Stats Bootcamp

MSc in Analytics

Dr. Anil D Chaturvedi

Email: anilchaturvedi@uchicago.edu

Phone: 301-299-2434

THE UNIVERSITY OF CHICAGO

Welcome to Stats Bootcamp

THE UNIVERSITY OF CHICAGO

Session 3: Topics

1. Hypothesis tests
2. Power of a Test
3. Simple Linear Regression the formulation
4. Simple Linear Regression – the solution
5. Simple Linear Regression – Assumptions and Inferences
6. The Error term. The Degrees of freedom
7. Simple Linear Regression – Diagnostics

Hypothesis Tests

	POPULATION STATUS NULL HYPOTHESIS FALSE	POPULATION STATUS NULL HYPOTHESIS TRUE
OUTCOME OF HYPOTHESIS TEST REJECT NULL	**POWER** **1-β**	**Type 1 Error - α**
OUTCOME OF HYPOTHESIS TEST FAIL TO REJECT NULL	**Type 2 Error - β**	**CONFIDENCE** **1-α**

Steps in Hypothesis Testing

1. Decide what you want to assess using Data (e.g. μ > 0) in population.
2. Set up the H0 in a direction opposite to what you want to demonstrate (e.g. μ <=0)
3. Decide the alpha level (confidence) – though this is OLD OLD fashioned – e.g. 1, .05, .01, .001
4. Determine the 'Rejection Region'
5. Compute the Test Statistic (e.g. M-0/(S/n**0.5))
6. Assess if it has a Normal, or a t distribution depending on your test statistic
7. Look up the appropriate theoretical distribution and locate the Test statisitc value.
8. If it falls in rejection region - Reject the Null . If it does NOT fall in rejection region, you "Fail to Reject Null". Remember – you DO NOT ACCEPT NULL.

Description	HO about Population	HA about Population	Calculated test Statistic in sample	Rejection Region
Test of Mean	$\mu = k$	$\mu <> k$	(x-k)/se	Left and right tails t
Test of Mean	$\mu <= k$	$\mu > k$	(x-k)/se	Right Tail t
Test of Mean	$\mu >= k$	$\mu < k$	(x-k)/se	Left Tail t
Test of proportion	$\pi = k$	$\pi <> k$	(p-k)/se	Left and Right tail Normal
Test of proportion	$\pi <= k$	$\pi > k$	(p-k)/se	Right Tail Normal
Test of proportion	$\pi >= k$	$\pi < k$	(p-k)/se	Left Tail Normal

Description	HO about Population	HA about Population	Calculated test Statistic in sample	Rejection Region
2-Group means	$\mu1 = \mu2$	$\mu1 <> \mu2$	(M1-M2)/se	Left and right tails t
2-Group means	$\mu1 <= \mu2$	$\mu1 > \mu2$	(M1-M2)/se	Right Tail t
2-Group means	$\mu1 >= \mu2$	$\mu1 < \mu2$	(M1-M2)/se	Left Tail t
2-Group proportions	$\pi1 = \pi2$	$\pi1 <> \pi2$	(p1-p2)/se	Left and Right tail Normal
2-Group proportions	$\pi1 <= \pi2$	$\pi1 > \pi2$	(p1-p2)/se	Right Tail Normal
2-Group proportions	$\pi1 >= \pi2$	$\pi1< \pi2$	(p1-p2)/se	Left Tail Normal

Illustration of Test

Analyzing Wages by Employee Health

Health	Wage
Good	75.04315
Vgood	70.47602
Good	130.98218
VGood	154.68529
Good	75.04315
VGood	127.11574

Analyzing Wages by Employee Health

Health	Wage
Good	75.04315
Vgood	70.47602
Good	130.98218
VGood	154.68529
Good	75.04315
VGood	127.11574

- n = 3000
- Good: 858, Mean - 101.7, sd = 35.2
- Vgood: 2142, Mean – 115.7, sd = 43.4
- Average variance = 39.3

- H0: Mean Wages for VGood <= Mean Wages for Good
- TS = Test Statistic: $\frac{(115.7-101.7)-0}{\sqrt{\frac{35.2^2}{858}+\frac{43.4^2}{2142}}}$ = 9.185

- Upper Tail rejection region: p(z>0.05) is any value > 1.64536
- qt(.95, df=3000-2) = 1.64536
- Since TS >>> 1.64536 – Reject H0. Significant at 95% level
- ALTERNATELY: Conclude that Almost certain that Mean Wage of VG > Mean Wage of G since 1-pt(9.185,df=3000-2) = 0

Analyzing Wages by Employee Health

Health	Wage
Good	75.04315
Vgood	70.47602
Good	130.98218
VGood	154.68529
Good	75.04315
VGood	127.11574

- n = 3000
  Good: 858, Mean - 101.7, sd = 35.2
- Vgood: 2142, Mean – 115.7, sd = 43.4
- Average variance = 39.3

- H0: Mean Wages for VGood <= Mean Wages for Good

- TS = Test Statistic: $\frac{(115.7-101.7)-0}{\sqrt{\frac{35.2^2}{858}+\frac{43.4^2}{2142}}}$ = 9.185

- Upper Tail rejection region: p(z>0.05) is any value > 1.64536
- qt(.95, df=3000-2) = 1.64536
- Since TS >>> 1.64536 – Reject H0. Significant at 95% level
- ALTERNATELY: Conclude that Almost certain that Mean Wage of VG > Mean Wage of G since 1-pt(9.185,df=3000-2) = 0

Analyzing Wages by Employee Health

Health	Wage
Good	75.04315
Vgood	70.47602
Good	130.98218
VGood	154.68529
Good	75.04315
VGood	127.11574

$$\text{d.f.} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}.$$

- n = 3000
- Good: 858, Mean - 101.7, sd = 35.2
- Vgood: 2142, Mean – 115.7, sd = 43.4
- Average variance = 39.3

- H0: Mean Wages for VGood <= Mean Wages for Good
- TS = Welch's Test Statistic: $\frac{(115.7-101.7)-0}{\sqrt{\frac{35.2^2}{858}+\frac{43.4^2}{2142}}}$ = 9.185

- Welch's df = 1931.75 ~ 1932
- Upper Tail rejection region: p(z>0.05) is any value > 1.64564
- qt(.95, df=1932) = 1.64564
- Since TS >>> 1.64564 – Reject H0. Significant at 95% level
- ALTERNATELY: Conclude that Almost certain that Mean Wage of VG > Mean Wage of G since 1-pt(9.185,df=1932) = 0

Analyzing Wages by Employee Health

Health	Wage
Good	75.04315
Vgood	70.47602
Good	130.98218
VGood	154.68529
Good	75.04315
VGood	127.11574

- n = 3000
- Good: 858, Mean - 113.7, sd = 35.2 **SUPPOSE**
- Vgood: 2142, Mean – 115.7, sd = 43.4
- H0: Mean Wages for VGood <= Mean Wages for Good
- TS = Welch's Test Statistic: $\frac{(115.7-113.7)-0}{\sqrt{\frac{35.2^2}{858}+\frac{43.4^2}{2142}}}$ = 1.312
- Upper Tail rejection region: p(t>0.05) is any value > 1.64564
- qt(.95, df=1932) = 1.64564
- Since TS < 1.64564 – Fail to Reject H0. Insignificant at 95%
- ALTERNATELY: Cannot Conclude that Mean Wage of VG > Mean Wage of G since 1-pt(1.312,df=1932) = 0.094838

Correlations

- Ranges between -1 and + 1
- Correlation does NOT imply causality
- Can be Hypothesis tested for significance using following formula statistic

$$r = \frac{1}{n-1}\sum_{i=1}^{n} x_{standardized} y_{standardized}$$

$$z = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

What Test to Use?

- Are the student final exam scores in classroom A significantly different from the average of 80% ?
- Is there a significant difference in the mean height between the Eastern and Western Conference National Basketball League players?
- Is there a significant difference in depression symptoms measured before and after attending a 10-week group therapy program?
- It took 10 minutes and 5 seconds for Hal to run a mile last week when it was sunny but windy. Hal wants to compare his time to his friend's running time of 9 minutes and 45 seconds, completed in perfect conditions. The day Hal ran, the mean time was 10 minutes and 50 seconds with a standard deviation of 15 seconds. The day his friend ran, the mean time was 10 minutes and 10 seconds with a standard deviation of 10 seconds. Who had the higher percentile score?
- Are students in MSCA getting significantly less sleep than the recommended 8 hours?
- Is there a significant difference between first year post college incomes of graduates from public versus private institutions?
- Did student SAT scores increase significantly after attending a SAT prep course?
- The GRE Verbal section mean = 500 and standard deviation = 113; Math section mean = 525 and standard deviation = 135. Class average scores are 600 on Verbal and 680 on Math: Which GRE section did the class do best on when comparing to others who took the same test?

Power of a Statistical Test

Factors Affecting Power

- Sample size - large
- Effect size -large
- Standard deviation -small
- α - large
- One or two tailed test – one tailed
- Dependent or independent samples – dependent samples

T Test with Related Samples

Patient	Before	After	D	D^2
1	9	3	-6	36
2	4	1	-3	9
3	5	0	-5	25
4	4	3	-1	1
5	7	2	-5	25
Sum	29	9	-20	96

- $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2 * r * \sigma_1 * \sigma_2}}$
- $D = \bar{x}_1 - \bar{x}_2$
- $t = \frac{\bar{D}}{\sqrt{\frac{\sum_{i-1}^{N} D^2 - \frac{(\sum D)^2}{N}}{N*(N-1)}}}$

H0: $\mu_A = \mu_B$
H1: $\mu_A \neq \mu_B$

$$t = \frac{1.8 - 5.8}{\sqrt{\frac{96 - 400/5}{5 * 4}}}$$

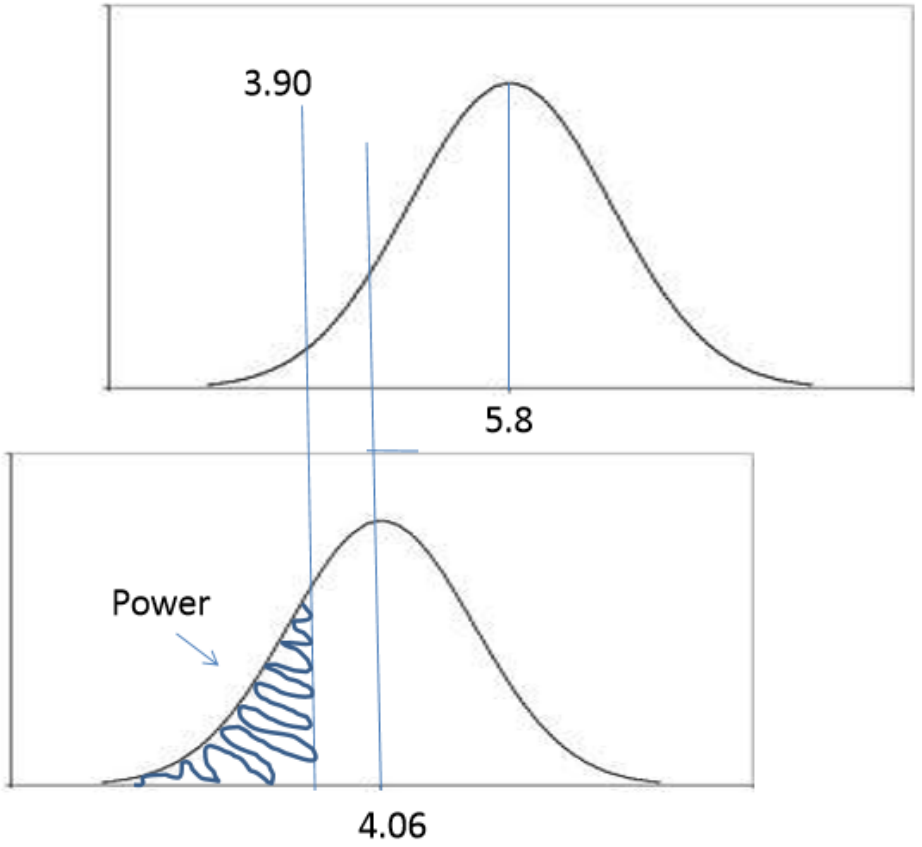
t = -4.49
df = N -1 = 5-1 = 4.
P(t(4)=-4.49) = 0.005

30% Decrease in Symptoms

Rejection region = 5.8 – 0.89 * 2.13 = 3.90

$t = \frac{3.90 - 4.06}{0.89} = -0.18$

P(t(4)<-0.18) = 0.43



Analyzing Wages by Employee Health

Health	Wage
Good	75.04315
Vgood	70.47602
Good	130.98218
VGood	154.68529
Good	75.04315
VGood	127.11574

- n = 3000
- Good: 858, Mean - 113.7, sd = 35.2 **SUPPOSE**
- Vgood: 2142, Mean – 115.7, sd = 43.4
- H0: Mean Wages for VGood <= Mean Wages for Good
- TS = Welch's Test Statistic: $\frac{(115.7-113.7)-0}{\sqrt{\frac{35.2^2}{858}+\frac{43.4^2}{2142}}}$ = 1.312
- Upper Tail rejection region: p(t>0.05) is any value > 1.64564
- qt(.95, df=1932) = 1.64564
- Since TS < 1.64564 – Fail to Reject H0. Insignificant at 95%
- ALTERNATELY: Cannot Conclude that Mean Wage of VG > Mean Wage of G since 1-pt(1.312,df=1932) = 0.094838

Ordinary Least Squares Regressions

OLS Regressions

1. Formulation of simple OLS
2. Solution, interpretation. Linkage to correlations
3. Model Fit and Coefficients
4. The Error term – Normality interpretations
5. Diagnostics - Heteroskedasticity
6. Outliers, High Leverage points, Influential observations
7. Geometric interpretations. Orthogonalization.
8. Modeling Non-linearities
9. Generalizations to multivariate. Covariance structures
10. Saturated models

Covariance

Covariance is a measure of linear association between two interval scaled variables

$$cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Correlations

Popular versions of Correlation formulae you will see in text books

$$r = \frac{cov(x, y)}{std(x) * std(y)}$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

Correlations

Correlation can also be measured as a multiplication of two standardized interval or ratio scaled variables

$$r = \frac{n-1}{n-1}\sum_{i=1}^{n}\left[\frac{(x_i - \bar{x})}{\sqrt{\sum(x_i - \bar{x})^2}}\right]\left[\frac{(y_i - \bar{y})}{\sqrt{\sum(y_i - \bar{y})^2}}\right]$$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left[\frac{(x_i - \bar{x})}{std_x}\right]\left[\frac{(y_i - \bar{y})}{std_y}\right]$$

$$r = \frac{1}{n-1}\sum_{i=1}^{n} x_{standardized} y_{standardized}$$

Correlations

- Ranges between -1 and + 1
- Correlation does **NOT imply causality**
- Can be Hypothesis tested for significance using following formula statistic

$$r = \frac{1}{n-1}\sum_{i=1}^{n} x_{standardized} y_{standardized}$$

Correlations

- Ranges between -1 and + 1
- Correlation does NOT imply causality
- Can be Hypothesis tested for significance using following formula statistic

$$r = \frac{1}{n-1}\sum_{i=1}^{n} x_{standardized} y_{standardized}$$

$$z = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

OLS Simple Regression: Model Formulation

Simple Linear Regression

Regression Equation in the Population

$$y_i = B_0 + B_1 x_i + u_i$$

- Remember that there is a regression equation in the population that the modeler does not know, and hopes to estimate

Simple Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

- Try to build a relationship between two variables y and x, in a sample of n observations
  - y is the dependent variable
  - x is the independent variable
  - e is the error term
  - β_0 and β_1 denote the unknown coefficients to be estimated
  - The subscript i denotes the i^{th} observation in the sample

OLS Regression Assumptions

- **Linearity**
- **Errors ϵ_i are iid and Normal N(0,σ^2)**
- **Homoskedasticity**

Linear Regression Assumptions

Assumptions of Linear Regression

- The $E(e_i) = 0$
- The variance of (e_i), denoted by σ^2, is same for all potential values of all y and x. **How can e_i have variance? After all, e_i is a single number?**
- The values of (e_i) are independent
- e_i are normally distributed. **How can that be?**
- For all predictors, covariance (predictor, e) $= 0$

Model Solution

Simple Linear Regression

Parameter Estimation

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

- Most often, by minimizing the error (e) using the OLS – Ordinary Least Squares function.
- If we let

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

- The OLS function is defined as minimizing:

$$SSE = e^2 = \sum^{n} (y_i - \hat{y}_i)^2$$

Simple Linear Regression

Simple Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + e_i$$
$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

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

- **OLS Estimates**

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Model Fit, Coefficients, and t-values

```
> lm(Wage2$wage~Wage2$age)

Call: lm(formula = Wage2$wage ~ Wage2$age)
Coefficients:   (Intercept)   Wage2$age
                    81.7047     0.7073

> summary(lm(Wage2$wage~Wage2$age))

Call: lm(formula = Wage2$wage ~ Wage2$age)

Residuals: Min 1Q Median 3Q Max -100.265 -25.115 -6.063 16.601 205.748

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.70474 2.84624     28.71    <2e-16 ***
Wage2$age    0.70728 0.06475     10.92    <2e-16 ***

--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.93 on 2998 degrees of freedom
Multiple R-squared: 0.03827, Adjusted R-squared: 0.03795
F-statistic: 119.3 on 1 and 2998 DF, p-value: < 2.2e-16

plot(lm(Wage2$wage~Wage2$age))
```

Simple Linear Regression

R^2: Measures of Model Fit

- Coefficient of Determination R^2
- Total Sum of Squares (TSS) = $\sum_{i=1}^{n}(y_i - \bar{y})^2$, where $\bar{y}$ is the average y across all n observations.
- Sum of Squares Regression (SSR) = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, where $\hat{y}_i$ is the predicted rating from the regression model for observation i
- Sum of Squared Errors (SSE) = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- **TSS= SSE + SSR**

Coefficient of Determination R^2

R^2 = Sum of squares regression/Total sum of squares = SSR/TSS

- R^2 = SSR/TSS = 1-SSE/TSS
- $0 < R^2 < 1$ or $0 < R^2 < 100\%$

OLS Regression Properties

- $\beta = \frac{cov(y,x)}{var(x)} = correlation(x,y) * \frac{std(y)}{std(x)}$
- $R^2 = cor(y, \hat{y})^2$ Why?

OLS Regression Properties

- $\beta = \frac{cov(y,x)}{var(x)} = correlation(x) * \frac{std(y)}{std(x)}$
- $R^2 = cor(y, \hat{y})^2$ Why?
- $R^2 = \frac{SSR}{TSS} = \frac{Var(\hat{y})}{Var(y)} = \frac{cov(y,x)^2}{var(x)*var(y)} = correlation^2$

OLS Regression Properties

- $\beta = \frac{cov(y,x)}{var(x)} = correlation(x) * \frac{std(y)}{std(x)}$
- $R^2 = cor(y, \hat{y})^2$ Why?
- $R^2 = \frac{SSR}{TSS} = \frac{Var(\hat{y})}{Var(y)} = \frac{cov(y,x)^2}{var(x)*var(y)} = correlation^2$

Also, since TSS = SSR + SSE,

SSE = TSS-SSR = TSS-TSS* R^2

This implies

SSE = TSS*(1-R^2) = Var(y)*(n-1)*(1-R^2)

OLS Regression Properties

- $\beta = \frac{cov(y,x)}{var(x)} = correlation(x) * \frac{std(y)}{std(x)}$
- $R^2 = cor(y, \hat{y})^2$ Why?
- $R^2 = \frac{SSR}{TSS} = \frac{Var(\hat{y})}{Var(y)} = \frac{cov(y,x)^2}{var(x)*var(y)} = correlation^2$

Also:

Since TSS = SSR + SSE, SSE = TSS-SSR = TSS-TSS* R^2

This implies SSE = TSS(1-R^2) = Var(y)(n-1)(1-R^2)

MSE = SSE/n-2 = $\hat{\sigma}^2 = \sigma_y(1 - R^2)\frac{n-1}{n-2}$

OLS Regression Properties

$$\text{var}\left(\hat{\mathbf{b}}\right) = \text{E}\left[\left(\mathbf{b}-\hat{\mathbf{b}}\right)\left(\mathbf{b}-\hat{\mathbf{b}}\right)'\right]$$

$$\begin{aligned}\mathbf{b}-\hat{\mathbf{b}} &= \mathbf{b}-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}\\ &= \mathbf{b}-\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{X}\mathbf{b}+\boldsymbol{\varepsilon}\right)\\ &= \mathbf{b}-\mathbf{b}+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}\boldsymbol{\varepsilon}\end{aligned}$$

or

$$\mathbf{b}-\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

$$\begin{aligned}\text{var}\left(\hat{\mathbf{b}}\right) &= \text{E}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]\\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\text{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\end{aligned}$$

or

$$\text{var}\left(\hat{\mathbf{b}}\right) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

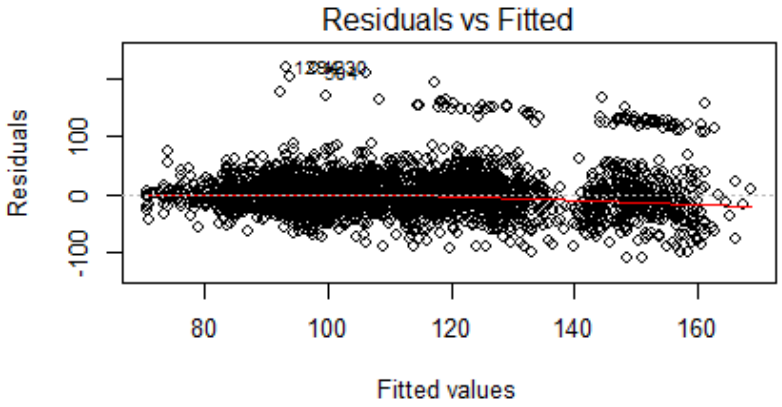
OLS Regression Properties

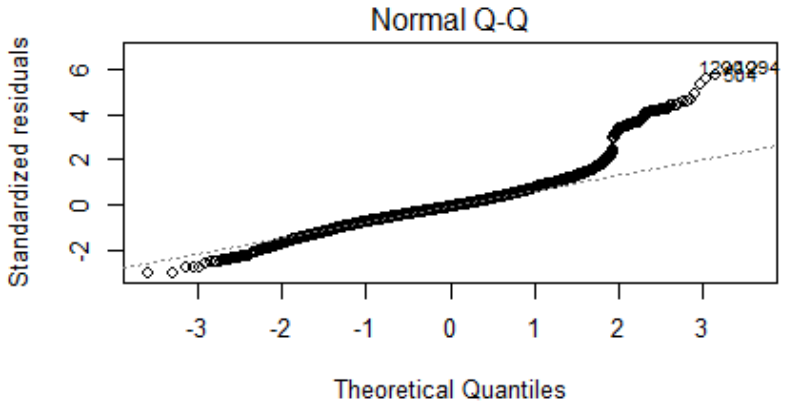
- Residuals and Predictors are orthogonal
- Mean of residuals is zero
- X's are GIVEN

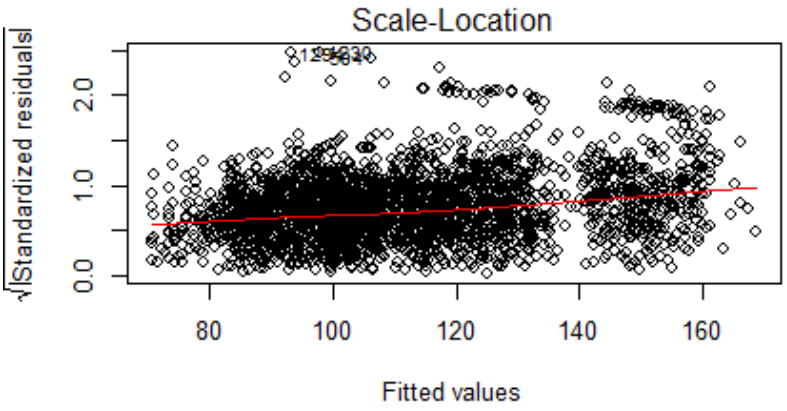
OLS Regression Properties

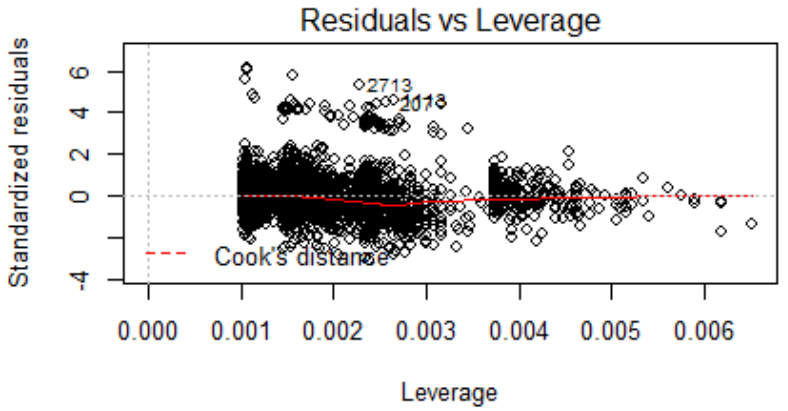
- The model yields $E(y|x) \sim x\beta$
- Regression towards the mean

• Linear Model Plots. QQ Plots for Residual Normality Check









Model Diagnostics

Model Diagnostics

- Examine the Errors. Errors are not necessarily white noise, even if they appear white noise
- Noise to one model can be structure to another.
- Just because errors are called “e” – be careful

- **Examine Errors for patterns. Question: What patterns?**

Model Diagnostics

- Non-normal errors
- Heteroskedasticity
- High leverage points, Outliers, and Influential Observations
- Correlated errors

Model Diagnostics: Non-normality

- One of the lesser important assumptions is assumption of Normality of errors.
- Simulations suggest it is the least important assumption
- However, prediction intervals may be incorrect

Model Diagnostics: Heteroskedasticity

- Signs of Heteroskedasticity : Systematic Patterns in errors correlated with predictor values or predicted y values
- Consequences:
  - Biased Standard Errors. Hence t-values cannot be relied on
    - Examine Standardized residuals vis-à-vis yhat – fitted values
    - Examine Studentized residuals

Model Diagnostics: Heteroskedasticity

Standardized Residuals

Fitted Values (yhat)

Outliers, Leverage points, Influential Observations

Regression and the HAT Matrix

- Since $\hat{\beta} = (X^T X)^{-1}(X^T y)$,
- $\hat{y} = X(X^T X)^{-1} X^T y = Hy$,

where H is the Hat Matrix

$$\mathrm{H} = X(X^T X)^{-1} X^T$$

Diagonal entries represent the "average distance of observation i from average of all the data on the respective variables

OLS Regression Properties

$$h_{ii} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$$

$$h_{ii} = \begin{bmatrix} 1 & x_i \end{bmatrix} \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ x_i \end{bmatrix}$$

$$h_{ii} = \begin{bmatrix} 1 & x_i \end{bmatrix} \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum x_i^2} \end{bmatrix} \begin{bmatrix} 1 \\ x_i \end{bmatrix} \text{ with centered } x$$

OLS Regression Properties

$$h_{ii} = \begin{bmatrix} 1 & x_i \end{bmatrix} \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum x_j^2} \end{bmatrix} \begin{bmatrix} 1 \\ x_i \end{bmatrix}$$

$$h_{ii} = \left(\frac{1}{n} + \frac{x_i^2}{\sum_{j=1}^{n} x_j^2} \right)$$

$$h_{ii} = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right) \text{ with centers}$$

OLS Regression Properties

$$h_{ii} = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right)$$

$$o \leq h_{ii} \leq 1 \; and \; \sum_i h_{ii} = p + 1$$

Simple Linear Regression

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

$$\mathrm{Var}(\boldsymbol{e}) = \mathrm{Var}((\mathrm{I}\text{-}\mathrm{H})\mathrm{y})$$

$$\mathrm{Var}(\mathrm{e}) = (\mathrm{I}\text{-}\mathrm{H})\mathrm{Var}(\mathrm{y})(\mathrm{I}\text{-}\mathrm{H})^{\mathrm{T}}$$

$$\mathrm{Var}(\mathrm{e}) = \boldsymbol{\sigma}^2(\mathrm{I}\text{-}\mathrm{H})^2 = \boldsymbol{\sigma}^2(\mathrm{I}\text{-}\mathrm{H})$$

Thus,

$$\mathrm{Var}(\mathrm{e}_i) = (1\text{-}\mathrm{h}_{ii})\,\boldsymbol{\sigma}^2$$

Outliers, Leverage points and Influential Observations

Mean Leverage: $$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{k+1}{n}$$

Standardized Residuals: $$r_i = \frac{e_i}{s(e_i)} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

Studentized Residuals: $$t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

$$t_i = r_i \left(\frac{n-k-2}{n-k-1-r_i^2} \right)^{1/2},$$

DFFITS and Cooks Distance

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{MSE_{(i)} h_{ii}}} \quad > \quad 2\sqrt{\frac{k+2}{n-k-2}}$$

Cooks Distance:

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(k+1) \times MSE} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right].$$

Prediction Intervals and Error Variances

OLS Regression Properties

$$h_{ii} = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right)$$

$$o \leq h_{ii} \leq 1 \; and \; \sum_i h_{ii} = p + 1$$

Confidence Intervals of Predictions

$$\mathrm{Var}(\hat{y}) = \mathrm{Var}(\mathrm{Hy})$$

$$\mathrm{Var}(\hat{y}) = \mathrm{HVar}(\mathrm{y})\mathrm{H}^{\mathrm{T}}$$

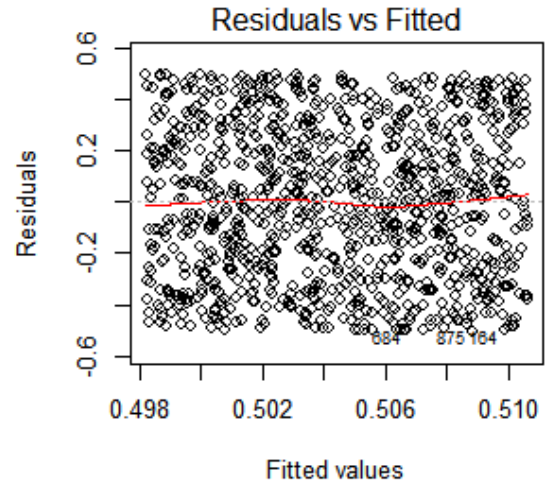
$$\mathrm{Var}(\hat{y}) = \boldsymbol{\sigma}^2\mathrm{H}^2 = \boldsymbol{\sigma}^2\mathrm{H}$$

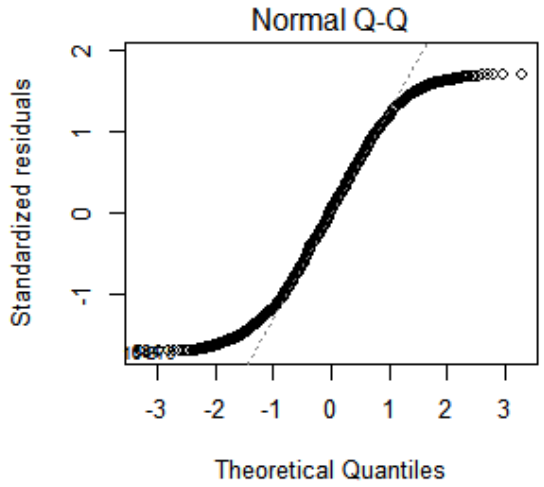
Thus, $\mathrm{Var}(\hat{y}_i) = \mathrm{h}_{ii}\,\boldsymbol{\sigma}^2$

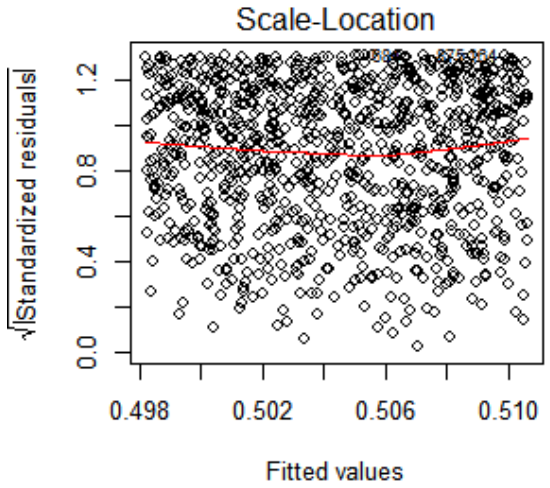
$$\mathrm{Var}(\hat{y}_i) = \boldsymbol{\sigma}^2\left(\frac{1}{\mathrm{n}} + \frac{(\mathrm{x}_i - \bar{\mathrm{x}})^2}{\sum_{\mathrm{j}=1}^{\mathrm{n}}(\mathrm{x}_\mathrm{j} - \bar{\mathrm{x}})^2}\right)$$

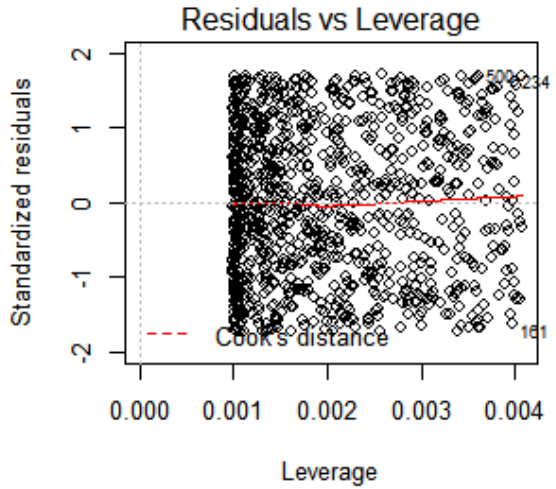
Regressing two Uniform Variables

```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.50369 -0.25569  0.00553  0.25225  0.49785

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.51067    0.01873  27.266   <2e-16 ***
x           -0.01249    0.03252  -0.384    0.701
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2925 on 998 degrees of freedom
Multiple R-squared:  0.0001477, Adjusted R-squared:  -0.0008542
F-statistic: 0.1474 on 1 and 998 DF,  p-value: 0.7011
```


Residuals vs Fitted
Residuals
Fitted values
0.498
0.502
0.506
0.510
-0.6
-0.2
0.2
0.6

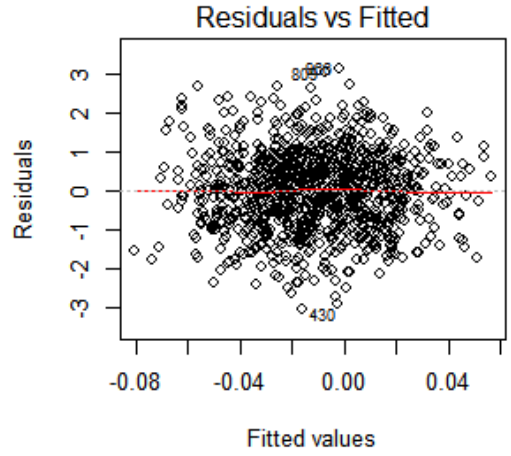

Normal Q-Q
Standardized residuals
Theoretical Quantiles
-3
-2
-1
0
1
2
3

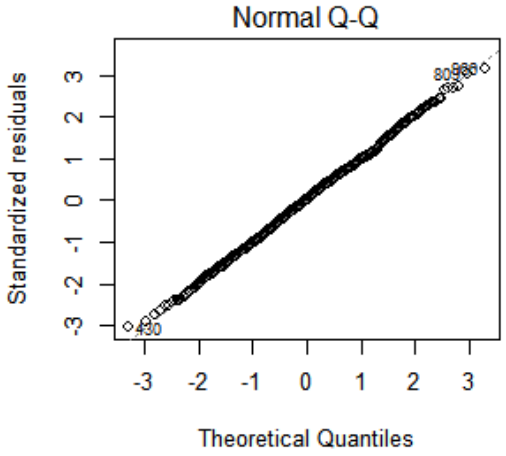

Scale-Location
√|Standardized residuals|
Fitted values
0.498
0.502
0.506
0.510
0.0
0.4
0.8
1.2

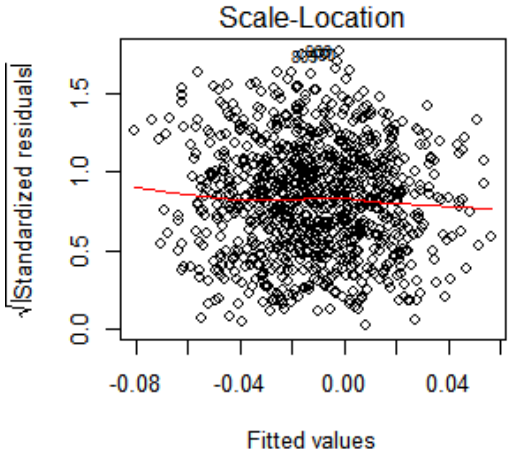

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.000
0.001
0.002
0.003
0.004
-2
-1
0
1
2

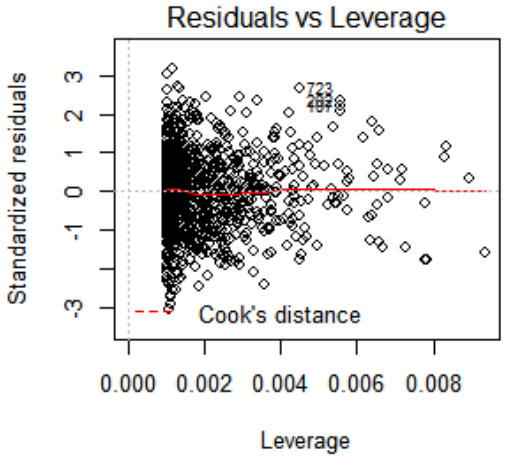
Regressing two Normal Variables

```
> y=rnorm(1000)
> x=rnorm(1000)
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-3.02673 -0.69237  0.02375  0.67264  3.12195

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01016    0.03144  -0.323    0.747
x            0.02470    0.03235   0.764    0.445

Residual standard error: 0.9934 on 998 degrees of freedom
Multiple R-squared:  0.0005839, Adjusted R-squared:  -0.0004175
F-statistic: 0.5831 on 1 and 998 DF,  p-value: 0.4453
```

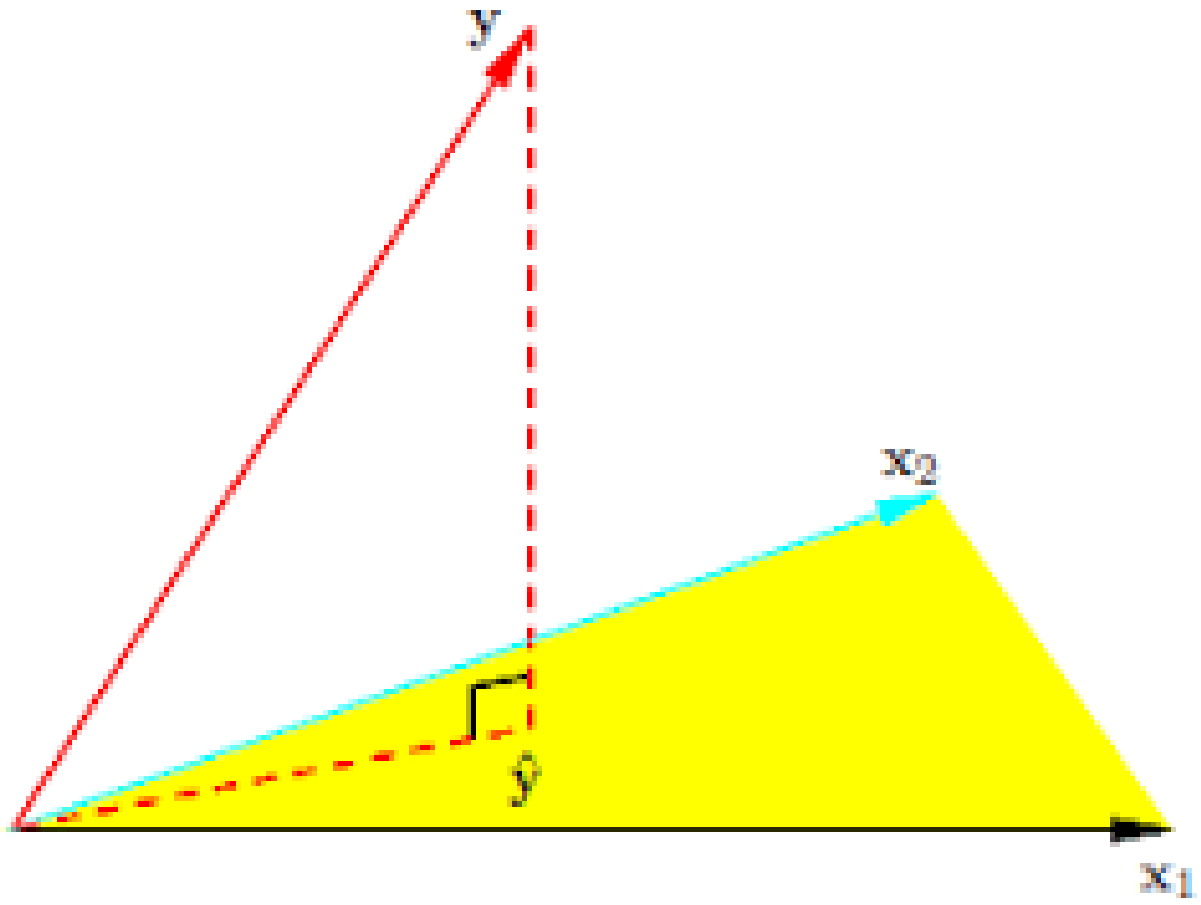

Residuals vs Fitted
Residuals
Fitted values
-0.08
-0.04
0.00
0.04
-3
-2
-1
0
1
2
3
430


Normal Q-Q
Standardized residuals
Theoretical Quantiles
-3
-2
-1
0
1
2
3
430


Scale-Location
√|Standardized residuals|
Fitted values
-0.08
-0.04
0.00
0.04
0.0
0.5
1.0
1.5


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.000
0.002
0.004
0.006
0.008
-3
-1
0
1
2
3

Geometric Interpretation Orthogonalization


y
x2
ŷ
x1

Modeling Non-Linearities

- Binning is Winning ☺

- How do you model in a bin?
  - Mean value of the responses
  - Linear function of the responses
  - Non-linear function in each bin.
    - Which?

      Polynomial

      Splines

Multiple Linear Regression

Multiple Linear Regression

- Multiple predictors (>1)
- Response variable is still Numeric
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots. \beta_p x_p + \epsilon$

- The following are **NOT** assumptions of MLR
  1. Response variable has to be normal
  2. Predictors have to be either multivariate normal or even normal.

DO NOT NEED TO TRANSFORM EITHER FOR DISTRIBUTIONAL REASONS

Multiple Linear Regression

The Solution

- Multiple predictors (>1)
- Response variable is still Numeric
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots. \beta_p x_p + \epsilon$
- $y = X\beta + \epsilon$ --- Linear Algebra Notation

- Optimal $\widehat{\beta} = (X^T X)^{-1}(X^T y)$
- $SE_\beta =$ Diagonals of $\sigma^2 (X^T X)^{-1}$

Saturated Regression Model

- What happens in a saturated model?
- $X = I$
- Optimal $\widehat{\beta} = (I^T I)^{-1}(I^T y) = y$
- SE_{β} = Diagonals of $\sigma^2 (I^T I)^{-1} = diagonals(\sigma^2 I) = \sigma^2$

Model Selection

Multiple Linear Regression

- Adjusted R-Square
- F-Statistic
- Chi-squares based on Likelihood Ratios
- Information Theoretic Criteria such as AIC, BIC etc
- Validation techniques

Multiple Linear Regression

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

$$Adjusted\ R^2 = 1 - \frac{SS_{residuals}/n-p-1}{SS_{total}/n-1}$$

$$F = \frac{R^2/p}{1-R^2/n-p-1} = \frac{SS_{regression}/p}{SS_{residuals}/n-p-1}$$

Multiple Linear Regression

$$F = \frac{\left(R^2_{Bigger} - R^2_{Smaller}\right) \Big/ \left(p_{bigger} - p_{smaller}\right)}{\left(1 - R^2_{bigger}\right) \Big/ \left(n - p_{bigger} - 1\right)}$$

$$= \frac{\left(SSE_{smaller} - SSE_{bigger}\right) \Big/ \left(p_{bigger} - p_{smaller}\right)}{SSE_{bigger} \Big/ \left(n - p_{bigger} - 1\right)}$$

Assumptions Regression does NOT make

The following are **NOT** assumptions of MLR

1. Response variable has to be normal
2. Predictors have to be either multivariate normal or even normal.

DO NOT NEED TO TRANSFORM EITHER FOR DISTRIBUTIONAL REASONS

WHY Oh WHY?

ANALYSIS OF VARIANCE

ANOVA: Purpose and Rationale

- What? Test differences in mean values across multiple groups >= 2
- How? By analyzing the variances in the Groups.
- Sounds strange? Let us see how.

ANOVA: Purpose and Rationale

Which of 3 Processes is Better?

	A	B	C
	58	58	48
	64	69	57
	55	71	59
	66	64	47
	67	68	49
Sample Mean	62	66	52
Sample Var	27.5	26.5	31.0
Sample Std	5.24	5.15	5.57

ANOVA: Purpose and Rationale

	A	B	C
	58	58	48
	64	69	57
	55	71	59
	66	64	47
	67	68	49
Sample Mean	**62**	**66**	**52**
Sample Var	27.5	26.5	31.0
Sample Std	5.24	5.15	5.57

- H0: $\boldsymbol{\mu_1 = \mu_2 = \mu_3}$
- HA: Not H0

- Assume:
- Normality within Populations
- Variance σ^2 equal in all populations
- Observations are independent

ANOVA: Purpose and Rationale

	A	B	C
	58	58	48
	64	69	57
	55	71	59
	66	64	47
	67	68	49
Sample Mean	**62**	**66**	**52**
Sample Var	27.5	26.5	31.0
Sample Std	5.24	5.15	5.57

H0: $\boldsymbol{\mu_1 = \mu_2 = \mu_3}$

HA: Not H0

Rationale: Under H0:
There are two estimates of population variance.

1. Within Group Variances
2. Between-Group Variance if H0 is true

ANOVA: Purpose and Rationale

	A	B	C
	58	58	48
	64	69	57
	55	71	59
	66	64	47
	67	68	49
Sample Mean	**62**	**66**	**52**
Sample Var	27.5	26.5	31.0
Sample Std	5.24	5.15	5.57

H0: $\mu_1 = \mu_2 = \mu_3$

HA: Not H0

Rationale: Under H0:
There are two estimates of population variance.

1. **Within Group estimate of Variance σ^2 = average of 27.5, 26.5, 31.0 = 28.33 (equal n)**
2. Between-Group Variance if H0 is true

ANOVA: Purpose and Rationale

	A	B	C
	58	58	48
	64	69	57
	55	71	59
	66	64	47
	67	68	49
Sample Mean	**62**	**66**	**52**
Sample Var	27.5	26.5	31.0
Sample Std	5.24	5.15	5.57

H0: $\mu_1 = \mu_2 = \mu_3$ HA: Not H0

Rationale: Under H0:

There are two estimates of population variance.

1. Within Group estimate of Variance σ^2 = average of 27.5, 26.5, 31.0 = 28.33 (equal n)
2. **Between-Group Variance if H0 is true: 3 samples come from same sampling distribution with means 62, 66, and 52. So sd of S.D. = (62-60)^2 + (66-60)^2 + (52-60)^2 /2 = 52. But this is sd of SD. So estimate of population is 52*5 = 260**

ANOVA: Purpose and Rationale

H0: $\mu_1 = \mu_2 = \mu_3$ HA: Not H0

Rationale: Under H0:

There are two estimates of population variance.

1. **Within Group estimate of Variance σ^2 = average of 27.5, 26.5, 31.0 = 28.33 (equal n)**
2. **Estimate of Between-Group based population variance is 52*5 = 260**
3. **Ratio of two independent estimates of variances is distributed Fisher (F).**
4. **Estimate (1) is much better than (2) since it does not depend on H0.**
5. **Estimate 2/Estimate 1 – large values indicate False H0.**
6. **Calculated F = 260/28.33 = 9.18**
7. **F(2,12) = 6.93 (Alpha = .01) and 5.10 (Alpha = .025)**

ANOVA Table

Source of variation	Sum of Squares	Degrees of Freedom	Mean Square Error	F	P-value
Groups	520	2	260	9.18	.004
Error	340	12	28.33		
Total	860	14			

ANOVA: Purpose and Rationale

Source	Sum of squares	Degree of Freedom	Mean squares	F	F-test
Treatment	SS_T	k-1	$MS_T = \frac{SS_T}{k-1}$	$F = \frac{MS_T}{MS_E}$	$F > F_{\alpha,k-1,N-k}$?
Error	SS_E	N-k	$MS_E = \frac{SS_E}{N-k}$		
Total	TotalSS	N-1			

Total Sum of Squares = Between Group + Within Group Sum of Squares

$$TSS = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n_1}(y_i - \bar{x})^2 + \sum_{i=n_1+1}^{n}(z_i - \bar{x})^2$$

$$\sum_{i=1}^{n_1}(y_i - \bar{x})^2 = \sum_{i=1}^{n_1}(y_i - \bar{y} + \bar{y} - \bar{x})^2 =$$

$$\sum_{i=1}^{n_1}(y_i - \bar{y})^2 + \sum_{i=1}^{n_1}(\bar{y} - \bar{x})^2 + 2\sum_{i=1}^{n}(y_i - \bar{y})(\bar{y} - \bar{x})$$

$$=$$

$$= TSS_{G1} + BGSS_{G1} + 0 = TSS_{G1} + BGSS_{G1}$$

$$\text{Similarly } \sum_{i=n_1+1}^{n}(z_i - \bar{x})^2 = TSS_{G2} + BGSS_{G2}$$

$$\text{Thus, } TSS = TSS_{G1} + TSS_{G2} + BGSS_{G1} + BGSS_{G2}$$

$$\text{Thus, } TSS \geq TSS_{G1} + TSS_{G2}$$