LECTURE 4 – CORRELATION AND REGRESSION

Book Chapter 12

LECTURE 4A – 1 – COVARIANCE AND CORRELATIONS

Multivariate Joint Distributions

- We have seen that variables such as Age, Income, Gender, etc., can be represented by different kinds of random variables.
- For example, Gender can be represented by Bernoulli and Binomial random variables, Education by a multinomial random , and Age and Income by Normal random variables.
- We dealt with these as individual random variables having probability distributions.
- For example, we calculated probabilities such as P(Income $\leq$ 30000) assuming a Normal Distribution.

Table 1

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Multivariate Joint Distributions

- We now begin to look at relationships *between random variables.*
- For example:
  - (Gender and Education) – Is Education independent of Gender or is there a dependence? (Contingency Table)
  - (Gender and Sales) – Is there a difference in mean sales between Males and Females in the population? (Two Sample test)
  - (Income and Sales) – What happens to Sales (in the population) when Income (in the population) increases? (simple Regression)
  - (Income, Net Worth and Sales) – What happens to Sales (in the population) when Income and Net Worth (in the population) increase? (Multiple Regression)
  - What happens to the odds that a person is Male, given that Income increases by 10,000? (Logistic Regression)
  - Is th Education? (ANC

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Table 1

Multivariate Joint Distributions

- A *joint distribution* gives the probability for the values of two or more random variables.
- We have already seen an example of this in the *contingency table* for two events. If X is the random variable that Carlos scores a Goal on his first attempt (X = 1) or Not (X = 0) and Y is the random variable that Carlos scores a Goal on his second attempt (Y = 1) or Not (Y = 0), then the contingency table can be modified to show it.
- The *joint probabilities* are the entries in the cell.
  - $P(X=1, Y=1) = 0.585$
  - $P(X=1, Y=0) = 0.065$
  - $P(X=0, Y=0) = 0.285$;
  - $P(X=0, Y=1) = 0.065$
- The marginal probabilities are the probabilities in the *margins.*
  - $P(X=1) = 0.65$
  - $P(X=0) = 0.35$
  - $P(Y=1) = 0.65$
  - $P(Y=0) = 0.35$
- Because $P(X=1, Y=1) = 0.585 \neq P(X=1) * P(Y=1) = 0.65*0.65$, X and Y are **not independent.**

<u>**The Carlos Goal Problem**</u>: A = the event Carlos is successful on his first attempt. $P(A) = 0.65$. B = the event Carlos is successful on his second attempt. $P(B) = 0.65$.

	A	**A^c**	
B	0.585	0.065	0.65
B^c	0.065	0.285	0.35
	0.65	0.35	

	X=1	**X=0**	
Y=1	0.585	0.065	0.65
Y=0	0.065	0.285	0.35
	0.65	0.35	

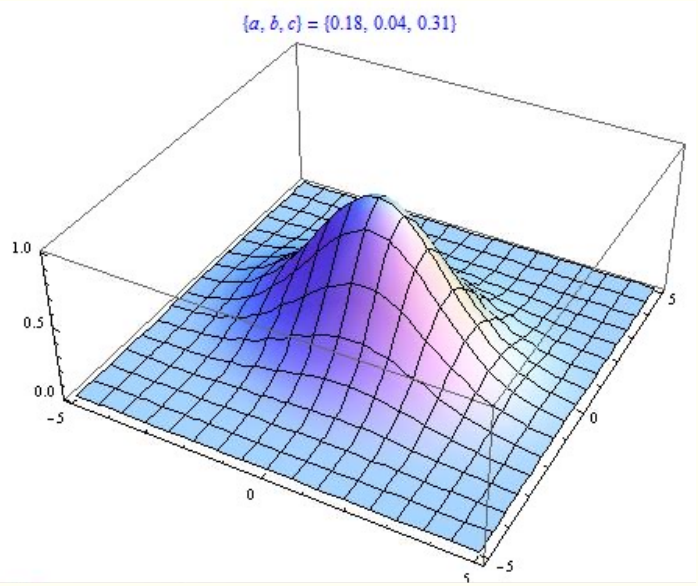
Multivariate Joint Distributions

- If both X and Y are continuous such as Age and Income (say normally distributed), the *joint distribution* can be expressed as:
  - P(Age ≤ 40 **and** Income ≤ 30000).

- The individual probabilities P(Age ≤ 40) and P(Income ≤ 30000) come from *marginal distributions*.

- For example:
  - P(Age ≤ 40) = 0.40, P(Income ≤ 30000) = 0.30 and P(Age ≤ 40 **and** Income ≤ 30000) = 0.20.
  - Because P(Age ≤ 40) * P(Income ≤ 30000) = 0.12 ≠ 0.20 = P(Age ≤ 40 **and** Income ≤ 30000), Age and Income are **not independent**.

- When the two continuous random variables are not independent, we say that they *covary* i.e., they have a relationship or association. One measure for this relationship is called *covariance*.



Covariance and Correlation

- The pairwise covariance between two random variables, as the name implies, measures how the two random variables covary and is computed as:
  - $\boldsymbol{\sigma}_{\mathbf{XY}} = E((X - \boldsymbol{\mu}_{\mathbf{X}})(Y - \boldsymbol{\mu}_{\mathbf{Y}}))$.
  - The Expectation is over the joint distribution of X & Y.
  - We do not need to know how to evaluate this expectation for the population.
  - We will just use sample formulas when needed.

- Sample Covariance: $s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$

- For the data set below, we can evaluate the covariance for every pair of the variables (Age, Income, NetWorth, Sales) as a Covariance Matrix using **R**.

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Covariance Matrix in R – CovCorr.R

- We first put the variables in a matrix (M) and then use the cov() function.
- The diagonals represent the sample variances and the off-diagonals represent the covariances
- It is a symmetric matrix, with the same values above and below the diagonals.
- For example, the covariance between (Age, Income) is -346651.11.
- The negative sign implies that as Age increases, Income decreases.
- The unit for this covariance is “years-dollars”.
- If the units for the covariances of different pairs are different, **we cannot compare covariances**.
- If we want to compare the relationships among pairs of variables, we have to use *correlations*.

```
> # Read csv file as a DataFrame
> #
> setwd("C:\\Users\\sarathy\\Documents\\2019-Teaching\\Fall2019\\Fall2019-MSIS5503\\MSIS-5503-Data")
> df <- read.table('ClassData.csv',
+                  header = TRUE, sep = ',')
>
> #Assign variable names to DataFrame Column objects
> id <- df$ID
> name <- df$Name
> age <- df$Age
> gender <-df$Gender
> education <- df$Education
> crediscore <-df$CreditScore
> income <- df$Income
> networth <-df$NetWorth
> sales <-df$Sales
> M <- cbind(age, income, networth, sales)
> #
> covar_mat <- cov(M, use="all.obs", method ="pearson")
> print(covar_mat)
                 age        income     networth         sales
age         226.7111     -346651.1     637788.5 -4.042667e+02
income   -346651.1111 13305218777.8 3235945770.0 -3.155592e+07
networth  637788.5333  3235945770.0 8743448161.3 -9.058230e+06
sales       -404.2667    -31555921.1   -9058229.9  5.067921e+05
> 
```

Covariance and Correlation – CovCorr.R

- **Correlation** which is a unitless measure of the relationship and has value between -1 and +1.
  - Correlation (population) $\boldsymbol{\rho_{XY}} = \boldsymbol{\sigma_{XY}}/(\boldsymbol{\sigma_X}^*\boldsymbol{\sigma_Y})$
  - Correlation (sample) $r_{XY} = s_{XY}/(s_X{}^*s_Y)$
- Dividing by the respective standard deviations of the two variables makes the measure unitless and have values between -1 and 1.
- *The closer it is to 1 (or -1), stronger the linear relationship*. Closer to 0 represents a weaker *linear* relationship.
- We can present the pairwise-correlations among the four variables as a ***symmetric correlation matrix.***
- In **R**, we use the cor() function.
- The correlation between Age and Income = -0.2 implying a relatively weak negative relationship. This relationship is stronger than that with Sales.

```
> #
> # We use the signif() function instead of round(), beccause round() will round to the nearest integer
> corr_mat <- cor(M, method = "pearson")
> print(signif(corr_mat),digits = 4)
              age  income networth     sales
age       1.00000 -0.1996   0.4530 -0.03772
income   -0.19959  1.0000   0.3000 -0.38429
networth  0.45300  0.3000   1.0000 -0.13608
sales    -0.03772 -0.3843  -0.1361  1.00000
> |
```

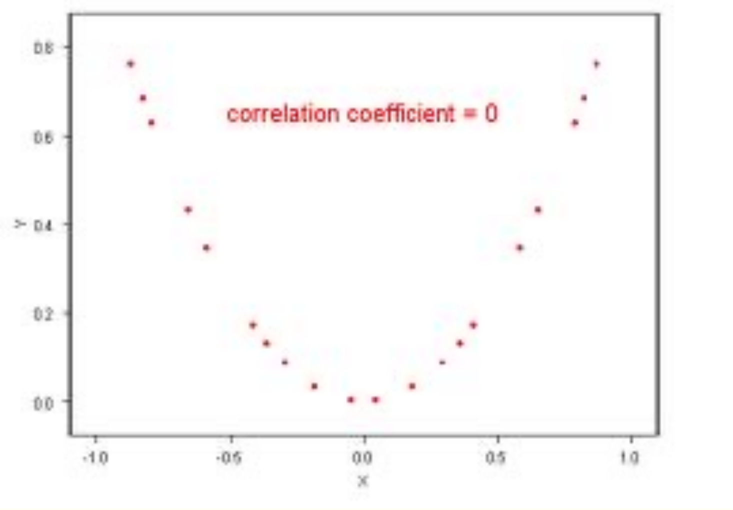
Correlation is a measure of a Linear Relationship

- Correlation measures the strength of *linear association* between two random variables

- Zero correlation *does not* imply that there is no association between the random variables. In fact, there may be a *strong non-linear association between them*

- *If two random variables are independent, they have zero correlation, because they have no relationship. However, if the correlation between them is 0, it does not mean that they are independent. They may still have a nonlinear relationship.*

- In the picture, the two variables have a strong non-linear relationship, but their correlation (linear relationship) would be almost zero.



Covariance and Correlation

- **Correlation** can also be viewed as the *covariance between the standardized versions* of the variables.
- Recall that we standardize variables by subtracting the mean of the variable from each observation, and then dividing by the standard deviation of the variable.
- i.e., $z\_x = (X - \overline{X})/\sigma_X$ for each variable. Note: σ_X has to be the <u>population standard deviation formula</u>, not the sample formula.
- The standardized variables are unitless, as are their covariances.
- Then, the *covariance* among these *standardized variables* is the *same as the correlation* among the *original variables.*

```
> corr_mat <- cor(M, method = "pearson")
> print(signif(corr_mat),digits = 4)
              age  income networth    sales
age       1.00000 -0.1996   0.4530 -0.03772
income   -0.19959  1.0000   0.3000 -0.38429
networth  0.45300  0.3000   1.0000 -0.13608
sales    -0.03772 -0.3843  -0.1361  1.00000
> #
> #
> z_age <- (age - mean(age))/sd(age)
> z_income <- (income - mean(income))/sd(income)
> z_networth <- (networth - mean(networth))/sd(networth)
> z_sales <- (sales - mean(sales))/sd(sales)
> #
> z_M <-  cbind(z_age, z_income, z_networth, z_sales)
> #
> covar_z_mat <- cov(z_M, use="all.obs", method ="pearson")
> print(signif(covar_z_mat),digits = 4)
              z_age z_income z_networth  z_sales
z_age       1.00000  -0.1996     0.4530 -0.03772
z_income   -0.19959   1.0000     0.3000 -0.38429
z_networth  0.45300   0.3000     1.0000 -0.13608
z_sales    -0.03772  -0.3843    -0.1361  1.00000
> 
```

Understanding Correlation

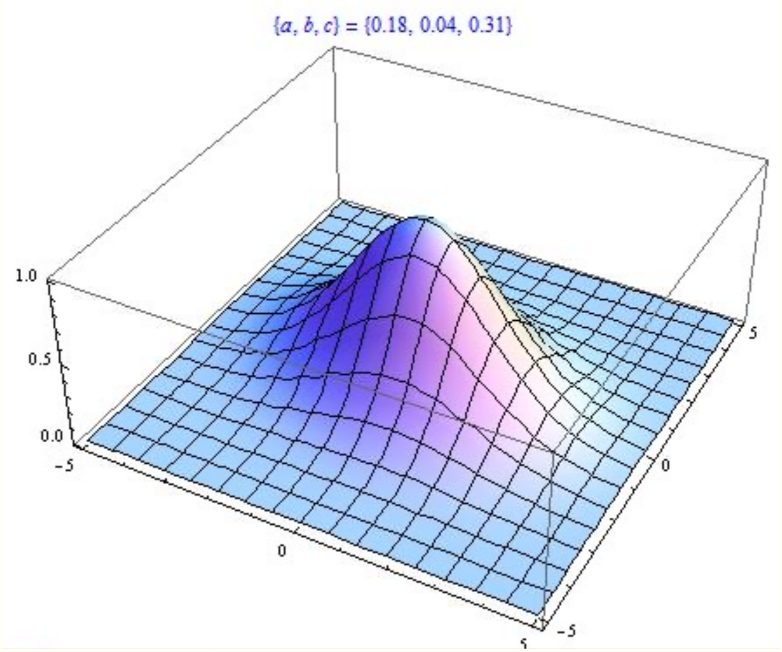
- One of the reasons correlation (or more correctly Pearson's Product Moment Correlation) is popular is because of the widespread use of the normal distribution
  - If X and Y have a joint (bivariate) normal distribution, then only a linear relationship between them is possible, and the strength of this relationship is correlation
  - If either or both are not normal, then their joint distribution is not-normal, and nonlinear relationships are possible. In fact, it is also possible that the linear relationship is weak and the nonlinear relationship is strong

- Pearson's product moment correlation is sometimes termed as *zero order correlation* because it does not account for the effect of other variables that are correlated to both X and Y

- It is possible that X and Y are correlated only because they both are related to a common variable Z. After *controlling* for the correlation of each with Z the correlation between X and Y may actually disappear.

- In other words, pairwise correlations are subject to misinterpretations.

- We will learn more about this in the next lecture.



	age	income	networth	sales
age	1.00000	-0.1996	0.4530	-0.03772
income	-0.19959	1.0000	0.3000	-0.38429
networth	0.45300	0.3000	1.0000	-0.13608
sales	-0.03772	-0.3843	-0.1361	1.00000

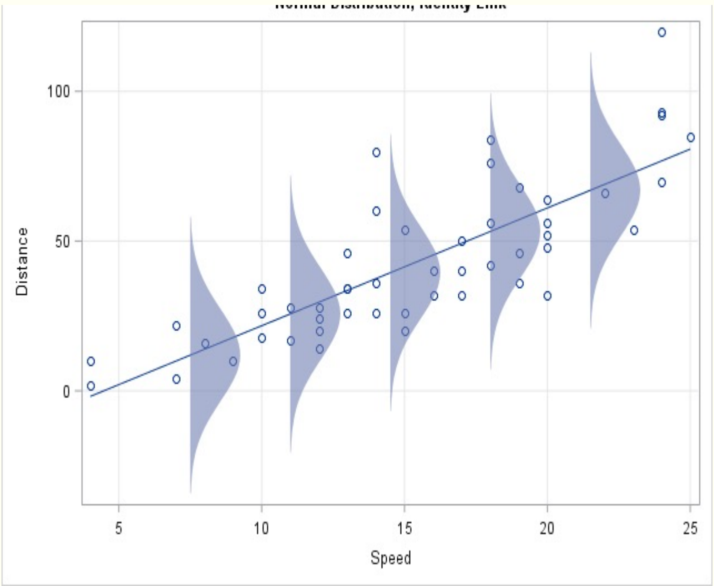
LECTURE 4A – 2 –CORRELATION AND SIMPLE REGRESSION

Simple Regression

- Simple Regression involves modeling the relationship between a *dependent* variable Y and ***one*** *independent* variable X.
- **Population model**:
- $Y = \boldsymbol{\alpha} + \boldsymbol{\beta} X + \boldsymbol{\epsilon}$, where
  - $\boldsymbol{\alpha}$ is called the *intercept parameter*,
  - $\boldsymbol{\beta}$ is called the *slope parameter* and
  - $\boldsymbol{\epsilon}$ is a called the *error* (or *noise* or *disturbance*) random variable independent of X, and assumed to be distributed as Normal($\mathbf{0}$, $\boldsymbol{\sigma}_{\varepsilon}$).
    - Here, $\boldsymbol{\sigma}_{\varepsilon}$ is also a parameter representing the **standard deviation** of the error term.

Visualizing the Regression Equation

- The regression model for the population is:
  - $Y = \boldsymbol{\alpha} + \boldsymbol{\beta}X + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, \boldsymbol{\sigma}_{\varepsilon})$.
- It says that for a particular value of $X = x$,
  - the value of Y = value of a point along a line ($\boldsymbol{\alpha} + \boldsymbol{\beta}x$)
  - +
  - a value drawn from a Normal distribution with mean = 0 and standard deviation $\boldsymbol{\sigma}_{\varepsilon}$.
- This is the same as
  - the value of Y = value of a point from a Normal distribution with mean = ($\boldsymbol{\alpha} + \boldsymbol{\beta}x$) and standard deviation $\boldsymbol{\sigma}_{\varepsilon}$. i.e., from $\text{Normal}(\boldsymbol{\alpha} + \boldsymbol{\beta}x, \boldsymbol{\sigma}_{\varepsilon})$.
- We can now see, that for a single value of $X=x$,
  - We will get many values for Y, each value drawn from $\text{Normal}(\boldsymbol{\alpha} + \boldsymbol{\beta}x, \boldsymbol{\sigma}_{\varepsilon})$.
- In other words, $Y \sim \text{Normal}(\boldsymbol{\alpha} + \boldsymbol{\beta}x, \boldsymbol{\sigma}_{\varepsilon})$ because $\boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, \boldsymbol{\sigma}_{\varepsilon})$.
- Notice that the mean of Y ($= \boldsymbol{\alpha} + \boldsymbol{\beta}x$) depends on the value of X,

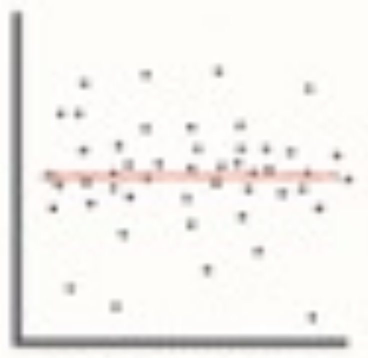


Visualizing the Regression Equation

- The *relationship* between X and Y is captured by the line Y = $\boldsymbol{\alpha} + \boldsymbol{\beta}X$.
- It is a *linear relationship*, because $\boldsymbol{\alpha} + \boldsymbol{\beta}X$ defines a line.
- As we saw, $\boldsymbol{\alpha} + \boldsymbol{\beta}X$ is the mean (or expected value) of Y, for each value of X.
  - If $\boldsymbol{\beta} > 0$, then the slope of the line is positive, and the mean of Y increases as X increases
  - If $\boldsymbol{\beta} < 0$, then the slope of the line is negative, and the mean of Y decreases as X increases
  - If $\boldsymbol{\beta} = 0$, then the slope of the line is positive, and the mean of Y stays the same as X increases (no relationship)

$\boldsymbol{\beta} > 0$

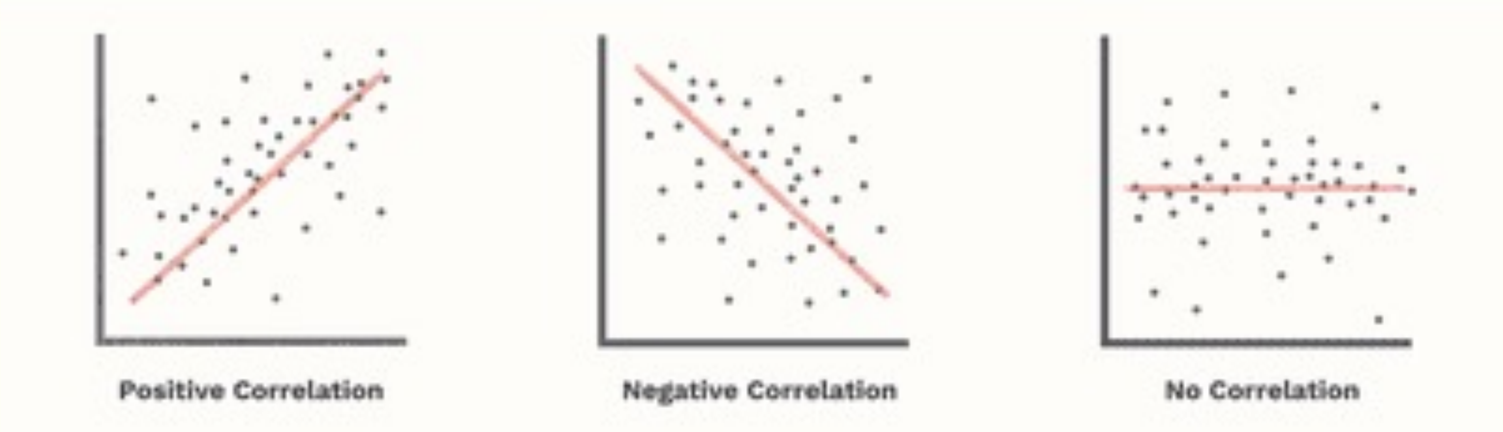
$\boldsymbol{\beta} < 0$



$\boldsymbol{\beta} = 0$

Correlation and Regression

- Correlation quantifies the strength of the linear relationship between a pair of variables, whereas regression expresses the relationship in the form of an equation.
- Clearly, they must be related.
- It turns out that **β** can be expressed as:
  - (correlation between X and Y)*Standard deviation of Y/Standard Deviation of X.
- Having an equation (instead of correlation) enables us to predict different values of Y for different values of X.



$\beta > 0$ $\beta < 0$ $\beta = 0$

Simple Regression on Sample Data

- In practice, we obtain sample data and fit a regression model to the sample, keeping in mind the underlying population model.

- **Population model**: Y = $\boldsymbol{\alpha}$ + $\boldsymbol{\beta}$X + $\boldsymbol{\epsilon}$, where $\boldsymbol{\alpha}$ is called the *intercept parameter*, $\boldsymbol{\beta}$ is called the *slope parameter* and $\boldsymbol{\epsilon}$ is a called the *error* (or *noise* or *disturbance*) random variable independent of X, and assumed to be distributed as Normal(**0**, $\boldsymbol{\sigma_\varepsilon}$). Here, $\boldsymbol{\sigma_\varepsilon}$ is also a parameter representing the **standard deviation** of the error term.

- **Sample model**: $\hat{y} = \hat{\alpha} + \hat{\beta}X$ where $\hat{\alpha}$ and $\hat{\beta}$ are the *estimators* of the corresponding population parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. Additionally, $\boldsymbol{\sigma_\varepsilon}^2$ will be estimated by the variance s_ε^2 of $\boldsymbol{\epsilon}$ the *residuals* of the sample regression model.

Parameter	Unbiased Estimator	Sampling Distribution	Standard Error	Population Distribution Assumption
μ	$\overline{X}$	$\overline{X}$ ~ Normal or t	$\frac{\sigma}{\sqrt{n}}$ or $\frac{s}{\sqrt{n}}$	Normal or General
σ²	$(n-1)s^2$	$\frac{(n-1)s^2}{\sigma^2}$ ~ Chi-square (n-1)	$\frac{\sqrt{2}s^2}{(n-1)}$	Normal
α	$\hat{\alpha} = \overline{Y} - \hat{\beta}{*}\overline{X}$	Not interesting	Not interesting	Not interesting
β	$\hat{\beta} = s_{XY}/s_X{}^2 = r_{XY}{}^*(s_Y/s_X)$	$\hat{\beta}$ ~ Normal or t	$\frac{S_\varepsilon}{\sqrt{n-1}(s_X)}$	Normal if $\boldsymbol{\epsilon}$ is Normal

Simple Regression – Example – **SimpleReg.R**

- Let us consider our data set as a *sample*. We will use simple regression to fit a linear model between Net Worth (Y, dependent variable) and Age (X, independent variable).

- To perform the regression in **R**, use the **lm(y~x)** Linear Models function.

- We will fit a model to predict Net Worth (Y) using Age (X).

- $\hat{y} = \hat{\alpha} + \hat{\beta}X$

- Net Worth = 81,106 + 2813.2*Age is the Estimation Equation or Prediction Equation.

```
> # Clear the Environment
> rm(list=ls())
>
> # Read csv file as a DataFrame
> #
> setwd("C:\\Users\\sarathy\\Documents\\2019-Teaching\\Fall2019\\Fall2019-MSIS5503\\MSIS-5503-Data")
> df <- read.table('ClassData.csv',
+                  header = TRUE, sep = ',')
>
> #Assign variable names to DataFrame Column objects
> id <- df$ID
> name <- df$Name
> age <- df$Age
> gender <-df$Gender
> education <- df$Education
> crediscore <-df$CreditScore
> income <- df$Income
> networth <-df$NetWorth
> sales <-df$Sales
> #
> # Simple Regression of NetWorth on Age using the Linear Models function lm(y~x)
> #
> df1 <- data.frame(networth, age)
> reg_NetWorth_Age <- lm(networth ~ age, data = df1)
> print(reg_NetWorth_Age)

Call:
lm(formula = networth ~ age, data = df1)

Coefficients:
(Intercept)          age
      81106         2813
```

- **Interpreting the Equation:**
  - The equation says that for 1 year increase in Age, the estimated increase in Net Worth is: $2813..

Simple Regression - Example

- The Expected or Predicted Net Worth at Age 40 is given by:
  - 81,106 + 2813*Age = 81,106 + 40*(28
  - $193,626 (approx.)

```
> # Clear the Environment
> rm(list=ls())
>
> # Read csv file as a DataFrame
> #
> setwd("C:\\Users\\sarathy\\Documents\\2019-Teaching\\Fall2019\\Fall2019-MSIS5503\\MSIS-5503-Data")
> df <- read.table('ClassData.csv',
+                  header = TRUE, sep = ',')
>
> #Assign variable names to DataFrame Column objects
> id <- df$ID
> name <- df$Name
> age <- df$Age
> gender <-df$Gender
> education <- df$Education
> crediscore <-df$CreditScore
> income <- df$Income
> networth <-df$NetWorth
> sales <-df$Sales
> #
> # Simple Regression of NetWorth on Age using the Linear Models function lm(y~x)
> #
> df1 <- data.frame(networth, age)
> reg_NetWorth_Age <- lm(networth ~ age, data = df1)
> print(reg_NetWorth_Age)

Call:
lm(formula = networth ~ age, data = df1)

Coefficients:
(Intercept)          age
      81106         2813

> #
> age_40 <- data.frame(age=40)
> age_40_networth <- predict(reg_NetWorth_Age, age_40)
> print(paste("Predicted Net Worth at Age 40 = ", round(age_40_networth,2)," dollars"))
[1] "Predicted Net Worth at Age 40 =  193635.23  dollars"
> |
```

Simple Regression – Estimating the Population Parameters

- As we saw earlier, $\hat{\alpha}$ and $\hat{\beta}$ are the *estimators* of the corresponding population parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively.

- The slope parameter $\hat{\beta}$ is calculated first as:
  - $\hat{\beta} = r_{XY}*(s_Y/s_X)$
  - i.e., sample correlation * sample standard deviation of Y divided by sample standard deviation of X.
  - $r_{XY} = 0.453$, $s_Y = 93506.4$ and $s_X = 15$
  - $\hat{\beta} = 0.453*93506.4/15.06 =$ 2813.
  - $\hat{\alpha} = \bar{Y} - \hat{\beta}*\bar{X}$
  - $\hat{\alpha}$ = 191947.3 - 2813.22*39.4=81106.38 (some round off error)

```
> samp_means <- c(mean(df1$networth), mean(df1$age))
> print(samp_means)
[1] 191947.3     39.4
> samp_sd <- c(sd(df1$networth), sd(df1$age))
> print(round(samp_sd,3))
[1] 93506.407     15.057
> samp_cor <- cor(networth, age)
> print(round(samp_cor,3))
[1] 0.453
> #
> # Beta slope coefficient = corr*sd(y)/sd(x)
> #
> beta_hat <- samp_cor*sd(df1$networth)/sd(df1$age)
> print(round(beta_hat),3)
[1] 2813
> alpha_hat <- mean(df1$networth) - beta_hat*mean(df1$age)
> print(round(alpha_hat,3))
[1] 81106.38
```

- **Important Notes:**
  - $\hat{\beta}$ is our point estimate of the slope of the population model $\boldsymbol{\beta}$.
  - Notice that $\hat{\beta}$ is a *statistic* because it is calculated completely from sample data. Therefore it has a *sampling distribution* and a *standard* error. For our example
  - We can conduct **hypothesis tests** and compute **confidence intervals** for the population slope parameter $\boldsymbol{\beta}$.
  - The intercept estimate $\hat{\alpha} = \bar{Y} - \hat{\beta}*\bar{X}$ is also a statistic that estimates the population intercept $\boldsymbol{\alpha}$.
  - In practice we are usually only interested in this *point estimate* of $\boldsymbol{\alpha}$.

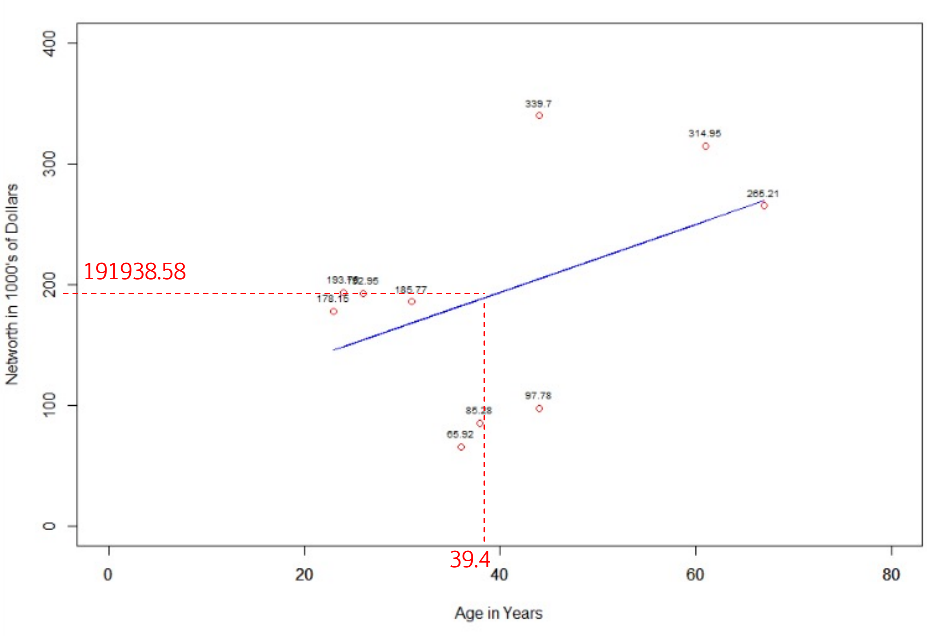
```
> print(reg_NetWorth_Age)

Call:
lm(formula = networth ~ age, data = df1)

Coefficients:
(Intercept)          age
      81106         2813
```

Simple Regression – Understanding the Predicted Equation

- Our equation for the predicted values is:
  - Predicted Net Worth = 81,106 + 2813*Age
  - The intercept 81,106 is obtained by setting Age = 0 in the equation.
- This is a line because the equation of a line is: $y = c + mx$, where c is the intercept and m is the slope
- The slope m gives the change in value of y for a unit change in x, regardless of the value of x.
  - Example:
    - Predicted Net Worth at Age = 10 = 81106 + 2813*10
    - Predicted Net Worth at Age = 11 = 81106 + 2813*11
    - The difference (predicted Net Worth for Age = 11 - predicted Net Worth for Age = 10) = 2813
    - Similarly, The difference (predicted Net Worth for Age = 6 - predicted Net Worth for Age = 5) = 2813
- The predicted line will always pass through the mean of X and the Mean of Y
  - Predicted Net Worth at Age = 39 (mean of X) = 81106.38 + 2813*39.4= 191938.58 (Mean of Y with some round off error)

```
> samp_means <- c(mean(df1$networth), mean(df1$age))
> print(samp_means)
[1] 191947.3     39.4
```

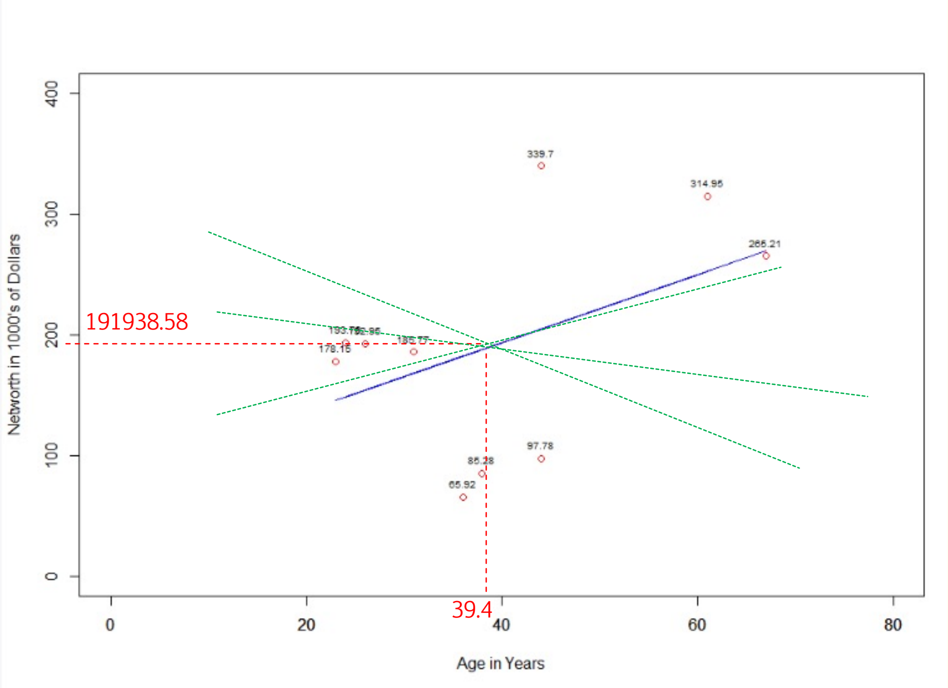
```
# Plot of Networth vs Age
#
plot(df1$age,df1$networth/1000,col="red",
     xlab="Age in Years",
     ylab="Networth in 1000's of Dollars",
     xlim = c(0, 80), ylim= c(0, 400))
text(df1$age, df1$networth/1000, round(df1$networth/1000, 2), cex=0.6, pos=3)
#
# Add Predicted Values to the Plot
#
par(new=TRUE)
plot(df1$age,df1$pred_networth/1000,type="l",
     yaxt='n', ann=FALSE, col="blue",  xlim = c(0, 80), ylim= c(0, 400))
```



LECTURE 4A – 3 – LEAST SQUARES ESTIMATION

Simple Regression – Understanding Least Squares Estimation

- We saw that, deciding a prediction line is the same as deciding what intercept and slope to use for prediction. i.e., what $\hat{\alpha}$ and $\hat{\beta}$ to use.
- What is the best way to decide which $\hat{\alpha}$ and $\hat{\beta}$ to use?
- Even if we decide that the line should pass through the means $(\overline{Y}, \overline{X})$, there are an infinite number of lines (i.e., infinite choices of $\hat{\alpha}$ and $\hat{\beta}$)
- To understand how we chose the particular:
  - $\hat{\beta} = r_{XY} * (s_Y / s_X)$
  - $\hat{\alpha} = \overline{Y} - \hat{\beta} * \overline{X}$
- we have to understand the concept of *residuals*.



Simple Regression – Least (Sum of) Squares (of Residuals) Principle

- To understand the Least Squares Principle we must first understand the following the concept of Residuals:
  - Residual = (Actual Net Worth – Predicted Net Worth) for each value of X.

- The red dotted lines, show the difference between the actual value of Y (Actual Net Worth) and the Predicted Value of Y (Predicted Net Worth $\hat{y}$) given by $=\hat{\alpha}+\hat{\beta}X$, where $\hat{\alpha}$ and $\hat{\beta}$ define the line. They are the residuals.

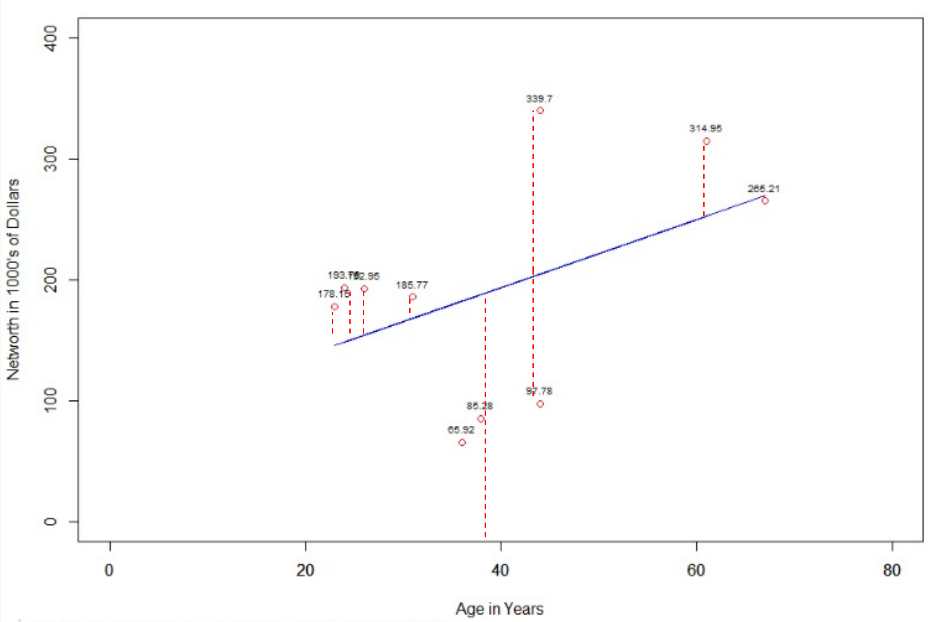
- Each line (i.e., each choice of $\hat{\alpha}$ and $\hat{\beta}$) defines a different set of residuals.

- We want that line (choice of $\hat{\alpha}$ and $\hat{\beta}$) which gives us the least "total amount" of residuals. The "amount" is obtained by squaring each residual and adding them.
  - The reason that we square the residuals before we add them up is to prevent canceling out of positive and negative values of residuals (see table) giving a false picture of the fit of the line.

- The **R** printout shows sum of squared residuals. While this may seem large, it is still the smallest value for any other line you can get from any other line.

- The line which gives us the least sum of squared residuals is called the **Least Squares Line**.

- The $\hat{\alpha}$ and $\hat{\beta}$ corresponding to this line are called **Ordinary Least Squares (OLS) or Least Squares Estimate (LSE)**. It can be shown that these are given by:




```
> # Obtaining Residuals Manually
> #
> df1$resid_networth = df1$networth - df1$pred_networth
> print(df1)
   networth age pred_networth resid_networth
1     65924  36      182382.3    -116458.347
2    178154  23      145810.5      32343.530
3    265209  67      269592.2      -4383.209
4     85277  38      188008.8    -102731.790
5    193760  24      148623.7      45136.308
6    314953  61      252712.9      62240.119
7    192946  26      154250.1      38695.866
8    339705  44      204888.1     134816.882
9    185767  31      168316.2      17450.759
10    97778  44      204888.1    -107110.118
> # Sum of Squared Residuals
> sum_sq_resid <- sum(df1$resid^2)
> print (sum_sq_resid)
[1] 62542870703
```

Simple Regression – The Sampling Distribution of $\hat{\beta}$

- We saw that **LSE** $\hat{\beta}$ is a statistic that estimates the slope $\boldsymbol{\beta}$ of the population model.

- Therefore, it has a sampling distribution and a standard error that can be used for confidence intervals and hypothesis testing.

- What is the Sampling Distribution of $\hat{\beta}$? Under the assumption that the population noise term $\boldsymbol{\epsilon \sim N(0, \sigma_\varepsilon)}$ (and consequently the sample residuals ϵ are $\mathbf{N(0, \sigma_\varepsilon)}$) or t with (n-1) degrees of freedom, it turns out that $\hat{\beta} \sim \mathbf{N(\beta, \sigma_\varepsilon / s_X)}$

- In practice, since we do not know $\boldsymbol{\sigma_\varepsilon^2}$ we estimate it from the sample as **Mean Square Error** = s_ε^2 ={**SSE**/ (n – k-1)}, where n is the sample size and (n- k - 1) is called the **error degrees of freedom** for the regression model, and k is the number of independent variables in the model (in our case, k = 1, so error degrees of freedom = 8).

- The **standard error** of $\hat{\beta}$, the standard deviation of the sampling distribution = $s_{\hat{\beta}} = \dfrac{s_\varepsilon}{\sqrt{n-1}(s_X)}$

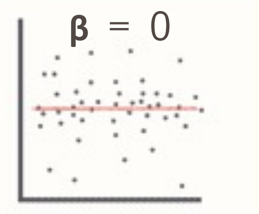
Parameter	Unbiased Estimator	Sampling Distribution	Standard Error	Population Distribution Assumption
α	$\hat{\alpha} = \overline{Y} - \hat{\beta} * \overline{X}$	Not interesting	Not interesting	Not interesting
β	$\hat{\beta} = s_{XY}/s_X^2 = r_{XY} * (s_Y/s_X)$	$\hat{\beta}$ ~ Normal or t	$\dfrac{s_\varepsilon}{\sqrt{n-1}(s_X)}$	Normal if $\boldsymbol{\epsilon}$ is Normal

Simple Regression – The Sampling Distribution of $\hat{\beta}$

- How do we understand the sampling distribution of $\hat{\beta}$?
- It is very much like understanding the sampling distribution of $\overline{X}$.
- If we keep taking samples of the same size and fit a regression line to each sample, we will get many $\hat{\beta}$s.
- The distribution of these $\hat{\beta}$s is the sampling distribution.
- The standard deviation of these $\hat{\beta}$s will be the standard error of $\hat{\beta}$ ($s_{\hat{\beta}}$) and is given by $\frac{s_{\varepsilon}}{\sqrt{n-1}(sX)}$, where $s_{\varepsilon}$is the RMSE (Root Mean Squared Residuals – will be defined later) and s_X is the sample standard deviation of X.
- The smaller the standard error, the closer the $\hat{\beta}$s of all samples with each other and with the true population slope, **β**.

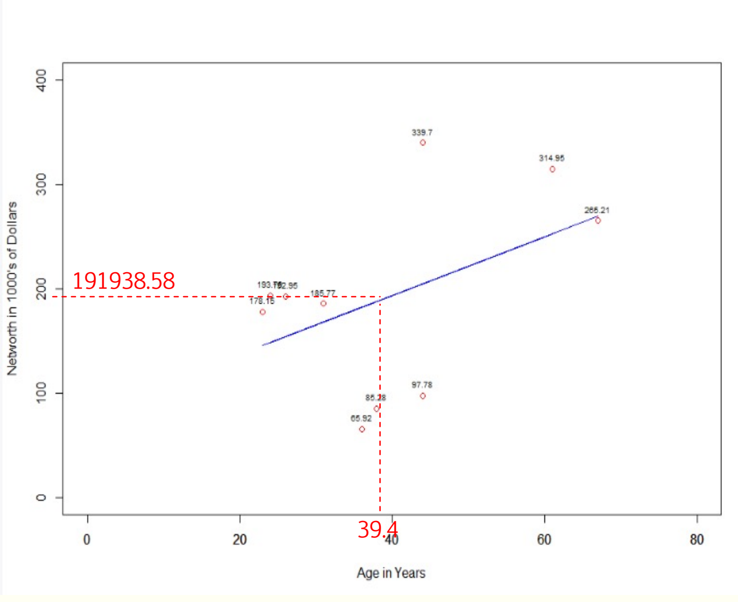
Simple Regression – The Hypothesis Test for the Slope β

- Once we have the Sampling Distribution of $\hat{\beta}$, we can do hypothesis tests and confidence intervals.



- For the Hypothesis Test, the Null Hypothesis is:
  - There is <u>no relationship</u> between X and Y
  - That is, H_0: $\boldsymbol{\beta} = 0$
- We want the sample to provide evidence of relationship
  - That is, H_a: $\boldsymbol{\beta} \neq 0$ (always two-tailed)
- Hence, the hypothesis test for the slope **β** is as follows:
  - H_0: $\boldsymbol{\beta} = 0$; H_a: $\boldsymbol{\beta} \neq 0$;
- The *test statistic*: $t = \hat{\beta}/s_{\hat{\beta}}$ (estimate/standard error) has (n-2) degrees of freedom.
- We saw that $\hat{\beta} = r_{XY} * (s_Y/s_X)$ and $s_{\hat{\beta}} = \dfrac{s_\varepsilon}{\sqrt{n-1}(s_X)}$
  - s_ε= Root Mean Square Error (RMSE) or the Standard Error of the Residuals
  - n = sample size
- In **R**, use the command **summary**(*model_name*) to obtain the results of the hypothesis test.

```
> #
> # Regression Model Summary
> #
> summary(reg_NetWorth_Age)

Call:
lm(formula = networth ~ age, data = df1)

Residuals:
    Min      1Q  Median      3Q     Max
-116458  -78145   24897   43526  134817

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    81106      82035   0.989    0.352
age             2813       1957   1.437    0.189

Residual standard error: 88420 on 8 degrees of freedom
Multiple R-squared:  0.2052,    Adjusted R-squared:  0.1059
F-statistic: 2.066 on 1 and 8 DF,  p-value: 0.1886
```



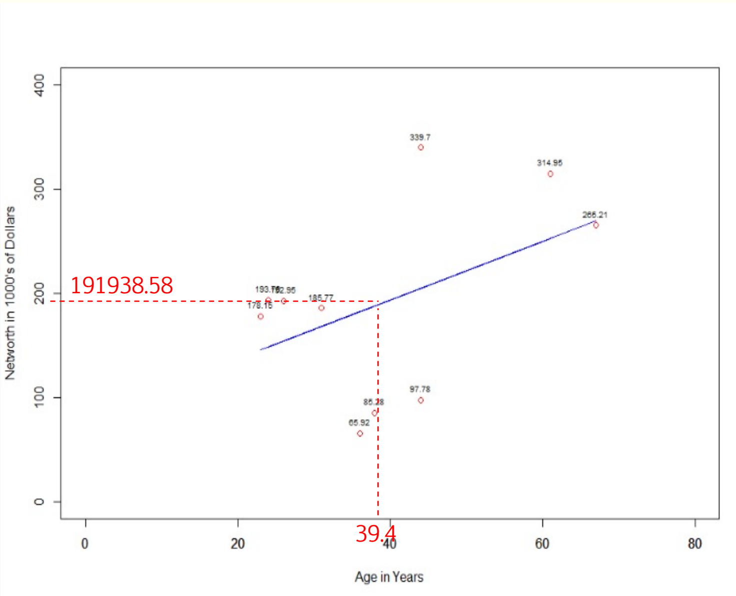
Simple Regression – The Hypothesis Test for the Slope β

- We saw that $\hat{\beta} = r_{XY} * (s_Y / s_X)$ and $s_{\hat{\beta}} = \dfrac{s_\varepsilon}{\sqrt{n-1}(s_X)} = \dfrac{88240}{\sqrt{9}(15.057)} = 1957$
  - s_ε = Root Mean Square Error (RMSE) or the Standard Error of the Residuals
  - n = sample size
- In **R**, use the command **summary**(*model_name*) to obtain the results of the hypothesis test.
- The p-value for $t = \hat{\beta}/s_{\hat{\beta}} = 1.437$ with 8 degrees of freedom = 0.1886

```
> print(paste("The p-value for the two-tailed t-test is ", 2*(1-pt(1.437, 8))))
[1] "The p-value for the two-tailed t-test is  0.188652236215373"
```

- We conclude that H_0: **β** = 0 **cannot be rejected** based on the sample.
- Our results show that we do NOT reject the null hypothesis of <u>no relationship</u> between Age and Net Worth, since p-value 0.189 is > α = 0.05.
- That is, there is <u>**no significant linear relationship**</u> between Age and Net Worth at α = 0.05, since the p-value 0.189 is > α = 0.05.

```
> #
> # Regression Model Summary
> #
> summary(reg_NetWorth_Age)

Call:
lm(formula = networth ~ age, data = df1)

Residuals:
    Min      1Q  Median      3Q     Max 
-116458  -78145   24897   43526  134817 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    81106      82035   0.989    0.352
age             2813       1957   1.437    0.189

Residual standard error: 88420 on 8 degrees of freedom
Multiple R-squared:  0.2052,	Adjusted R-squared:  0.1059 
F-statistic: 2.066 on 1 and 8 DF,  p-value: 0.1886
```



Simple Regression – Predictions

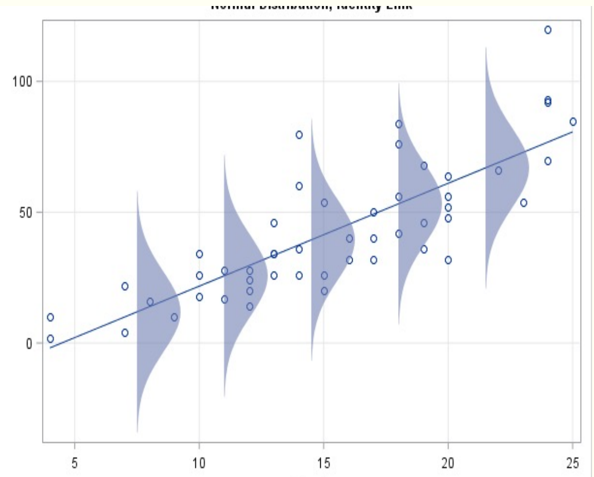
- Our Prediction Equation from the Prediction Line is:
  - $\hat{y} = \hat{\alpha} + \hat{\beta}X$
  - **Predicted Net Worth = 81,106 + 2813.2*Age**

- This is also the Expected Value of the Mean of Y, for different values of X. That is:
  - **Expected Net Worth = 81,106 + 2813.2*Age**

- We note that the prediction line is therefore the mean (or expected) value of Y for different values of X.
  - The Expected value of Net Worth for Age = 61 is $252,712.88 and at Age = 67 is $269,952.2.
  - The expected value of $\hat{y}$ increases along a line with Age.

- Individual values of Y are spread around this (mean) line. That is, for a particular value of X, there are many actual values of Y.

```
> # Obtaining Residuals Manually
> #
> df1$resid_networth = df1$networth - df1$pred_networth
> print(df1)
   networth age pred_networth resid_networth
1     65924  36      182382.3    -116458.347
2    178154  23      145810.5      32343.530
3    265209  67      269592.2      -4383.209
4     85277  38      188008.8    -102731.790
5    193760  24      148623.7      45136.308
6    314953  61      252712.9      62240.119
7    192946  26      154250.1      38695.866
8    339705  44      204888.1     134816.882
9    185767  31      168316.2      17450.759
10    97778  44      204888.1    -107110.118
```



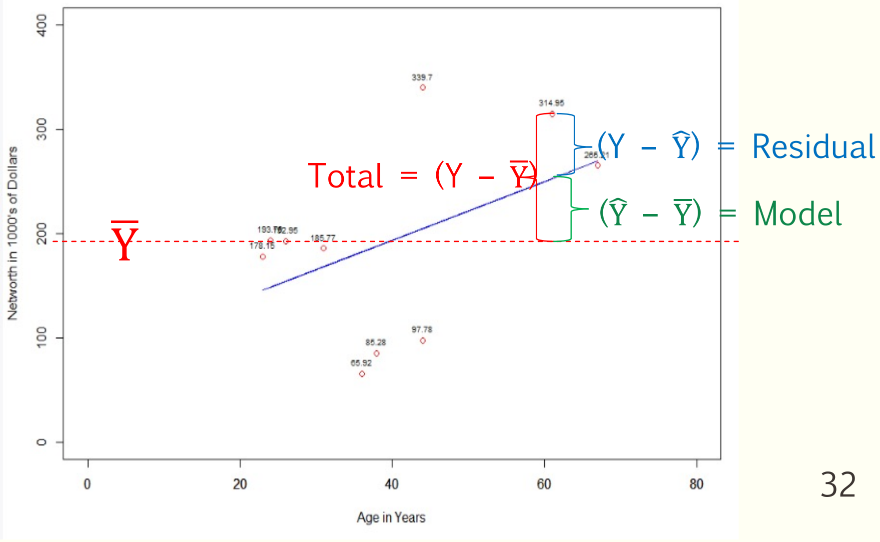
LECTURE 4A – 4 – R-SQUARE AND MODEL FIT

Regression as “Explaining Variability in Y”

- We can also look at regression as explaining the variability in Y using X. We can split the variability in Y into:
  - Variability that can be assigned to the X variables
  - Variability that is *unexplained* (left to error).
- We can show that:
  - Sum $(Y - \bar{Y})^2$ = Sum $(\hat{Y} - \bar{Y})^2$ + Sum $(Y - \hat{Y})^2$ where
  - $(Y - \bar{Y})$ = Deviation of Y from its mean
  - $(\hat{Y} - \bar{Y})$ = Deviation of Predictions from the mean of Y
  - $(Y - \hat{Y})$ = Prediction error (or residual e)
- This very important expression shows that:
  - Variability in Y (or Variance of Y) or *Total Sum of Squares = Total SS*
  - the variability of the predicted values of Y from the mean of Y (called sum of squares due to regression model or *Model Sum of Squares) = Model SS*
  - + the variability of the residuals)or *Residual Sum of Squares = Residuals SS.*

```
> #
> df1$sq_YfromMean <- (df1$networth - mean(df1$networth))^2
> df1$sqpredFromMean <- (df1$pred_networth - mean(df1$networth))^2
> df1$Sq_resid <- (df1$resid)^2
> print(df1)
   networth age pred_networth resid_networth sq_YfromMean sqpredFromMean    Sq_resid
1     65924  36      182382.3    -116458.347 1.588187e+10       91488317 13562546698
2    178154  23      145810.5      32343.530 1.902551e+08     2128607064  1046103920
3    265209  67      269592.2      -4383.209 5.367277e+09     6028731845    19212518
4     85277  38      188008.8    -102731.790 1.137855e+10       15511860 10553820705
5    193760  24      148623.7      45136.308 3.285881e+06     1876935051  2037286342
6    314953  61      252712.9      62240.119 1.513040e+10     3692455799  3873832449
7    192946  26      154250.1      38695.866 9.974017e+05     1421076310  1497370031
8    339705  44      204888.1     134816.882 2.183234e+10      167464773 18175591641
9    185767  31      168316.2      17450.759 3.819611e+07      558426957   304528995
10    97778  44      204888.1    -107110.118 8.867857e+09      167464773 11472577402
> #
> #
> print(sum(df1$sqpredFromMean) + sum(df1$Sq_resid))
[1] 78691033452
> print(sum(df1$sq_YfromMean))
[1] 78691033452
> print(paste("R-Squared is ", sum(df1$sqpredFromMean)/sum(df1$sq_YfromMean)))
[1] "R-Squared is  0.20520969214341"
> 
```

Sum $(\hat{Y} - \bar{Y})^2$ + Sum $(Y - \hat{Y})^2$ = Sum$(Y - \bar{Y})^2$



Variability Analysis

- In **R**, the *anova(model_name)* function provides the variability analysis.

- We can see that:
  - Model SS = Sum $(\hat{Y} - \bar{Y})^2$ = 16148162749 with 1 degree of freedom (df = 1)
  - Residuals SS = Sum $(Y - \hat{Y})^2$ = 62542870703 with (n-2) or 8 degrees of freedom (df = 8)
  - So Total SS = Total variability in Y = 161448162749 + 62542870703 = 78619033452

- The greater the Model Sum of Squares (Model SS) relative to the Total Sum of Squares (Total SS), smaller the total prediction errors (Residual sum of squares) and better the model.

```
> # Variability Analysis - Using the anova() function
> #
> anova_networth_age <- anova(reg_NetWorth_Age)
> print(round(anova_networth_age),12)
Analysis of Variance Table

Response: networth
          Df      Sum Sq     Mean Sq F value    Pr(>F)    
age        1 16148162749 16148162749       2 < 2.22e-16 ***
Residuals  8 62542870703  7817858838                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression as "Explaining Variability in Y" and R^2

- Using the variability analysis, we can develop a normalized measure (between 0 and 1) of the *proportion of variability in Y explained by X called* R^2

- **R^2** = Variability in Y due to X's (Model Variability) / Total Variability in Y = Model Sum of Squares/Total Sum of Squares

- i.e., $R^2 = \frac{\text{Model SS}}{\text{Total SS}}$ and has a value between 0 and 1.
  - In our case, $R^2 = \frac{16148162749}{78619033452} = 0.2052$

- **Greater the R^2, better the model fit to the data**. Higher R^2 the better the "variability explained by the model" and hence higher the "explanatory power" of X with respect to Y.
  - In our case, Age explains 20.5% of the variability in Net Worth. All the other variables that explain the variability in Net Worth are in the error term and explain the remaining 79.5% of the variability in Y.

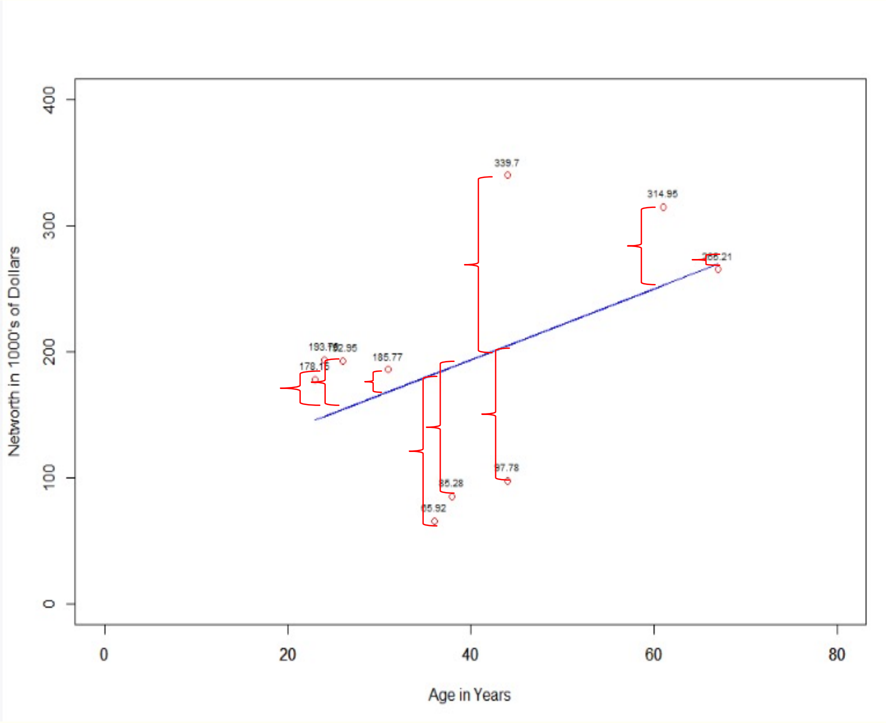
- The R^2 calculated from the sample estimates a corresponding population R^2

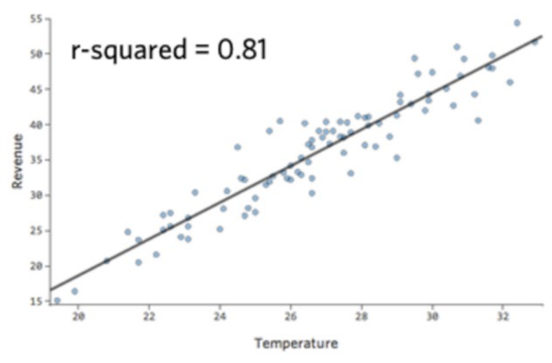
```
> # Variability Analysis - Using the anova() function
> #
> anova_networth_age <- anova(reg_NetWorth_Age)
> print(round(anova_networth_age),12)
Analysis of Variance Table

Response: networth
          Df      Sum Sq     Mean Sq F value    Pr(>F)
age        1 16148162749 16148162749       2 < 2.22e-16 ***
Residuals  8 62542870703  7817858838
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

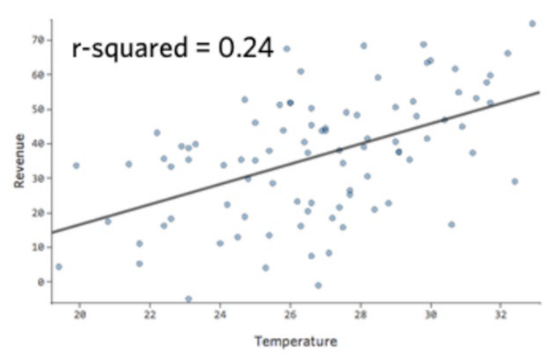
```
> #
> # Regression Model Summary
> #
> summary(reg_NetWorth_Age)

Call:
lm(formula = networth ~ age, data = df1)

Residuals:
    Min      1Q  Median      3Q     Max
-116458  -78145   24897   43526  134817

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    81106      82035   0.989    0.352
age             2813       1957   1.437    0.189

Residual standard error: 88420 on 8 degrees of freedom
Multiple R-squared:  0.2052,    Adjusted R-squared:  0.1059
F-statistic: 2.066 on 1 and 8 DF,  p-value: 0.1886
```

Understanding R-Squared Visually

- In our example, the R^2 tells us that Age explains only 20.5% of the variability in Net Worth.
- This means that 79.5% of the variability in Y is unexplained (i.e., is explained by error)
- We can see visually that the residuals on the fitted line are going to be large, resulting in high variability in error relative to the variability in Y.
- This results in lower R^2.







Simple Regression – Relationship Between Correlation and R^2

- In Simple Regression, the R^2 is simply the square of the correlation between the independent variable (X) and the dependent variable (Y).

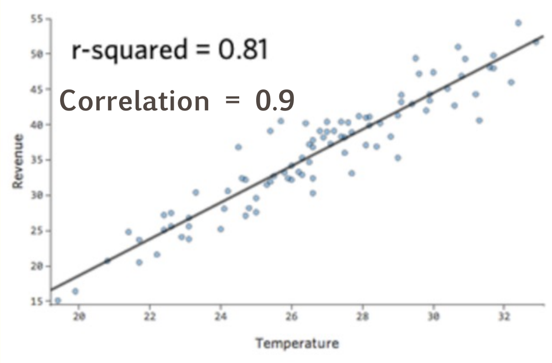
- In our example, the $R^2 = 0.2052$ = the square of the sample correlation between Age and Net Worth = $(0.453)^2$.

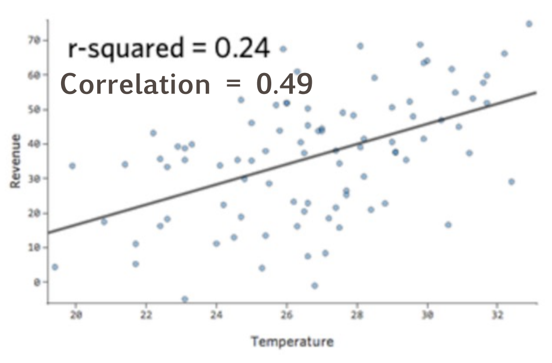
- As the correlation increases, the scatter of Y values around the prediction line will be smaller, leading to higher R^2.

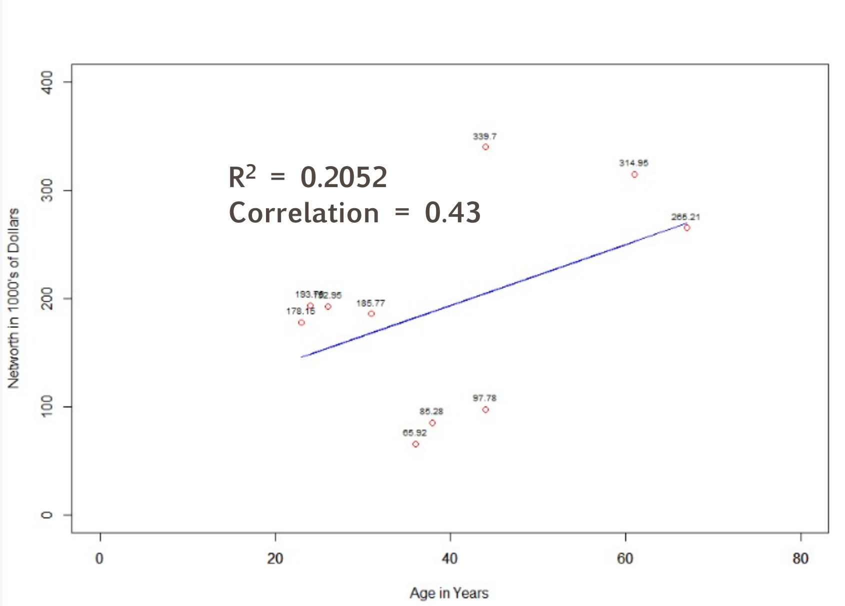
- To summarize R^2, is a measure of how well the model (prediction line) fits the data (scatter).

- Higher the R^2, the better the model fit to the data, leading to better predictions. The linear relationship between the dependent and independent variable is stronger.

```
> samp_means <- c(mean(df1$networth), mean(df1$age))
> print(samp_means)
[1] 191947.3     39.4
> samp_sd <- c(sd(df1$networth), sd(df1$age))
> print(round(samp_sd,3))
[1] 93506.407    15.057
> samp_cor <- cor(networth, age)
> print(round(samp_cor,3))
[1] 0.453
> #
> # Beta slope coefficient = corr*sd(y)/sd(x)
> #
> beta_hat <- samp_cor*sd(df1$networth)/sd(df1$age)
> print(round(beta_hat),3)
[1] 2813
> alpha_hat <- mean(df1$networth) - beta_hat*mean(df1$age)
> print(round(alpha_hat,3))
[1] 81106.38
```







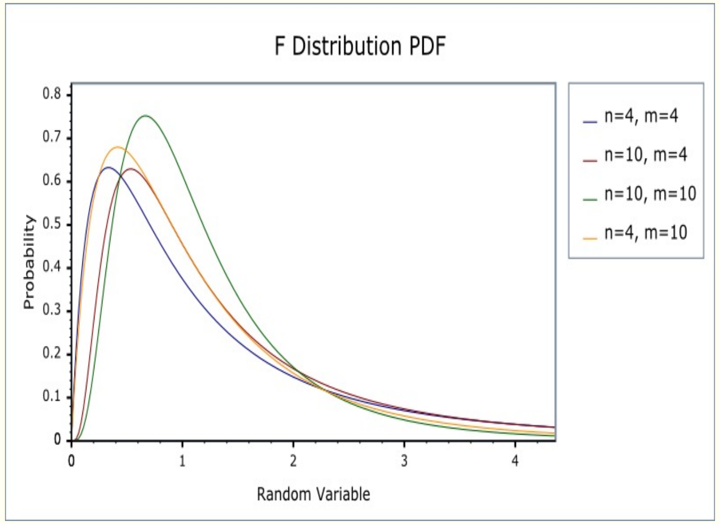
Statistical Test for Regression Model Fit – The F-distribution

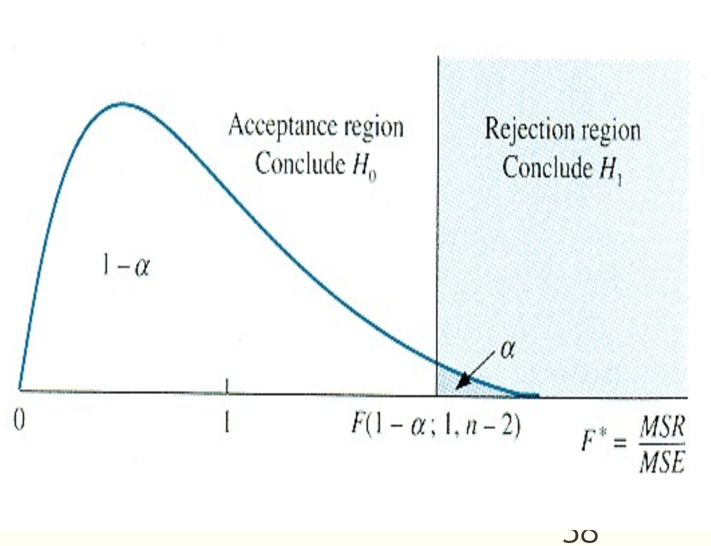
- There is also a statistical test, that will assess whether the fit of the model to data is *statistically significant*.
- To conduct this test, we will compare the variability in Y explained by the model, to that not explained by the model (i.e., explained by error).
- We can divide the Sum of Squares of Model and Error by their respective degrees of freedom to get the Mean Squares of Model and Error.
- In a simple regression model, the Model Degrees of Freedom is 1 (because there is only one independent variable).
- The Total Degrees of Freedom is (n-1) where n is the sample size, and the Error (or Residual) degree of freedom is n-2.
- So:
  - Total Degrees of Freedom (n-1) = (Model Degrees of Freedom (= 1) + Residual (or Error) Degrees of Freedom (n-2)
- Dividing each sum of Squares by its respective degrees of freedom gives us "Mean Squares"
  - MS_{Model} = Model SS/1 = Model SS (in Simple Regression) = 16148162749
  - MS_{Error} or MSE = Residual SS/(n-2) = 7817858838

```
> # Variability Analysis - Using the anova() function
> #
> anova_networth_age <- anova(reg_NetWorth_Age)
> print(round(anova_networth_age),12)
Analysis of Variance Table

Response: networth
          Df      Sum Sq     Mean Sq F value    Pr(>F)    
age        1 16148162749 16148162749       2 < 2.22e-16 ***
Residuals  8 62542870703  7817858838                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Statistical Test for Regression Model Fit – The F-distribution

- The ratio MS_{Model}/MS_{Error} is a statistic that follows an "F-distribution" whose p-value can be used in a hypothesis test.
- The F-distribution has parameters (numerator degrees of freedom, denominator degrees of freedom).
- The Null Hypothesis is H_0: $\boldsymbol{\beta} = 0$.
- This ratio is derived from the Null Hypothesis H_0: $\boldsymbol{\beta} = 0$ as follows:
  - F* (test-statistic) = Explained Variability in Y by Regression Model with X ("Full Model") / Explained Variability in Y by Error alone, without X ("Reduced Model")
  - = Variance of $(\hat{y} = \hat{\alpha} + \hat{\beta}X + \epsilon)$ / Variance of $(\hat{y} = \hat{\alpha} + \epsilon)$
  - = MS_{Model}/MS_{Error}
- Under Null Hypothesis H_0: $\boldsymbol{\beta} = 0$, this ratio should be 1, so *the greater the F test statistic is than 1, the more extreme its value*
- The alternative hypothesis will work out to be H_a: $\boldsymbol{\beta} \neq 0$





Statistical Test for Regression Model Fit – The F-distribution

- Even though alternative hypothesis is equivalent to H_a: $\boldsymbol{\beta} \neq 0$, in terms of the F-test we have a right-tailed test because even if $\boldsymbol{\beta} < 0$, the variances are always positive and we reject the null hypothesis when F is more extreme than 1.

```
> #
> # Variability Analysis - Using the anova() function
> #
> anova_networth_age <- anova(reg_Networth_Age)
> print(round(anova_networth_age),12)
Analysis of Variance Table

Response: networth
          Df       Sum Sq     Mean Sq F value     Pr(>F)
age        1 16148162749 16148162749       2 < 2.22e-16 ***
Residuals  8 62542870703  7817858838
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> print((anova_networth_age))
Analysis of Variance Table

Response: networth
          Df     Sum Sq    Mean Sq F value Pr(>F)
age        1 1.6148e+10 1.6148e+10  2.0655 0.1886
Residuals  8 6.2543e+10 7.8179e+09
```

- For our model the test-statistic
  - $F^* = MS_{model}/MS_{error} = 16148162749/7817858838 = 2.0655$
  - The p-value = Prob ($F^* > 2.0655$) =

```
> print(paste("P-value for F=2.055, numerator df =1 and denominator df = 8 is ", 1 - pf(2.0655, 1, 8)))
[1] "P-value for F=2.055, numerator df =1 and denominator df = 8 is  0.188601399168585"
```

  - The critical F-value = $F_{(0.95,\ 1,\ 8)}$ = F.INV(0.95,1,8) = 5.318

```
> print(paste("Critical value for alpha = 0.05, numerator df =1 and denominator df = 8 is ", qf(0.95, 1, 8)))
[1] "Critical value for alpha = 0.05, numerator df =1 and denominator df = 8 is  5.31765507157871"
```

- We **fail to reject** the null hypothesis H_0: $\boldsymbol{\beta} = 0$ and conclude that the “Full Regression Model” does not significantly explain more than what the error term does.

- **Note**: In simple linear regression $F^* = (t^*)^2$, the test-statistic for the t-test i.e., $2.0655 = 1.438^2$ and the p-vaues will be the same for both tests

```
> #
> # Regression Model Summary
> #
> summary(reg_NetWorth_Age)

Call:
lm(formula = networth ~ age, data = df1)

Residuals:
    Min      1Q  Median      3Q     Max
-116458  -78145   24897   43526  134817

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    81106      82035   0.989    0.352
age             2813       1957   1.437    0.189

Residual standard error: 88420 on 8 degrees of freedom
Multiple R-squared:  0.2052,    Adjusted R-squared:  0.1059
F-statistic: 2.066 on 1 and 8 DF,  p-value: 0.1886
```

Simple Regression – Relationship between Test for Beta and the Test for Model Fit.

- In simple regression, the direct test for H_0: $\boldsymbol{\beta} = 0$ vs H_a: $\boldsymbol{\beta} \neq 0$ uses the t-statistic with (n-2) degrees of freedom:
  - $t = \hat{\beta}/s_{\hat{\beta}}$
  - In our example the t-value was 1.437 with 8 degrees of freedom.
  - The p-value is: 0.1886

```
> print(paste("The p-value for the two-tailed t-test is ", 2*(1-pt(1.437, 8))))
[1] "The p-value for the two-tailed t-test is  0.188652236215373"
```

```
> #
> # Regression Model Summary
> #
> summary(reg_NetWorth_Age)

Call:
lm(formula = networth ~ age, data = df1)

Residuals:
    Min      1Q  Median      3Q     Max 
-116458  -78145   24897   43526  134817 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    81106      82035   0.989    0.352
age             2813       1957   1.437    0.189

Residual standard error: 88420 on 8 degrees of freedom
Multiple R-squared:  0.2052,	Adjusted R-squared:  0.1059 
F-statistic: 2.066 on 1 and 8 DF,  p-value: 0.1886
```

- The F-test for Model fit also tests H_0: $\boldsymbol{\beta} = 0$ vs H_a: $\boldsymbol{\beta} \neq 0$ using the F-statistic with (1, n-2) degrees of freedom:
  - $F^* = MS_{model}/MS_{error} = 2.0655$
  - The p-value = Prob $(F^* > 2.0655) = 0.1886$

```
> print(paste("Critical value for alpha = 0.05, numerator df =1 and denominatior df = 8 is ", qf(0.95, 1, 8)))
[1] "Critical value for alpha = 0.05, numerator df =1 and denominatior df = 8 is  5.31765507157871"
```

- Thus, in simple regression (1 independent variable), both test give the same information with the same p-value

- In fact $F^* = t^2$. i.e., $2.0655 = (1.437)^2$

The Standardized Model

- We can also perform regression on Y using X by removing units from both, very similar to removing units from variance using standard deviation and from covariance using correlation.

- In other words, we *standardize X and Y before performing the regression*.

- We replace X by $z_x = (X - \overline{X})/\ \sigma_x$ and Y by $z_y = (Y - \overline{Y})/\ \sigma_y$ and perform the regression $z_y = \hat{a} + \hat{b}z_x + \epsilon$, where $\hat{a}$ and $\hat{b}$ are called ***standardized regression coefficients***.

- The units of z_x and z_y are the *standard deviations* of X and Y respectively

- $\hat{b}$ is interpreted as the number of standard deviations change in Y for one standard deviation increase in X

- In simple linear regression, $\boldsymbol{\hat{b}}$ **will become equal to the correlation between X and Y**.

- In our example, the sample correlation between Net Worth and Age was 0.453. You can see that $\boldsymbol{\hat{b}}$ from the printout of the standardized model summary is also 0.453.

```
> #
> # The Standardized Regression Model
> #
> z_networth <- (networth - mean(networth))/sd(networth)
> z_age <- (age - mean(age))/sd(age)
> std_model <- lm(z_networth ~ z_age)
> summary(std_model)

Call:
lm(formula = z_networth ~ z_age)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2455 -0.8357  0.2663  0.4655  1.4418 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.079e-17  2.990e-01   0.000    1.000
z_age       4.530e-01  3.152e-01   1.437    0.189

Residual standard error: 0.9456 on 8 degrees of freedom
Multiple R-squared:  0.2052,	Adjusted R-squared:  0.1059 
F-statistic: 2.066 on 1 and 8 DF,  p-value: 0.1886
```

LECTURE 4A – 5 – ANOTHER SIMPLE REGRESSION EXAMPLE

Driver Age vs Visibility Distance – **AgeDistance.R**

- The file Driver-Age.csv contains data on the age of a driver and the distance they can see (in feet).
- We will conduct a simple regression analysis.
- We calculate the sample means, sample standard deviations and Correlation, for reference.
- We notice that the correlation is fairly high (-.801)
- The negative sign tells us that as Age *increases* driver distance visibility *decreases.*

```
> # Read csv file as a DataFrame
> #
> setwd("C:\\Users\\sarathy\\Documents\\2019-Teaching\\Fall2019\\Fall2019-MSIS5503\\MSIS-5503-Data")
> df1 <- read.table('Driver-Age.csv',
+                   header = TRUE, sep = ',')
>
> #Assign variable names to DataFrame Column objects
> age <- df1$Age
> distance <- df1$Distance
> print(df1)
   Age Distance
1   18      510
2   20      590
3   22      560
4   23      510
5   23      460
6   25      490
7   27      560
8   28      510
9   29      460
10  32      410
11  37      420
12  41      460
13  46      450
14  49      380
15  53      460
16  55      420
17  63      350
18  65      420
19  66      300
20  67      410
21  68      300
22  70      390
23  71      320
24  72      370
25  73      280
26  74      420
27  75      460
28  77      360
29  79      310
30  82      360
> #
> #
> samp_means <- c(mean(distance), mean(age))
> print(samp_means)
[1] 423.3333  51.0000
> samp_sd <- c(sd(distance), sd(age))
> print(round(samp_sd,3))
[1] 81.720 21.776
> samp_cor <- cor(distance, age)
> print(round(samp_cor,3))
[1] -0.801
```

Driver Age vs Visibility Distance

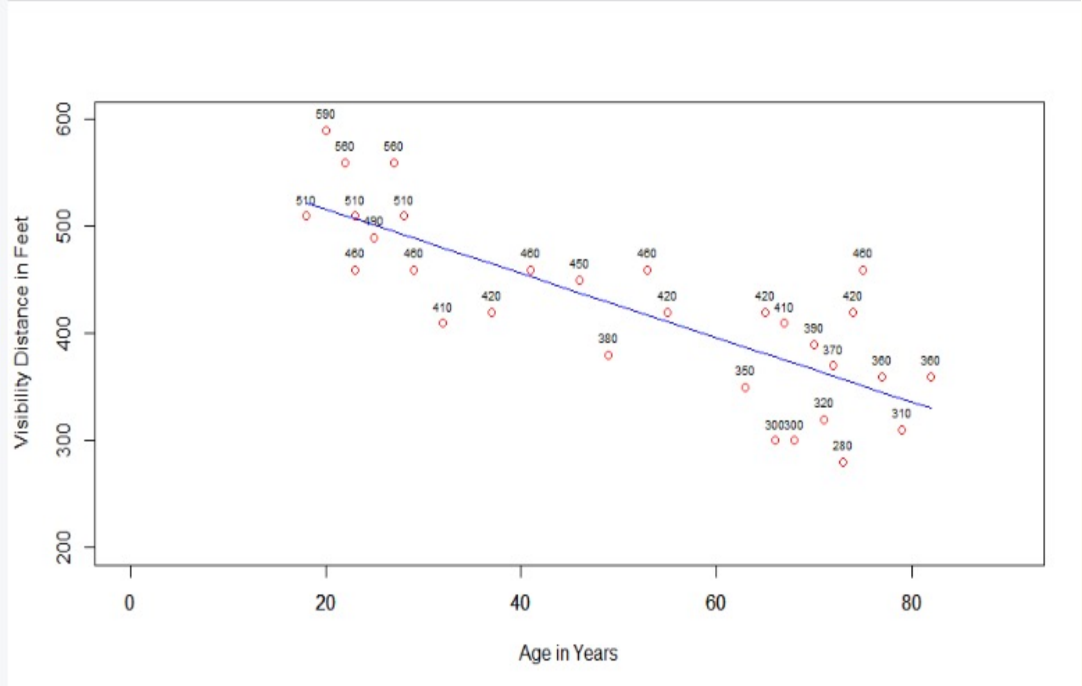
- Next we run the regression model, generate predicted values and residuals.
- The estimated regression equation (prediction equation) is:
  - Predicted Distance = 576.68 – 3.0068Age
- The equation tells us that for each year increase in Age, visibility distance reduces by 3 feet.
- From printout, the predicted visibility distance at age = 49 is 429.35 feet.
- We can also calculate it as:
  - 576.68 – 3.0068(49) = 429.45 feet
- The actual sample distance value for Age = 49 is 380, so the *residual* is 380 – 429.35 = -49.35

```
> #
> reg_mod <- lm(distance ~ age)
> summary(reg_mod)

Call:
lm(formula = distance ~ age)

Residuals:
    Min      1Q  Median      3Q     Max 
-78.231 -41.710   7.646  33.552 108.831 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 576.6819    23.4709  24.570  < 2e-16 ***
age          -3.0068     0.4243  -7.086 1.04e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49.76 on 28 degrees of freedom
Multiple R-squared:  0.642,	Adjusted R-squared:  0.6292 
F-statistic: 50.21 on 1 and 28 DF,  p-value: 1.041e-07

> #Generate Predicted Values and Residuals and Add them to the Data Frame df1
> #
> df1$p_distance <- predict(reg_mod)
> df1$r_distance <- residuals.lm(reg_mod)
> print (df1)
   Age Distance p_distance r_distance
1   18      510   522.5589 -12.558901
2   20      590   516.5452  73.454770
3   22      560   510.5316  49.468441
4   23      510   507.5247   2.475276
5   23      460   507.5247 -47.524724
6   25      490   501.5111 -11.511053
7   27      560   495.4974  64.502618
8   28      510   492.4905  17.509453
9   29      460   489.4837 -29.483711
10  32      410   480.4632 -70.463205
11  37      420   465.4290 -45.429029
12  41      460   453.4017   6.598313
13  46      450   438.3675  11.632490
14  49      380   429.3470 -49.347004
15  53      460   417.3197  42.680337
16  55      420   411.3060   8.694008
17  63      350   387.2513 -37.251309
18  65      420   381.2376  38.762362
19  66      300   378.2308 -78.230803
20  67      410   375.2240  34.776033
21  68      300   372.2171 -72.217132
22  70      390   366.2035  23.796539
23  71      320   363.1966 -43.196626
24  72      370   360.1898   9.810209
25  73      280   357.1830 -77.182955
26  74      420   354.1761  65.823880
27  75      460   351.1693 108.830716
28  77      360   345.1556  14.844386
29  79      310   339.1419 -29.141943
30  82      360   330.1214  29.878563
> 
```

Driver Age vs Visibility Distance

- We next plot the values of Distance vs Age as well as the Prediction Line
- First, we notice that the variability in Y across the different X-values are not too high.
- We can therefore expect that the correlation would not be low (it was – 0.801)
- We can also see that the values are relatively well crowded around the prediction line. The R-square will also not be low. i.e., the model Sum of Squares would be a relatively high proportion of Total Sum of Squares (Total variability in Y).
- In fact, R-square should be $(-0.801)^2 = 0.64$

```
> #
> # Plot of Networth vs Age
> #
> plot(age,distance,col="red",
+      xlab="Age in Years",
+      ylab="Visibility Distance in Feet",
+      xlim = c(0, 90), ylim= c(200, 600))
> text(age, distance, distance, cex=0.6, pos=3)
> #
> # Add Predicted Values to the Plot
> #
> par(new=TRUE)
> plot(age,df1$p_distance,type="l",
+      yaxt='n', ann=FALSE, col="blue",  xlim = c(0, 90), ylim= c(200, 600))
> |
```



Driver Age vs Visibility Distance

- We will next check the value of the estimated slope coefficient and intercept:
  - $\hat{\beta} = r_{XY}{}^{*}(s_Y/s_X) = (-.801)^{*}81.72/21.776 = -3.005$
  - $\hat{\alpha} = \bar{Y} - \hat{\beta}^{*}\bar{X} = 423.333 - (-.3.005)^{*}(51) = 576.588$
- The RMSE (or Residual Standard Error s_ε) is given in the printout as: 49.76 on 28 degrees of freedom.
- The Standard Error of the $\hat{\beta}$ of Age $= s_{\hat{\beta}} = \dfrac{s_\varepsilon}{\sqrt{n-1}(s_X)} = 0.4243.$
  - $s_\varepsilon = 49.76$, $s_X = 21.776$, $n-1 = 29$, so $\dfrac{49.76}{\sqrt{29}(21.776)} = s_{\hat{\beta}} = 0.4243$
- The t-statistic, therefore $= (\hat{\beta}/s_{\hat{\beta}}) = (-.3.005/0.4243) =$ -7.086 with 28 degrees of freedom.
- The p-value for this t-statistic

```
> print(paste("p-value for t= -7.086, with 28 df, for two.tailed test is ", 2*pt(-7.086,28)))
[1] "p-value for t= -7.086, with 28 df, for two.tailed test is  1.04087676568375e-07"
```

- We therefore **Reject** the Null Hypothesis H_0: $\boldsymbol{\beta} = 0$ (i.e., there is no linear relationship between Age and Distance Visibility) at $\alpha = 0.05$.
- We conclude (at $\alpha = 0.05$) that there is a significant linear relationship between Age and Distance Visibility. For each unit increase in Age, distance visibility decreases by 3 feet.

```
> #
> reg_mod <- lm(distance ~ age)
> summary(reg_mod)

Call:
lm(formula = distance ~ age)

Residuals:
    Min      1Q  Median      3Q     Max 
-78.231 -41.710   7.646  33.552 108.831 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 576.6819    23.4709  24.570  < 2e-16 ***
age          -3.0068     0.4243  -7.086 1.04e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49.76 on 28 degrees of freedom
Multiple R-squared:  0.642,	Adjusted R-squared:  0.6292 
F-statistic: 50.21 on 1 and 28 DF,  p-value: 1.041e-07
```

```
> samp_means <- c(mean(distance), mean(age))
> print(samp_means)
[1] 423.3333  51.0000
> samp_sd <- c(sd(distance), sd(age))
> print(round(samp_sd,3))
[1] 81.720 21.776
> samp_cor <- cor(distance, age)
> print(round(samp_cor,3))
[1] -0.801
```

Driver Age vs Visibility Distance

- **Model Fit:**
- The R^2 for the model is: 0.642
- Age explains 64.2% of the variability in Distance Visibility; the remaining 35.8% is explained by error (noise).
- We can check the $R^2 = \frac{\text{Model SS}}{\text{Total SS}}$.
- From the ANOVA output
  - Model SS = 124333, Residual SS = 69334, so Total SS = 124333 + 69334 = 193667
  - $R^2 = \frac{\text{Model SS}}{\text{Total SS}} = R^2 = \frac{124333}{193667} = 0.642$
- F-statistic = $= \frac{\text{MS(Model)}}{\text{MS(Error)}} = \frac{\text{Model SS/(df=1)}}{\text{Residual SS/(df=28)}} = = \frac{124333}{69334/28} = 50.21$ with (1, 28) degrees of Freedom
- The p-value is:

```
> print(paste("P-value for F = 50.21 with 1 and 28 df = ",1 - pf(50.21, 1, 28)))
[1] "P-value for F = 50.21 with 1 and 28 df =  1.04114158627766e-07"
```

- We therefore **Reject** the Null Hypothesis H_0: $\beta = 0$ (i.e., there is no linear relationship between Age and Distance Visibility) at $\alpha = 0.05$.
- We conclude (at $\alpha = 0.05$) that there is a significant linear relationship between Age and Distance Visibility. For each unit increase in Age, distance visibility decreases by 3 feet.

```
> #
> reg_mod <- lm(distance ~ age)
> summary(reg_mod)

Call:
lm(formula = distance ~ age)

Residuals:
    Min      1Q  Median      3Q     Max 
-78.231 -41.710   7.646  33.552 108.831 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 576.6819    23.4709  24.570  < 2e-16 ***
age          -3.0068     0.4243  -7.086 1.04e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49.76 on 28 degrees of freedom
Multiple R-squared:  0.642,	Adjusted R-squared:  0.6292 
F-statistic: 50.21 on 1 and 28 DF,  p-value: 1.041e-07
```

```
> # Obtaining Sums of Squares using ANOVA
> #
> anova_mod <- anova(reg_mod)
> print(anova_mod)
Analysis of Variance Table

Response: distance
          Df Sum Sq Mean Sq F value    Pr(>F)    
age        1 124333  124333  50.211 1.041e-07 ***
Residuals 28  69334    2476                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
```

LECTURE 4A – 6 – SIMPLE REGRESSION WITH CATEGORICAL PREDICTOR

Simple Regression with Categorical Predictors

- Sometimes the predictor (independent variable) X is a categorical (nominal or ordinal) variable

- For example, we may be interested in Gender vs Income.

- In this case, we convert the categorical variable into a *dummy variable* by assigning a numerical value to the classes of the categorical variable. We will first deal with Gender, which has only 2 classes; Male and Female

- We use a Dummy variable d_gender as follows: d_gender=1 if gender is Male, and d_gender=0 if Female.

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Simple Regression with Categorical Predictors

- The population model becomes: $Y = \boldsymbol{\alpha} + \boldsymbol{\beta} d\_gender + \boldsymbol{\epsilon}$.
- But, since d_gender = 0 or 1, we get two equations: one for Males at d_gender=1 and one for females at d_gender= 0

- We have for the population model:
  - $Y = \boldsymbol{\alpha} + \boldsymbol{\beta} + \boldsymbol{\epsilon}$ (Males); $Y = \boldsymbol{\alpha} + \boldsymbol{\epsilon}$ (Females).
  - $E(Y_{Males}) = E(\boldsymbol{\alpha} + \boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{\alpha} + \boldsymbol{\beta}$, since $E(\boldsymbol{\epsilon}) = 0$
  - $E(Y_{Females}) = E(\boldsymbol{\alpha} + \boldsymbol{\epsilon}) = \boldsymbol{\alpha}$

- Therefore, we have:
  - $\boldsymbol{\alpha} = E(Y_{Females})$
  - $\boldsymbol{\beta} = E(Y_{Males}) - E(Y_{Females}) = \boldsymbol{\mu_{males} - \mu_{Females}}$

- Correspondingly, in the *sample*:
  - $\widehat{\alpha} = E(\hat{y}_{Females})$
  - $\widehat{\beta} = E(\hat{y}_{Males}) - E(\hat{y}_{Females})$

- We can then see that when we test for H_0: $\boldsymbol{\beta} = 0$, we are testing for the *difference between the mean of Y for the first group (Males) and the mean of Y for the second group (Females), i.e.,* ***differences in means of two groups***

Simple Regression with Categorical Predictors

- We will perform a regression where Y is Income and X is Gender.
- From Regression output:
  - $E(\hat{y}_{Females}) = \hat{\alpha}$ = \$177,000 = Mean Income for Females
  - $E(\hat{y}_{Males}) = \hat{\beta} + E(\hat{y}_{Females})$ = -1863.33 + 177,000 = \$158,316.67
- Is the difference between Mean Income for Males and Females significant?
  - Test for H_0: $\boldsymbol{\beta}$ = 0 becomes = $\boldsymbol{\mu_{males} - \mu_{Females} = 0}$ i.e., ***no difference in means of two groups***
  - Test t-statistic = -0.2374
  - p-value = 0.8183
  - We ***fail to reject*** the null hypothesis of ***no difference in means of two groups***
- Proportion of variability in Income explained by Gender:
  - R^2 = 0.006, that is Gender explains only 0.6% of variability in Income and is a poor "explainer" or "Predictor"
- Analysis of Variance F-test of H_0: $\boldsymbol{\beta}$ = 0 or $\boldsymbol{\mu_{males} - \mu_{Females} = 0}$ i.e., ***no difference in means of two groups***
  - Test F-statistic = 0.056
  - p-value = 0.8183
  - We ***fail to reject*** the null hypothesis of ***no difference in means of two groups***

```
> #
> # Regression Using categorical variable Gender
> #
> # Convert gender to a Dummy Variable
> #
> d_gender <- ifelse(gender=="M", 1,0)
> #
> gender_mod <- lm(income ~ d_gender)
> summary(gender_mod)

Call:
lm(formula = income ~ d_gender)

Residuals:
    Min      1Q  Median      3Q     Max
-119417  -73488  -26158   46013  275683

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   177000      60958   2.904   0.0198 *
d_gender      -18683      78697  -0.237   0.8183
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 121900 on 8 degrees of freedom
Multiple R-squared:  0.006996,	Adjusted R-squared:  -0.1171
F-statistic: 0.05636 on 1 and 8 DF,  p-value: 0.8183
```

```
> male_income <- c(66000, 82000, 111000, 218000, 38900, 434000)
> female_income <- c(112000, 182000, 242000, 172000)
> mean_male_income <- mean(male_income)
> mean_female_income <- mean(female_income)
> print(mean_male_income)
[1] 158316.7
> print(mean_female_income)
[1] 177000
```

Comparing Simple Regression with Categorical Variable & the t-test

- If $\boldsymbol{\mu_1}$ is the mean of population A with standard deviation $\boldsymbol{\sigma_1}$ and $\boldsymbol{\mu_2}$ is the mean of population B with standard deviation $\boldsymbol{\sigma_2}$, then the t-test for two population mean differences is as follows:
  - $\mathbf{H_0}$: $\boldsymbol{\mu_1}$- $\boldsymbol{\mu_2}$ = 0 (No difference in population means)
  - $\mathbf{H_a}$: $\boldsymbol{\mu_1}$- $\boldsymbol{\mu_1}$ $\neq$ 0 (Population means are different)

- Under the assumptions that:
  1. The two **<u>independent</u>** samples are simple random samples of size n_1 and n_2 respectively from A and B.
  2. Either
     - distributions are normal (small sample sizes n)
     - distributions are any (large sample sizes n)

- The difference is sample means ($\bar{x}_1$ - $\bar{x}_2$) ~ N($\boldsymbol{\mu_1}$- $\boldsymbol{\mu_2}$, $\boldsymbol{\sigma^2_1}$ + $\boldsymbol{\sigma^2_2}$), where $\bar{x}_1$ & $\bar{x}_2$ are the means of the independent samples from Populations A and B respectively.

- The test statistic for the hypothesis test (the populations standard deviations are not assumed known) has a standard error: $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^3}{n_2}}$ and is given by a t-di
  - $t^* = [(\bar{x}_1 - \bar{x}_2) - (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})]/\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^3}{n_2}}$

Degrees of freedom

$$df = \frac{\left(\frac{(s_1)^2}{n_1}+\frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1-1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2+\left(\frac{1}{n_2-1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2}$$

Comparing Simple Regression with Categorical Variable & the t-test

- For example, we have the Income of Males and Females.
- We can use the *t.test()* function in **R** and compare it with the regression results.
- The results of the 2-tailed test show a p-value of 0.784 for the test-statistic. At a significance level of $\alpha = 0.05$, we fail to reject the Null Hypothesis that the difference in Mean Income for Males and Females in the population is 0.
- Note that this t-test is based on Lecture 3D two-sample test where we do not know the population standard deviations.

```
> #
> # Compare wioth t-test
> #
> male_income <- c(66000, 82000, 111000, 218000, 38900, 434000)
> female_income <- c(112000, 182000, 242000, 172000)
> t.test(male_income, female_income, paired=FALSE, alternative = "two.sided")

        Welch Two Sample t-test

data:  male_income and female_income
t = -0.28203, df = 6.6964, p-value = 0.7864
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -176783.2  139416.6
sample estimates:
mean of x mean of y 
 158316.7  177000.0 
```

Comparing Simple Regression with Categorical Variable & the t-test

- When we compare the results from t-test and Regression we get the same results.

- The test of hypotheses in Regression were:
  - Is the difference between Mean Income for Males and Females significant?
  - Test for H_0: $\boldsymbol{\beta} = 0 = \boldsymbol{\mu_{males} - \mu_{Females} = 0}$ i.e., ***no difference in means of two groups***
  - Test t-statistic = -0.2374
  - p-value = 0.8183
  - We ***fail to reject*** the null hypothesis of ***no difference in means of two groups at*** $\alpha = 0.05$.

- The test of hypotheses in t-test were:
  - If $\boldsymbol{\mu_{Male}}$ is the mean income of Males with standard deviation $\boldsymbol{\sigma_{Male}}$ and $\boldsymbol{\mu_{Female}}$ is the mean income of Females with standard deviation $\boldsymbol{\sigma_{Female}}$, the test is as follows:
  - $\boldsymbol{H_0}$: $\boldsymbol{\mu_{Males}} - \boldsymbol{\mu_{Females}} = 0$ (No difference in population means of Income)
  - $\boldsymbol{H_a}$: $\boldsymbol{\mu_{Males}} - \boldsymbol{\mu_{Females}} \neq 0$ (Population means of Income are different)
  - Test F-statistic (square of regression t-statistic) = 0.06
  - p-value = 0.784
  - We ***fail to reject*** the null hypothesis of ***no difference in means of two groups at*** $\alpha = 0.05$.

```
> #
> # Regression Using categorical variable Gender
> #
> # Convert gender to a Dummy Variable
> #
> d_gender <- ifelse(gender=="M", 1,0)
> #
> gender_mod <- lm(income ~ d_gender)
> summary(gender_mod)

Call:
lm(formula = income ~ d_gender)

Residuals:
    Min      1Q  Median      3Q     Max
-119417  -73488  -26158   46013  275683

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   177000      60958   2.904   0.0198 *
d_gender      -18683      78697  -0.237   0.8183
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 121900 on 8 degrees of freedom
Multiple R-squared:  0.006996,	Adjusted R-squared:  -0.1171
F-statistic: 0.05636 on 1 and 8 DF,  p-value: 0.8183

> #
> # Compare wioth t-test
> #
> male_income <- c(66000, 82000, 111000, 218000, 38900, 434000)
> female_income <- c(112000, 182000, 242000, 172000)
> t.test(male_income, female_income, paired=FALSE, alternative = "two.sided")

	Welch Two Sample t-test

data:  male_income and female_income
t = -0.28203, df = 6.6964, p-value = 0.7864
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -176783.2  139416.6
sample estimates:
mean of x mean of y
 158316.7  177000.0
```