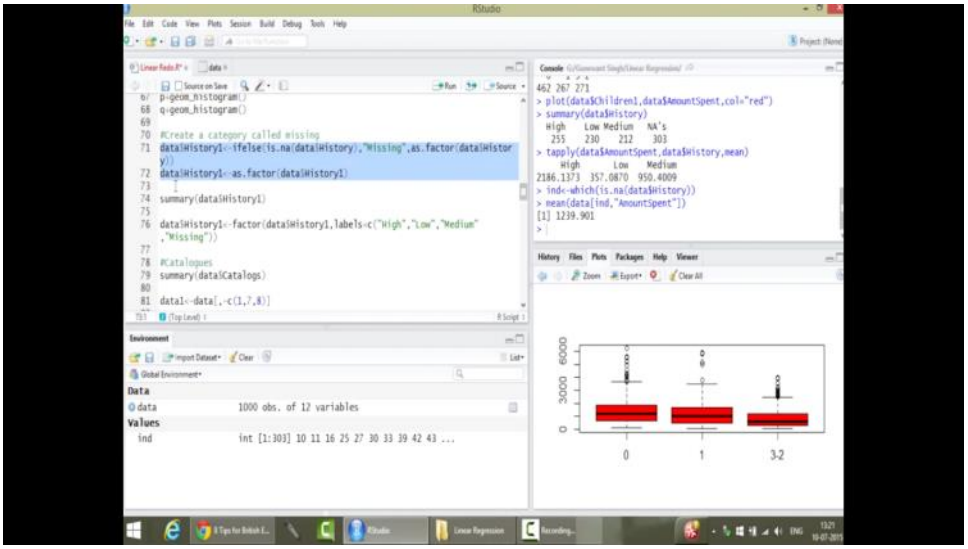


Now, lets look at purchase history
I see missing values and need to decide what needs to be done with them:



Summary(df$history)
#input missing values:

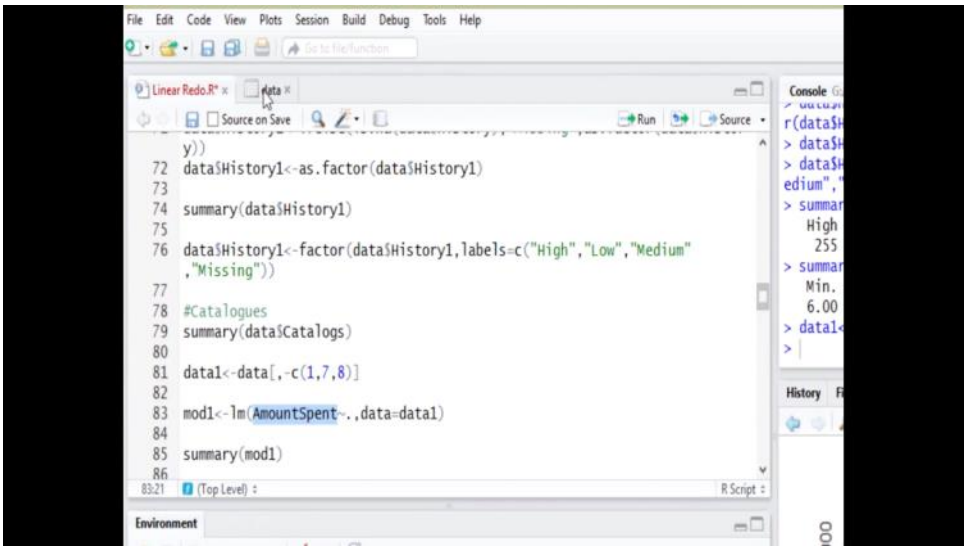
- Tapply(df$amountspent, df$history, mean) --> This will group data based on history and calculate the mean of amount spent for each category
- Ind<-which(is.na(df$history)) --> find the index of the cells with no values
- Mean(df[ind,"AmountSpent"]) --> this will give the means for the data with no history
- Since the mean for missing values is very different than the one for others, we will create a new category called as "Missing" and add that to the history column,



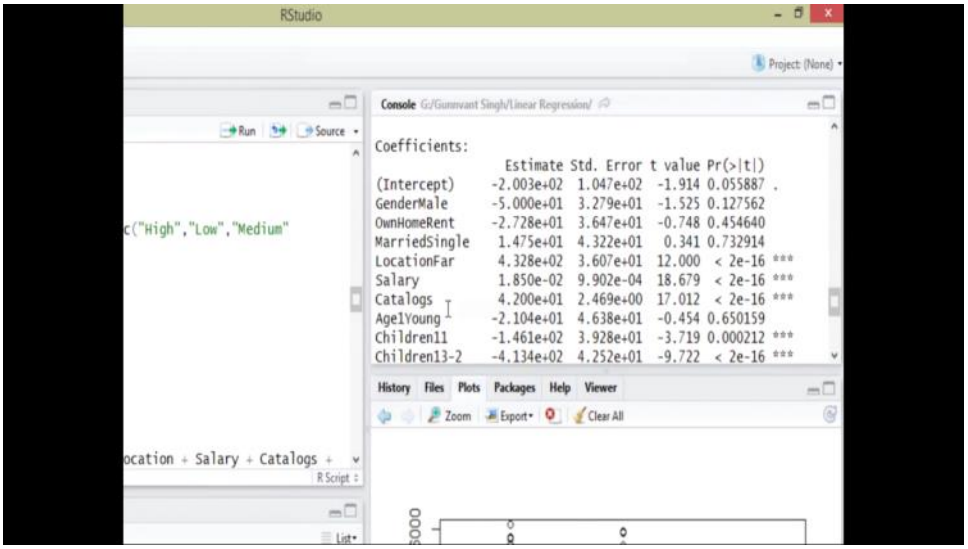
- Df$history1<-ifelse(is.na(df$history,"Missing", as.factor(df$history))
- Df$history1<-as.factor(data$history1)

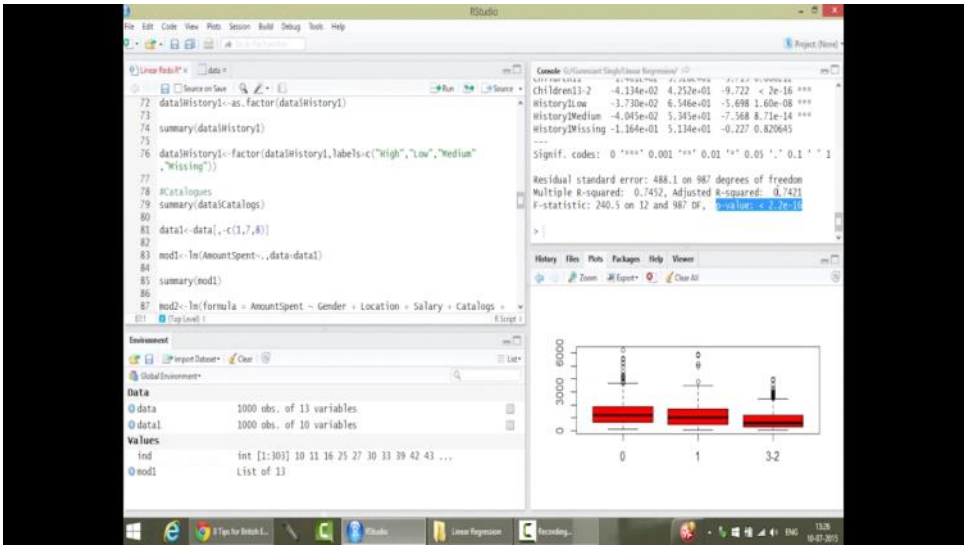
Lets look at catelogues:

- Summary(df$catelogues)



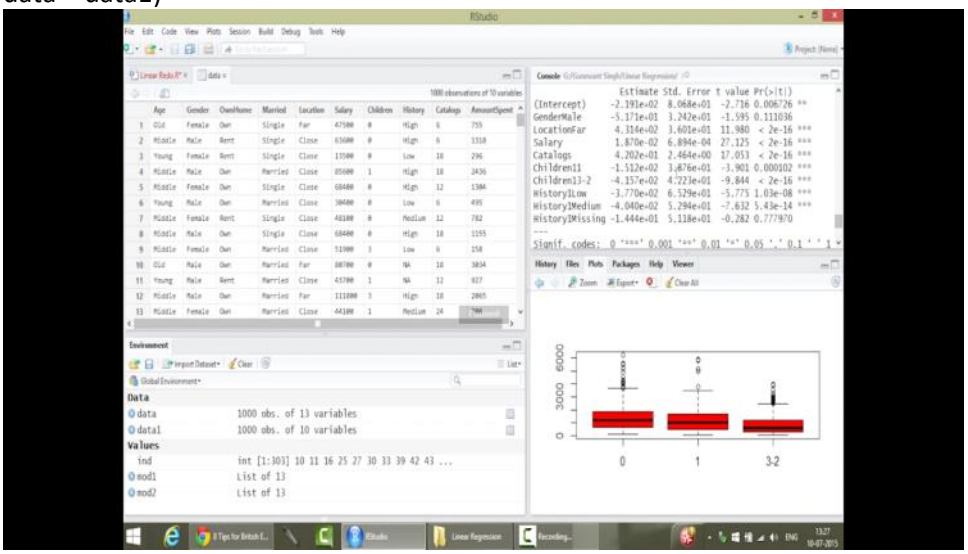
If you see lm(amountspent~.,data=data1)
The ~. Signifies --> use all variables other than amountspent in the evvaluation

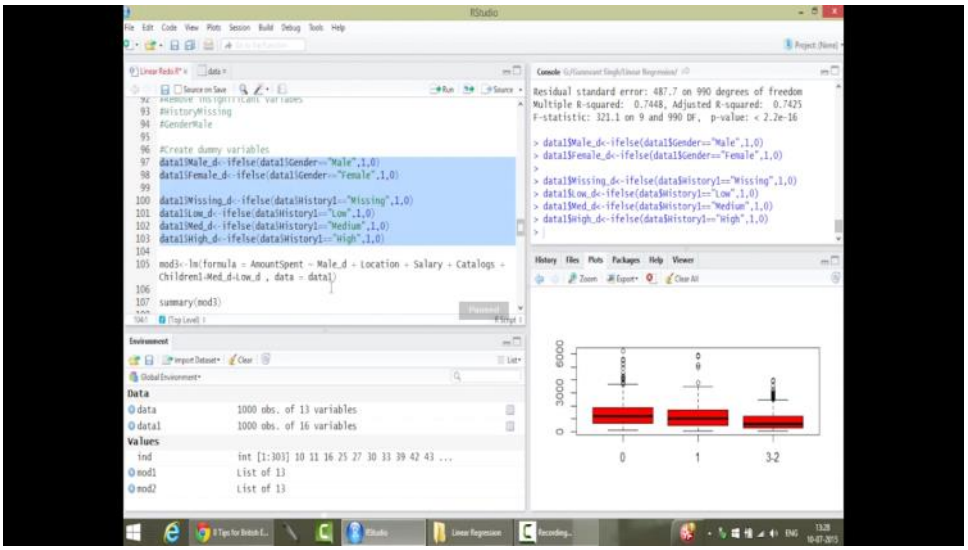




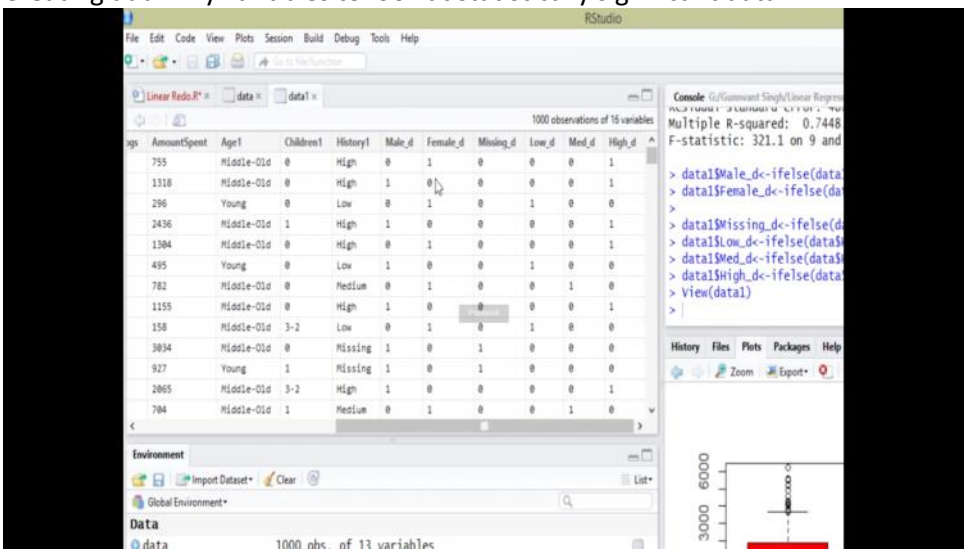
The pvalue --> represents anova for regression. It means adding the predictors is statistically significant. Good indicator

Lets rebuild the model by removing the variables that are not significant:
Mod2<-lm(formula = AmountSpent~Gender + Location + Salary + Catalogues + Children1 + History1, data = data1)

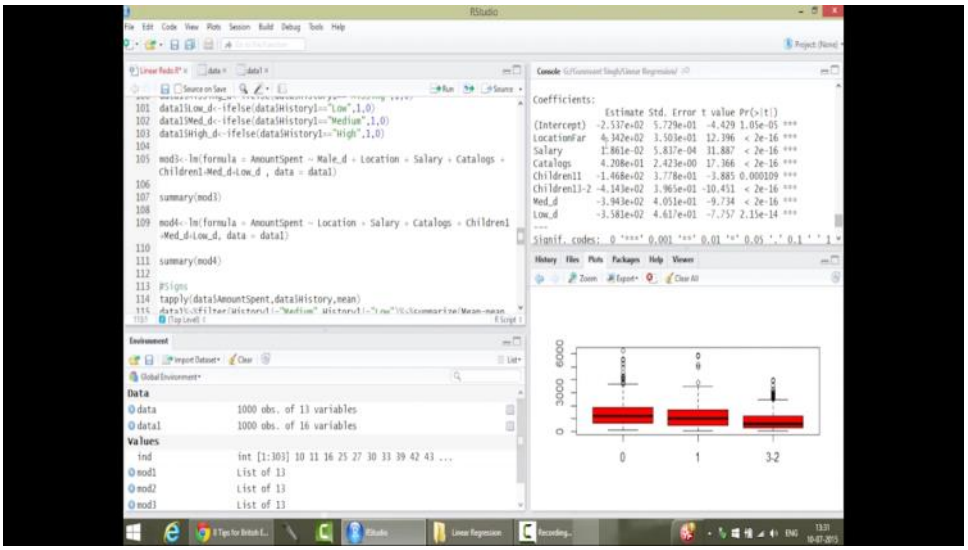




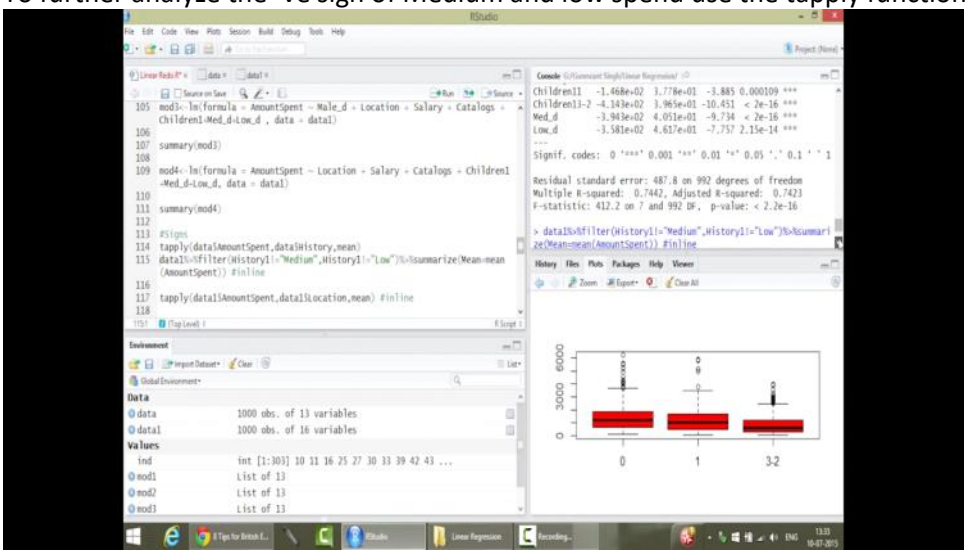
Creating ddummy variables to look at statistically significant data:



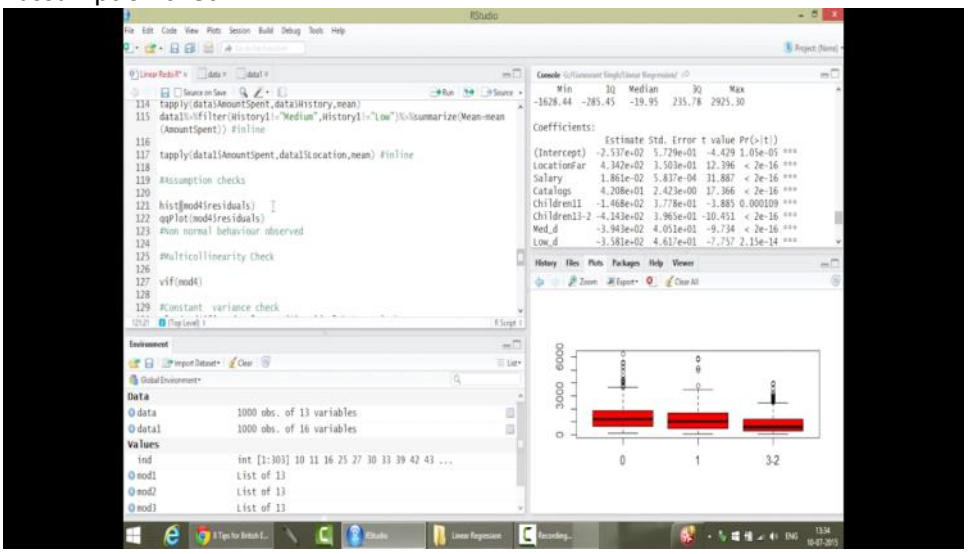
Let's run another model by including the dummy variables -->
Mod3<-lm(formula = AmountSpent~Male_d+<add all other variables needed> , data = data1)



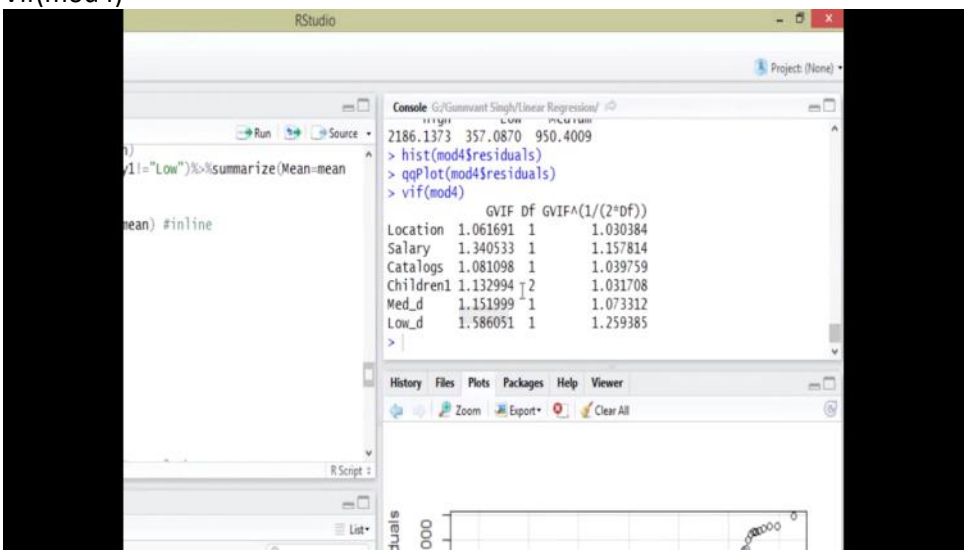
To further analyze the -ve sign of Medium and low spend use the tapply function to look at them:



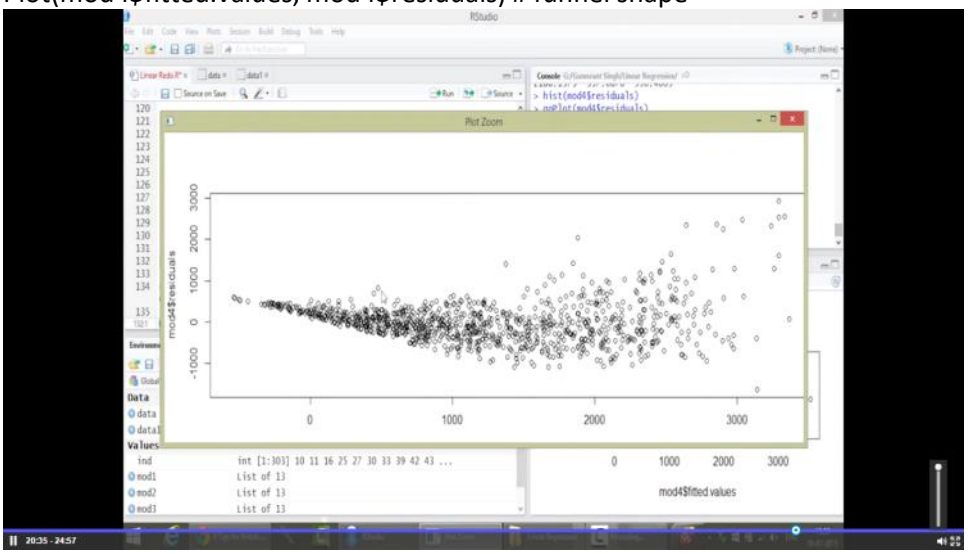
Once this is done, we need to validate our assumptions:
#assumption check



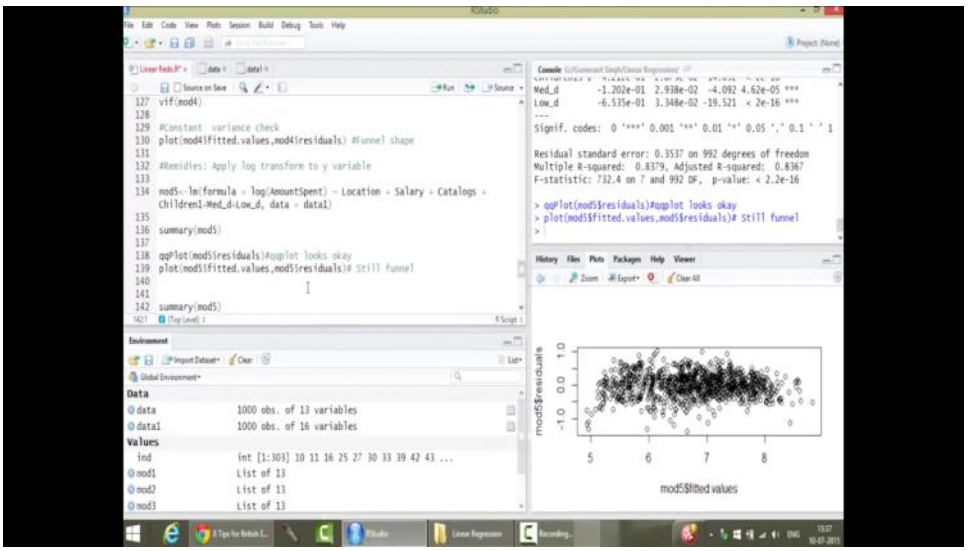
Then check for multicollenearity:
#multicollenear check
Vif(mod4)

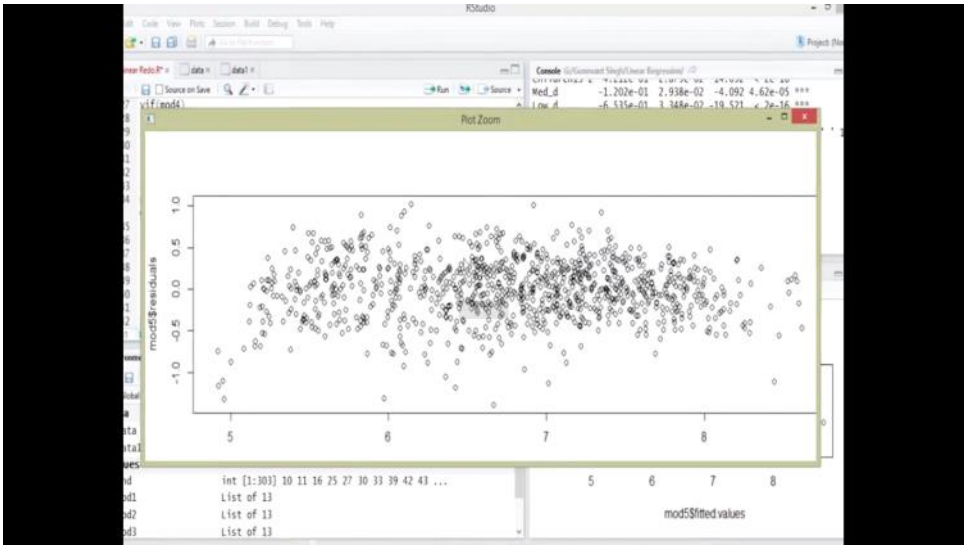


Next, check for constant variance check:
Plot(mod4$fitted.values, mod4$residuals) # funnel shape

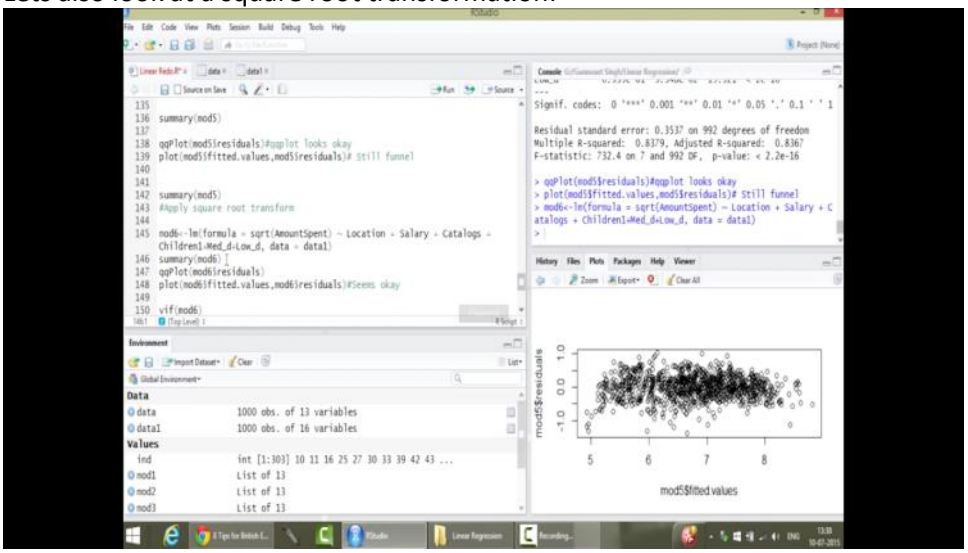


How can these be corrected??
#Remedies: log transformation to Y variable
Mod5<-lm(formula=log(AmountSPent)~<all Indepednent variables>, data=data1)





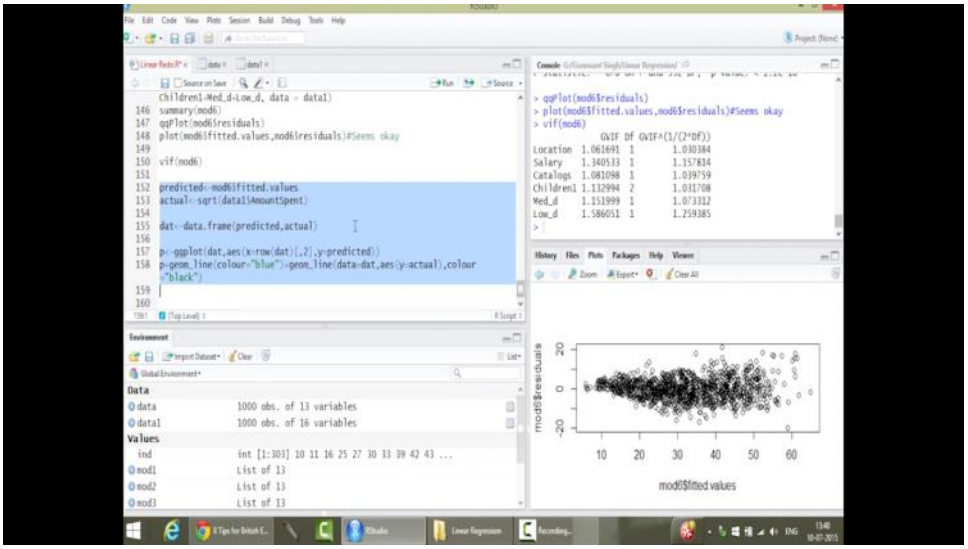
Lets also look at a square root transformation:

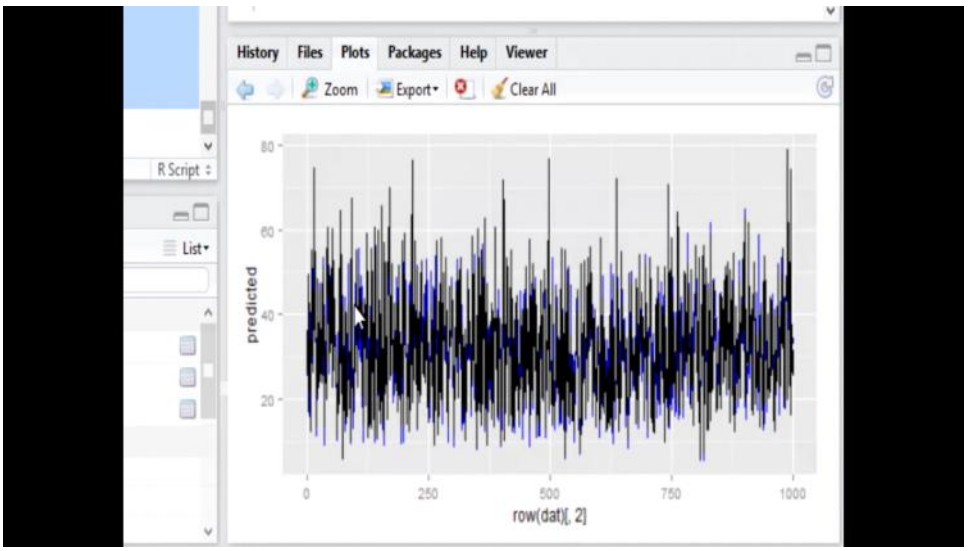


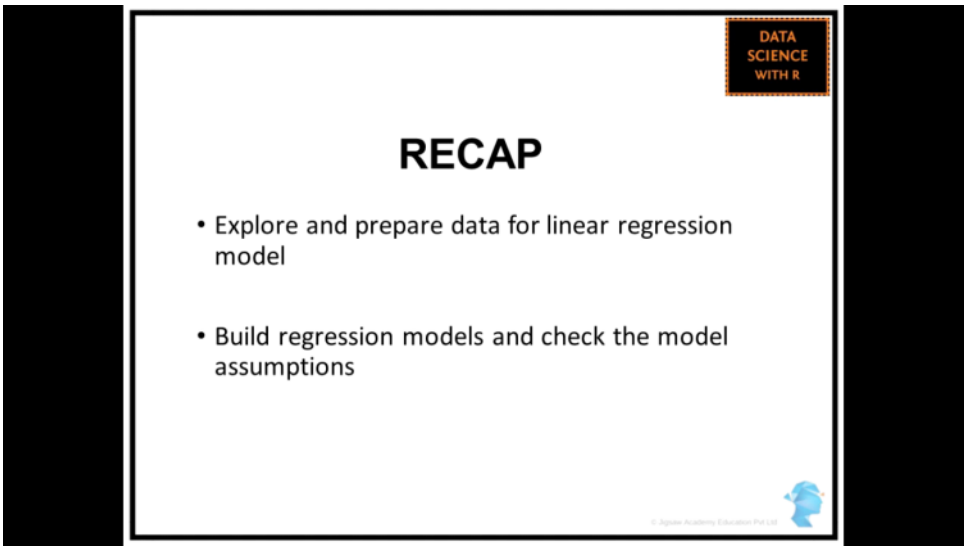
Check VIF values and they need to be less than 10 .
Vif(mod5)

The next step after finallizing the lm is to find the preddicted and actuals

Predicted<-mod5$fitted.values
Actual<-sqrt(data1$amountspent)
Dat<-data.frame(predicted,actual)
p<-qqplot(dat, aes(x=row(dat)[,2],y=predicted))


predicted<-mod6$fitted.values
actual<-sqrt(data1$AmountSpent)
dat<-data.frame(predicted,actual)
p<-ggplot(dat,aes(x=row(dat)[,2],y=predicted))
p+geom_line(colour="blue")+geom_line(data=dat,aes(y=actual),colour="black")
> qqPlot(mod6$residuals)
> plot(mod6$fitted.values,mod6$residuals)#Seems okay
> vif(mod6)
mod6$residuals
mod6$fitted.values


History
Files
Plots
Packages
Help
Viewer
Zoom
Export
Clear All
predicted
row(dat)[, 2]


DATA SCIENCE WITH R
RECAP
• Explore and prepare data for linear regression model
• Build regression models and check the model assumptions