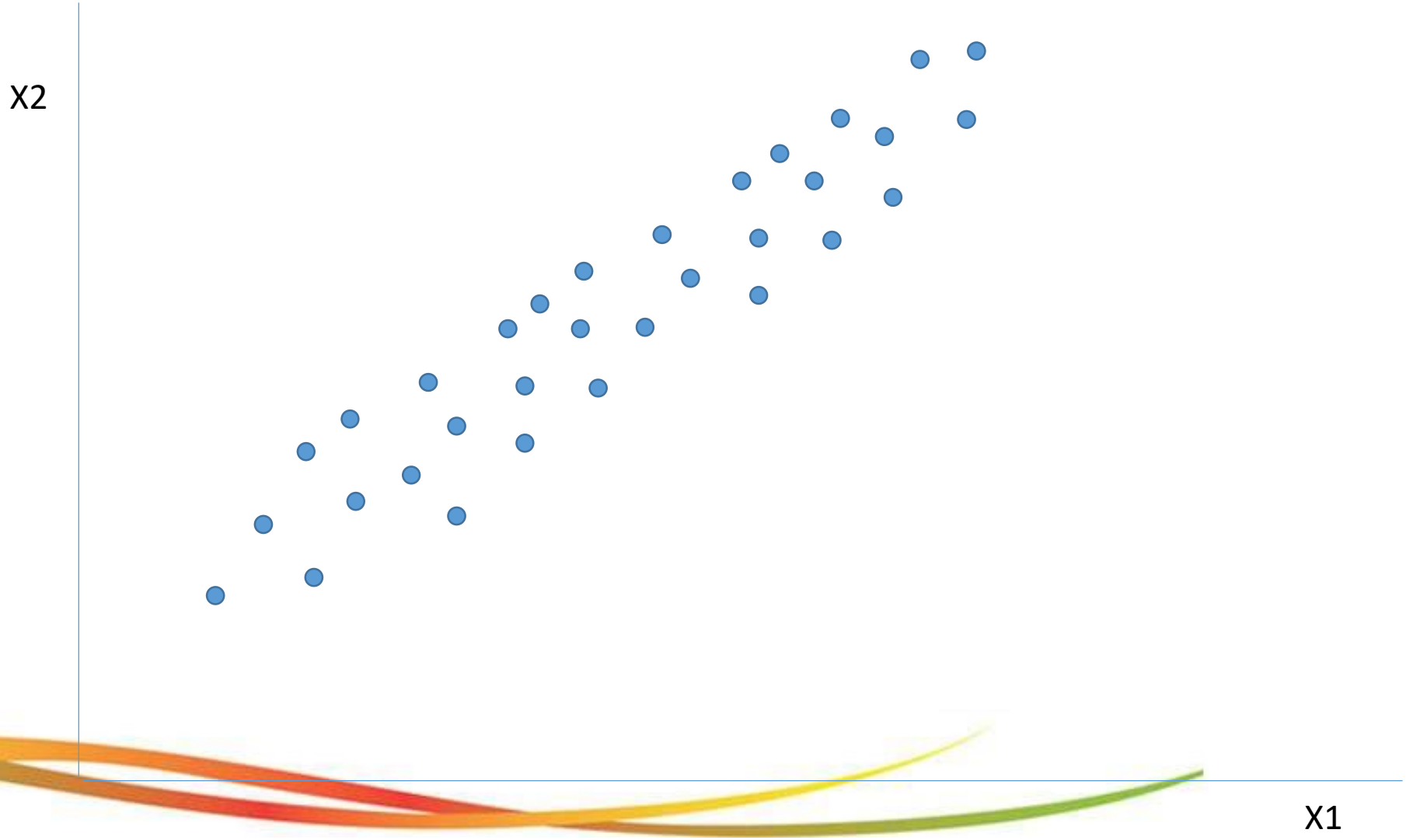
ANALYTICS PATH

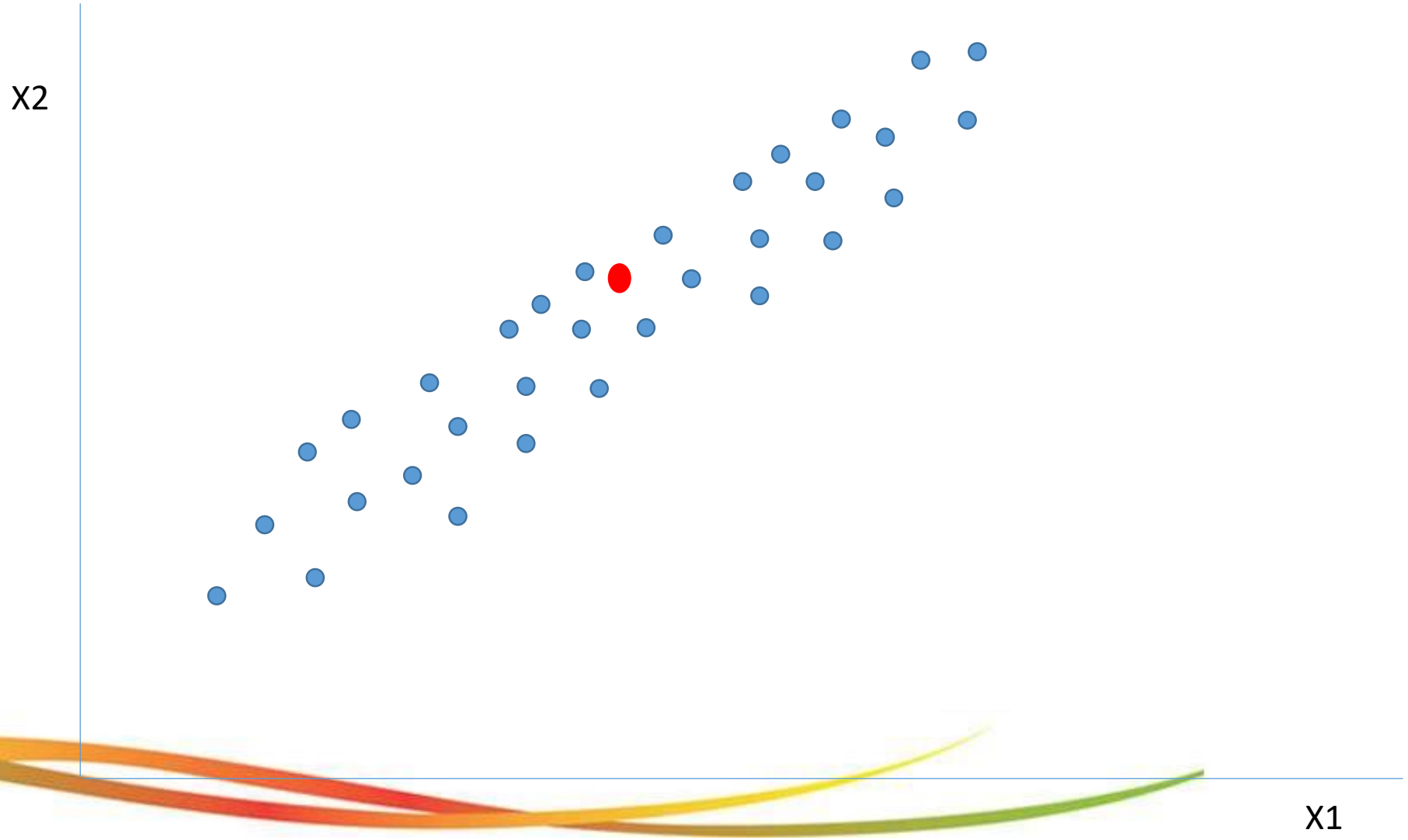
Linear Regression

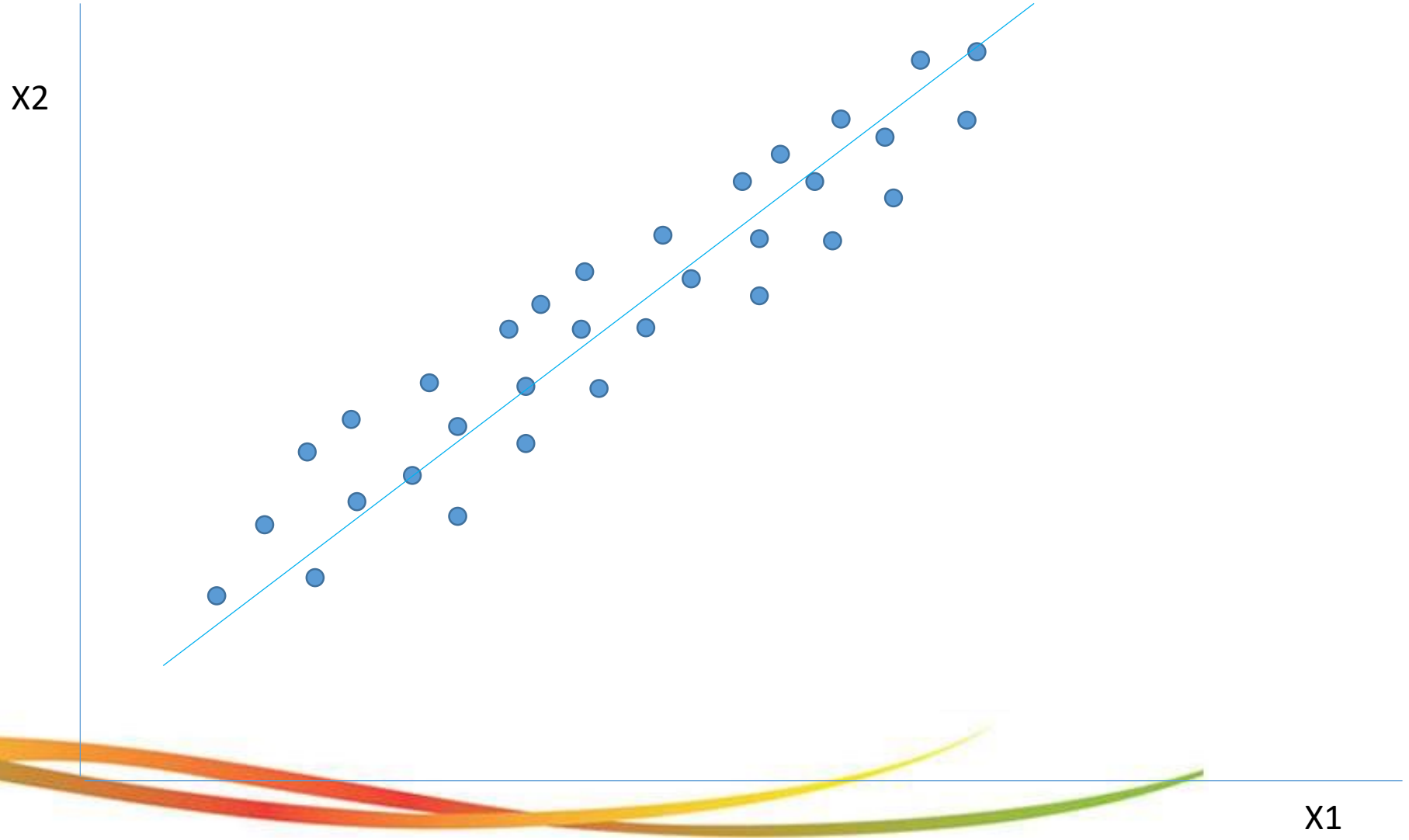
Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City,
Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

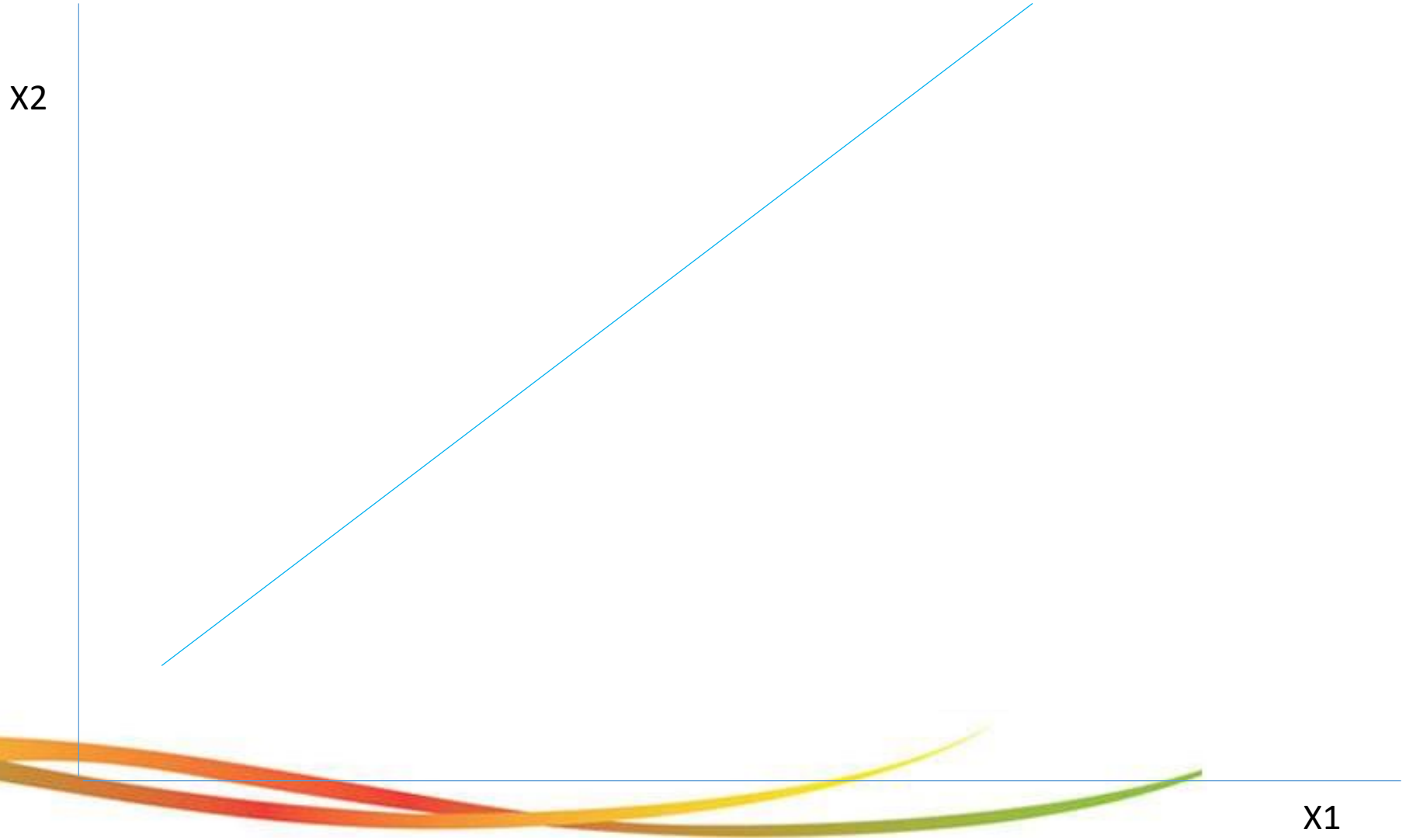
What is simple linear regression?

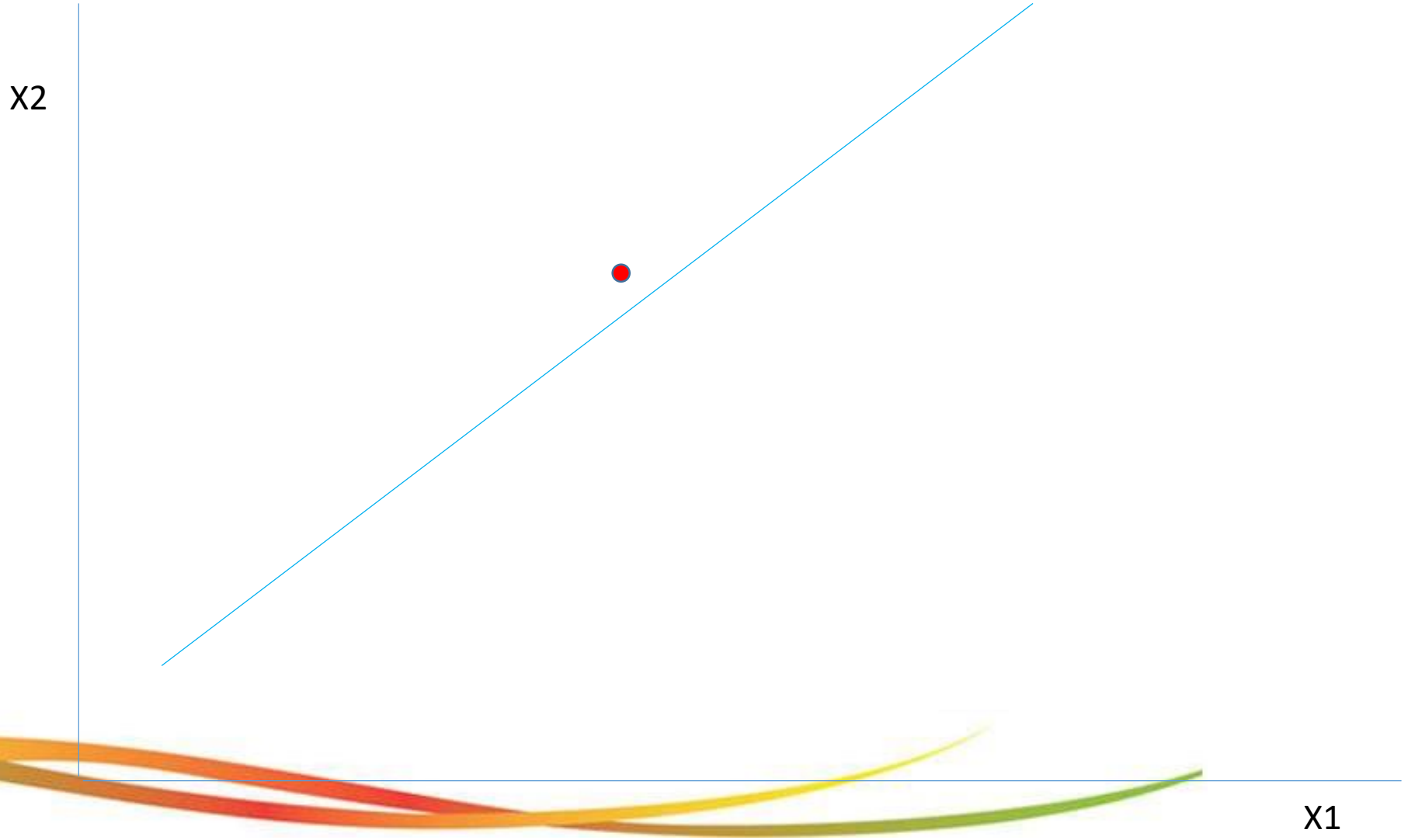
- A statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables – dependent (response, target) and independent (predictor, explanatory)

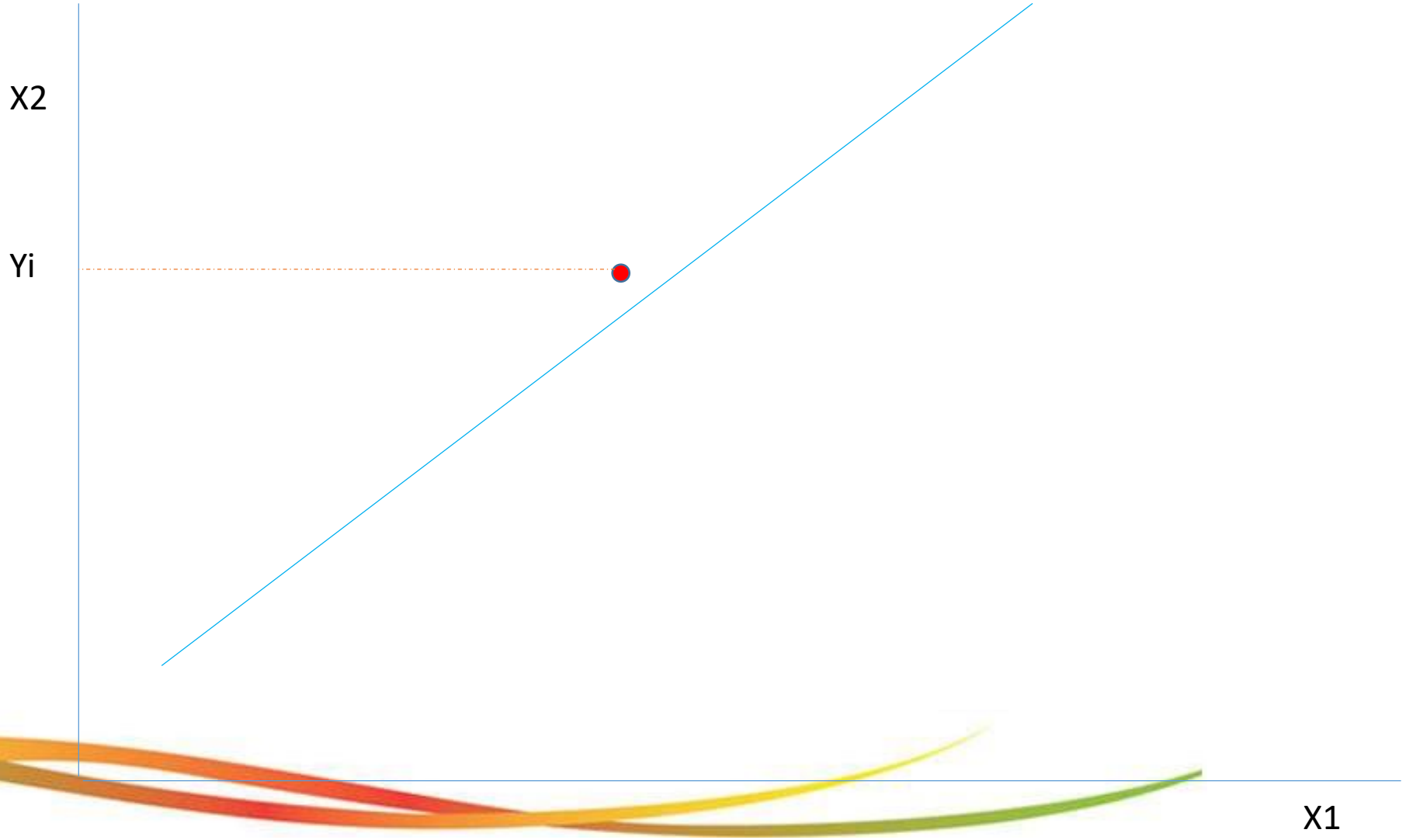

X2
X1

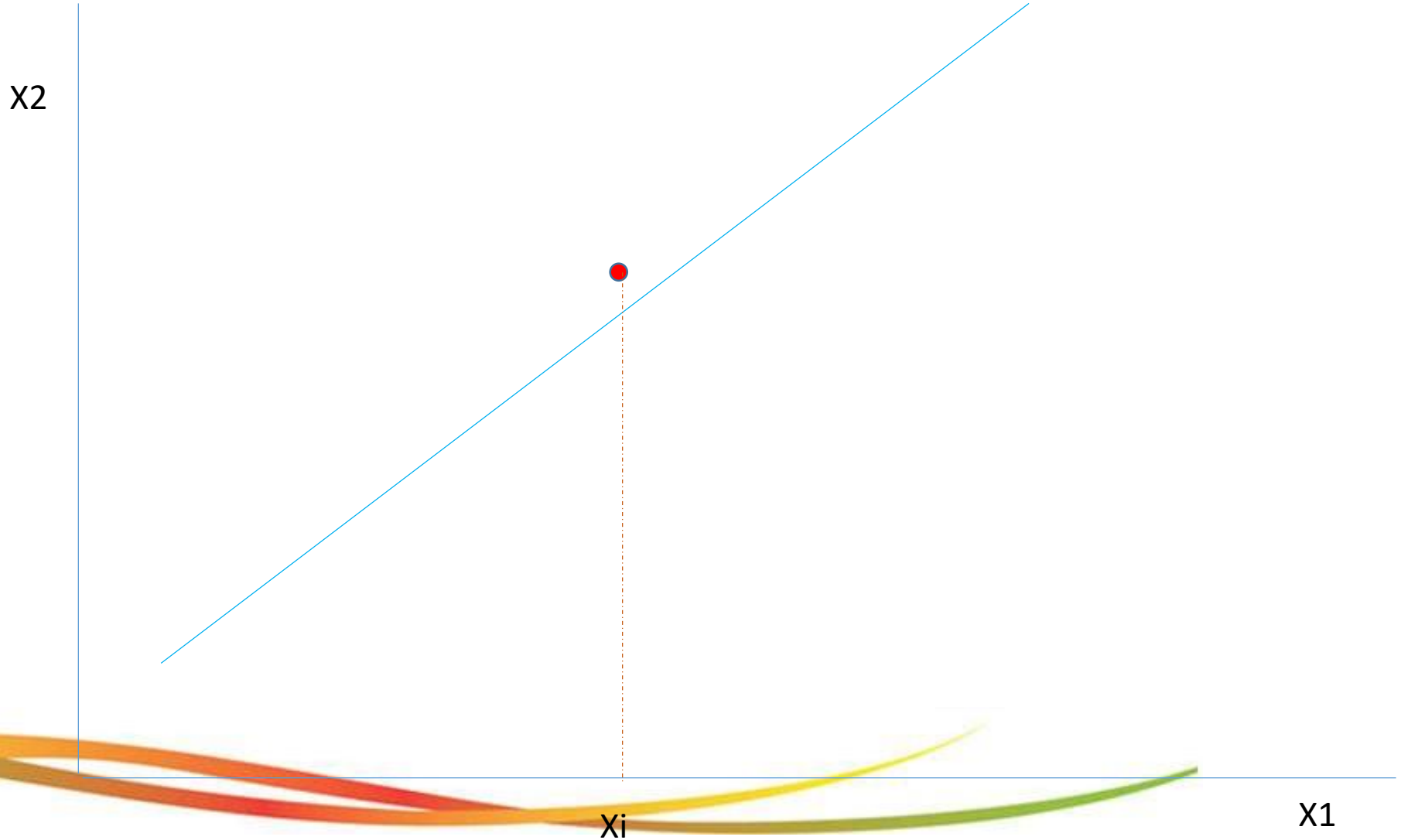

X2
X1

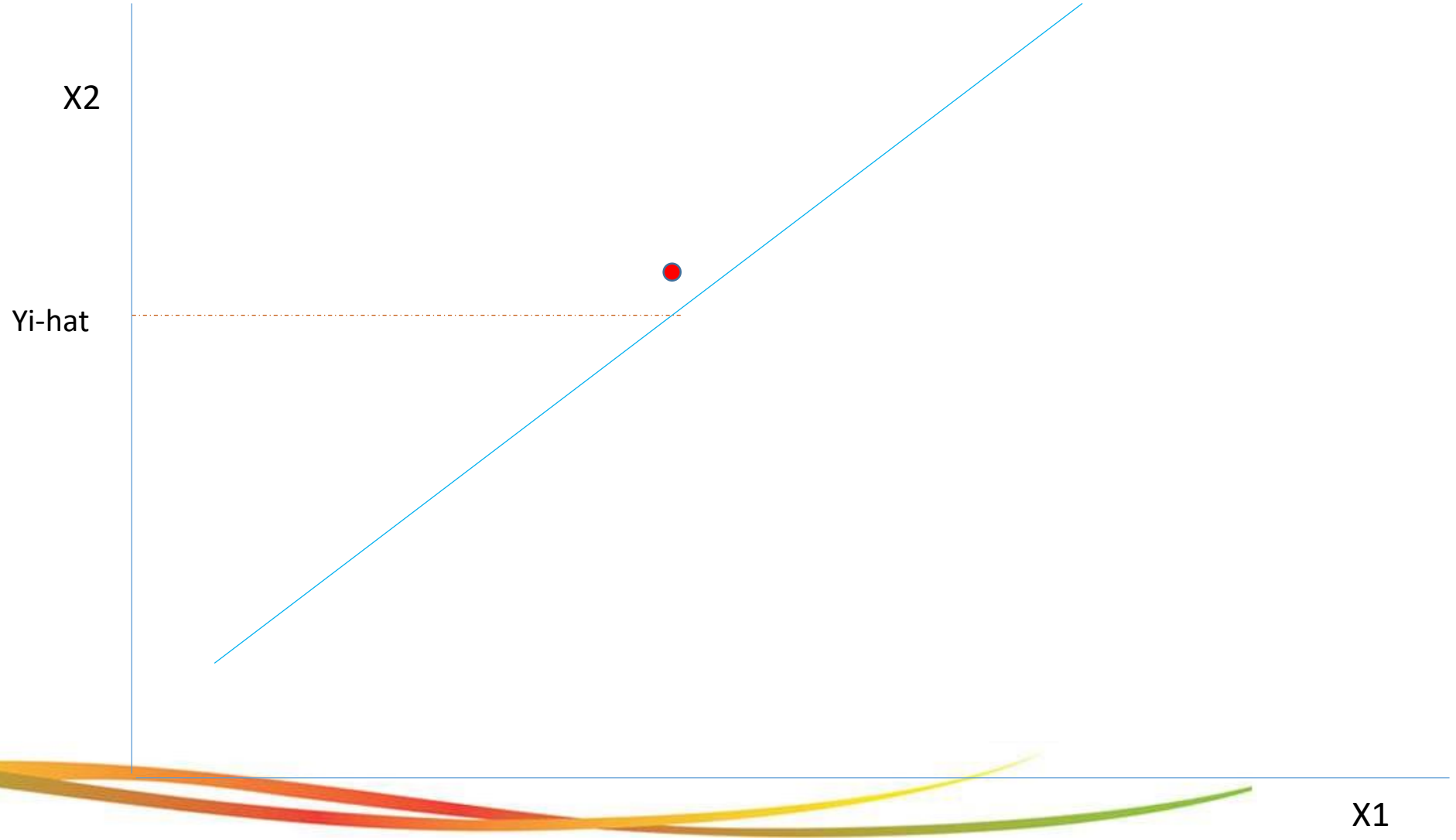

X2
X1


X2
X1

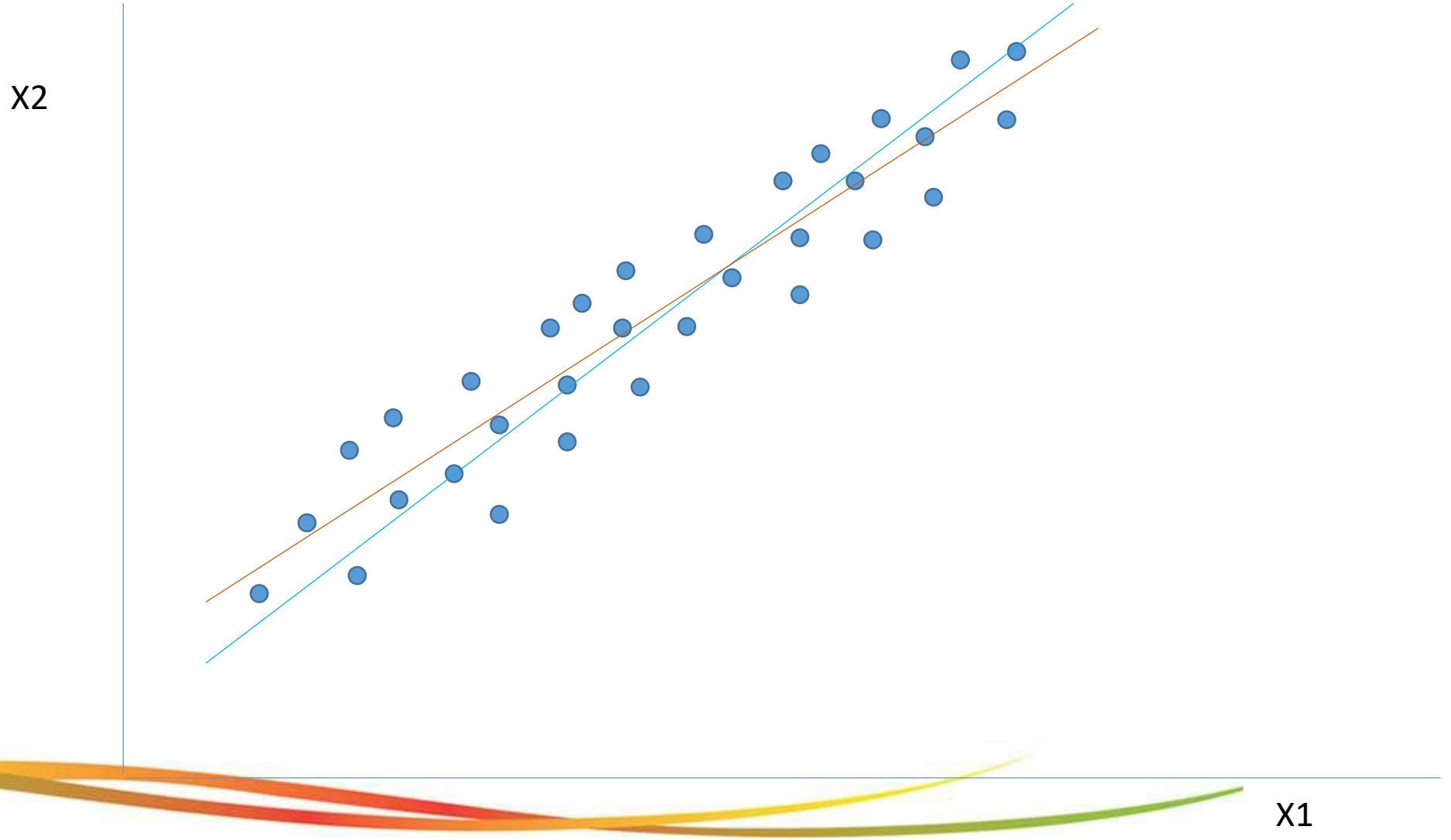

X2
X1


X2
Yi
X1


X2
Xi
X1

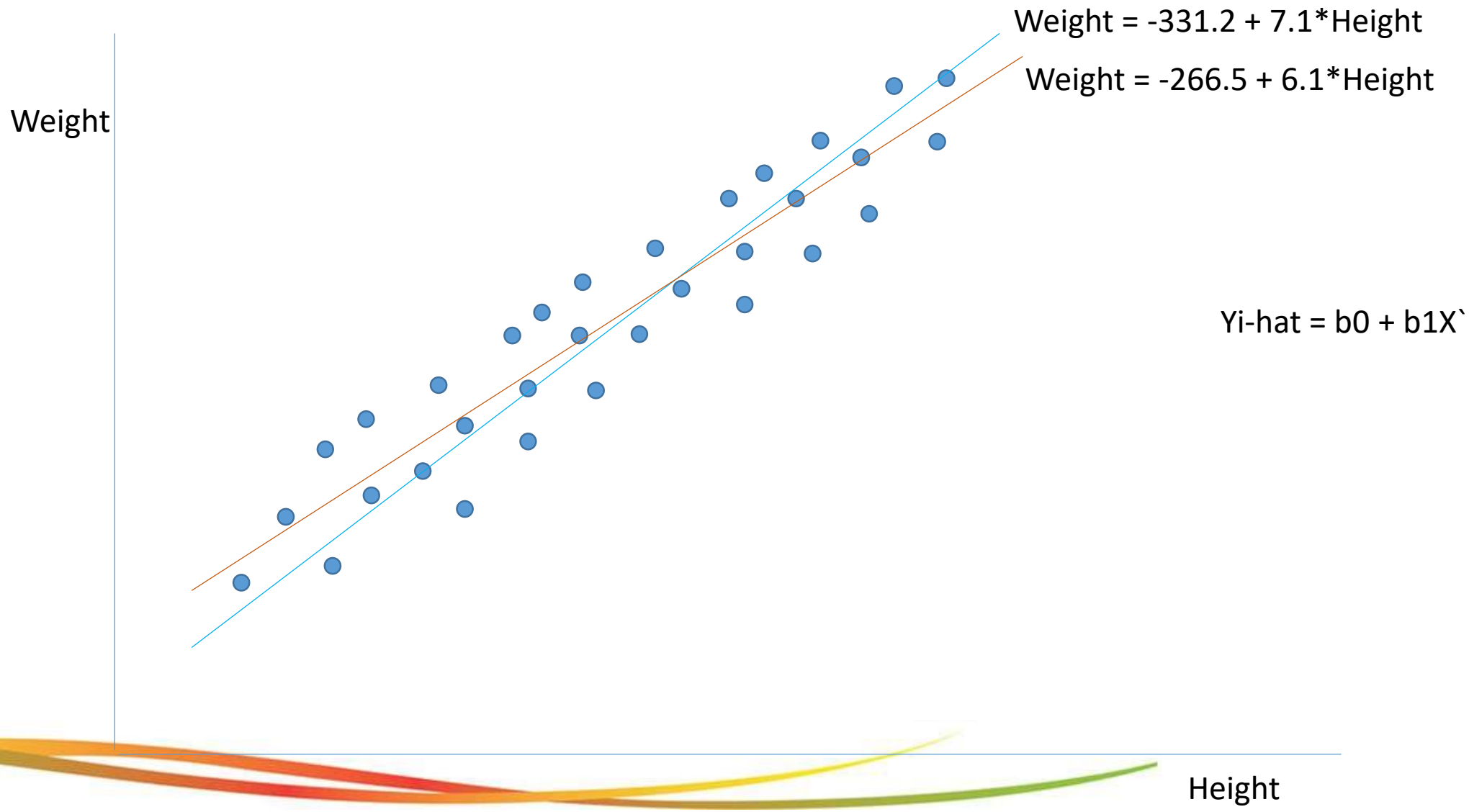

X2
Yi-hat
X1

But How Do You Decide The Best Line?



Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City,
Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

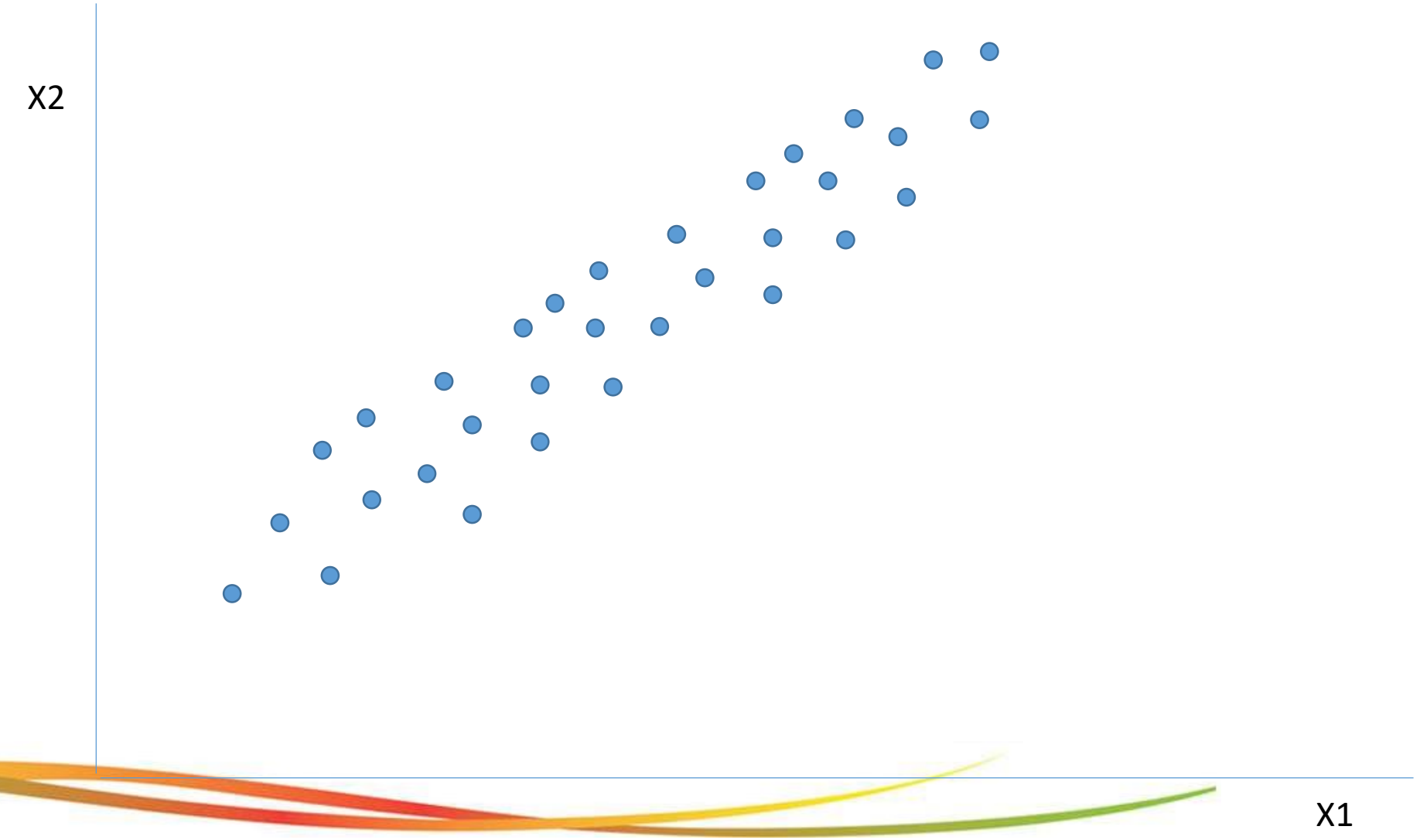
But How Do You Decide The Best Line?



Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City,
Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

$w = -331.2 + 7.1\,h$ (the dashed line)					
i	x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	63	127	116.1	10.9	118.81
2	64	121	123.2	-2.2	4.84
3	66	142	137.4	4.6	21.16
4	69	157	158.7	-1.7	2.89
5	69	162	158.7	3.3	10.89
6	71	156	172.9	-16.9	285.61
7	71	169	172.9	-3.9	15.21
8	72	165	180.0	-15.0	225.00
9	73	181	187.1	-6.1	37.21
10	75	208	201.3	6.7	44.89
					766.5

$w = -266.53 + 6.1376\,h$ (the solid line)					
i	x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	63	127	120.139	6.8612	47.076
2	64	121	126.276	-5.2764	27.840
3	66	142	138.552	3.4484	11.891
4	69	157	156.964	0.0356	0.001
5	69	162	156.964	5.0356	25.357
6	71	156	169.240	-13.2396	175.287
7	71	169	169.240	-0.2396	0.057
8	72	165	175.377	-10.3772	107.686
9	73	181	181.515	-0.5148	0.265
10	75	208	193.790	14.2100	201.924
					597.4


X2
X1

Lets implement this in R…

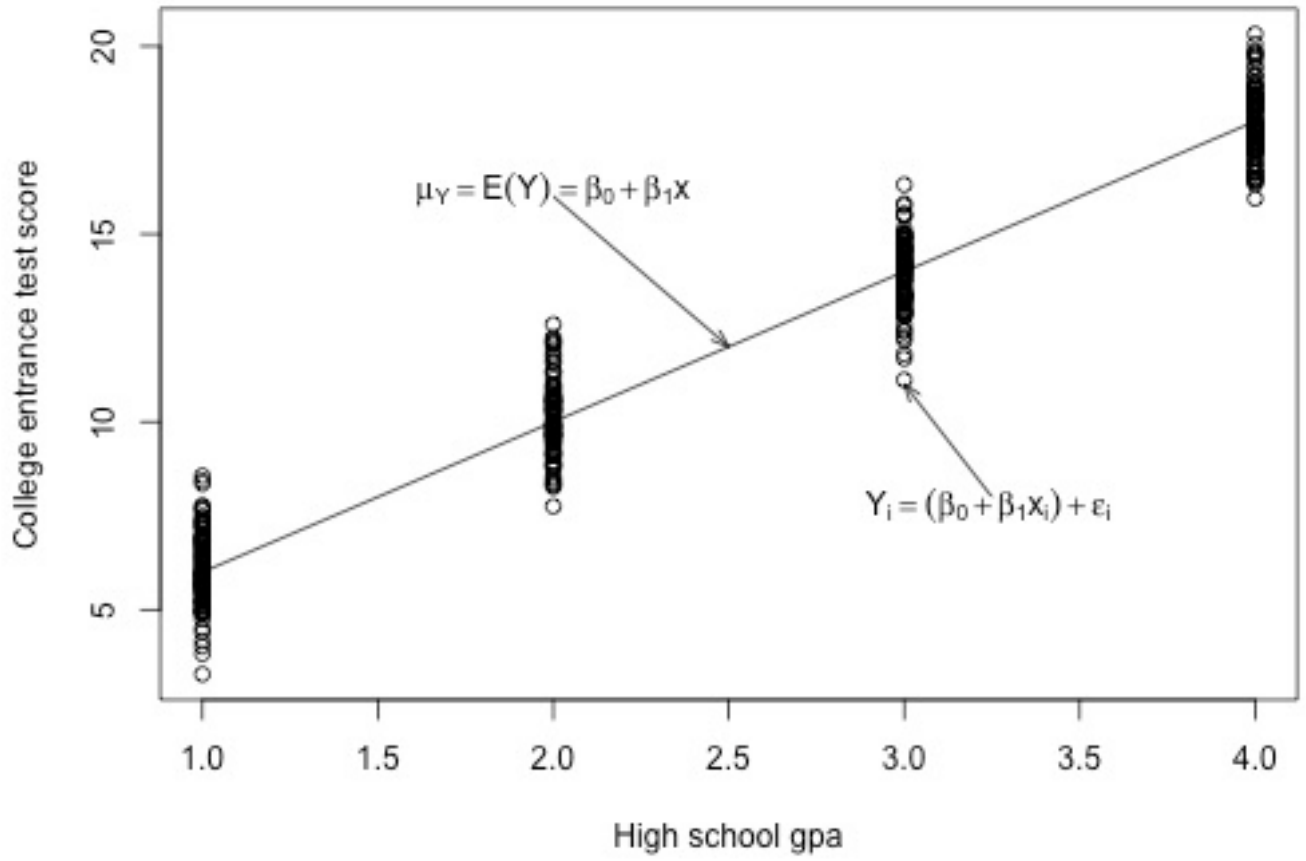
R output

```
Call:
lm(formula = Salary ~ YearsExperience, data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max
-7325.1 -3814.4   427.7  3559.7  8884.6

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)         25592       2646   9.672 1.49e-08 ***
YearsExperience      9365        421  22.245 1.52e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5391 on 18 degrees of freedom
Multiple R-squared:  0.9649,    Adjusted R-squared:  0.963
F-statistic: 494.8 on 1 and 18 DF,  p-value: 1.524e-14
```

```
> anova(regressor)
Analysis of Variance Table

Response: Salary
                Df     Sum Sq    Mean Sq F value    Pr(>F)
YearsExperience  1 1.4379e+10 1.4379e+10  494.84 1.524e-14 ***
Residuals       18 5.2305e+08 2.9058e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

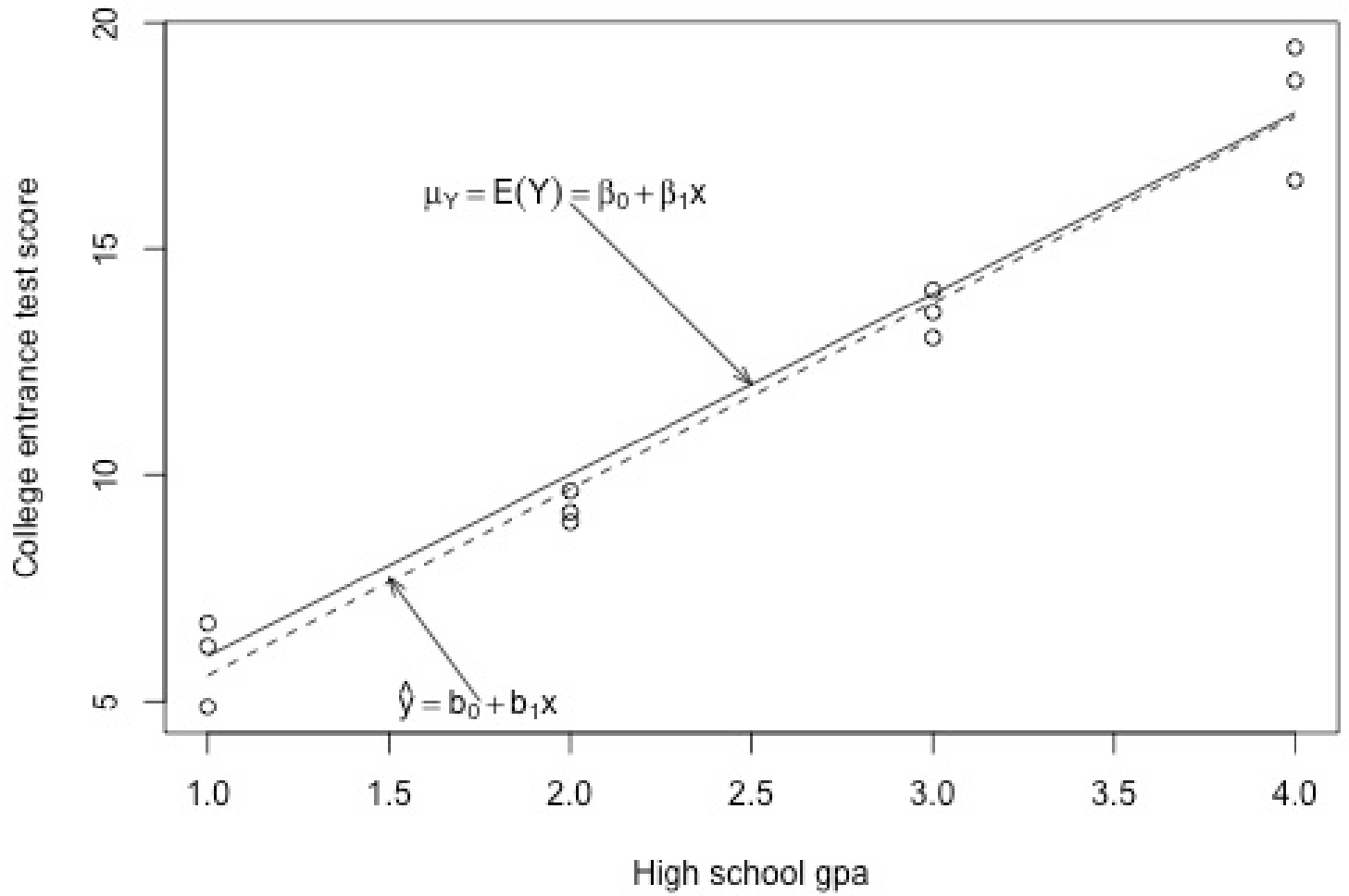
Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

What do b0 and b1 estimate?



Population Data

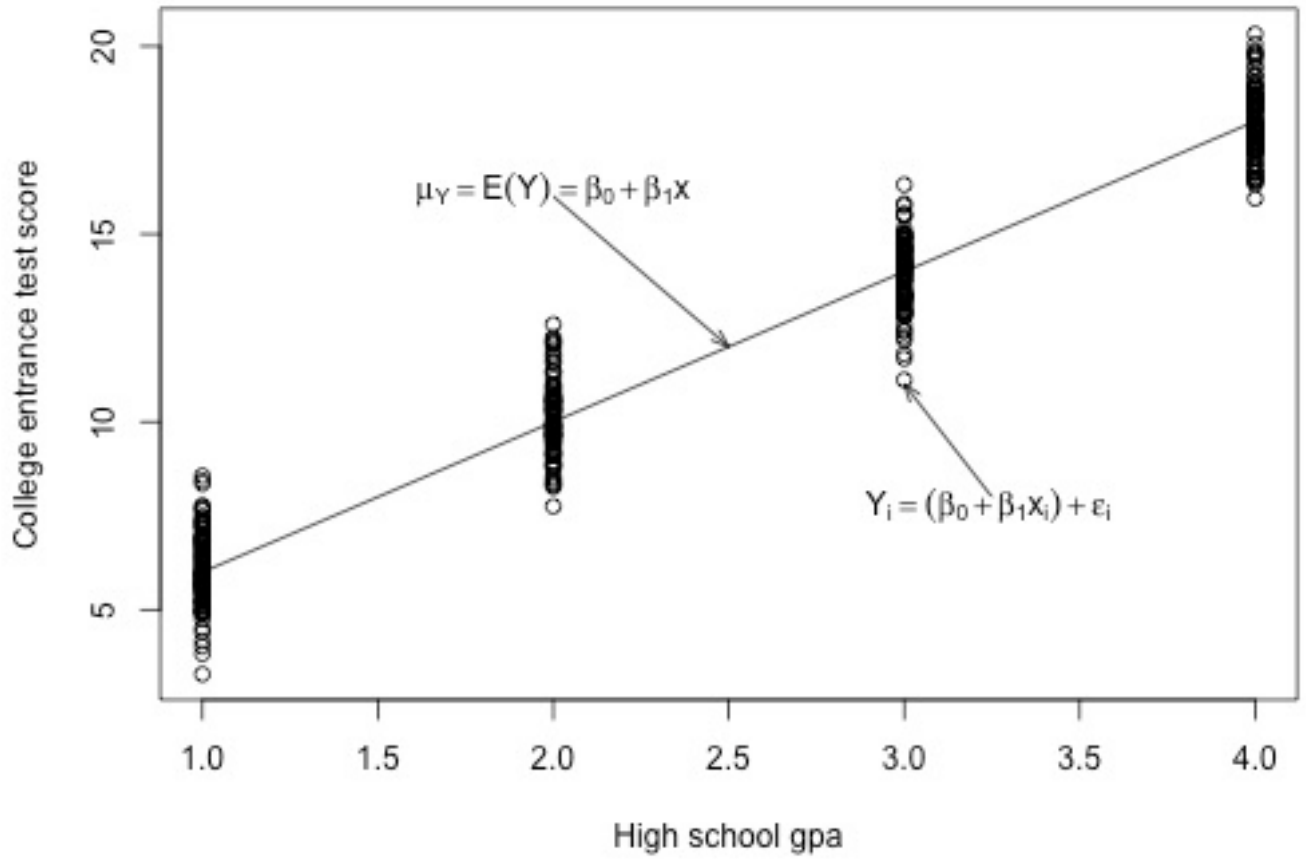
What do b0 and b1 estimate?



Sample Data

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

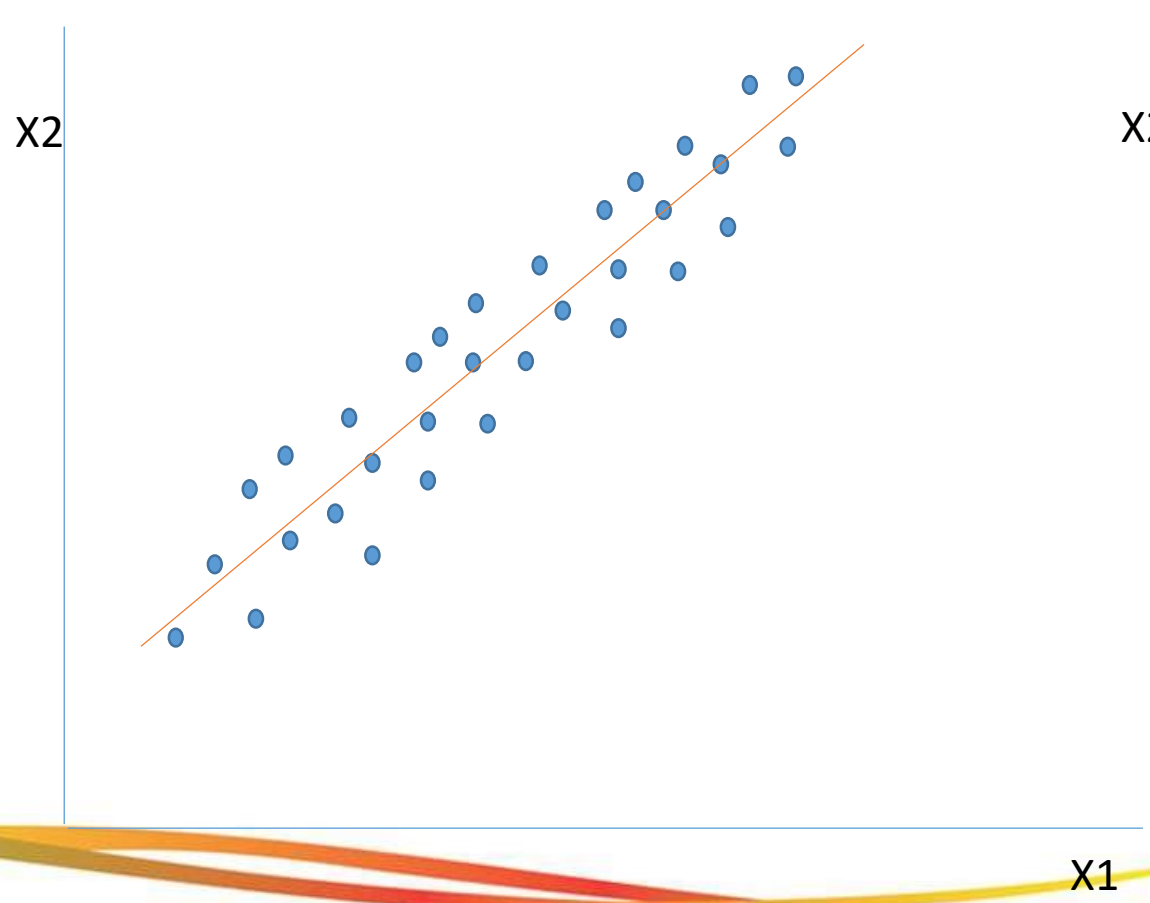
Assumptions of Linear Regression

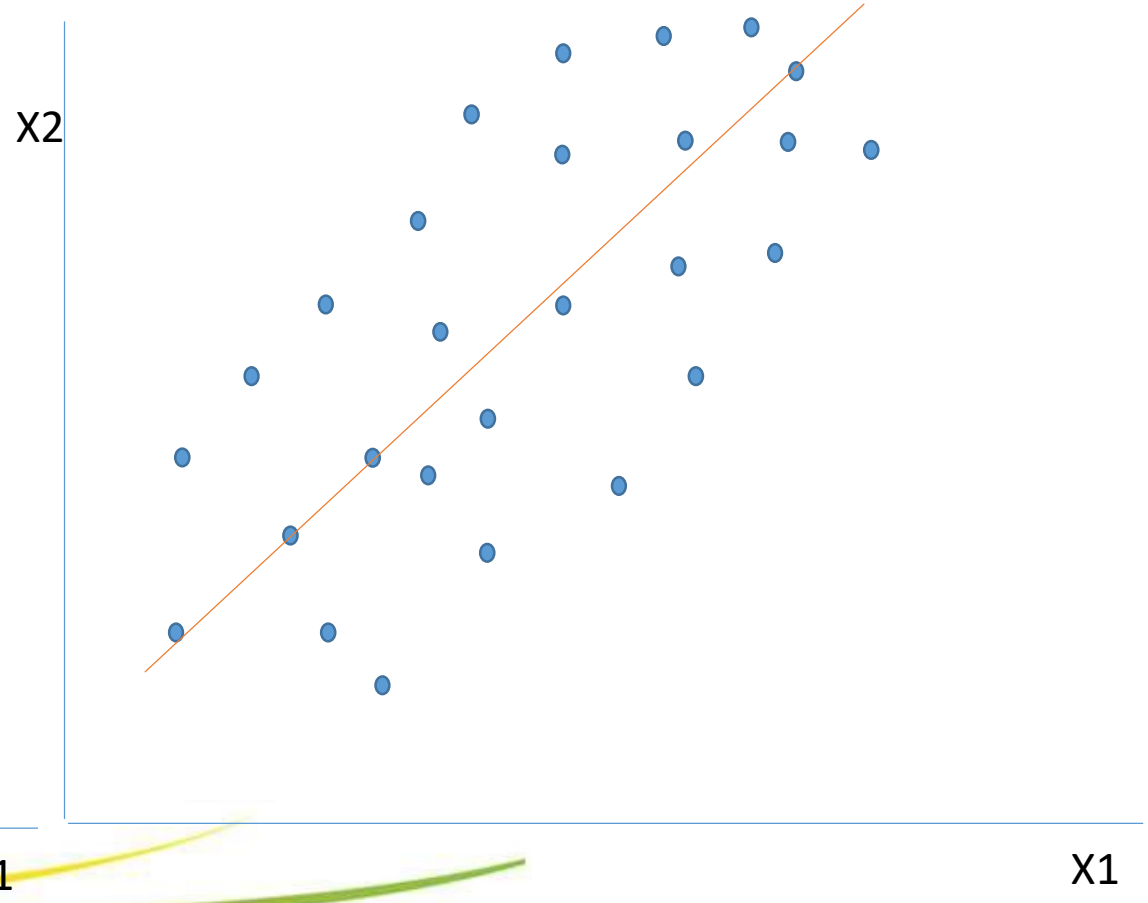


Population Data

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Common Error Variance (σ^2)





Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Common Error Variance (σ^2)

$$\sigma^2 \sim \text{MSE}$$

Population

For Sample

$$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Common Error Variance (σ^2)

```
Call:
lm(formula = Salary ~ YearsExperience, data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max
-7325.1 -3814.4   427.7  3559.7  8884.6

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)        25592       2646   9.672 1.49e-08 ***
YearsExperience     9365        421  22.245 1.52e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5391 on 18 degrees of freedom
Multiple R-squared:  0.9649,    Adjusted R-squared:  0.963
F-statistic: 494.8 on 1 and 18 DF,  p-value: 1.524e-14
```

```
> anova(regressor)
Analysis of Variance Table

Response: Salary
                Df     Sum Sq    Mean Sq F value    Pr(>F)
YearsExperience  1 1.4379e+10 1.4379e+10  494.84 1.524e-14 ***
Residuals       18 5.2305e+08 2.9058e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Sqrt(MSE)

MSE

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Coefficient of Determination (R^2)

```
Call:
lm(formula = Salary ~ YearsExperience, data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max
-7325.1 -3814.4   427.7  3559.7  8884.6

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)        25592       2646   9.672 1.49e-08 ***
YearsExperience     9365        421  22.245 1.52e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5391 on 18 degrees of freedom
Multiple R-squared:  0.9649,    Adjusted R-squared:  0.963
F-statistic: 494.8 on 1 and 18 DF,  p-value: 1.524e-14
```

SSR

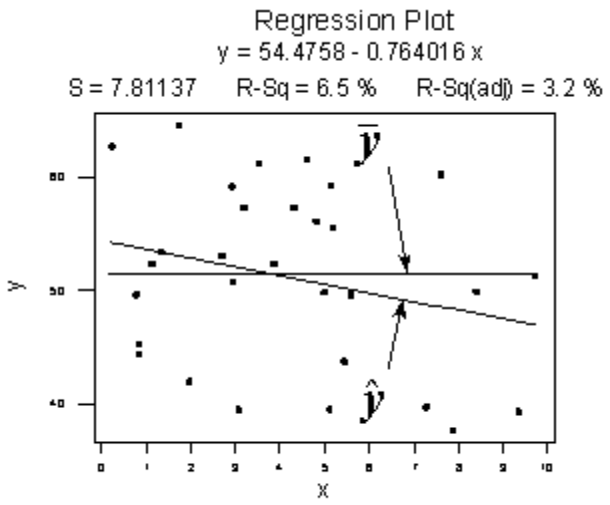
```
> anova(regressor)
Analysis of Variance Table

Response: Salary
                Df     Sum Sq    Mean Sq F value    Pr(>F)
YearsExperience  1 1.4379e+10 1.4379e+10  494.84 1.524e-14 ***
Residuals       18 5.2305e+08 2.9058e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSE

$$R^2 = 1\text{-}SSE/SSTO$$

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

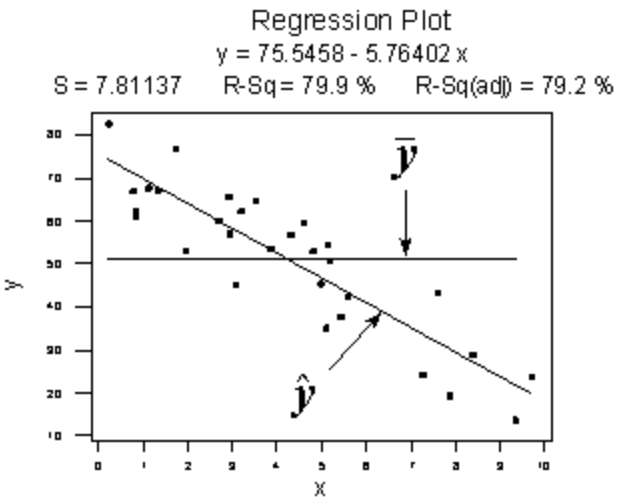
Coefficient of Determination (R^2)



$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 119.1$

$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 1708.5$

$SSTO = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 1827$



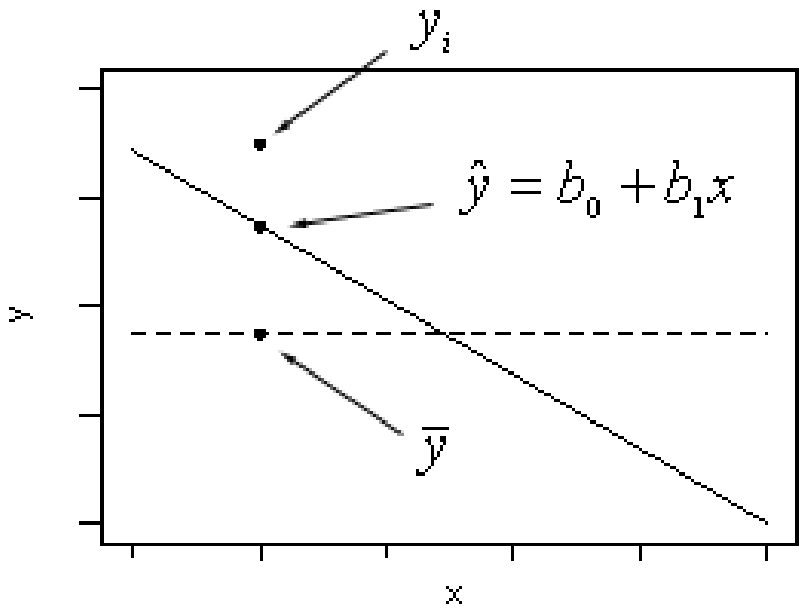
$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 6679.3$

$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 1708.5$

$SSTO = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 8487.8$

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Coefficient of Determination (R^2)



$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

SSTO Total sum of squares

SSR Regression sum of squares

SSE Error sum of squares

$$SSTO = SSR + SSE$$

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Coefficient of Determination (R^2)

- What is the range of values of R^2?
- How to interpret R^2?

Examples - r^2 and r

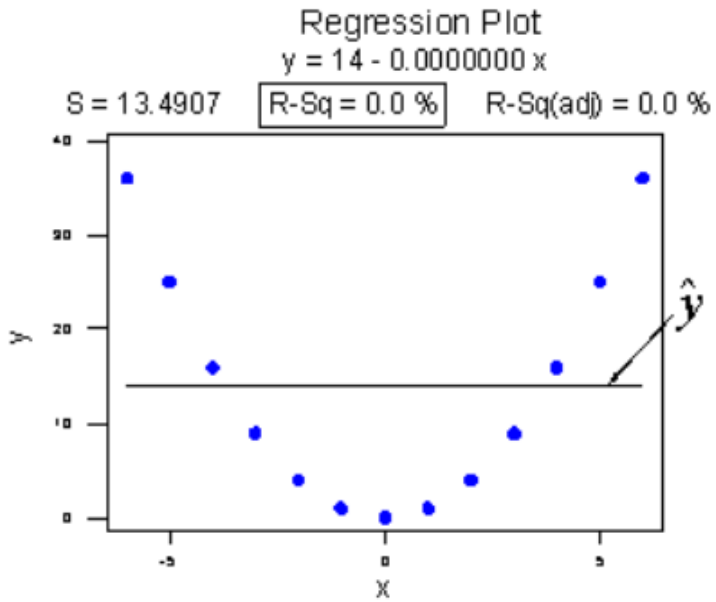
Be Cautious While Interpreting r^2 and r!!!

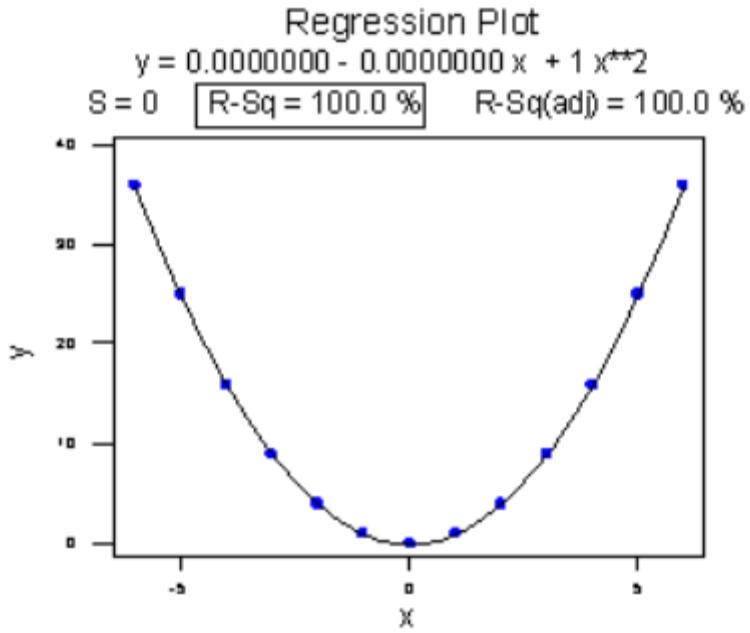
Be Cautious While Interpreting r^2 and r!!!

- **Caution 1**: r^2 and r quantify the strength of a *linear* relationship. It is possible that $r^2 = 0\%$ and $r = 0$, suggesting there is no linear relation between *x* and *y*, and yet a perfect curved (or "curvilinear" relationship) exists.

Be Cautious While Interpreting r^2 and r!!!

- **Caution 1**: r^2 and r quantify the strength of a *linear* relationship. It is possible that r^2 = 0% and r = 0, suggesting there is no linear relation between x and y, and yet a perfect curved (or "curvilinear" relationship) exists.





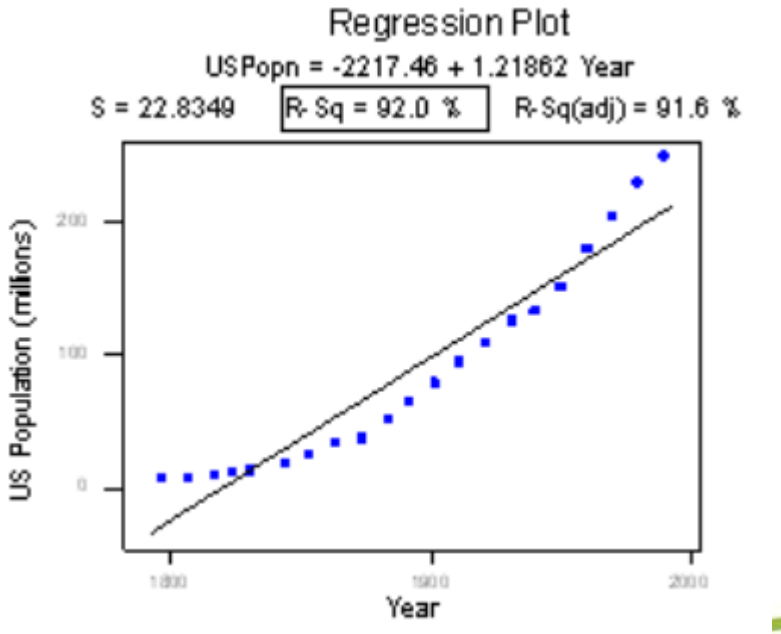
Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Be Cautious While Interpreting r^2 and r!!!

- **Caution 2**: A large r^2 value should not be interpreted as meaning that the estimated regression line fits the data well. Another function might better describe the trend in the data

Be Cautious While Interpreting r^2 and r!!!

- **Caution 2**: A large r^2 value should not be interpreted as meaning that the estimated regression line fits the data well. Another function might better describe the trend in the data

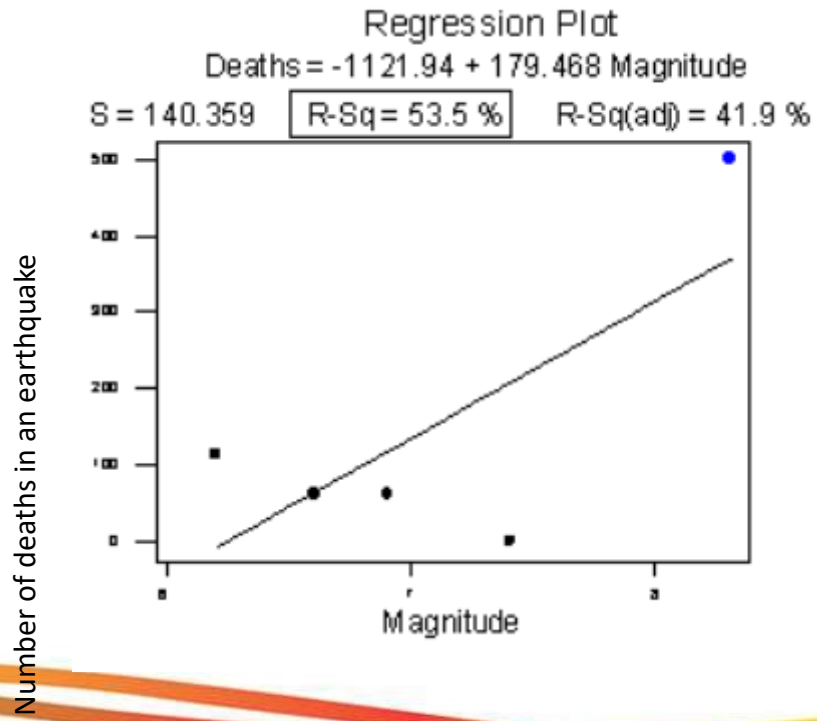


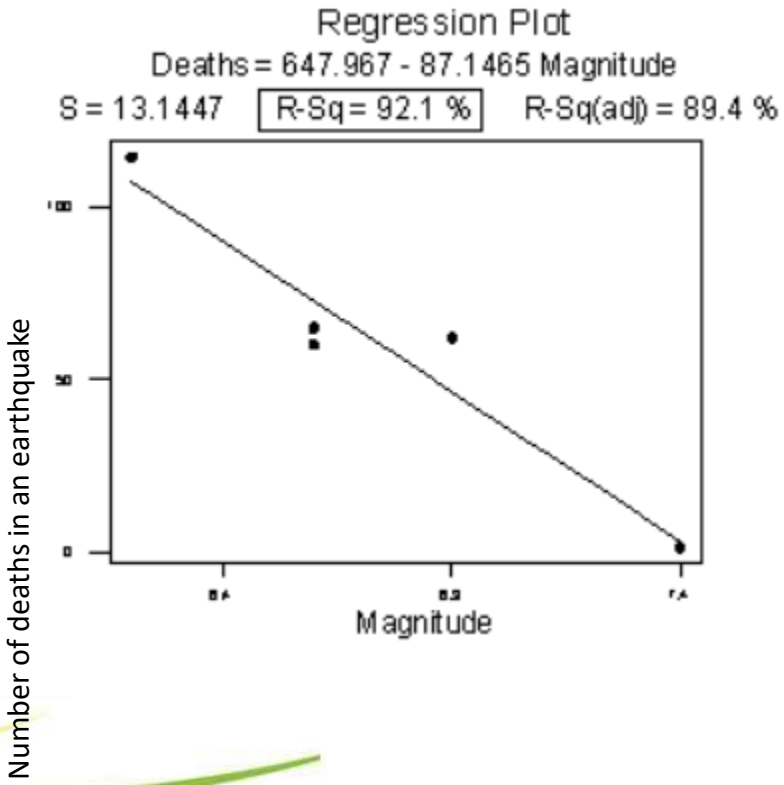
Be Cautious While Interpreting r^2 and r!!!

- **Caution 3**: The coefficient of determination r^2 and the correlation coefficient *r* can both be greatly affected by just one data point (or a few data points)

Be Cautious While Interpreting r^2 and r!!!

- **Caution 3**: The coefficient of determination r^2 and the correlation coefficient *r* can both be greatly affected by just one data point (or a few data points)





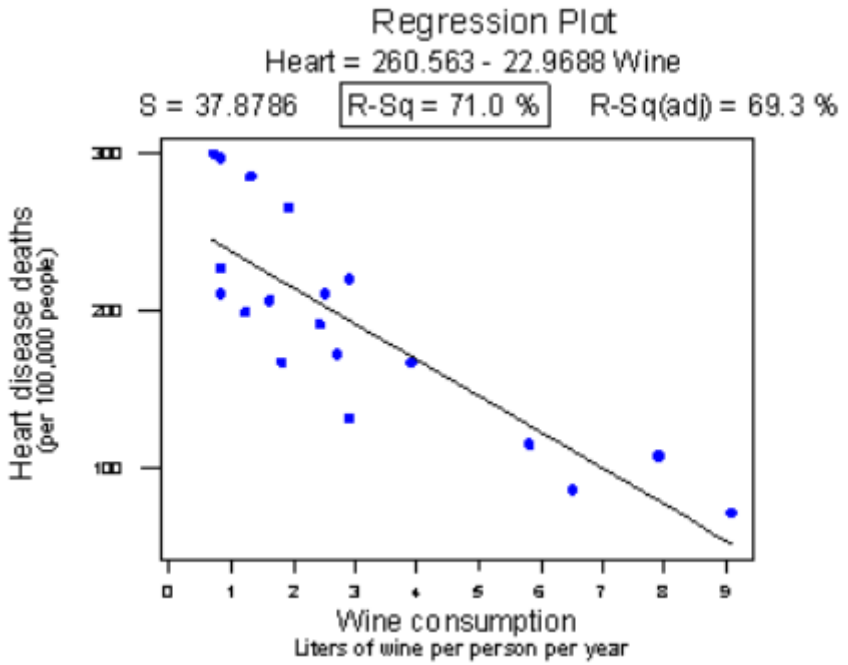
Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Be Cautious While Interpreting r^2 and r!!!

- **Caution 4**: Correlation (or association) does not imply causation
  - The predictor *x* does indeed cause the changes in the response *y*.
  - The causal relation may instead be reversed. That is, the response *y* may cause the changes in the predictor *x*.
  - The predictor *x* is a contributing but not sole cause of changes in the response variable *y*.
  - There may be a "lurking variable" that is the real cause of changes in *y* but also is associated with *x*, thus giving rise to the observed relationship between *x* and *y*.
  - The association may be purely coincidental

Be Cautious While Interpreting r^2 and r!!!

- **Caution 4**: Correlation (or association) does not imply causation



Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City,
Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Be Cautious While Interpreting r^2 and r!!!

- **Caution 5**: A large r^2 doesn't mean an accurate prediction can be made for a new data point. It is still possible to get prediction intervals or confidence intervals that are too wide to be useful. We will learn more on this later…

Examples - Caution in Interpreting r^2

- A large r^2 value should not be interpreted as meaning that the estimated regression line fits the data well

The American Automobile Association has published data (Defensive Driving: Managing Time and Space, 1991) that looks at the relationship between the average stopping distance (*y = distance*, in feet) and the speed of a car (*x = speed*, in miles per hour). The data set carstopping.txt contains 63 such data points.

- Create a fitted line plot of the data. Does a line do a good job of describing the trend in the data?
- Interpret the r^2 value. Does car speed explain a large portion of the variability in the average stopping distance? That is, is the r^2value large?
- Summarize how the title of this section is appropriate.

Examples - Caution in Interpreting r^2

- One data point can greatly affect the r^2 value

The mccoo.txt data set contains data on the running back Eric McCoo's rushing yards (*mccoo*) for each game of the 1998 Penn State football season. It also contains Penn State's final score (*score*).

- Create a fitted line plot. Interpret the r^2 value, and note its size.
- Remove the one data point in which McCoo ran 206 yards. Then, create another fitted line plot on the reduced data set. Interpret the r^2 value. Upon removing the one data point, what happened to the r^2 value?

Examples - Caution in Interpreting R^2

- Association is not causation

"Time" is often a lurking variable. If two things (e.g. road deaths and chocolate consumption) just happen to be increasing over time for totally unrelated reasons, a scatter plot will suggest there is a relationship, regardless of it existing only because of the lurking variable "time." The data set drugdea.txt contains data on drug law expenditures and drug-induced deaths. The data set gives figures from 1981 to 1991 on the U.S. Drug Enforcement Agency budget (*budget*) and the numbers of drug-induced deaths in the United States (*deaths*).

- Create a fitted line plot treating *deaths* as the response *y* and *budget* as the predictor *x*. Do you think the budget caused the deaths?
- Create a fitted line plot treating *budget* as the response *y* and *deaths* as the predictor *x*. Do you think the deaths caused the budget?
- Create a fitted line plot treating *budget* as the response *y* and *year* as the predictor *x*.
- Create a fitted line plot treating *deaths* as the response *y* and *year* as the predictor *x*.
- What is going on here? Summarize the relationships between *budget*, *deaths*, and *year* and explain why it might appear that as drug-law expenditures increase, so do drug-induced deaths.

Model Evaluation

```
Call:
lm(formula = Salary ~ YearsExperience, data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max
-7325.1 -3814.4   427.7  3559.7  8884.6

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)         25592       2646   9.672 1.49e-08 ***
YearsExperience      9365        421  22.245 1.52e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5391 on 18 degrees of freedom
Multiple R-squared:  0.9649,    Adjusted R-squared:  0.963
F-statistic: 494.8 on 1 and 18 DF,  p-value: 1.524e-14
```

```
> anova(regressor)
Analysis of Variance Table

Response: Salary
                Df     Sum Sq    Mean Sq F value    Pr(>F)
YearsExperience  1 1.4379e+10 1.4379e+10  494.84 1.524e-14 ***
Residuals       18 5.2305e+08 2.9058e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Evaluation

- How do we draw conclusions about the population from the sample we observed?
- Answer - Confidence Interval

```
Call:
lm(formula = Salary ~ YearsExperience, data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max
-7325.1 -3814.4   427.7  3559.7  8884.6

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)        25592       2646   9.672 1.49e-08 ***
YearsExperience     9365        421  22.245 1.52e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5391 on 18 degrees of freedom
Multiple R-squared:  0.9649,    Adjusted R-squared:  0.963
F-statistic: 494.8 on 1 and 18 DF,  p-value: 1.524e-14
```

95% confidence interval about β_1= (9365 + 2.101 * 421), (9365 - 2.101 * 421) i.e. (10249.521, 8480.479)

Example 1 - Model Evaluation

Is there a *positive* relationship between sales of leaded gasoline and lead burden in the bodies of newborn infants? Researchers (Rabinowitz, *et al*, 1984) who were interested in answering this research question compiled data (leadcord.txt) on the monthly gasoline lead sales (in metric tons) in Massachusetts and mean lead concentrations (µl/dl) in umbilical-cord blood of babies born at a major Boston hospital over 14 months in 1980-1981.

Lets evaluate the model using p-value and the CI…

Example 2 - Model Evaluation

Is there a (linear) relationship between height and grade point average (heightgpa.txt)?

Lets evaluate the model using p-value and the CI…

Getting Back To SLR Assumptions

How do we know if the model we are using is good?

- Assess whether the assumptions underlying the simple linear regression model seem reasonable when applied to the dataset in question
- Since the assumptions relate to the (population) prediction errors, we do this through the study of the (sample) estimated errors, the residuals

Why Do We Need To Care About Assumptions?

- Remember the LINE assumptions??
  - All of the estimates, intervals, and hypothesis tests arising in a regression analysis have been developed assuming that the model is correct. That is, all the formulas depend on the model being correct!
  - If the model is incorrect, then the formulas and methods we use are at risk of being incorrect.
- We will learn about what happens if any of the LINE assumptions are violated
- Use diagnostics as an art!

Residual Analysis

Using sample residuals to estimate about population errors

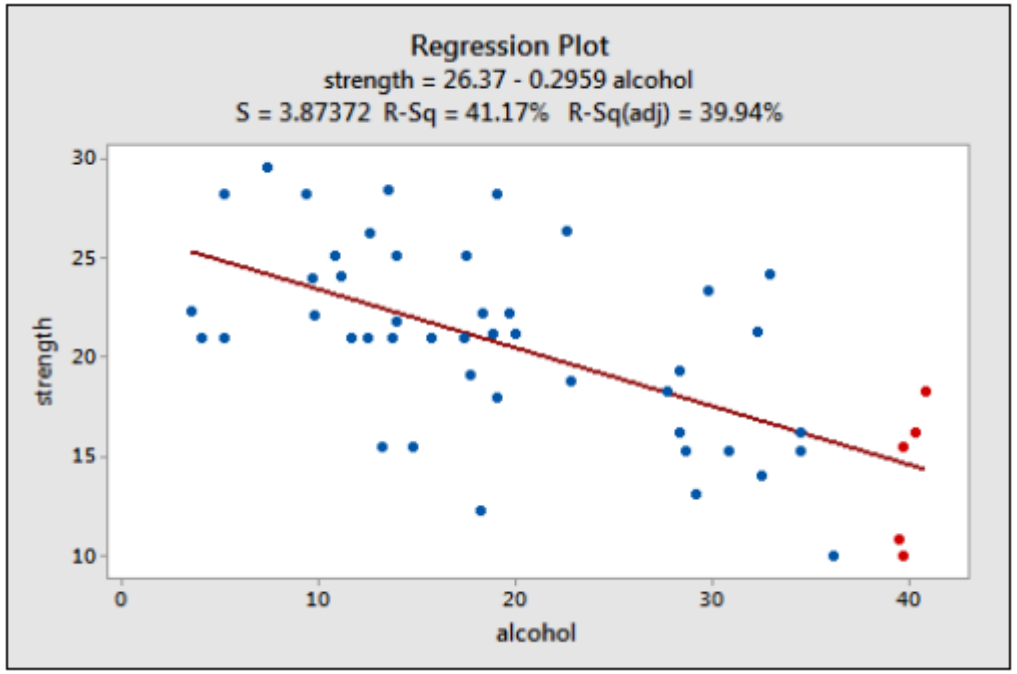
$e_i = y_i - \hat{y}_i$ (Sample residual)

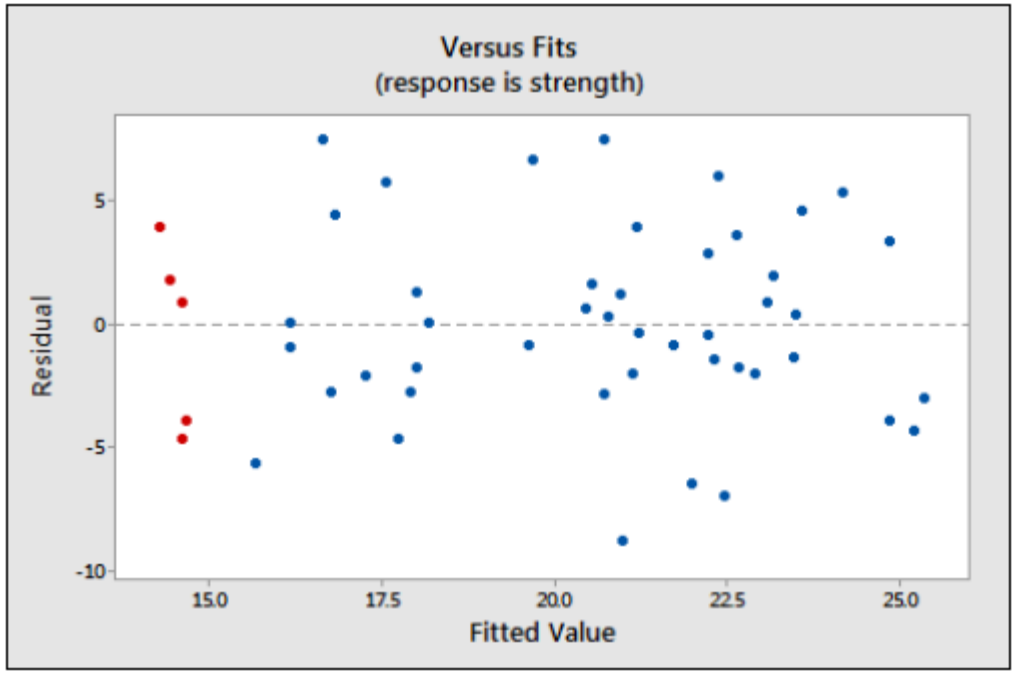
$\epsilon_i = Y_i - E(Y_i)$ (Actual true error)

Diagnosis Tool 1 – Residuals vs Fits Plot

- Used to detect non-linearity, unequal error variances, and outliers
- Lets check out how an ideal plot should look like, using an example
- Some researchers were interested in determining whether or not alcohol consumption was linearly related to muscle strength. The researchers measured the total lifetime consumption of alcohol (x) on a random sample of $n = 50$ alcoholic men. They also measured the strength (y) of the deltoid muscle in each person's nondominant arm (alcoholarm.txt)

Diagnosis Tool 1 – Residuals vs Fits Plot





CAUTION – DO NOT OVER-INTERPRET THE PLOT!!!

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City,
Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

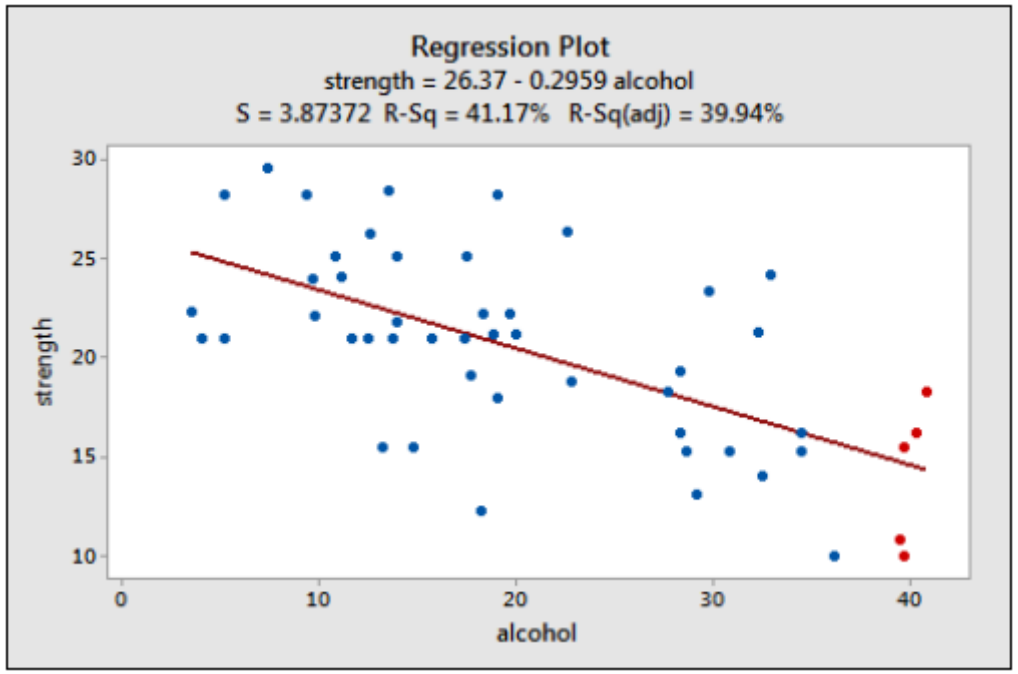
Diagnosis Tool 2 – Residuals vs Predictor Plot

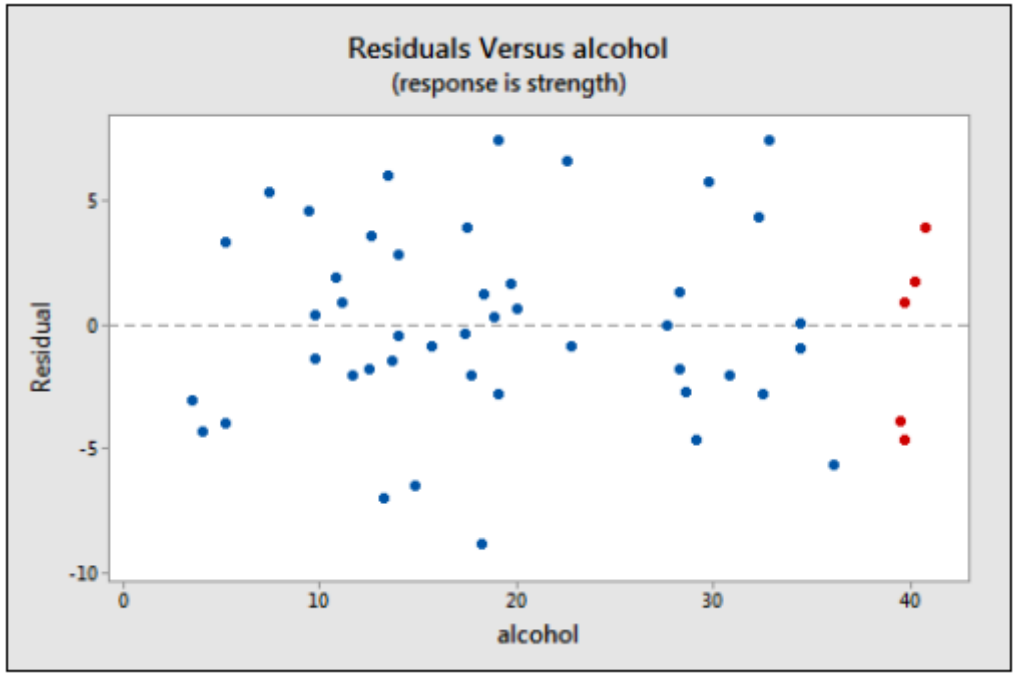
Diagnosis Tool 2 – Residuals vs Predictor Plot

- If the predictor on the *x* axis is the same predictor that is used in the regression model, the residuals vs. predictor plot offers no new information to that which is already learned by the residuals vs. fits plot
- If the predictor on the *x* axis is a new and different predictor, the residuals vs. predictor plot can help to determine whether the predictor should be added to the model (and hence a multiple regression model used instead)

Diagnosis Tool 2 – Residuals vs Predictor Plot

- Lets first consider the case when the same predictor is used in the regression model



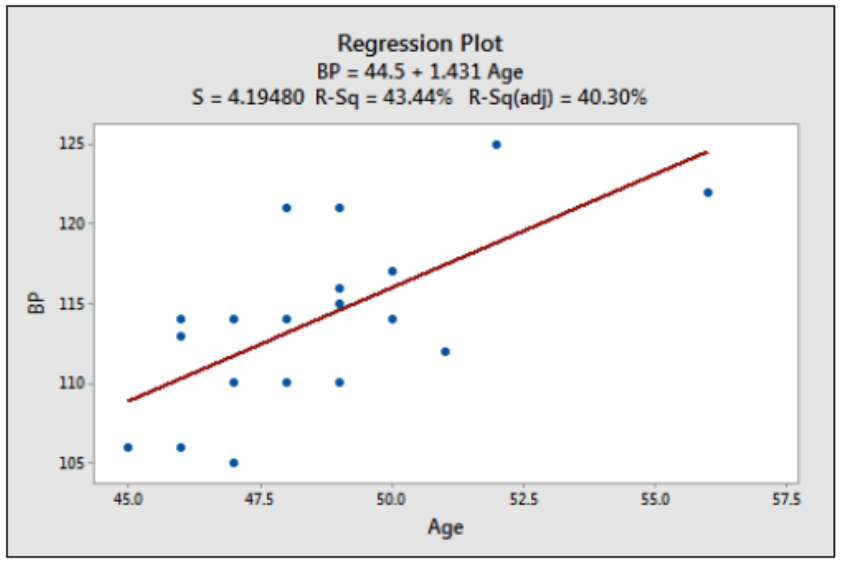


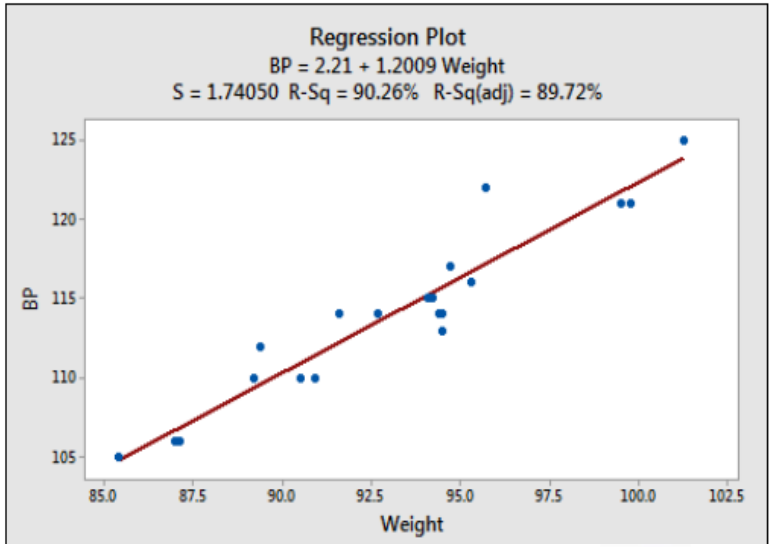
Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

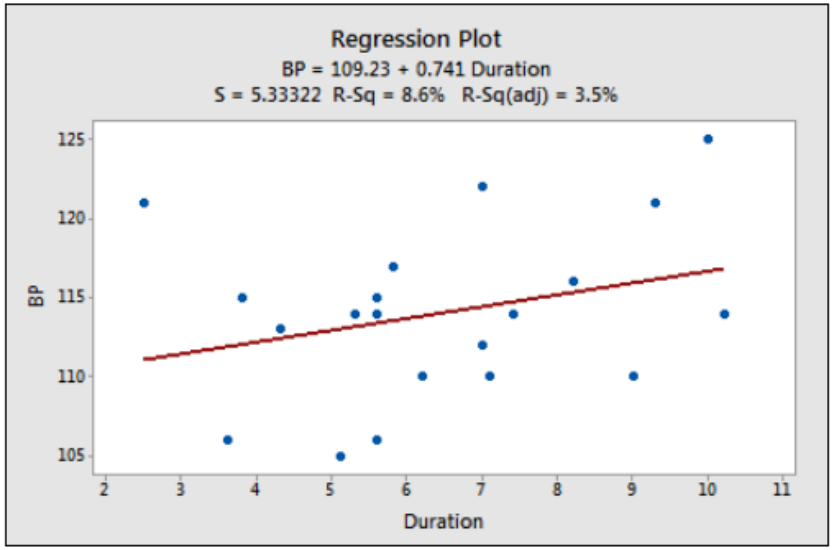
Diagnosis Tool 2 – Residuals vs Predictor Plot

- Lets now consider the case when the question is to add a new predictor or not
- A researcher is interested in determining which of the following — age, weight, and duration of hypertension — are good predictors of the diastolic blood pressure of an individual with high blood pressure. The researcher measured the age (in years), weight (in pounds), duration of hypertension (in years), and diastolic blood pressure (in mm Hg) on a sample of *n* = 20 hypertensive individuals (bloodpress.txt).

Diagnosis Tool 2 – Residuals vs Predictor Plot



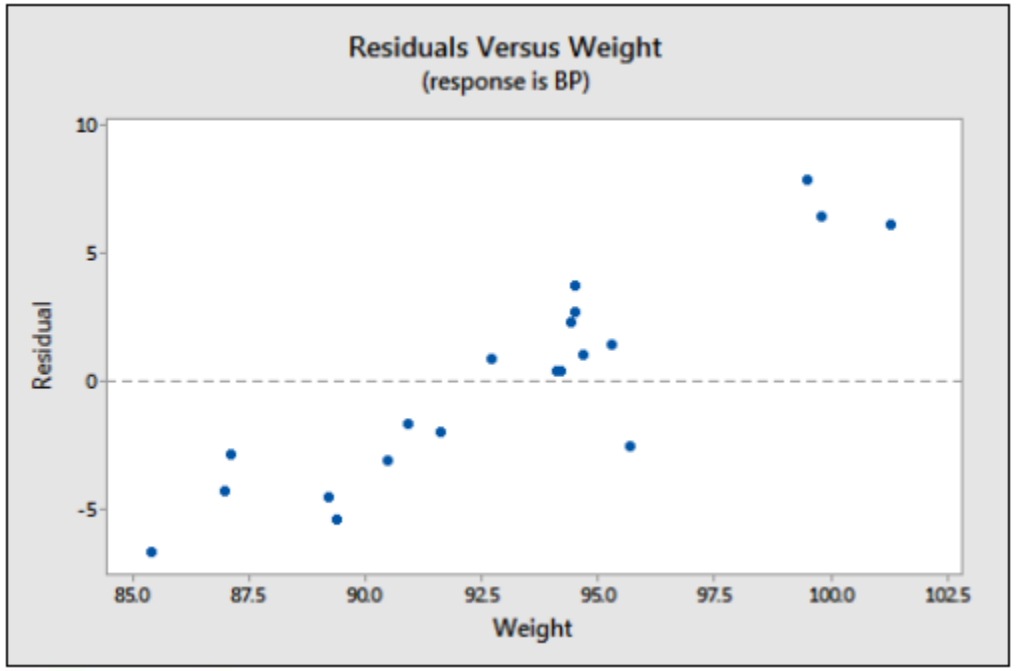


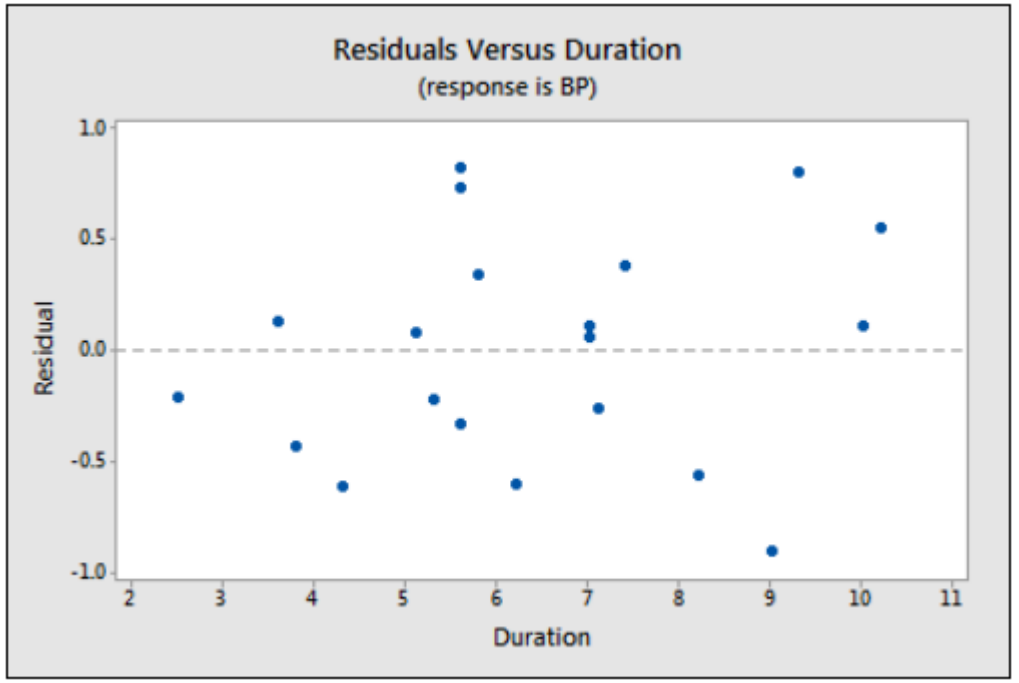


Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Diagnosis Tool 2 – Residuals vs Predictor Plot

Let's investigate various residuals vs. predictors plots to learn whether adding predictors to any of the above three simple linear regression models is advised. Upon regressing blood pressure on age, obtaining the residuals, and plotting the residuals against the predictor weight, we obtain the following "residuals versus weight" plot





Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Lets now understand how a non-ideal residual vs fits plot look like…

Things we will learn:

- How a non-linear regression function shows up on a residuals vs. fits plot
- How unequal error variances show up on a residuals vs. fits plot
- How an outlier show up on a residuals vs. fits plot

How a non-linear regression function shows up on a residuals vs. fits plot?

- The residuals depart from 0 in some *systematic manner*
- Lets understand this through an example…

How a non-linear regression function shows up on a residuals vs. fits plot?

- Is tire tread wear linearly related to mileage? A laboratory conducted an experiment in order to answer this research question. As a result of the experiment, the researchers obtained a data set (treadwear.txt) containing the mileage (*x*, in 1000 miles) driven and the depth of the remaining groove (*y*, in mils)

How a non-linear regression function shows up on a residuals vs. fits plot?

- Did you notice that the r^2 value is very high (95.26%)? This is an excellent example of the caution "a large r^2 value should not be interpreted as meaning that the estimated regression line fits the data well."
- The large r^2 value tells you that if you wanted to predict groove depth, you'd be better off taking into account mileage than not.
- The residuals vs. fits plot tells you, though, that your prediction would be better if you formulated a non-linear model rather than a linear one.

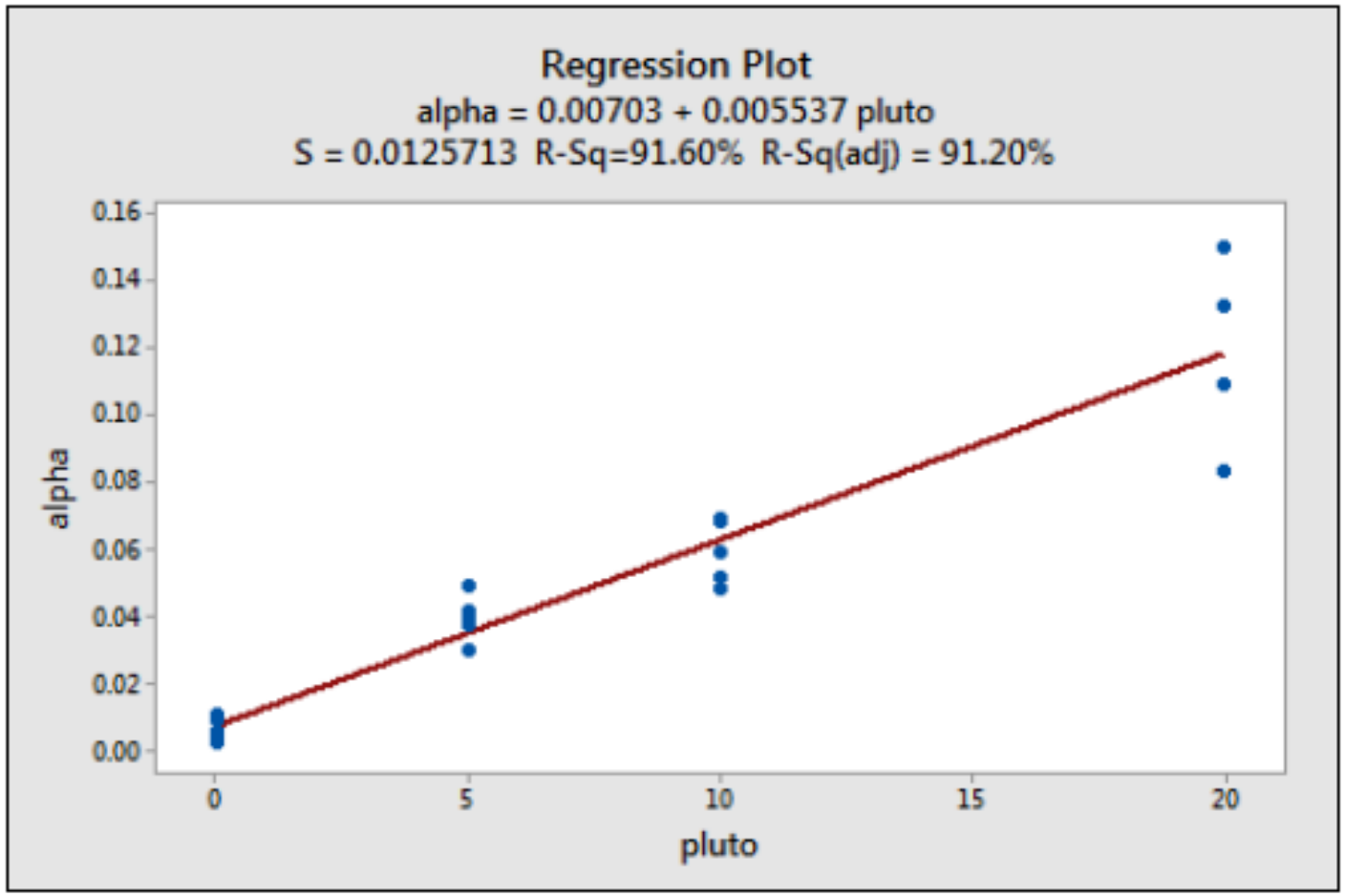
How does non-constant error variance (heteroscedasticity) show up on a residual vs. fits plot?

- The plot has a "**fanning**" effect. That is, the residuals are close to 0 for small *x* values and are more spread out for large *x* values.
- The plot has a "**funneling**" effect. That is, the residuals are spread out for small *x* values and close to 0 for large *x* values.
- Or, the spread of the residuals in the residuals vs. fits plot varies in some complex fashion
- Again, lets understand this through an example…

How does non-constant error variance (heteroscedasticity) show up on a residual vs. fits plot?

- How is plutonium activity related to alpha particle counts?
- Plutonium emits subatomic particles — called alpha particles. Devices used to detect plutonium record the intensity of alpha particle strikes in counts per second.
- To investigate the relationship between plutonium activity (*x*, in pCi/g) and alpha count rate (*y*, in number per second), a study was conducted on 23 samples of plutonium (alphapluto.txt)

How does non-constant error variance (heteroscedasticity) show up on a residual vs. fits plot?



Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

How does an outlier show up on a residuals vs. fits plot?

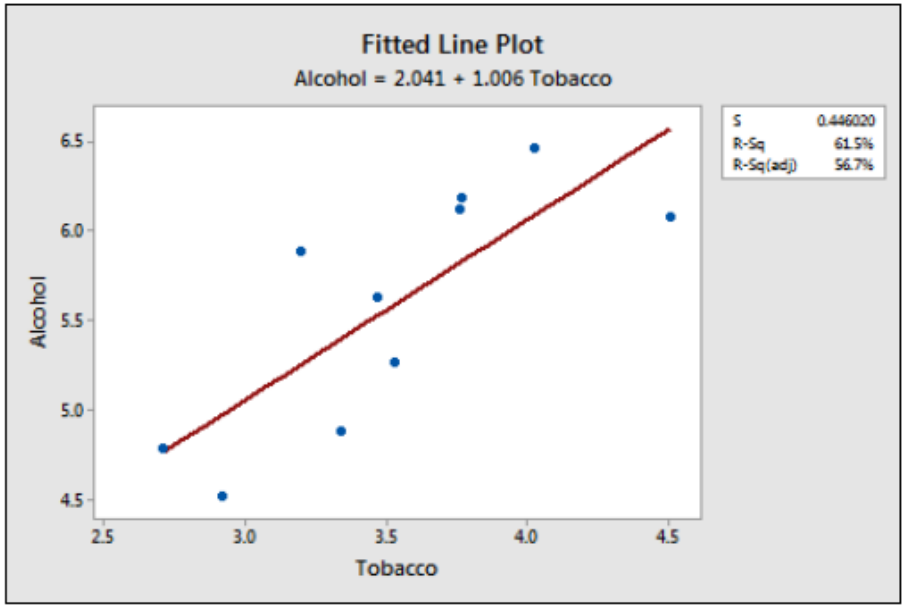
- The observation's residual stands apart from the basic random pattern of the rest of the residuals.
- The random pattern of the residual plot can even disappear if one outlier really deviates from the pattern of the rest of the data
- Another example…

How does an outlier show up on a residuals vs. fits plot?

- Is there a relationship between tobacco use and alcohol use?
- The British government regularly conducts surveys on household spending. One such survey determined the average weekly expenditure on tobacco (*x*, in British pounds) and the average weekly expenditure on alcohol (*y*, in British pounds) for households in *n* = 11 different regions in the United Kingdom (alcoholtobacco.txt)

How does an outlier show up on a residuals vs. fits plot?

- This is an excellent example of the caution that the "coefficient of determination r^2 can be greatly affected by just one data point." r^2 is only 5%.
- Removing Northern Ireland's data point from the data set, and refitting the regression line, we obtain



Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

How does an outlier show up on a residuals vs. fits plot?

- You might be wondering how large a residual has to be before a data point should be flagged as being an outlier.
- The answer is not straightforward, since the magnitude of the residuals depends on the units of the response variable. That is, if your measurements are made in pounds, then the units of the residuals are in pounds. And, if your measurements are made in inches, then the units of the residuals are in inches.
- Therefore, there is no one "rule of thumb" that we can define to flag a residual as being exceptionally unusual

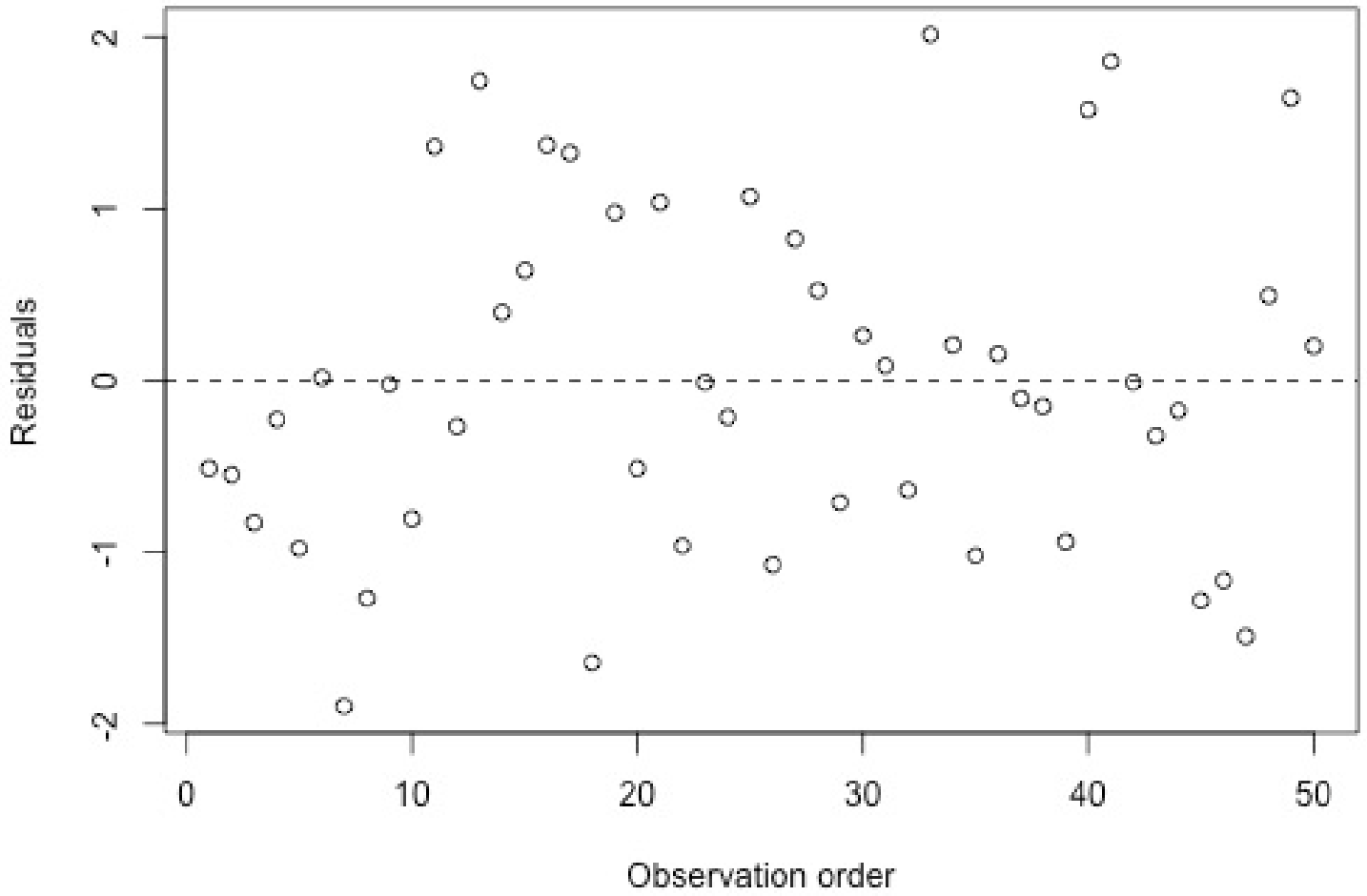
How does an outlier show up on a residuals vs. fits plot?

- There's a solution to this problem!
- We can make the residuals "unitless" by dividing them by their standard deviation. In this way we create what are called "**standardize/studentized residuals**."

Diagnosis Tool 3 – Residuals vs Order Plot

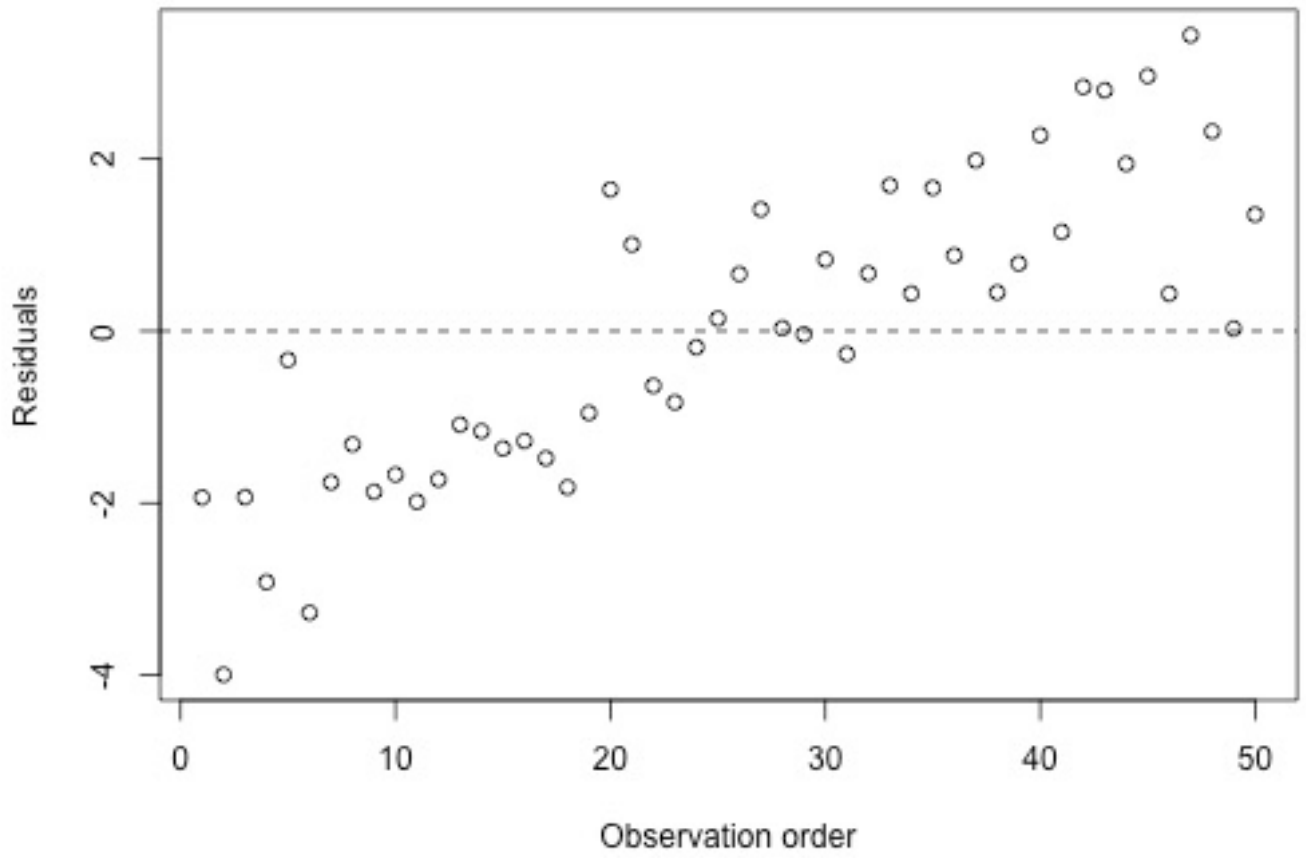
- Used to test the “I” (Independence of residuals) criteria of the model
- If the data are obtained **in a time (or space) sequence**, a residuals vs. order plot helps to see if there is any correlation between the error terms that are near each other in the sequence

Diagnostic Tools: Residuals vs Order Plot



Example of a well behaved plot

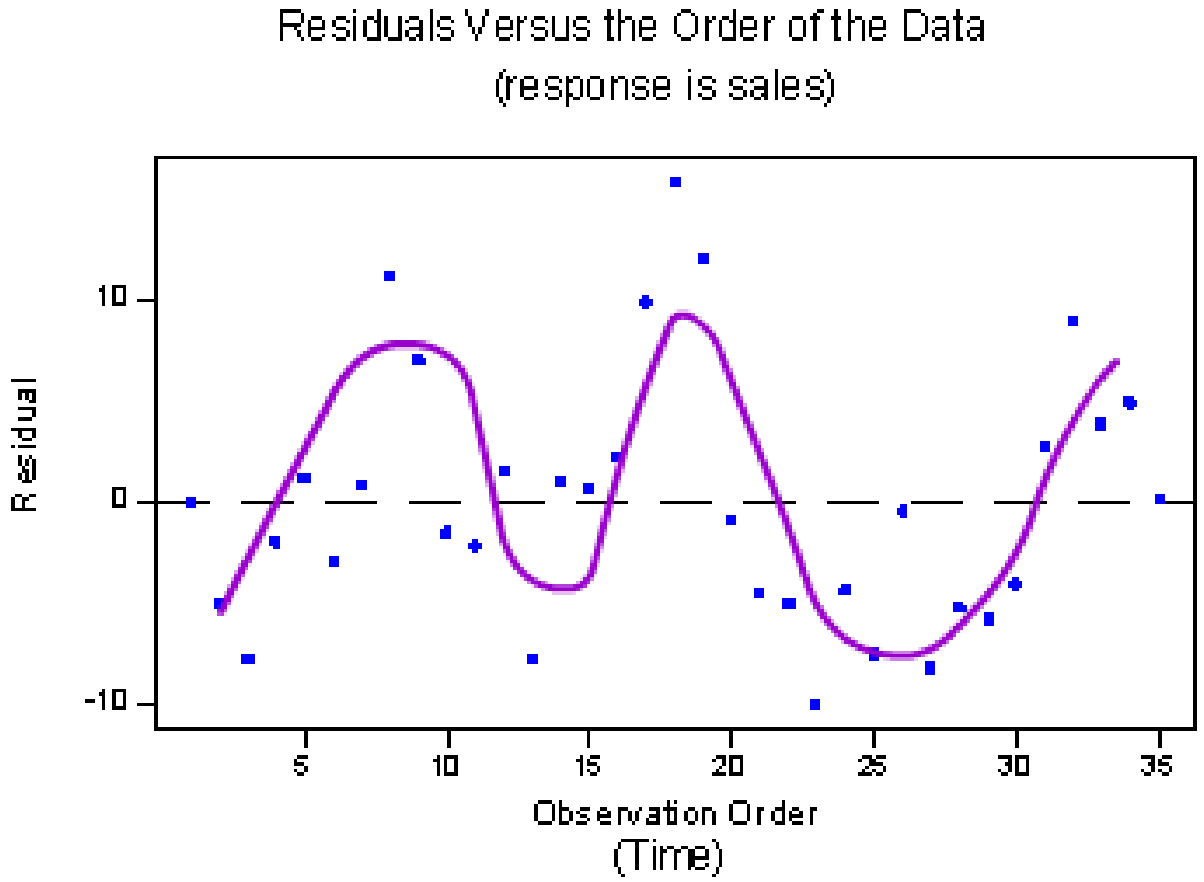
Diagnostic Tool 2 – Residuals vs Order Plot



Add the predictor "time" to the model

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City,
Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

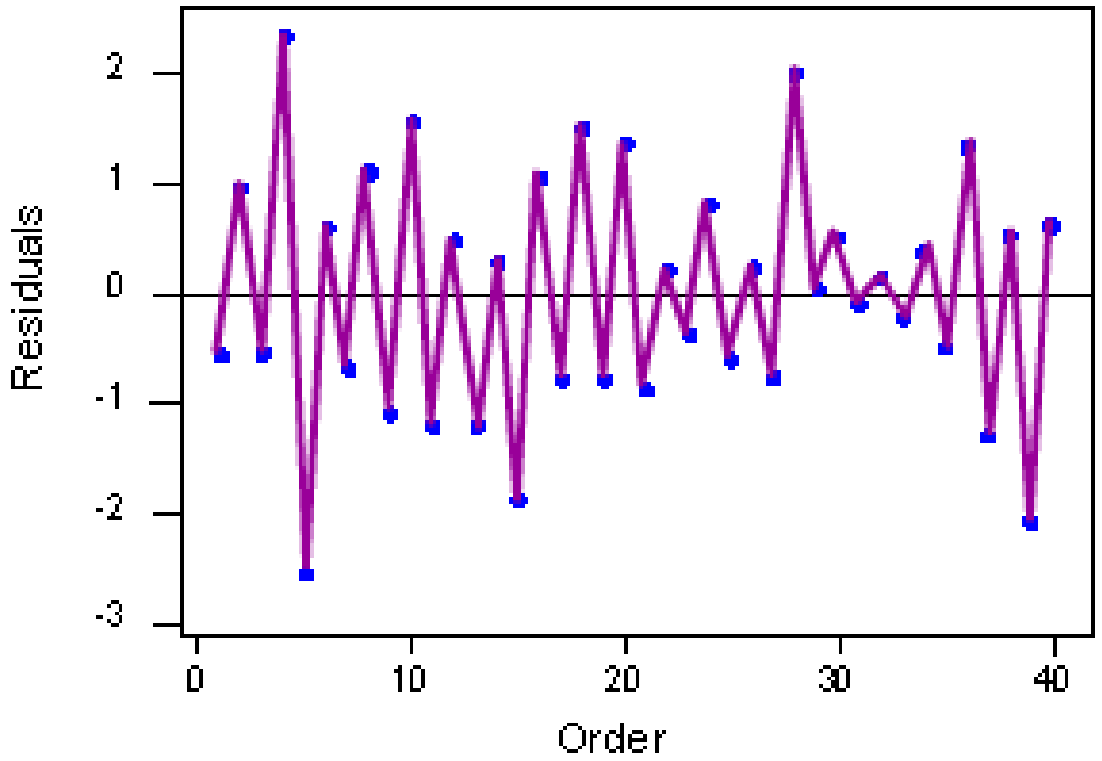
Diagnosis Tool 3 – Residuals vs Order Plot



Positive Serial Correlation

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Diagnosis T[illegible] Residuals Order Plot



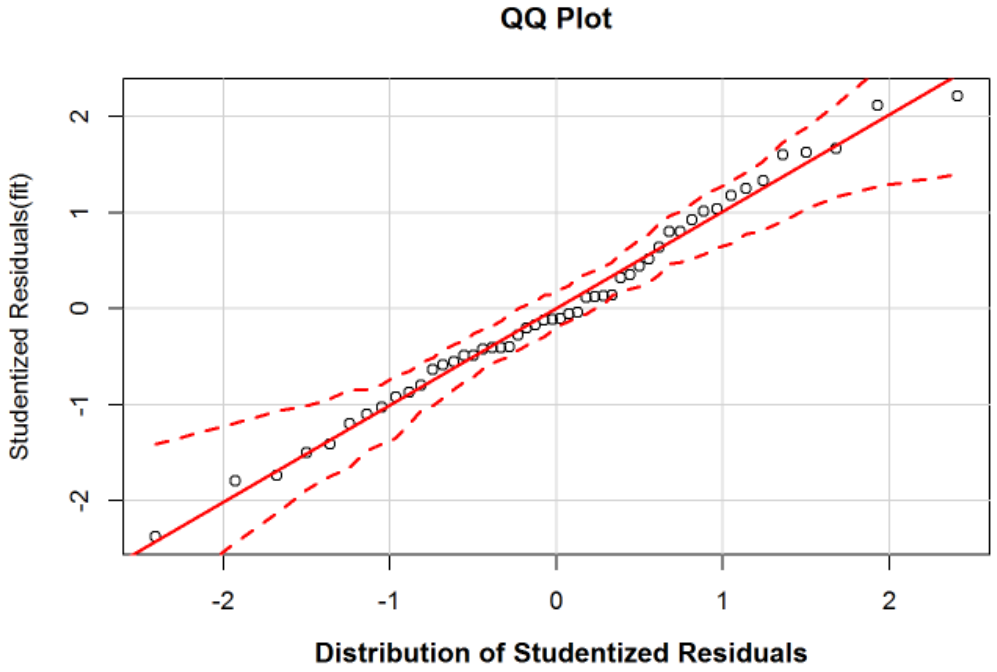
Negative Serial Correlation

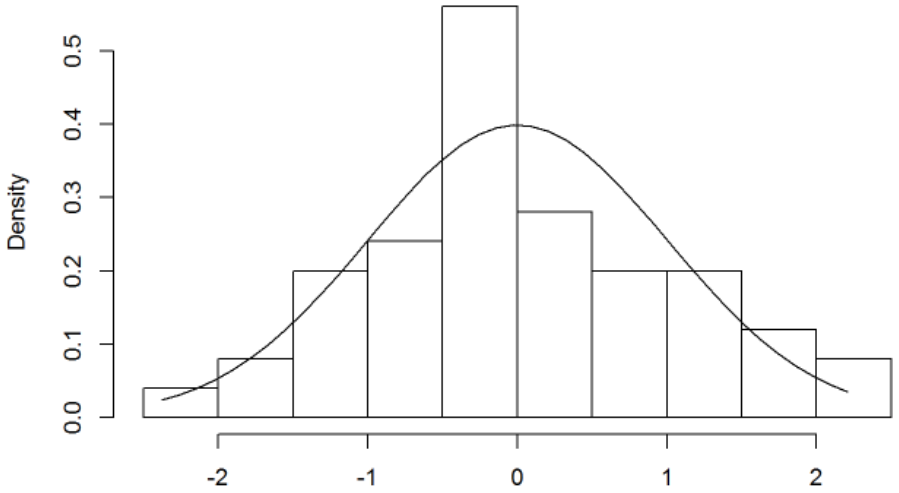
Diagnosis Tool 4 – Normal Probability Plot (QQ Plot) of Residuals

- Used to test the “N” (Normal distribution of residuals) criteria of the model
- Quantile-Quantile (Q-Q) plots are used to assess whether the model’s distribution of residuals (represented on the Y axis) roughly approximate a normal distribution of residuals (represented on the X axis).

Diagnosis Tool 4 – Normal Probability Plot (QQ Plot) of Residuals

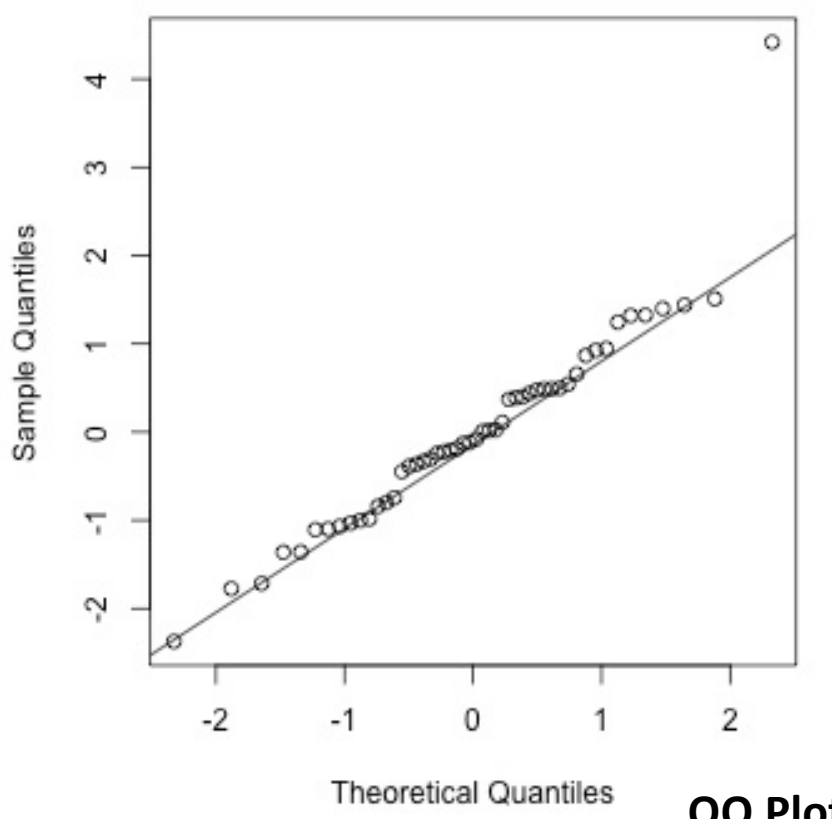
- The points should mostly fall on the diagonal line in the middle of the plot. If this assumption is violated, the points will fall in some sort of curve shape, such as an S, or will form two separate, variable lines
- qqPlot(residuals) [part of the car library]

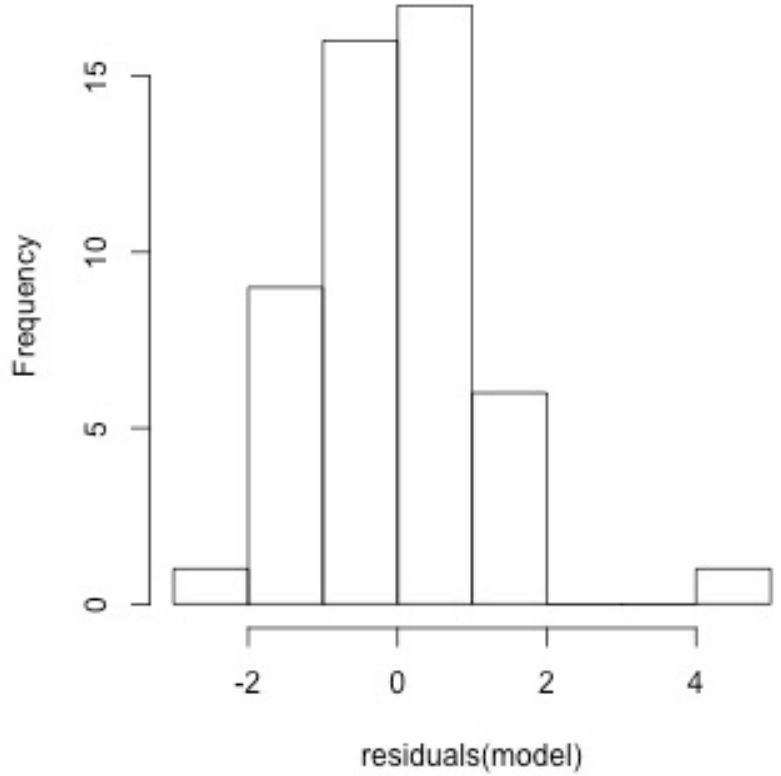




Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Diagnosis Tool 4 – Normal Probability Plot (QQ Plot) of Residuals

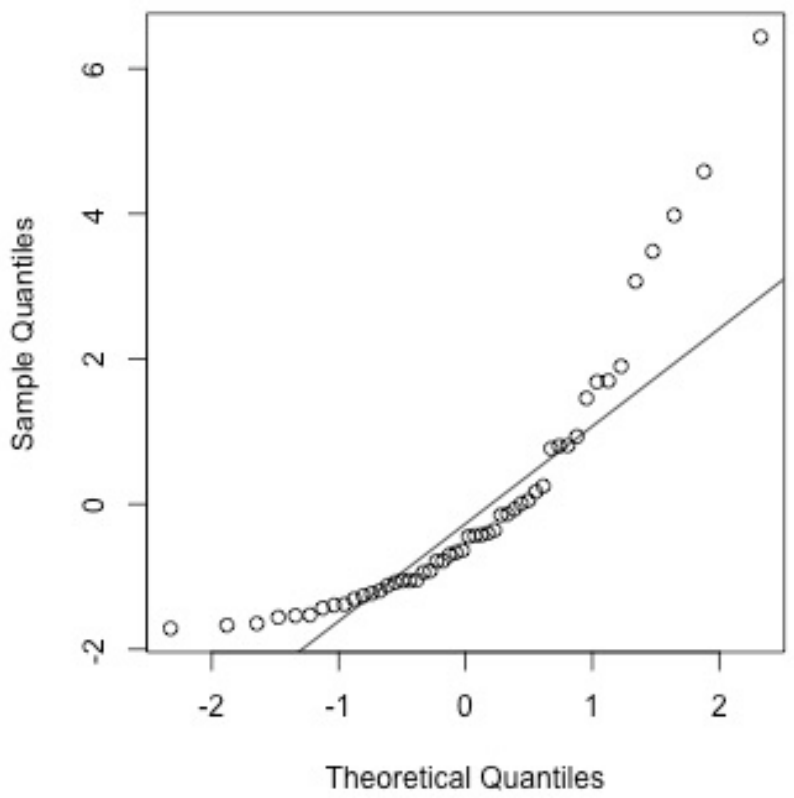


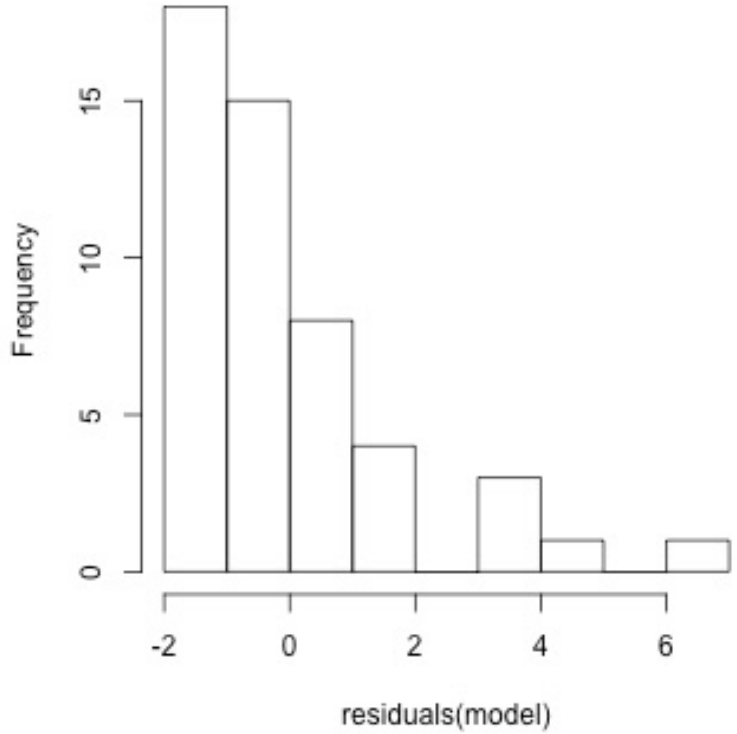


QQ Plot with an outlier

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

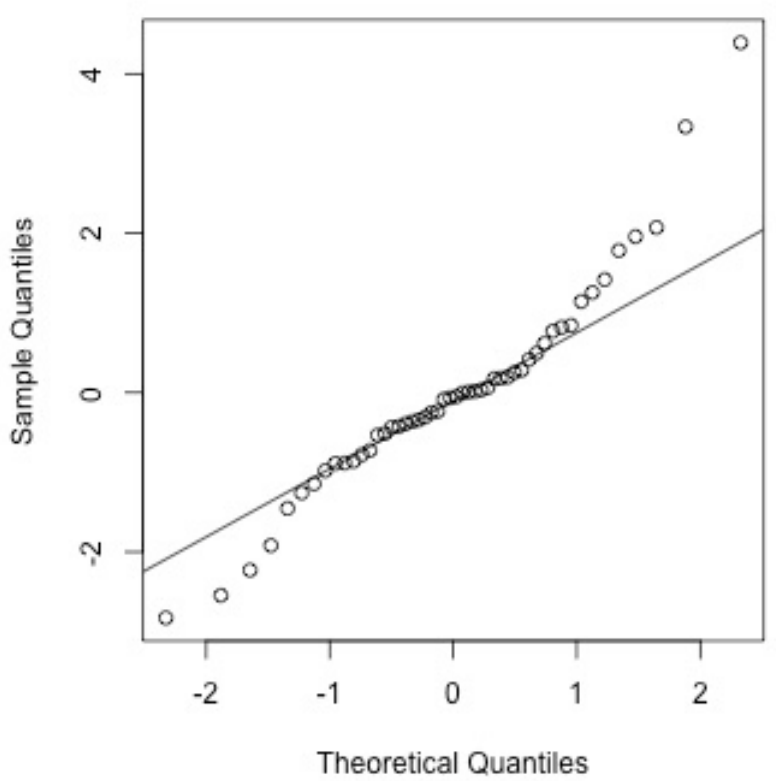
Diagnosis Tool 4 – Normal Probability Plot (QQ Plot) of Residuals

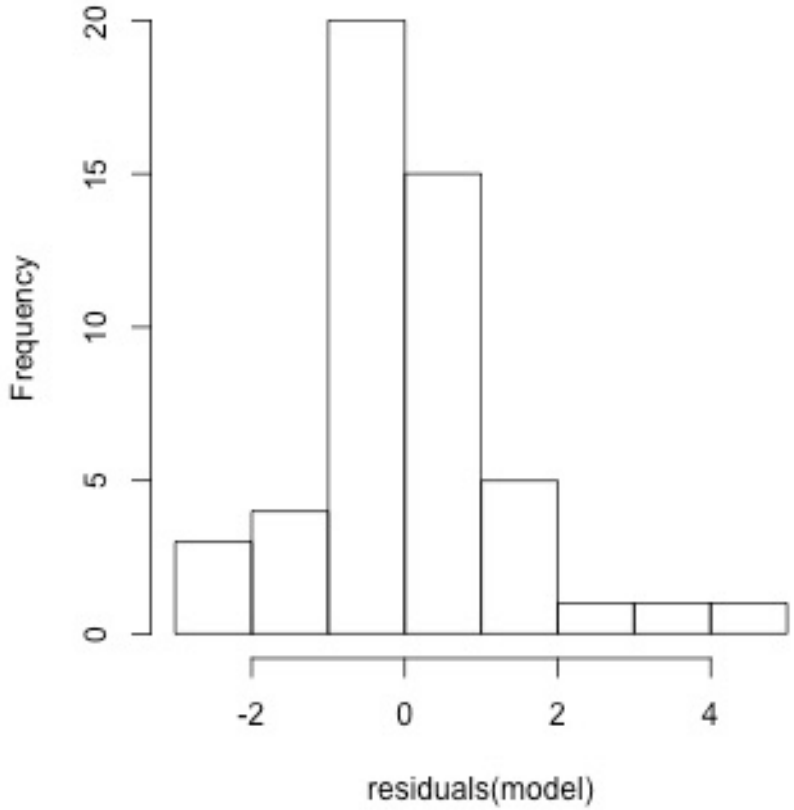




QQ Plot with skewed residuals

Diagnosis Tool 4 – Normal Probability Plot (QQ Plot) of Residuals

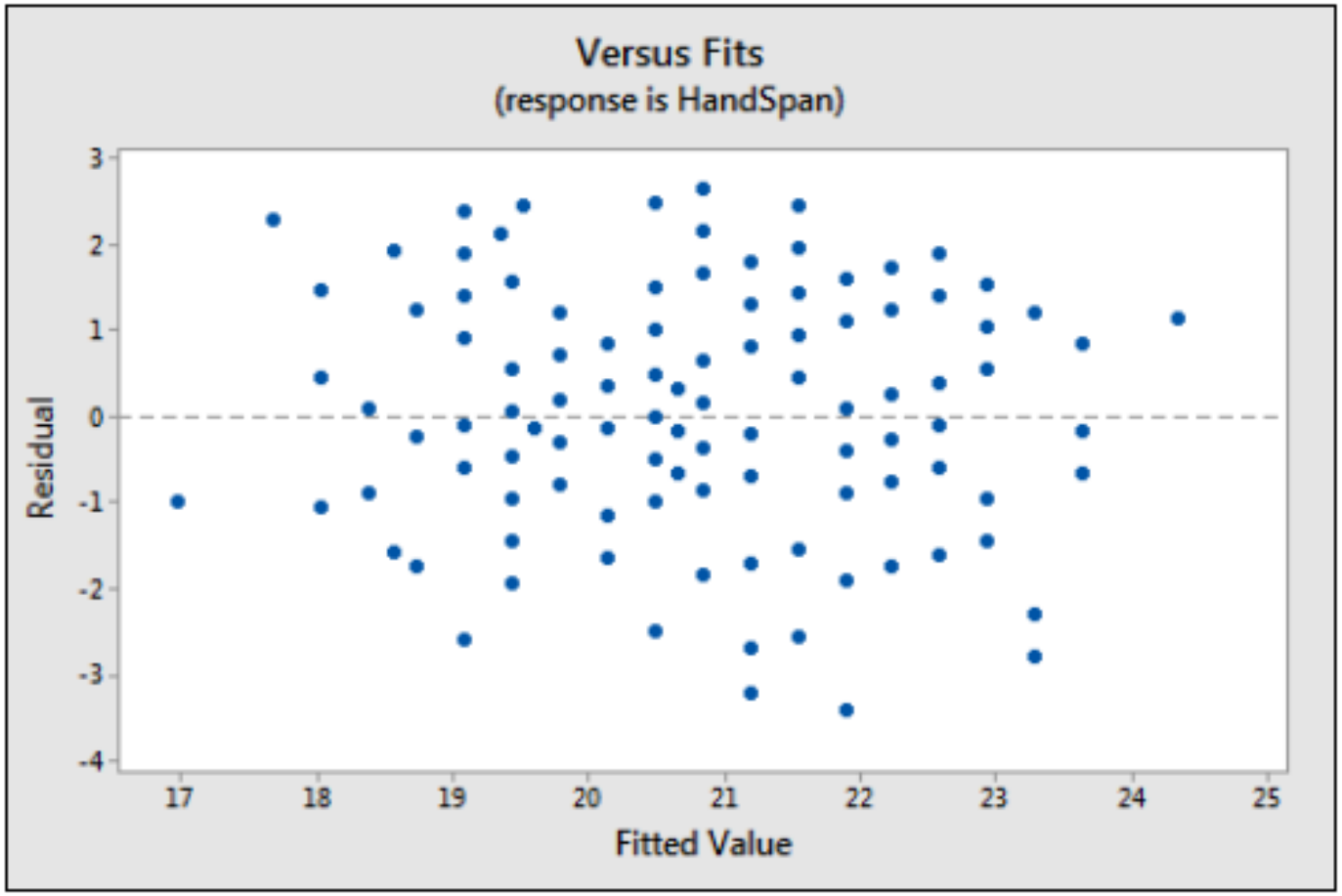




QQ Plot with heavy/long tails

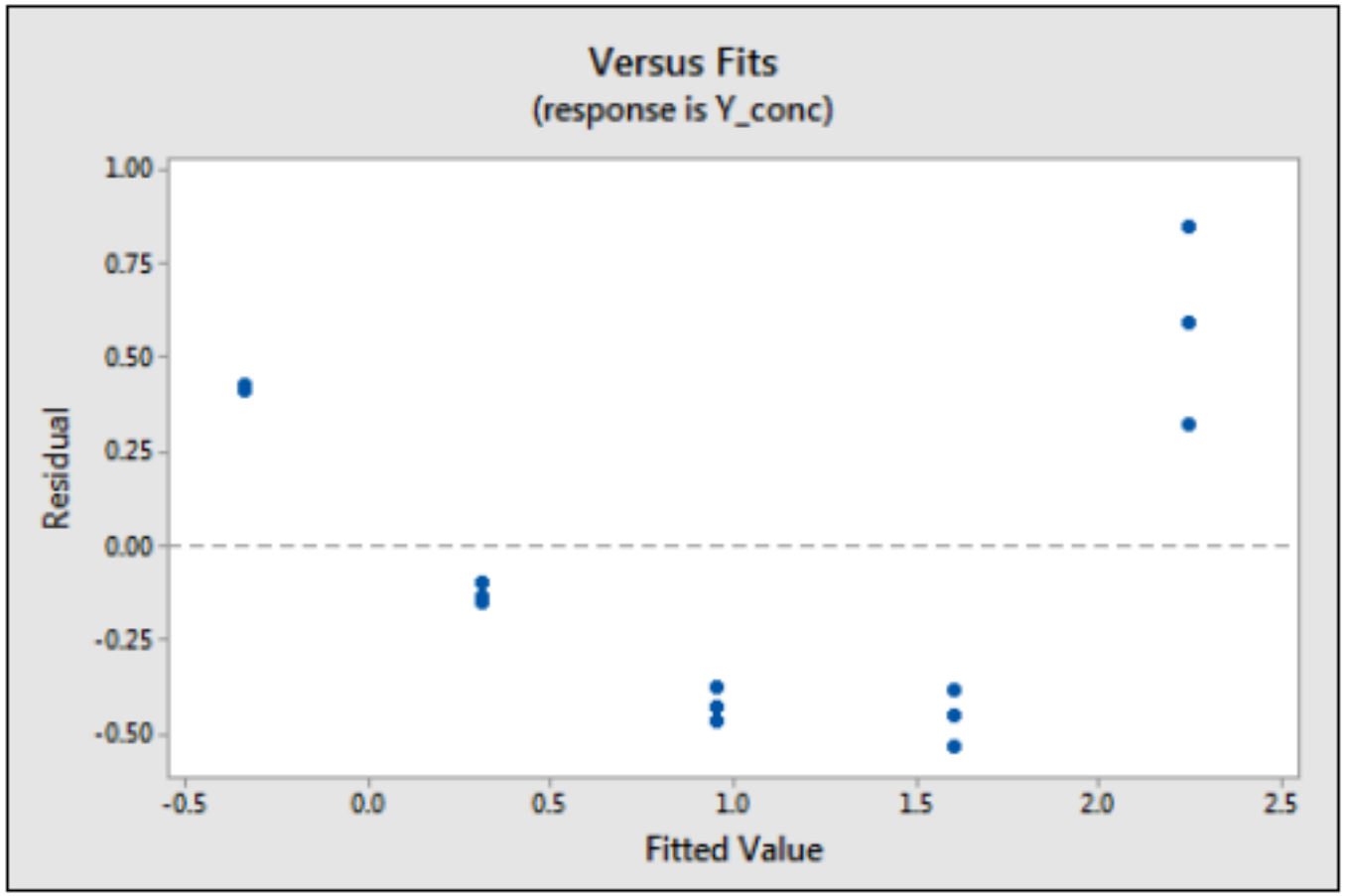
Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Problems in residual plots – More Examples



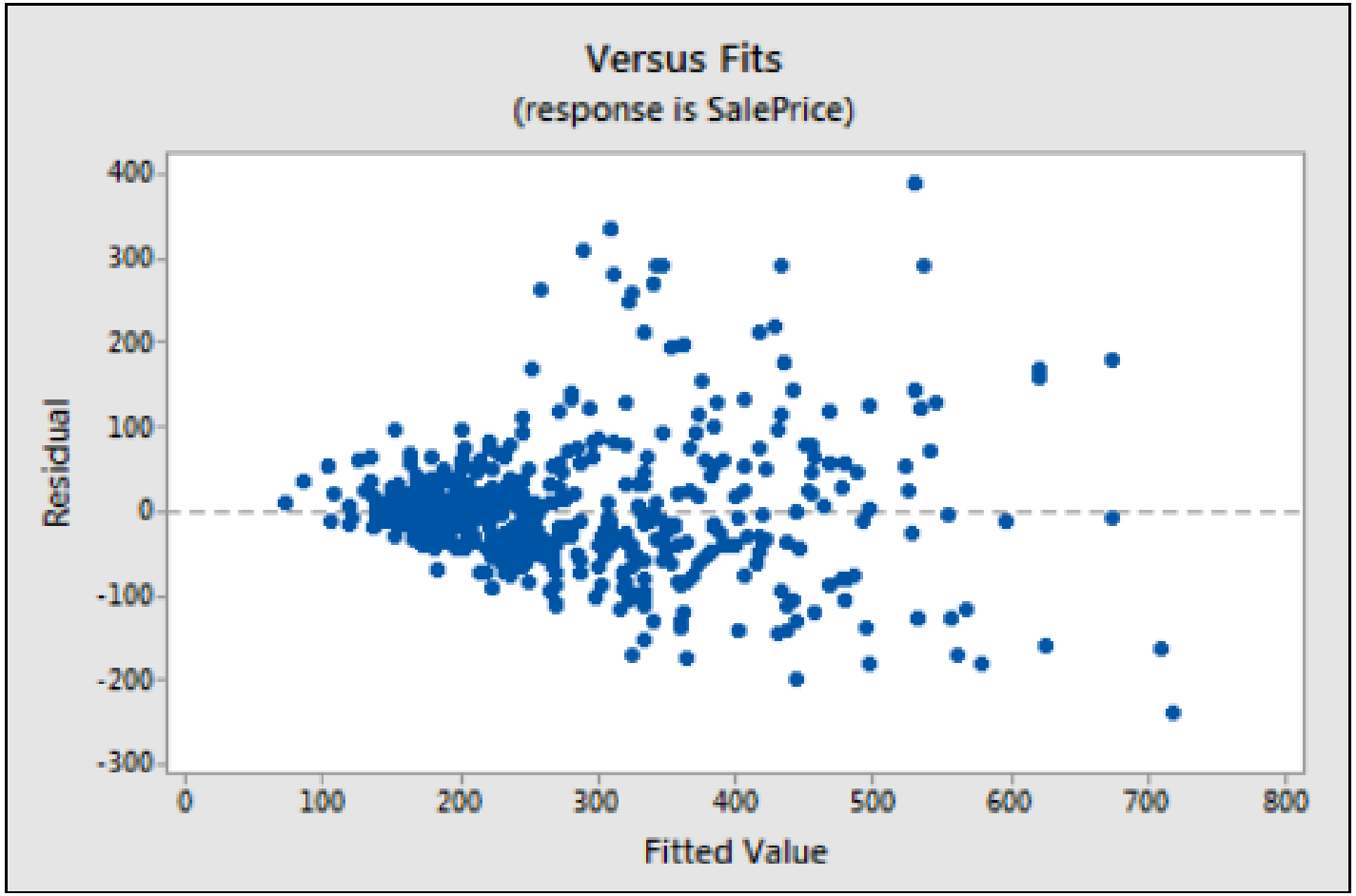
Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Problems in residual plots – More Examples



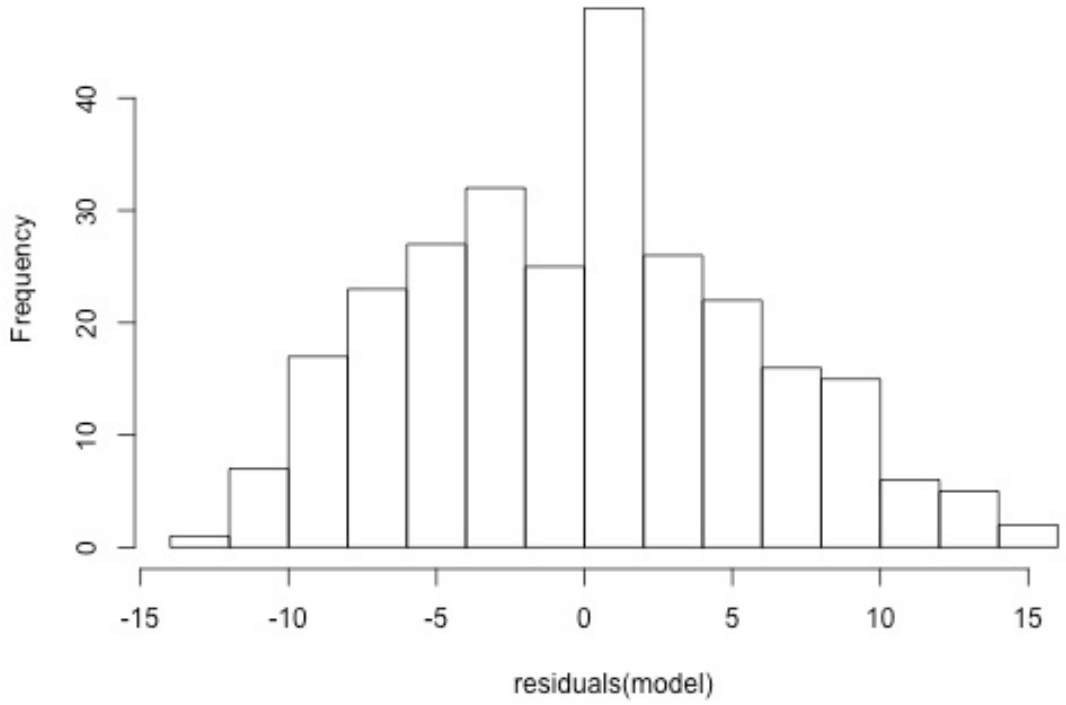
Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City,
Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

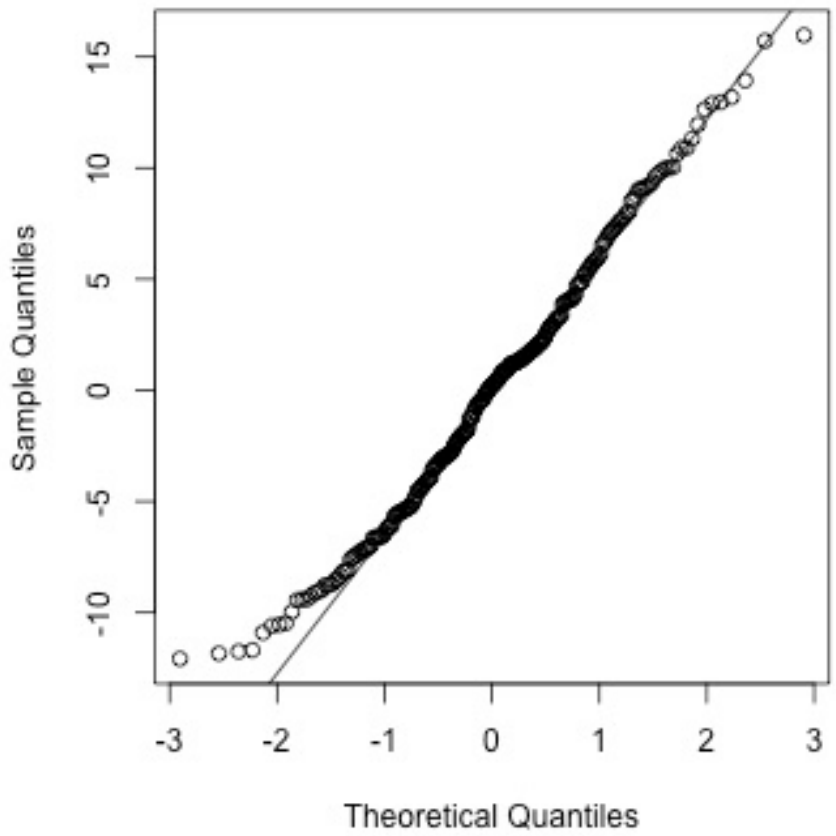
Problems in residual plots – More Examples



Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City,
Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

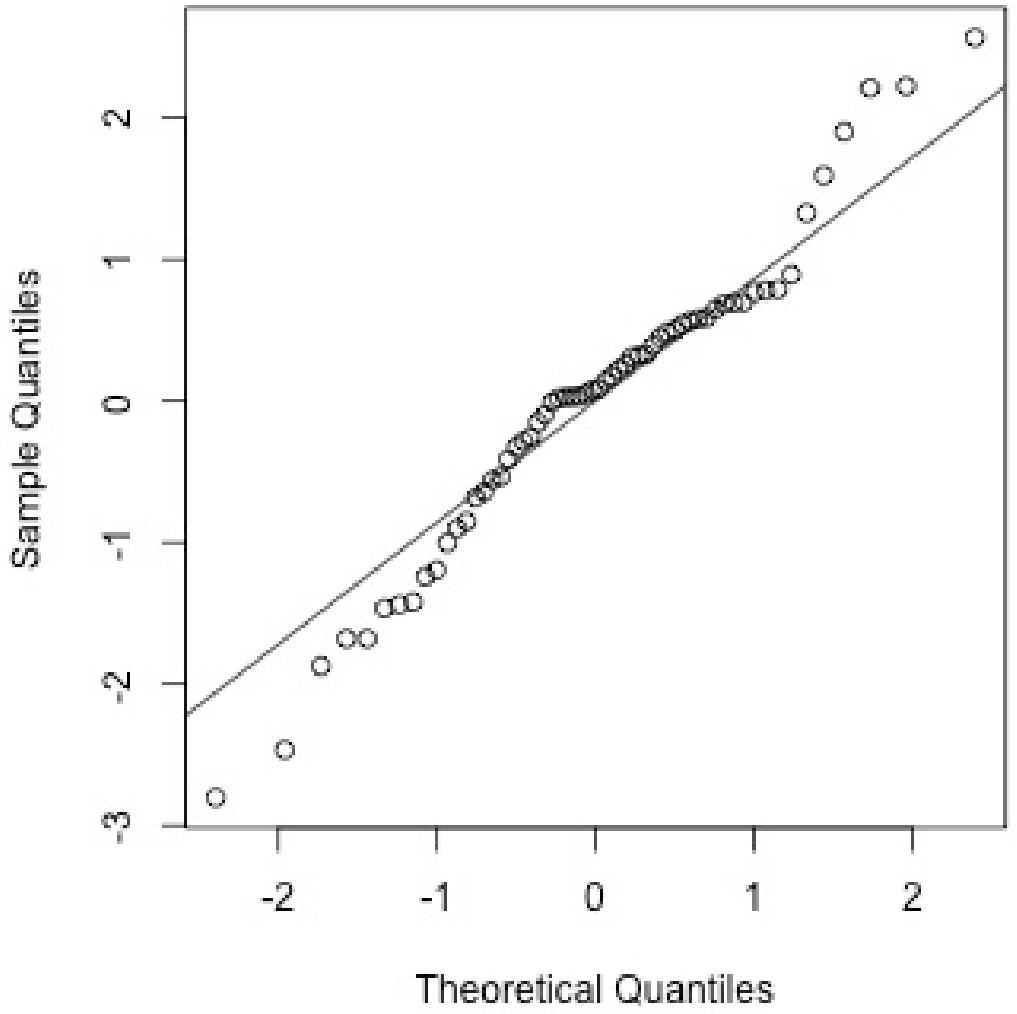
Problems in residual plots – More Examples





Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Problems in residual plots – More Examples



Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

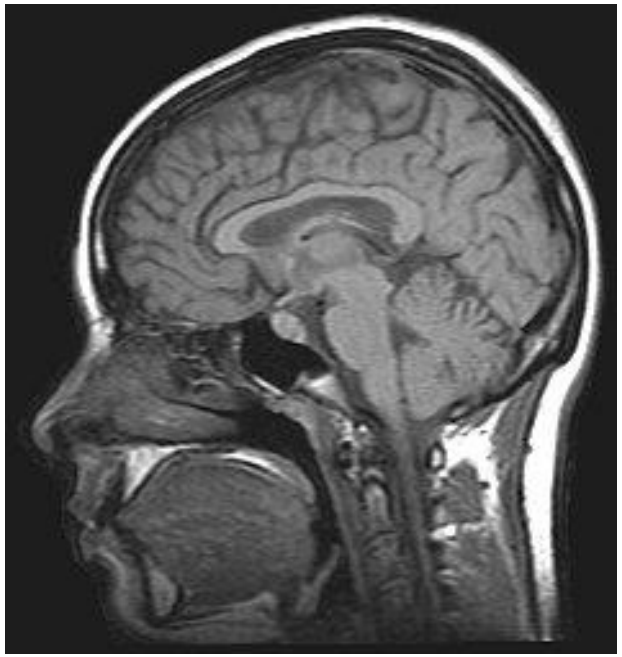
Multiple Linear Regression

Multiple Linear Regression

- Good News! All our learnings for SLR are still valid, with slight modifications
  - The models have similar "LINE" assumptions (this time for all predictors)
  - We will use adjusted R^2 instead of R^2
  - We can use Confidence Intervals to make estimations of the population

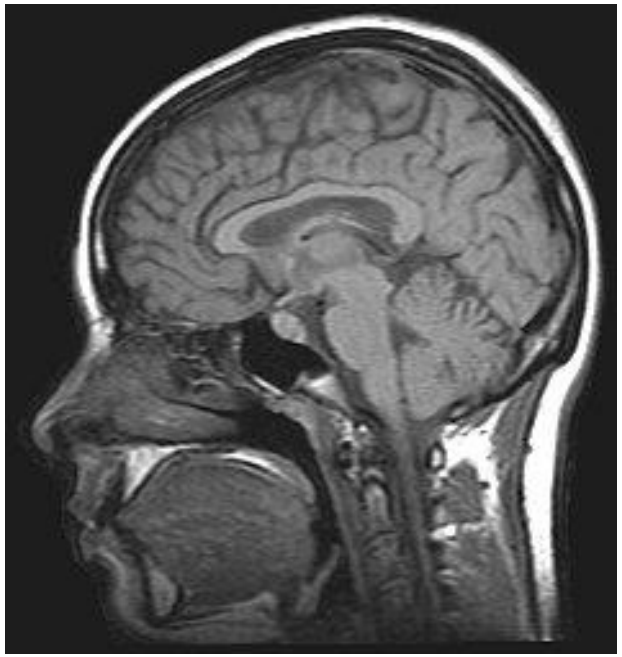
Lets Understand MLR through an example…

Are a person's brain size and body size predictive of his or her intelligence?

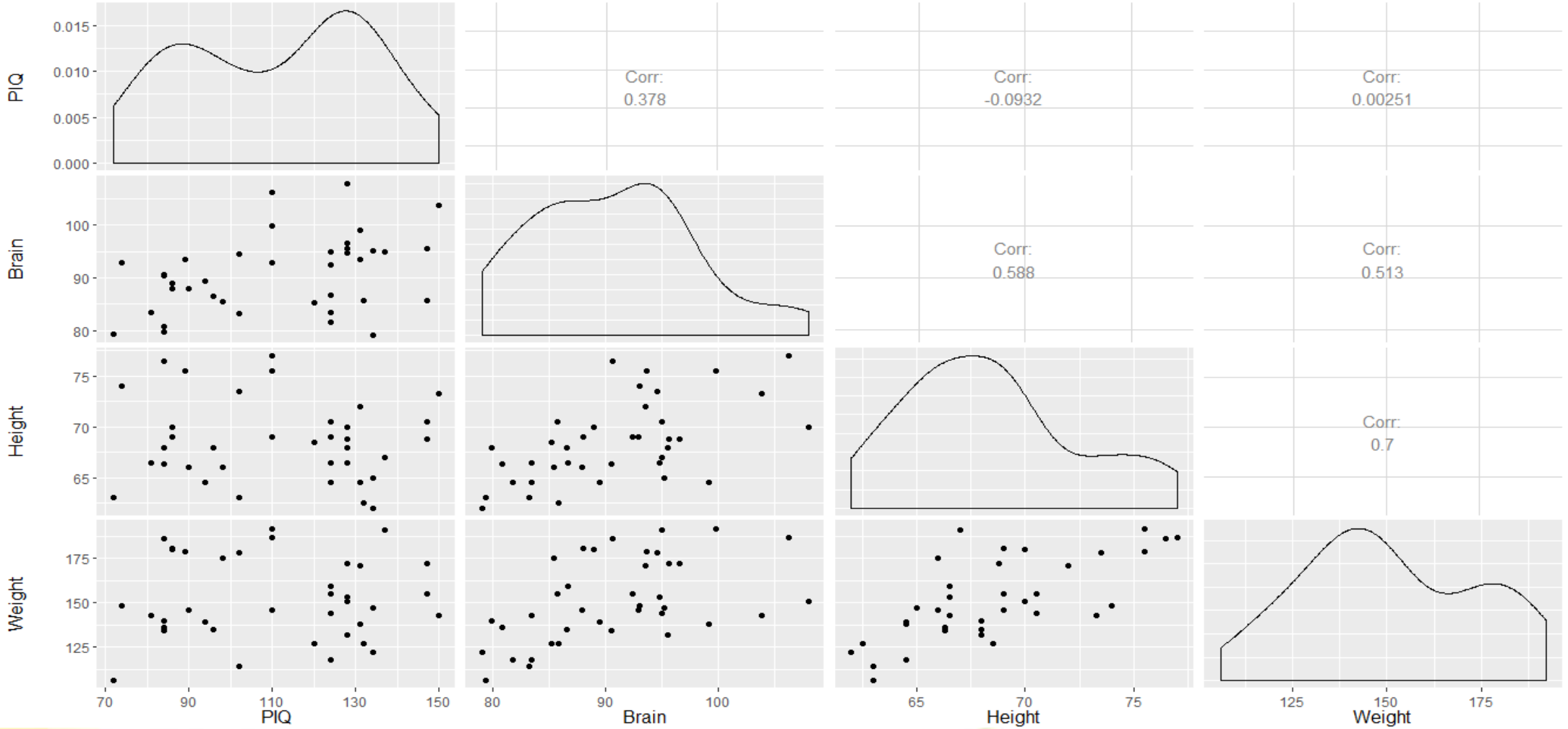


Lets Understand MLR through an example…

- Interested in answering this question, some researchers collected the data (iqsize.txt) on a sample of n = 38 college students:
- Response (y): Performance IQ scores (**PIQ**) from the revised Wechsler Adult Intelligence Scale. This variable served as the investigator's measure of the individual's intelligence.
- Potential predictor (x_1): Brain size based on the count obtained from **MRI** scans (given as count/10,000).
- Potential predictor (x_2): **Height** in inches.
- Potential predictor (x_3): **Weight** in pounds.



Lets Understand MLR through an example…



Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Lets Understand MLR through an example…

- We can formulate our model as follows:

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i$$

where:

- y_i is the intelligence (**PIQ**) of student *i*
- x_{i1} is the brain size (**MRI**) of student *i*
- x_{i2} is the height (**Height**) of student *i*
- x_{i3} is the weight (**Weight**) of student *I*

and the **independent** error terms ε_i follow a **normal** distribution with mean 0 and **equal variance** σ^2

Lets Understand MLR through an example…

```
Call:
lm(formula = PIQ ~ ., data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-32.74 -12.09  -3.84  14.17  51.69

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.114e+02  6.297e+01   1.768 0.085979 .
Brain        2.060e+00  5.634e-01   3.657 0.000856 ***
Height      -2.732e+00  1.229e+00  -2.222 0.033034 *
Weight       5.599e-04  1.971e-01   0.003 0.997750
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.79 on 34 degrees of freedom
Multiple R-squared:  0.2949,    Adjusted R-squared:  0.2327
F-statistic: 4.741 on 3 and 34 DF,  p-value: 0.007215

> anova(regressor)
Analysis of Variance Table

Response: PIQ
          Df  Sum Sq Mean Sq F value  Pr(>F)
Brain      1  2697.1 2697.09  6.8835 0.01293 *
Height     1  2875.6 2875.65  7.3392 0.01049 *
Weight     1     0.0    0.00  0.0000 0.99775
Residuals 34 13321.8  391.82
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City,
Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Lets Understand More About MLR…

- Population Model that relates a y variable to p-1 x variables is represented as:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i.$$

- We assume that the ϵi have a normal distribution with mean 0 and constant variance σ2
- Each *x*-variable can be a predictor variable or a transformation of predictor variables (such as the square of a predictor variable or two predictor variables multiplied together)

Lets Understand More About MLR…

- The estimates for the β coefficients are the values that minimize the sum of squared errors for the sample
- The letter b is used to represent the sample estimate of the β coefficient. Thus, b0 is the sample estimate of β0, b1 is the sample estimate of β1, and so on…
- MSE=SSE/(n−p) estimates σ2, the variance of the errors. In the formula, n = sample size, p = number of β coefficients in the model (including the intercept) and SSE = sum of squared errors
- S=√MSE estimates σ and is known as the regression standard error or the residual standard error

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City,
Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Lets Understand More About MLR…

- Each β coefficient represents the change in the mean response, E(y), per unit increase in the associated predictor variable when all the other predictors are held constant.
- For example, β1 represents the change in the mean response, E(y), per unit increase in x1 when x2, x3, ..., xp−1 are held constant.
- The intercept term, β0, represents the mean response, E(y), when all the predictors x1, x2, ..., xp−1, are all zero (which may or may not have any practical meaning).

Lets Understand More About MLR…

- A predicted value is calculated as y^i=b0+b1*xi,1+b2*xi,2+…+bp−1*xi,p−1, where the b values come from R and the x-values are specified by us
- A residual (error) term is calculated as ei=yi−y^i, the difference between an actual and a predicted value of y
- A plot of residuals versus predicted values ideally should resemble a horizontal random band. Departures from this form indicates difficulties with the model and/or data
- R^2 and adjusted R^2
- Hypothesis testing for MLR

MLR Assumptions…

- The four conditions ("LINE") applies to MLR as well:
  - The mean of the response , E(Yi), at each set of values of the predictors, (x1i,x2i,…), is a **L**inear function of the predictors.
  - The errors, εi, are **I**ndependent.
  - The errors, εi, at each set of values of the predictors, (x1i,x2i,…), are **N**ormally distributed.
  - The errors, εi, at each set of values of the predictors, (x1i,x2i,…), have **E**qual variances (denoted σ2).
- Alternatively, the errors, ϵi, are independent normal random variables with mean zero and constant variance, σ2.

How To Test These Assumptions?

- Diagnostic 1 (Residual vs Fits plot) - Create a scatterplot with the residuals, ei, on the vertical axis and the fitted values, y^i, on the horizontal axis and visual assess whether:
  - the (vertical) average of the residuals remains close to 0 as we scan the plot from left to right (this affirms the "L" condition);
  - the (vertical) spread of the residuals remains approximately constant as we scan the plot from left to right (this affirms the "E" condition);
  - there are no excessively outlying points
  - violation of any of these three may necessitate remedial action (such as transforming one or more predictors and/or the response variable), depending on the severity of the violation (more on this later...)

How To Test These Assumptions?

- Diagnostic 2 (Residual vs Order Plot) - If the data observations were collected over time (or space) create a scatterplot with the residuals, ei, on the vertical axis and the time (or space) sequence on the horizontal axis and visual assess whether there is no systematic non-random pattern (this affirms the "I" condition).
- Violation may suggest the need for a time series model

How To Test These Assumptions?

- Diagnostic 3a (Residuals vs Predictors plot) - Create a series of scatterplots with the residuals, ei, on the vertical axis and each of the predictors in the model on the horizontal axes and visual assess whether:
  - the (vertical) average of the residuals remains close to 0 as we scan the plot from left to right (this affirms the "L" condition);
  - the (vertical) spread of the residuals remains approximately constant as we scan the plot from left to right (this affirms the "E" condition);
  - violation of either of these for at least one residual plot may suggest the need for transformations of one or more predictors and/or the response variable

How To Test These Assumptions?

- Diagnostic 3b (Residuals vs Predictors plot) - Create a series of scatterplots with the residuals, ei, on the vertical axis and each of the available predictors that have been omitted from the model on the horizontal axes and visual assess whether:
  - there are no strong linear or simple nonlinear trends in the plot;
  - violation may indicate the predictor in question (or a transformation of the predictor) might be usefully added to the model

How To Test These Assumptions?

- Diagnostic 4 (QQ Plot) - Create a histogram, boxplot, and/or normal probability plot of the residuals, ei to check for approximate normality (the "N" condition)

Example - How To Test These Assumptions?

- Lets test these assumptions on the IQ data

Example - How To Test These Assumptions?

- Lets test these assumptions on the IQ data

Test for Error Normality

- Few more tests to aid graphical analysis:
  - Anderson-Darling Test
  - Shapiro-Wilk Test
  - Kolmogorov-Smirnov (Lillie) Test
- In each of these tests, we hope to *fail to reject* the null hypothesis as this would mean that the errors follow a normal distribution. Hence a high p-value is desired
  - install.packages('nortest')
  - library('nortest')
  - ad.test(stdres(regressor))
  - shapiro.test(stdres(regressor))
  - lillie.test(stdres(regressor))

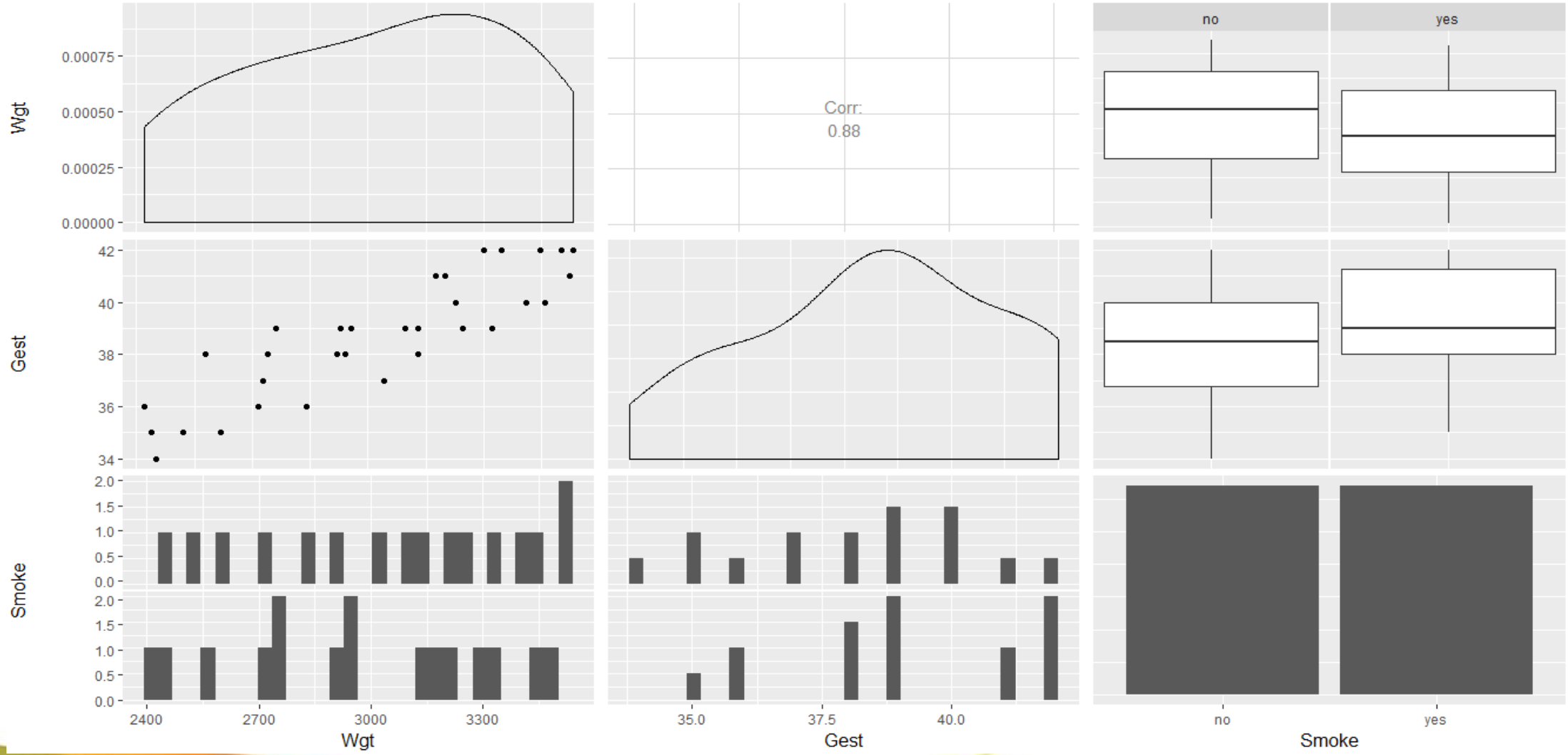
How To Deal With Categorical Variables?

- 50 Startups

How To Deal With Categorical Variables?

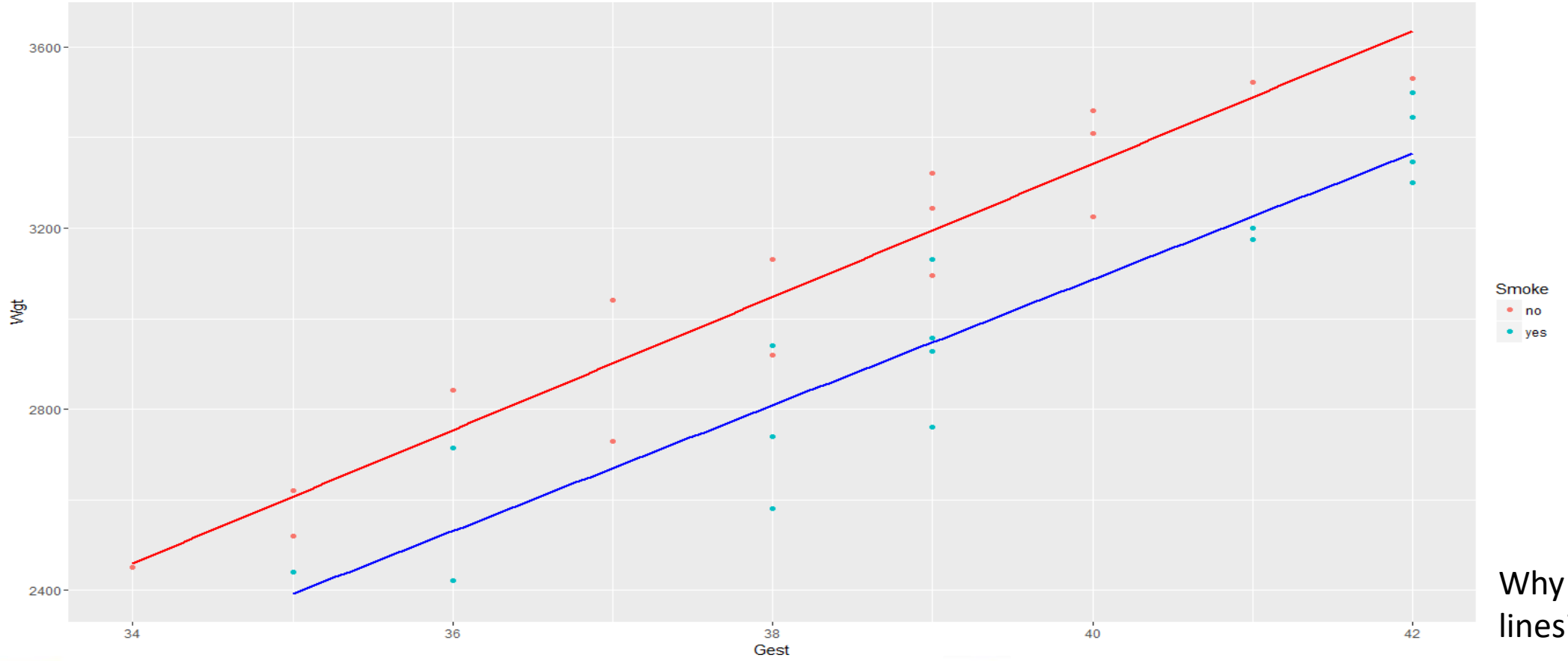
- Lets consider this with another example…
- Is a baby's birth weight related to the mother's smoking during pregnancy?
- Researchers interested in answering the above research question collected the data (birthsmokers.txt) on a random sample of $n = 32$ births:
  - Response (y): birth weight (**Weight**) in grams of baby
  - Potential predictor (x_1): **Smoking** status of mother (yes or no)
  - Potential predictor (x_2): length of gestation (**Gest**) in weeks

How To Deal With Categorical Variables?



Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

How To Deal With Categorical Variables?



Why parallel lines?

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

How To Deal With Categorical Variables?

- So, is baby's birth weight related to smoking during pregnancy, after taking into account length of gestation?

How To Deal With Categorical Variables?

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) + \epsilon_i$$

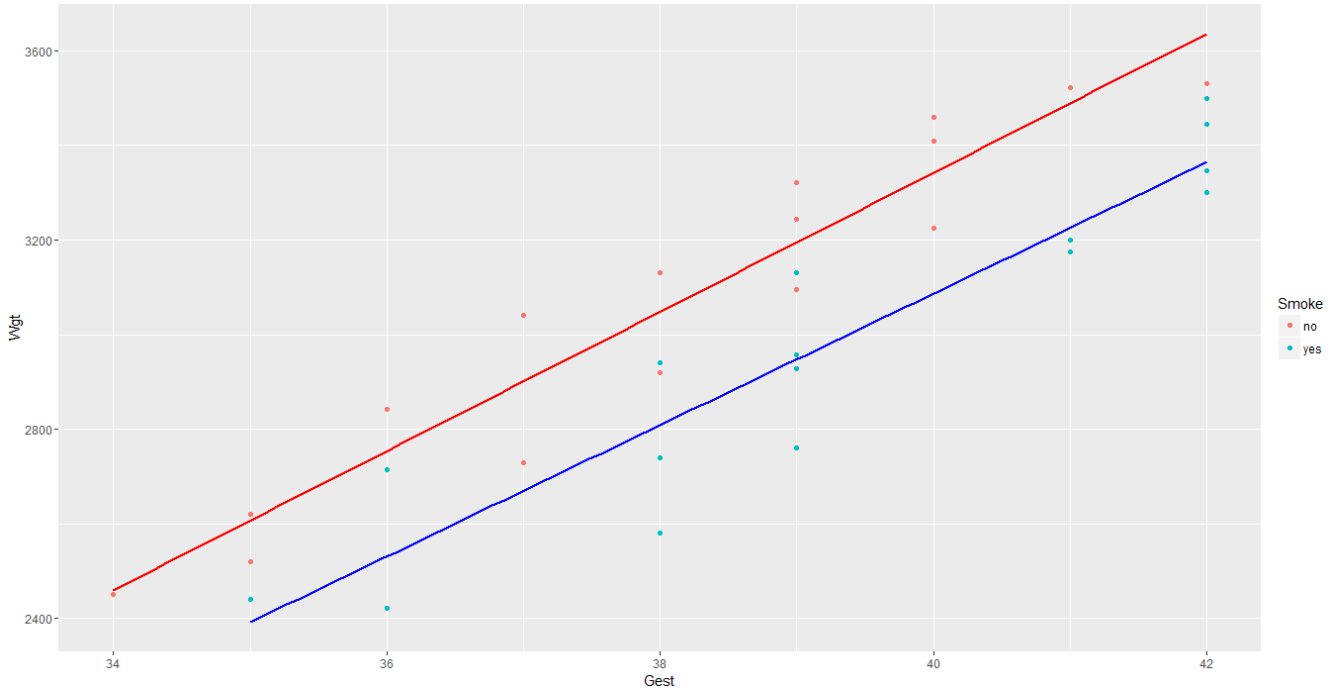
where:

yi is birth weight of baby i in grams
xi1 is the length of gestation of baby i in weeks
xi2 = 1, if the mother smoked during pregnancy, and xi2 = 0, if she did not

How do we interpret the slopes in this case?

How To Deal With Categorical Variables?

- Lets revisit the plot of baby weight vs gestation period for smoking and non-smoking mothers
- Does the effect of the gestation length on mean birth weight depend on whether or not the mother is a smoker?
- Does the effect of smoking on mean birth weight depend on the length of gestation?



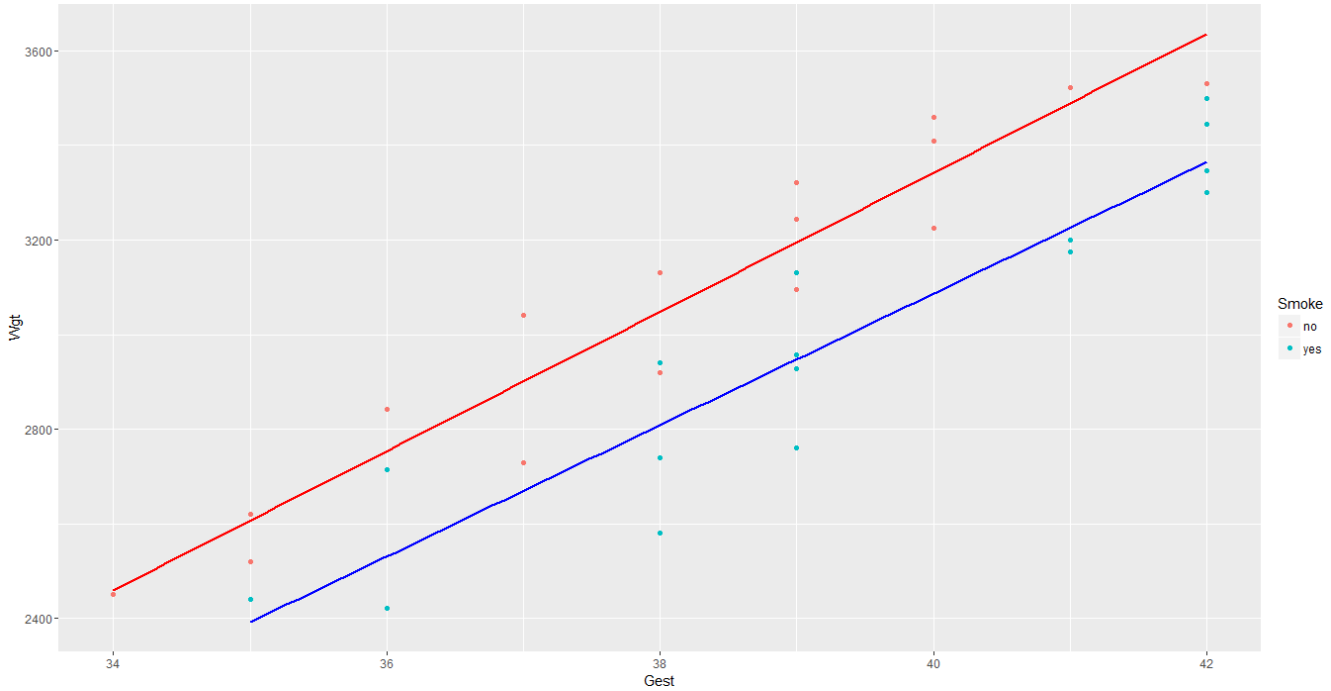
How To Deal With Categorical Variables?

- When two predictors do not interact, we say that each predictor has an "additive effect" on the response.
- More formally, a regression model contains additive effects if the response function can be written as a sum of functions of the predictor variables:

$$\mu y = f1(x1) + f2(x2) + ... + fp-1(xp-1)$$

- For example, our regression model for the birth weights of babies contains additive effects, because the response function can be written as a sum of functions of the predictor variables:

$$\mu y = (\beta 0) + (\beta 1 xi1) + (\beta 2 xi2)$$



Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

How To Deal With Categorical Variables?

- Lets consider another example
- Some researchers were interested in comparing the effectiveness of three treatments for severe depression. For the sake of simplicity, we denote the three treatments A, B, and C. The researchers collected the data (depression.txt) on a random sample of $n = 36$ severely depressed individuals:

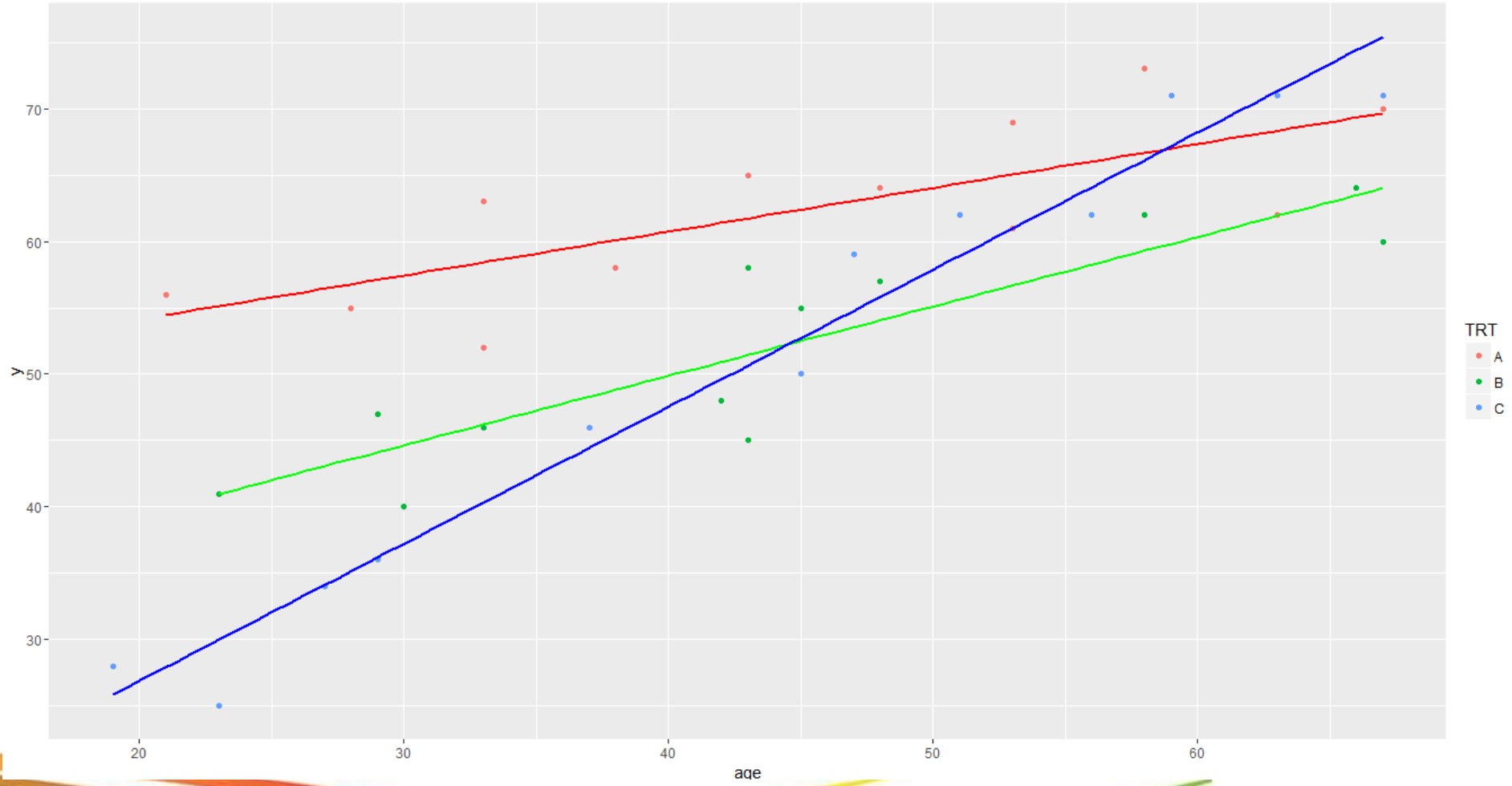
  y_i = measure of the effectiveness of the treatment for individual i

  x_{i1} = age (in years) of individual i

  x_{i2} = 1 if individual i received treatment A and 0, if not

  x_{i3} = 1 if individual i received treatment B and 0, if not

How To Deal With Categorical Variables?



Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City,
Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

How To Deal With Categorical Variables?

- In this case, we need to include what are called "**interaction terms**" in our formulated regression model
- A (second-order) multiple regression model with interaction terms is:

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_{12} x_{i1}x_{i2}+\beta_{13} x_{i1}x_{i3}+\epsilon_i$$

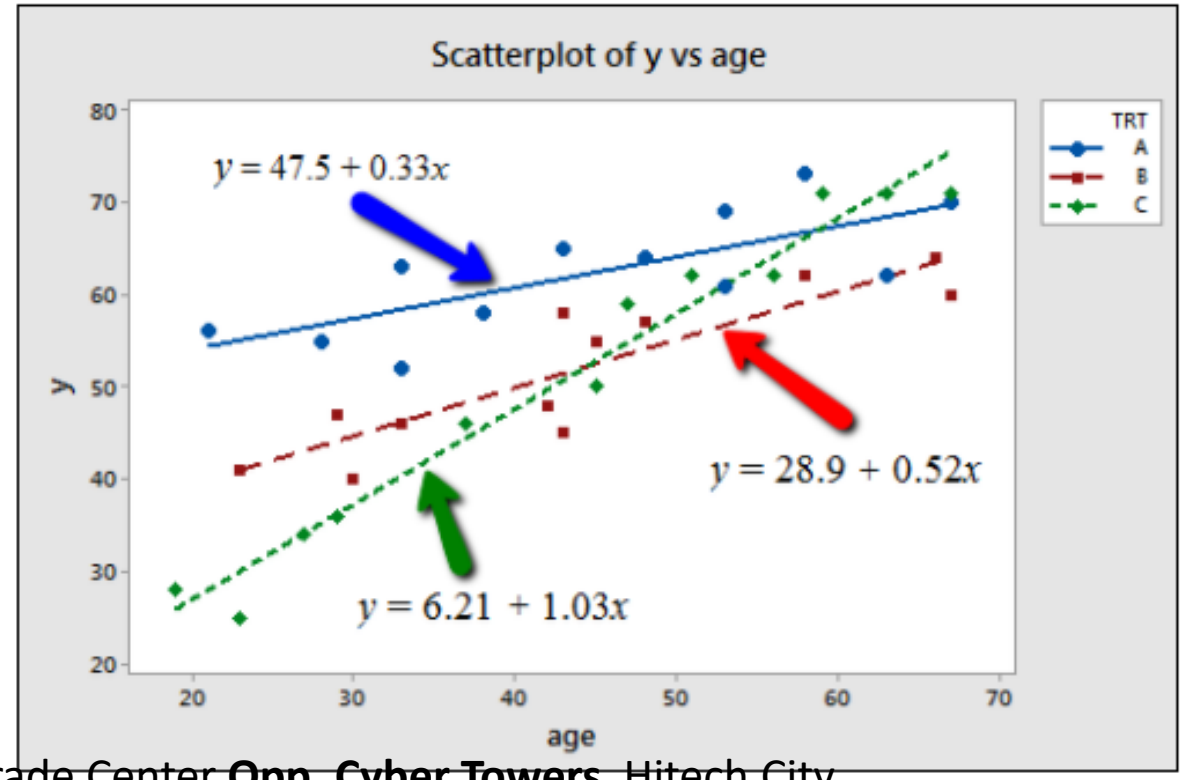
How To Deal With Categorical Variables?

- In general, then, what does it mean for two predictors "**to interact**"?
  - Two predictors interact if the effect on the response variable of one predictor **depends on the value of the other**.
  - A slope parameter can no longer be interpreted as the change in the mean response for each unit increase in the predictor, while the other predictors are held constant.

How To Deal With Categorical Variables?

How to deal with interaction effects?

- Create separate Linear Regression outputs for each of the categories



Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Data Transformations

Data Transformations

- Why do we need to transform the data?
- Options available:
  - We transform the predictor (x) values only
  - We transform the response (y) values only
  - We transform both the predictor (x) values and response (y) values
- Remember, data transformation requires trial and error!!

Log Transformation of only the predictor

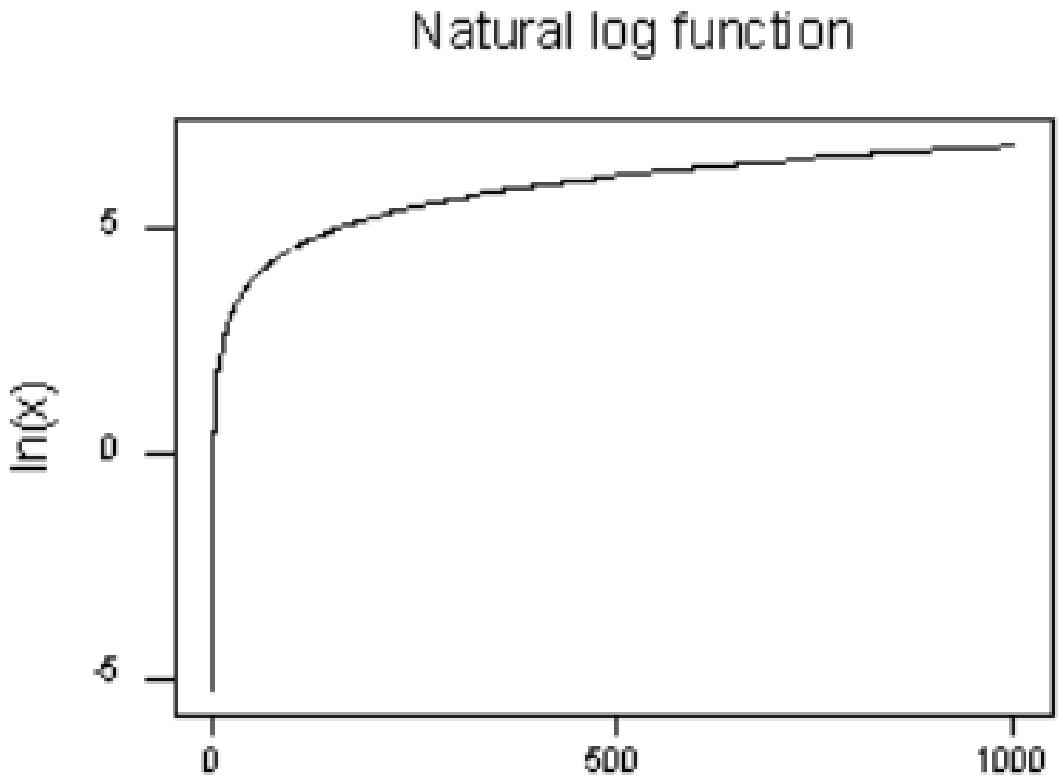
- Transforming the *x* values is appropriate **when non-linearity is the only problem** — the independence, normality and equal variance conditions are met
- It may be necessary to correct the non-linearity before you can assess the normality and equal variance assumptions
- While some assumptions may appear to hold prior to applying a transformation, they may no longer hold once a transformation is applied
- For SLR we will use scatter plot of the independent and dependent variables; for MLR we will use the scatter plot of the residuals vs each of the predictors

Log Transformation of only the predictor

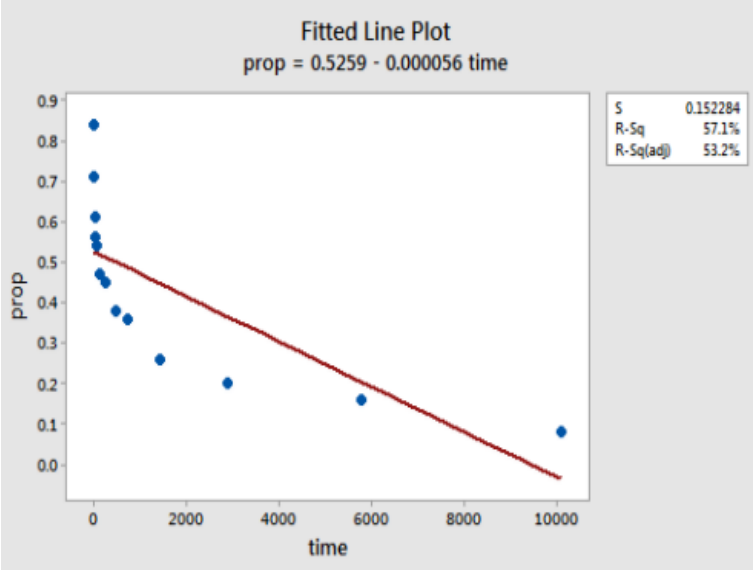
- Lets consider an example…
- Let's consider the data from a memory retention experiment in which 13 subjects were asked to memorize a list of disconnected items.
- The subjects were then asked to recall the items at various times up to a week later. The proportion of items (*y = prop*) correctly recalled at various times (*x = time*, in minutes) since the list was memorized were recorded (wordrecall.txt).

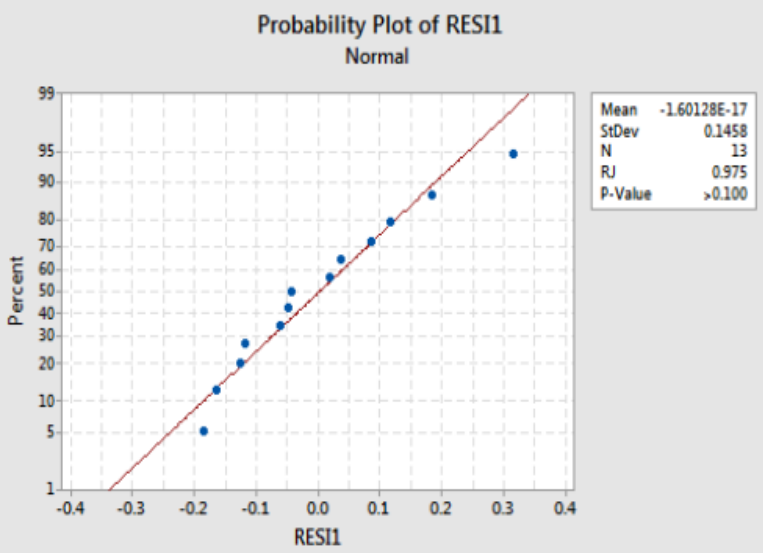
Log Transformation of only the predictor

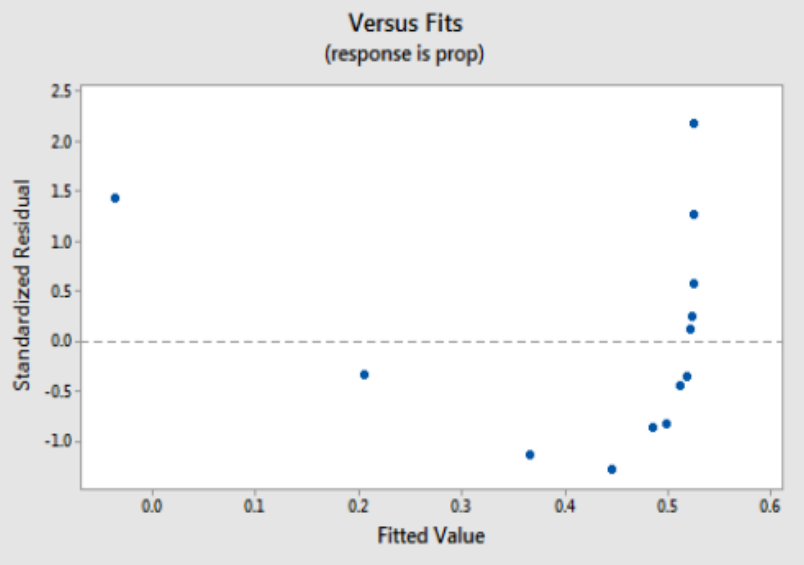
- What is log transformation?
- How does it work?



Log Transformation of only the predictor







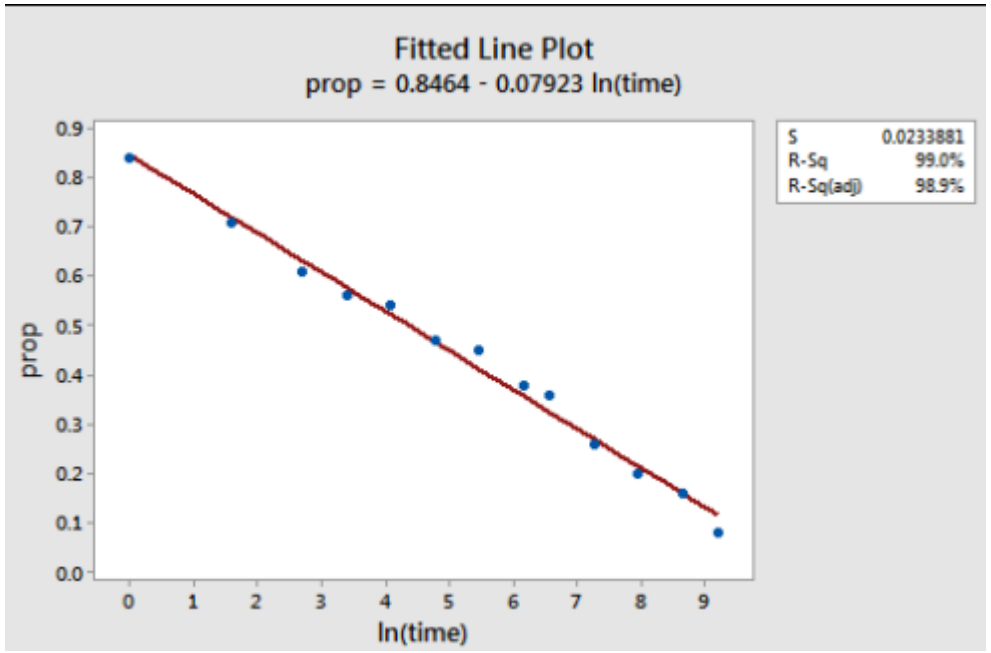
Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

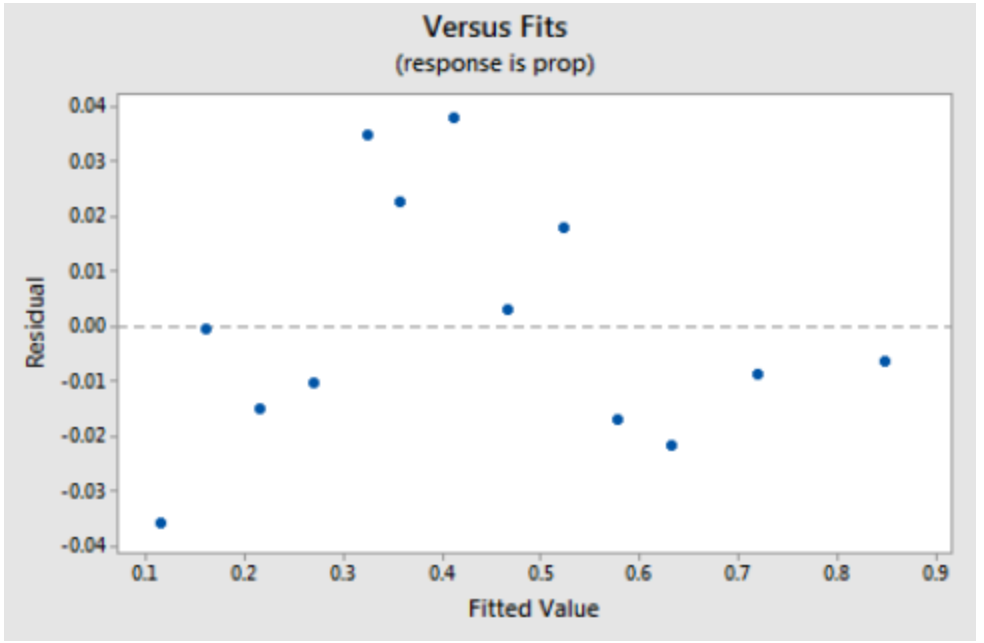
Log Transformation of only the predictor

- Lets see what happens if we take a natural logarithm of the predictor ‘time’…

Log Transformation of only the predictor

- Lets see what happens if we take a natural logarithm of the predictor ‘time’…





Log Transformation of only the predictor

- What is the nature of the association between time since memorized and the effectiveness of recall?
- Is there an association between time since memorized and effectiveness of recall?
- What proportion of words can we expect a randomly selected person to recall after 1000 minutes?
- How much does the expected recall change if time increases ten-fold?

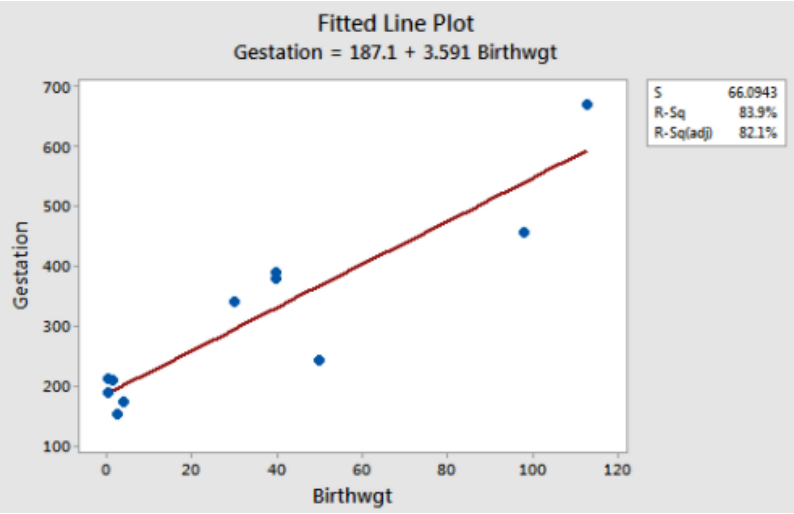
Log Transformation of only the response

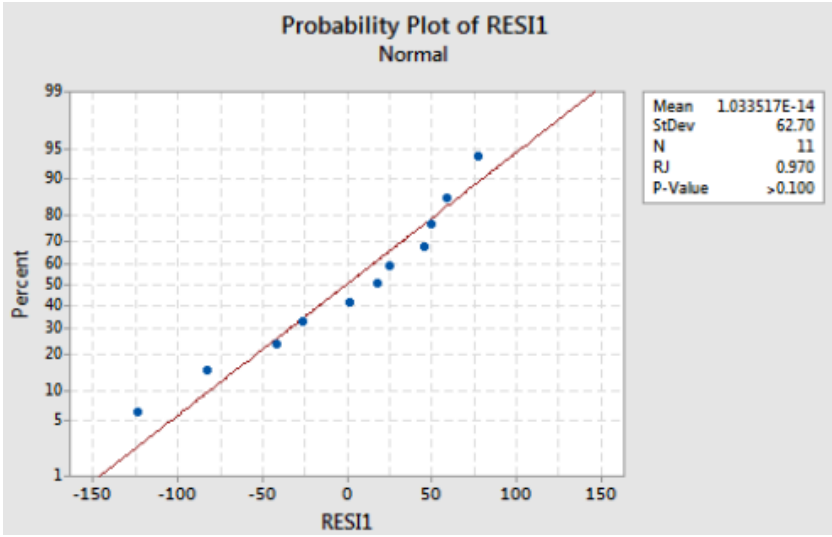
- Transforming the *y* values should be considered when non-normality and/or unequal variances are the problems with the model
- As an added bonus, the transformation on *y* may also help to "straighten out" a curved relationship

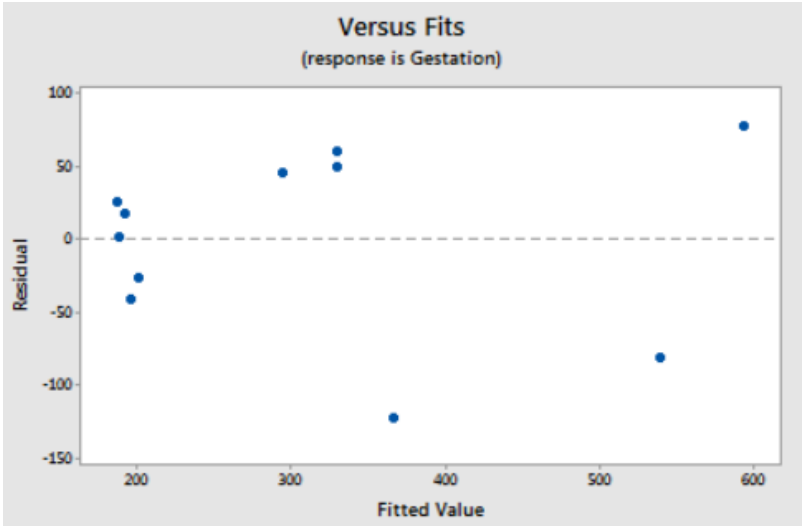
Log Transformation of only the response

- Lets consider another example…
- Let's consider data (mammgest.txt) on the typical birthweight and length of gestation for various mammals. We treat the birthweight (*x*, in kg) as the predictor and the length of gestation (*y*, in number of days until birth) as the response

Log Transformation of only the response





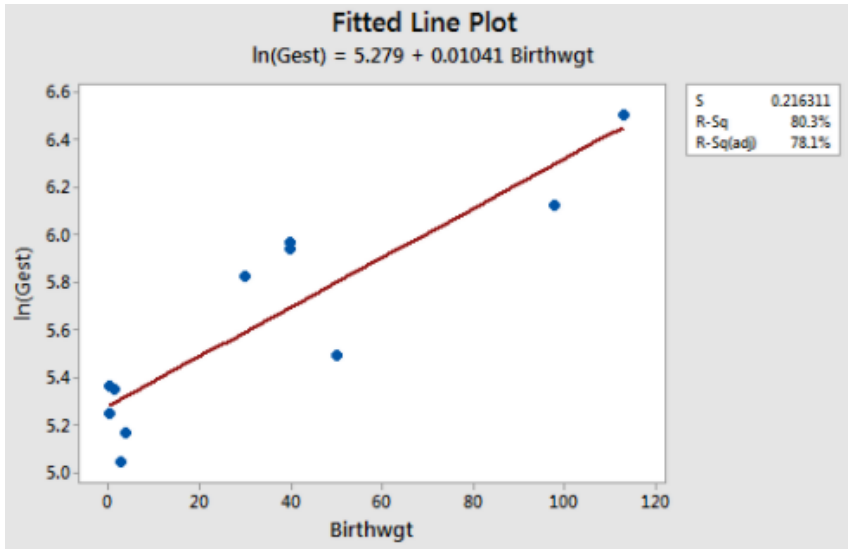


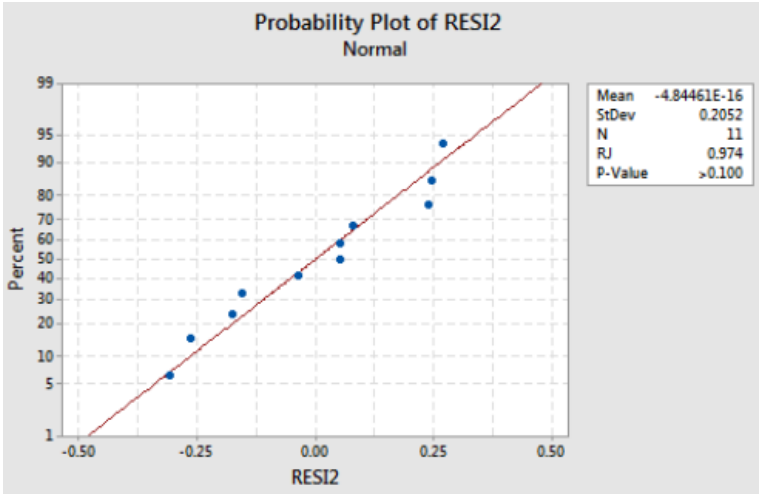
Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

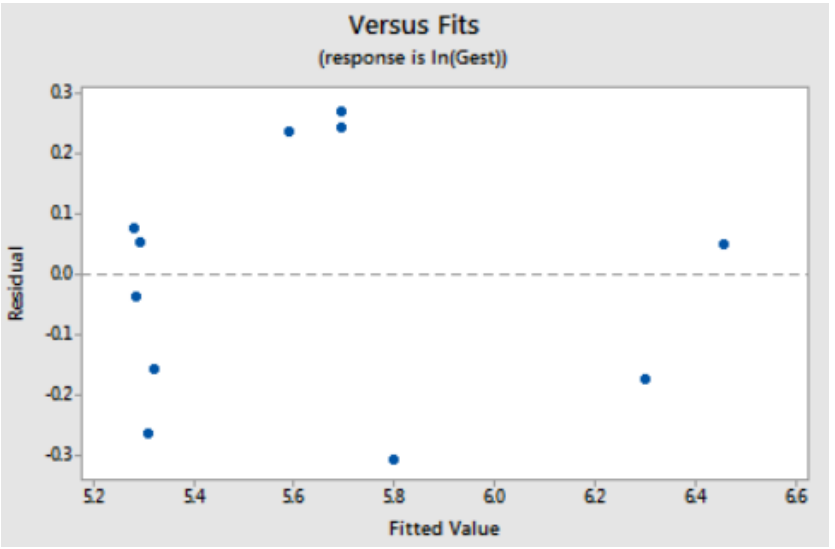
Log Transformation of only the response

- Lets see what happens if we transform the response y in this case…

Log Transformation of only the response







Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Log Transformation of only the response

- What is the nature of the association between mammalian birth weight and length of gestation?
- Is there an association between mammalian birth weight and length of gestation?
- What is the expected gestation length of a new 50 kg mammal?
- What is the expected change in length of gestation for each one pound increase in birth weight? What is the expected change in length of gestation for each one pound increase in birth weight?

Log Transformation of both predictor and response

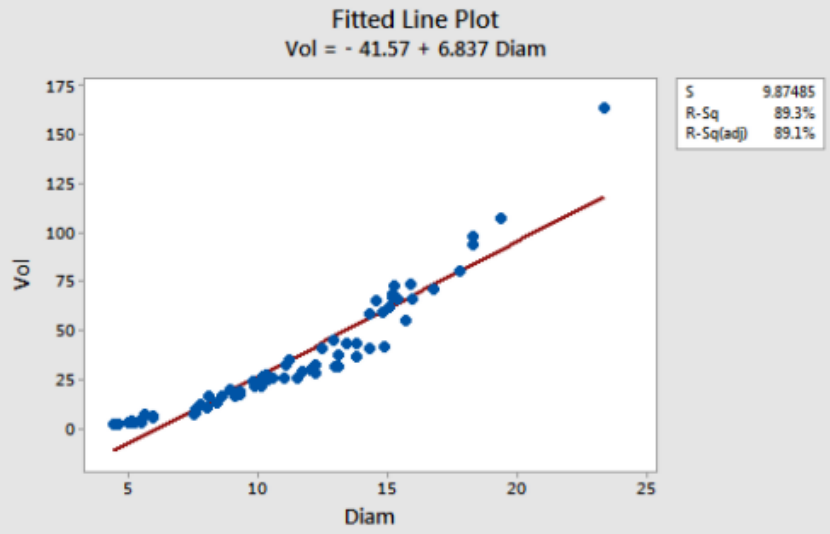
Log Transformation of both predictor and response

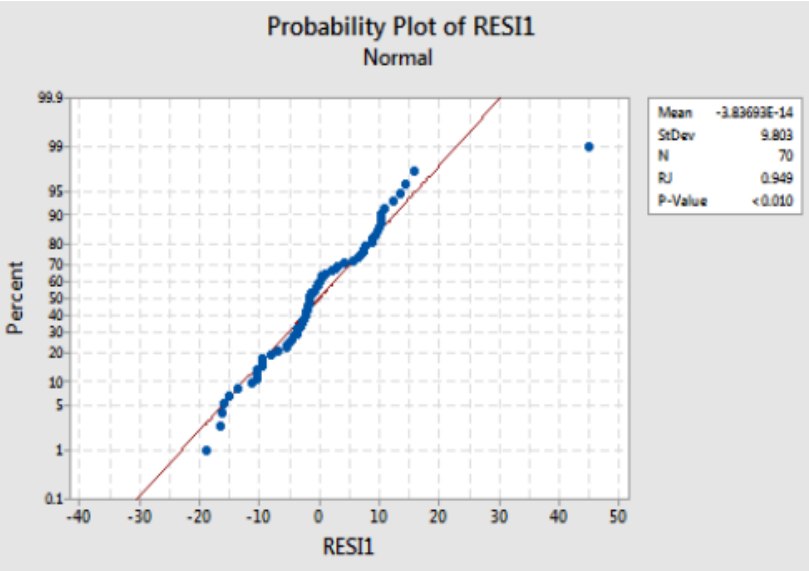
- You might have to do this when everything seems wrong — when the regression function is not linear and the error terms are not normal and have unequal variances
- In general (although not always!):
  - Transforming the y values corrects problems with the error terms (and may help the non-linearity).
  - Transforming the x values primarily corrects the non-linearity.

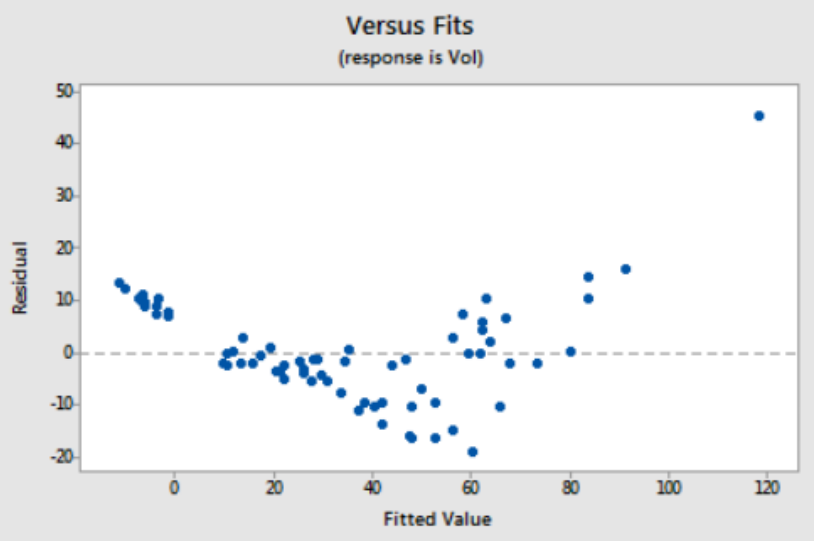
Log Transformation of both predictor and response

- Lets again consider an example to understand this…
- Many different interest groups — such as the lumber industry, ecologists, and foresters — benefit from being able to predict the volume of a tree just by knowing its diameter. One classic data set (shortleaf.txt) concerned the diameter (*x*, in inches) and volume (*y*, in cubic feet) of *n* = 70 shortleaf pines

Log Transformation of both predictor and response

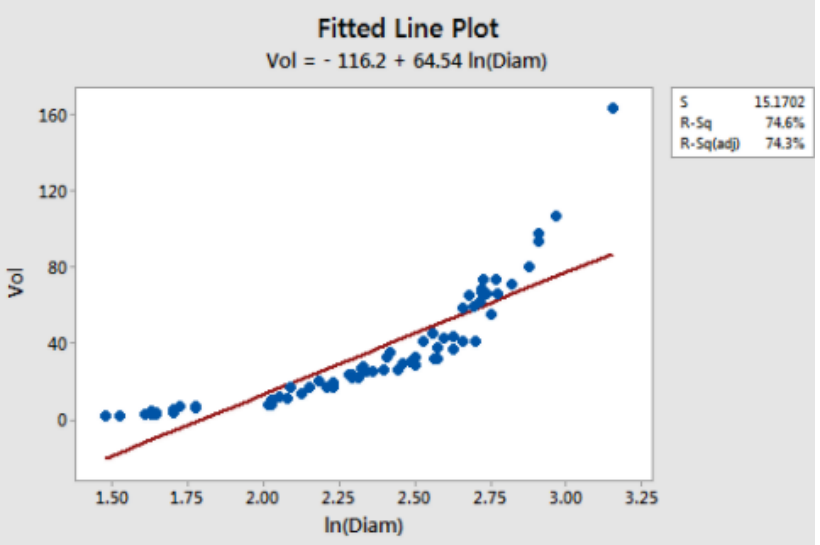


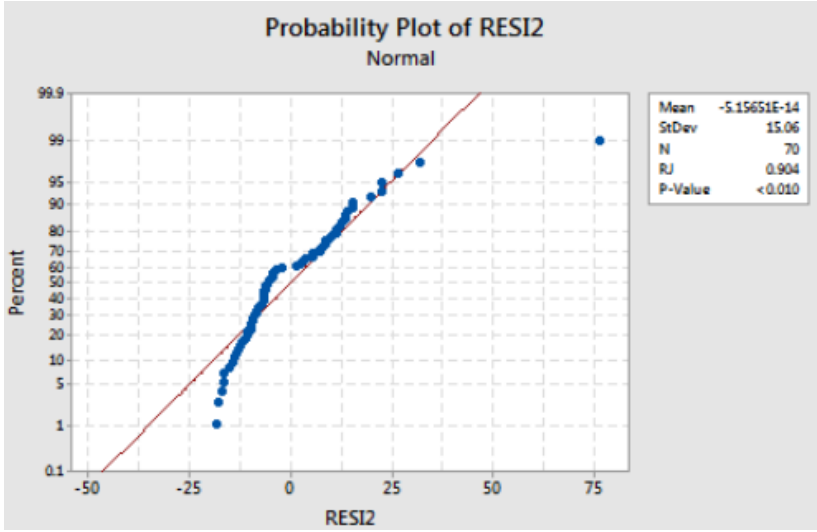


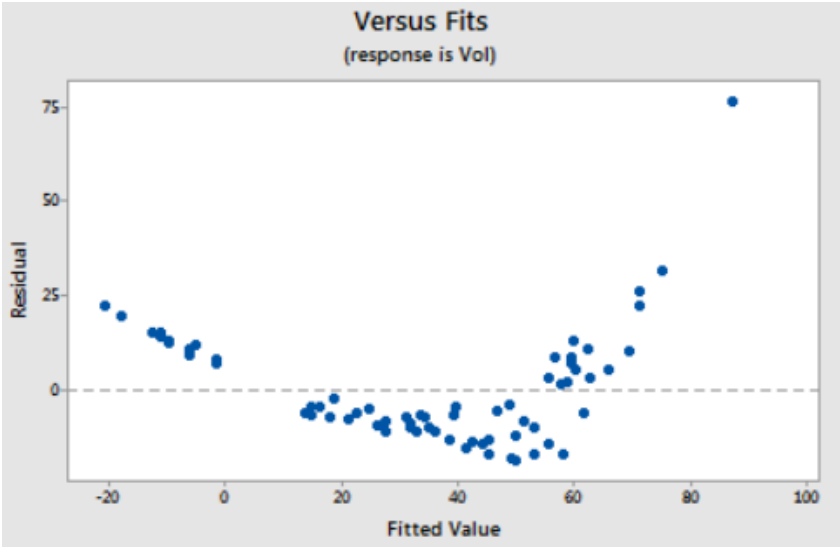


Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Log Transformation of both predictor and response



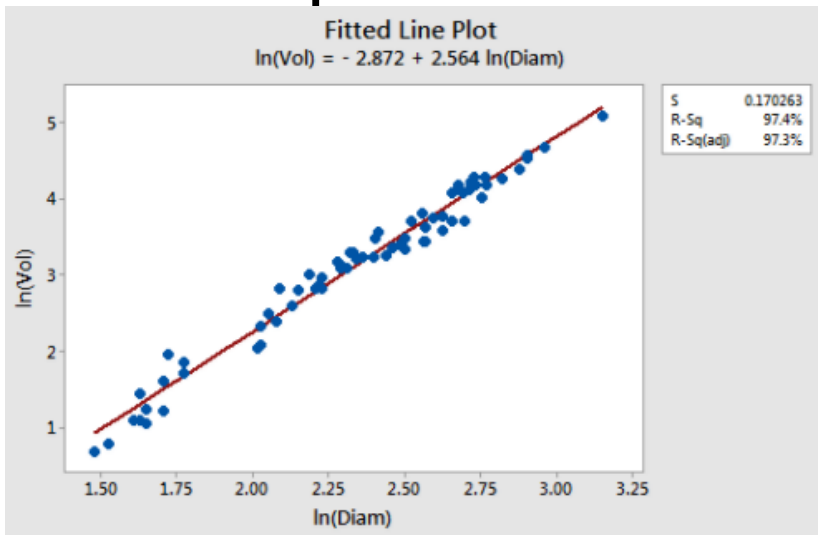


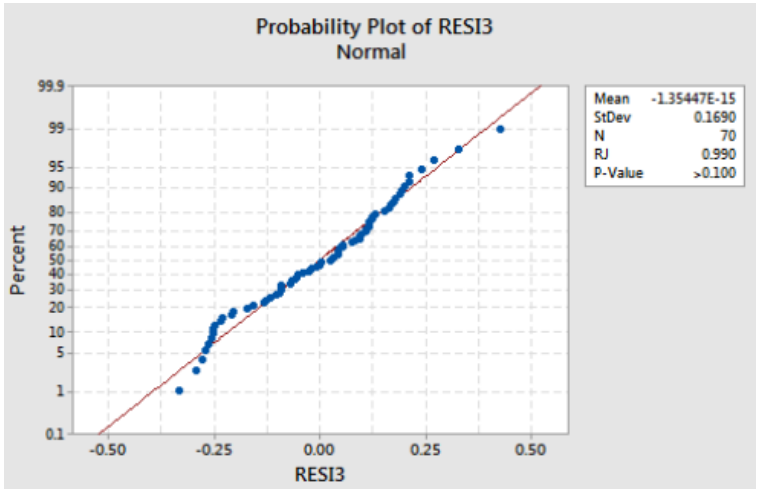


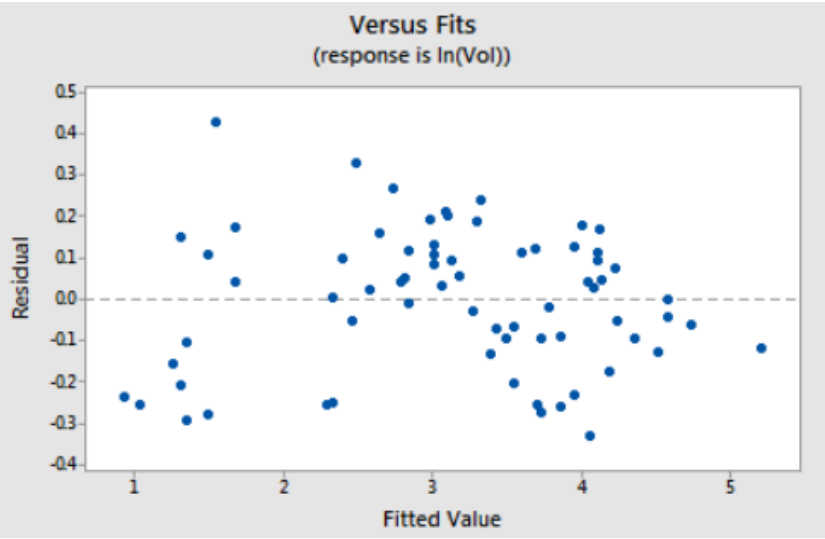
Transforming the x values (lnDiam)

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Log Transformation of both predictor and response







Transforming the y values (lnVol)

Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Log Transformation of both predictor and response

- What is the nature of the association between diameter and volume of shortleaf pines?
- Is there an association between diameter and volume of shortleaf pines?

Other Data Transformations

- Remember – you will need lots of trial & error!
- Therefore, the best we can do is offer advice and hope that you find it helpful!

Other Data Transformations

- Advice 1: If the primary problem with your model is non-linearity, look at a scatter plot of the data to suggest transformations that might help

Other Data Transformations

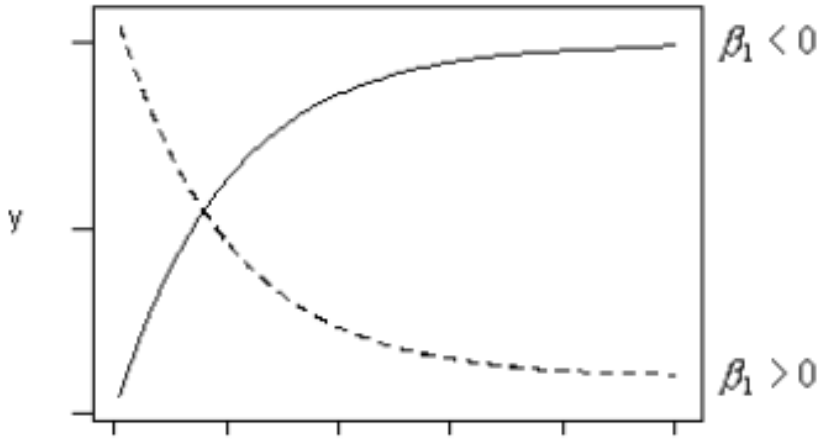
- If the trend in your data follows either of these patterns, you could try fitting this regression function:

$$\mu_Y = \beta_0 + \beta_1 e^{-x}$$

OR

$$\mu_Y = \beta_0 + \beta_1 (1/x)$$

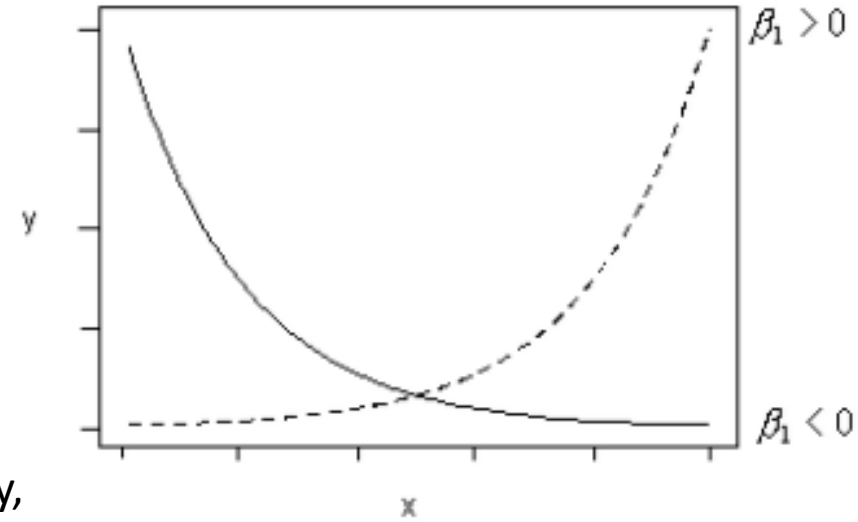
to your data.



- If the trend in your data follows either of these patterns, you could try fitting this regression function:

$$\mu_{\ln Y} = \beta_0 + \beta_1 x$$

to your data.



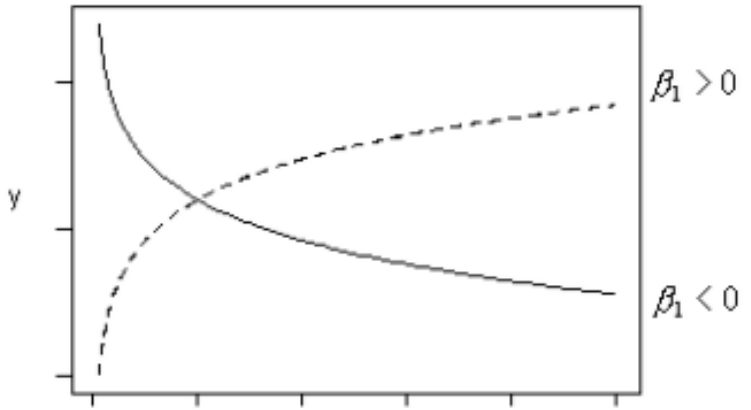
Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Other Data Transformations

- If the trend in your data follows either of these patterns, you could try fitting this regression function:

$$\mu Y=\beta 0+\beta 1\ln(x)$$

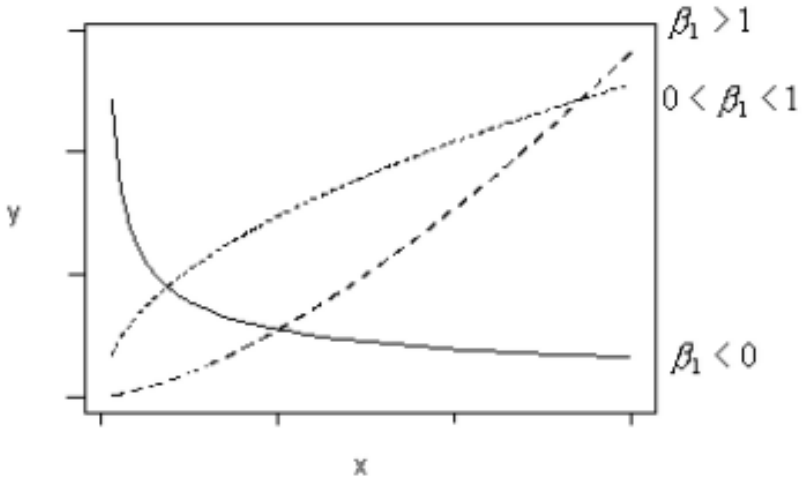
to your data.



- If the trend in your data follows either of these patterns, you could try fitting this regression function:

$$\mu \ln Y=\beta 0+\beta 1\ln(x)$$

to your data.



Plot No. 28, 4th Floor, Suraj Trade Center **Opp. Cyber Towers,** Hitech City, Hyderabad. Tel: 040 - 66828899Email: info@analyticspath.com

Other Data Transformations

- Advice 2: If the variances are unequal and/or error terms are not normal, try a "power transformation" on y.
- A power transformation on y involves transforming the response by taking it to some power λ. That is y∗=y^λ.
- Most commonly, for interpretation reasons, λ is a "meaningful" number between -1 and 2, such as -1, -0.5, 0, 0.5, (1), 1.5, and 2 (i.e., it's rare to see λ=1.362, for example
- One procedure for estimating an appropriate value for λ is the so-called Box-Cox Transformation, which we'll explore further…

Other Data Transformations

- Advice 2: If the variances are unequal and/or error terms are not normal, try a "power transformation" on y.
- A power transformation on y involves transforming the response by taking it to some power λ. That is y∗=y^λ.
- Most commonly, for interpretation reasons, λ is a "meaningful" number between -1 and 2, such as -1, -0.5, 0, 0.5, (1), 1.5, and 2 (i.e., it's rare to see λ=1.362, for example

Other Data Transformations

- Advice 3: It's not really okay to remove some data points just to make the transformation work better, but if you do, make sure you report the scope of the model.
- Advice 4: It's better to give up some model fit than to lose clear interpretations. Just make sure you report that this is what you did.