**CSC591**: Foundations of Data Science  
HW3: Regression Analysis

Released: 10/21/15  
Due: **10/27/15 (23:55pm);** (One day late: -25%; -100% after that).

Student Name: Parth Satra   
Student ID: pasatra (200062999)

**Notes**

* Submit single zip file containing: (1) all solutions as single pdf file (Filename: Lastname\_StudentID.pdf); (2) separate R code file for Q4-Q5 with appropriately named readme files.
* You can also submit scanned hand written solution (should be legible, TA’s interpretation is final).
* This h/w is worth 5% of total grade + **Bonus questions account for 3% of grade**
* You can discuss with your friends, but solution should be yours.
* Any kind of copying will result in 0 grade (minimum penalty), serious cases will be referred to appropriate authority.
* All submissions must be through Moodle (you can email to TA with cc to Instructor – only if these is a problem – if not received on time, then standard late submission rules apply)
* No makeups; for regarding policies, refer to syllabus and 1st day lecture slides.

|  |  |  |
| --- | --- | --- |
| Q# | Max Points | Your Score |
| 1 | 55 |  |
| 2 | 15 |  |
| 3 | 1% of grade (Bonus) |  |
| 4 | 30 (R mini project) |  |
| 5 | 2% of grade (Bonus) |  |

**Note**:

1. All questions and sub-parts (Q1-Q3) of this h/w require hand calculations using the formulas that you learned from the course materials. You can use any calculator.
2. For all project/programming questions, you need implementations in R and submit codes and plots (as requested). Please provide comments (in the code), and separate readme.txt file describing steps to run the code.

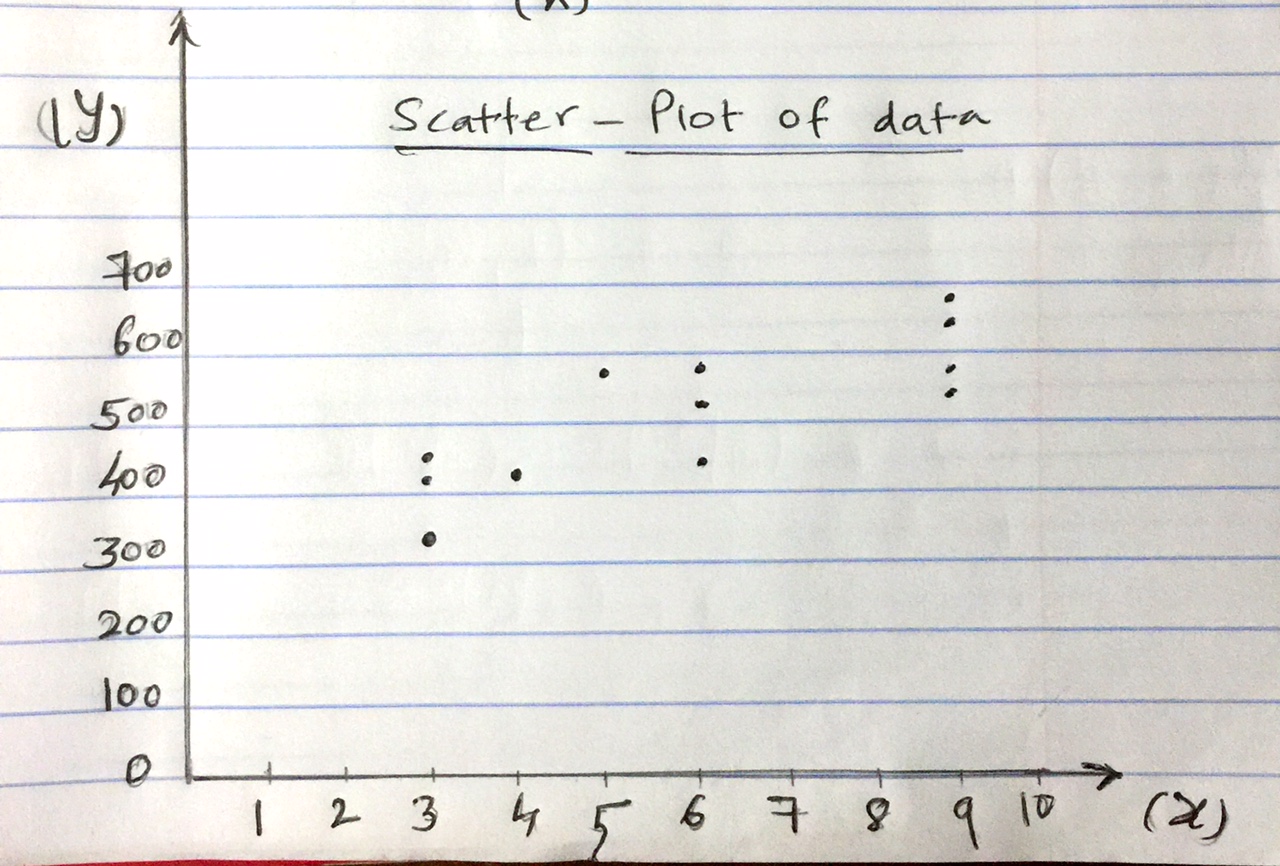
**Q1**. Simple Linear Regression (**55 points**)

Following table 1 show the data required to answer this question.

|  |  |
| --- | --- |
| **x** | **y** |
| 6 | 520 |
| 4 | 421 |
| 6 | 581 |
| 9 | 630 |
| 3 | 412 |
| 9 | 562 |
| 6 | 434 |
| 3 | 443 |
| 9 | 590 |
| 5 | 570 |
| 3 | 348 |
| 9 | 672 |

1. Draw 2-d scatter plot (choose appropriate scaling for x and y axis). (5 points)

Answer



1. Compute the slope and intercept of the simple linear regression equation (show all computations. **Hint**: use tabular format to compute intermediate quantities)   
   (10 points)

Answer

The following table computes all the required statistics to find the slope and intercept of the simple linear regression. The last row in the table provides the summation of each column. The formula are as follows:

y’ = β0 + β1x

β0 = ( (∑y)( ∑x2) – (∑x)( ∑xy)) / ( n(∑x2) –(∑x)2 )

β1 = ( n(∑xy) – (∑x)( ∑y)) / ( n(∑x2) –(∑x)2 )

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **x** | **y** | **xy** | **x2** | **y2** |
| 6 | 520 | 3120 | 36 | 270400 |
| 4 | 421 | 1684 | 16 | 177241 |
| 6 | 581 | 3486 | 36 | 337561 |
| 9 | 630 | 5670 | 81 | 396900 |
| 3 | 412 | 1236 | 9 | 169744 |
| 9 | 562 | 5058 | 81 | 315844 |
| 6 | 434 | 2604 | 36 | 188356 |
| 3 | 443 | 1329 | 9 | 196249 |
| 9 | 590 | 5310 | 81 | 348100 |
| 5 | 570 | 2850 | 25 | 324900 |
| 3 | 348 | 1044 | 9 | 121104 |
| 9 | 672 | 6048 | 81 | 451584 |
| **72** | **6183** | **39439** | **500** | **3297983** |

Thus using the above table we get

β0 = ((6183)(500) – (72)(39439)) / (12(500) – (72)2)

β0 = 308.691

β1 = ((12)(39439) – (72)(6183)) / (12(500) – (72)2)

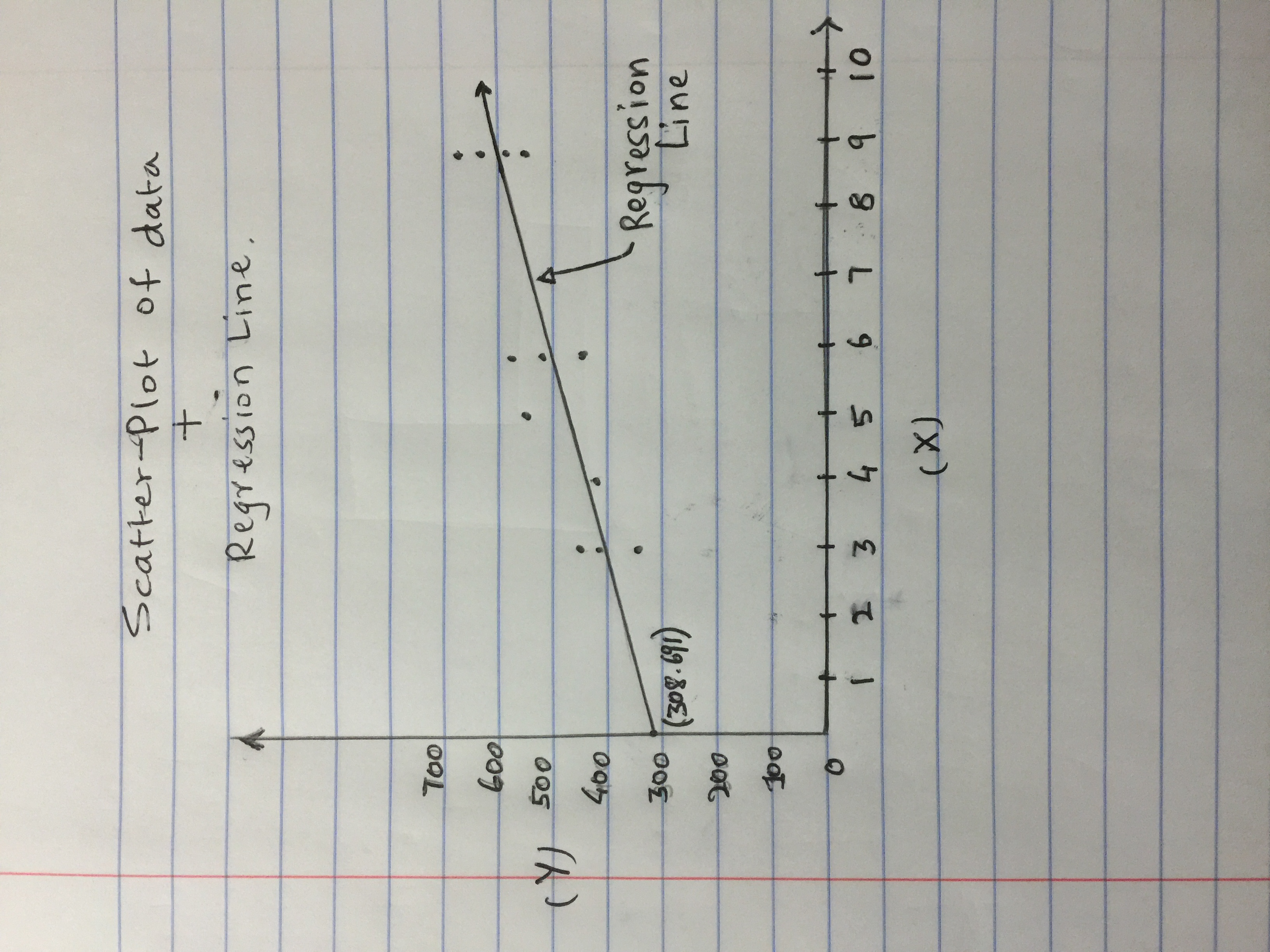
β1 = 34.4265

Thus,

**y’ = 308.691 + 34.4265x**

1. Draw resulting regression line on the 2-d scatter plot (you can copy initial plot from (a)). (5 points)

**Answer**

****

1. Compute the fitted values and residuals for each observation and verify that the residuals sum to zero (or approximately zero). (5 points)

**Answer**

In the following table as you can see the value of the residuals for each observation and sum is almost zero.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **x** | **y** | **y'** | **Residual = y - y'** |
|  | 6 | 520 | 515.25 | 4.75 |
|  | 4 | 421 | 446.397 | -25.397 |
|  | 6 | 581 | 515.25 | 65.75 |
|  | 9 | 630 | 618.5295 | 11.4705 |
|  | 3 | 412 | 411.9705 | 0.0295 |
|  | 9 | 562 | 618.5295 | -56.5295 |
|  | 6 | 434 | 515.25 | -81.25 |
|  | 3 | 443 | 411.9705 | 31.0295 |
|  | 9 | 590 | 618.5295 | -28.5295 |
|  | 5 | 570 | 480.8235 | 89.1765 |
|  | 3 | 348 | 411.9705 | -63.9705 |
|  | 9 | 672 | 618.5295 | 53.4705 |
| **SUM** |  |  |  | **7.38964E-13 ≈ 0** |

1. How much of variation in y is explained by x? (5 points)

Answer:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **x** | **y** | **y'** | **y' - mean(y)** | **(y' - mean(y))2** |
|  | 6 | 520 | 515.25 | 0 | 0 |
|  | 4 | 421 | 446.397 | -68.853 | 4740.735609 |
|  | 6 | 581 | 515.25 | 0 | 0 |
|  | 9 | 630 | 618.5295 | 103.2795 | 10666.65512 |
|  | 3 | 412 | 411.9705 | -103.2795 | 10666.65512 |
|  | 9 | 562 | 618.5295 | 103.2795 | 10666.65512 |
|  | 6 | 434 | 515.25 | 0 | 0 |
|  | 3 | 443 | 411.9705 | -103.2795 | 10666.65512 |
|  | 9 | 590 | 618.5295 | 103.2795 | 10666.65512 |
|  | 5 | 570 | 480.8235 | -34.4265 | 1185.183902 |
|  | 3 | 348 | 411.9705 | -103.2795 | 10666.65512 |
|  | 9 | 672 | 618.5295 | 103.2795 | 10666.65512 |
| **SUM** | **72** | **6183** |  |  | **80592.50535** |
| **MEAN** | **6** | **515.25** |  |  |  |

Thus **80592.51** is the total variance explained by x.

1. Compute the standard error of the estimate (5 points)

Answer:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **x** | **y** | **y'** | **y - y'** | **(y - y')2** |
|  | 6 | 520 | 515.25 | 4.75 | 22.5625 |
|  | 4 | 421 | 446.397 | -25.397 | 645.007609 |
|  | 6 | 581 | 515.25 | 65.75 | 4323.0625 |
|  | 9 | 630 | 618.5295 | 11.4705 | 131.5723703 |
|  | 3 | 412 | 411.9705 | 0.0295 | 0.00087025 |
|  | 9 | 562 | 618.5295 | -56.5295 | 3195.58437 |
|  | 6 | 434 | 515.25 | -81.25 | 6601.5625 |
|  | 3 | 443 | 411.9705 | 31.0295 | 962.8298703 |
|  | 9 | 590 | 618.5295 | -28.5295 | 813.9323702 |
|  | 5 | 570 | 480.8235 | 89.1765 | 7952.448152 |
|  | 3 | 348 | 411.9705 | -63.9705 | 4092.22487 |
|  | 9 | 672 | 618.5295 | 53.4705 | 2859.09437 |
| **SUM** |  |  |  |  | **31599.88235** |

Sest =

Sest = = **56.214**

1. What are the predicted values for x = 2, 4, 6, 7, 10. (5 points)

Answer

The following table shows the corresponding predicted values using the equation

**y’ = 308.691 + 34.4265x**

|  |  |
| --- | --- |
| **x** | **y'** |
| 2 | 377.544 |
| 4 | 446.397 |
| 6 | 515.25 |
| 7 | 549.6765 |
| 10 | 652.956 |

1. Test for significance of “r” (linear relationship) at α = 0.05. (10 points)

Answer

The correlation coefficient ‘r’ can be calculated as

r =

The values of these parameters can be obtained from the table in answer to part(b). Thus replacing the values we get

r =

r = 28092 / 33144.93 = **0.84755**

The obtained ‘r’ is a sample correlation coefficient and an estimate of the population correlation coefficient ρ. To verify the significance of this estimate we need to test for significance. The hypothesis testing procedure is as follows:

**STEP 1**: Determine Null hypothesis and the claim.

H0: ρ = 0 and H1: ρ ≠ 0

**STEP 2**: Calculate the critical value

Given α = 0.05 and conducting a two-tail t-test we get

**Critical Value = 2.228**

**STEP 3:** Calculate test Value

t =

t =

t = 0.84755 (5.958)

t = **5.05**

**STEP 4**: Compare values and make decision

The value 5.05 > 2.228 and hence we reject the null hypothesis.

**STEP 5**: Summary

Thus, there is enough evidence to support the claim that there is a strong between the variables.

1. Compute the prediction interval for same α(5 points)

Answer

The prediction interval can be given by

y’ ± tα/2 Sest

The values of these variables can be found in table for part b and previous parts.

y’ ± 2.228 \* 56.214

y’ ± 125.24

**Q2**. (20 points)

1. Sate the assumptions of Simple, Multiple, and Logistic regression (if you can’t find answer in slides, you should look at any standard book and/or online; clearly cite the reference). (3x4 = 12 points).

Answer

Assumptions of Simple Linear Regression:

1. **Normality Assumption**: For a given value of x, the value of dependent variable y is normally distributed.
2. **Equal Variance Assumption**: For each value of the independent variable x, the variance for all the corresponding dependent variable y are the same. So the variance for all the y variables remains the same.
3. **Linearity Assumption**: The independent variable x, is linearly related to the dependent variable y.
4. **Independence Assumption**: The values of dependent variable y are independent of each other.
5. **Independence of Errors**: The errors are of the response variables are unrelated to each other.

Assumptions of Multiple Linear Regression:

All the assumptions of the simple linear regression hold good. Apart from these we have following additional assumptions for multiple Linear Regression:

1. **Non-multicollinearity Assumption**: Multiple linear regression has multiple independent variables on which the dependent variable y is related to. So all these independent variables should be unrelated to each other.

Assumptions of Logistic Linear Regression:

Following are the assumptions of Logistic Linear Regression:

1. **Non-multicollinearity Assumption**: In case of multiple independent variables, each independent variable should be unrelated to each other. This does not apply in case there exists only one independent variable.
2. **Independence Assumption:** The values of dependent variable y are independent of each other.
3. **Linearity Assumption with Log odds**: Even though logistic regression does not require the dependent variables to be linearly related to independent variables, it requires the independent variables and log odds to be linearly related.
4. **Independence of Error Term**: The errors are of the response variables are unrelated to each other.
5. **Asymptotically better:** The accuracy increases with increase in the number of samples since the theory is asymptotically based.

**Note**: All the assumptions are taken from Lecture Slides shared by professor and the references [1][2][3] as stated at the end.

1. Pickup any one assumption from each group (i.e., simple, multiple, and logistic) and state what happens if that assumption is violated (3x2 = 6 points). (These three assumptions should be different from each other).

**Answer**

1. **Linearity Assumption**: In simple linear regression if this assumption is violated then we will not have any value of β0 and β1 that correctly fits y and it will be impossible to predict the values to y with a linear model. The accuracy will be bad and the errors will keep increasing. This also happens in case of multiple linear regression.
2. **Equal Variance Assumption:** In case of simple and multiple linear regression if the all the predicted values of y do not have the same variance, then for each predicted value variance will be different and it will not be possible to determine the confidence level of the prediction since the normal distribution of that predicted value will have unknown variance. Also since the sample values can also have difference variance, each sample will have different importance in building the model.
3. **Linearity Assumption with log odds:** This says that there should be a linear relationship between the independent variables and log odds. If not the case then the model underestimates the strength of the relationship and thus resulting in not rejecting Null hypothesis by considering the relationship not significant while it actually is significant and thus predicting wrong results.
4. **Errors are Independent:** There needs to be no correlation between errors. If there exists such a correlation among the error terms, then the model underestimates the value of standard error resulting in inaccurate confidence level and prediction level.

**Note**: All the assumptions are taken from Lecture Slides shared by professor and the references [1][2][3] as stated at the end.

**Q3.** Bonus Question (1% of grade)

In logistic regression analysis (Slide 8), we mentioned that

1. Errors can’t be normally distributed. With proper analysis (or sound arguments) show why this is the case? (5 points)

**Answer**

As stated in class discussion forum by professor, This question actually says that errors can’t be normally distributed in case we use linear regression instead of logistic regression.

Logistic regression is used when the response variable is categorical and not continuous. Assuming we have a binary response variable then such that y = 1 or 0. This can also be represented as probability

P(Yi = 1) = πi

P(Yi = 0) = (1 – πi)

Thus,

E(Yi) = 1 \* πi + 0 \* (1 – πi) = πi

The binary response can thus be in general be represented as

**E(Yi|Xi) = βo + β1Xi = πi**

Now if we try to fit the linear model to this, then we have a constraint on the response variable since the probability is always between 0 and 1. But additionally, binary nature of the response also creates difficulties in how the individual values are distributed around the mean. Since there are only two responses these cannot be normally distributed around the mean and thus violating the linear regression assumptions. Thus even the errors generated will not be normally distributed as there are only 2 error values possible.

Reference from [3].

1. Error variance is not constant. Show why? (10 points)

**Answer**

As stated above if linear regression is applied on a binary response variable instead of logistic regression then the expected value in general can be represented as follows.

E(Yi|Xi) = βo + β1Xi = πi

The variance of the binary response function can be represented as

**Var(Yi) = πi \* (1 – πi)**

Now since ‘πi’ is itself a function of Xi the variance cannot be constant. Thus the again violating the assumption of linear regression.

Reference from [3].

**Mini R project (30 points; 10 points each for Q4-Q6)**

**Note:** You should provide answers (along with inline R code) in the pdf file. In addition submit R scripts.

**Q4**. Using “Advertising” data answer (a) and (b)

**(a)** Fit simple linear regression (separately) for each covariate. Provide scatter plots with fitted regression line. Which covariate provides best prediction?

**Answer**

**R Code**

*#Question 4(a)*

*# Clean the environment*

*rm(list = ls())*

*# Load data*

*adv\_data = read.csv("Advertising.csv")*

*adv\_data = adv\_data[,-1]*

*#Simple linear Regression for TV covariate*

*plot(adv\_data$TV, adv\_data$Sales, xlab = "Sales", ylab = "TV - Advertising",*

*main = "Scatter Plot of TV - Sales", col = "firebrick",*

*type = "p", pch = 16)*

*model\_TV = lm(adv\_data$Sales ~ adv\_data$TV)*

*abline(model\_TV, col = "steelblue", lw = 3)*

*print(model\_TV$coefficients)*

*print(summary(model\_TV))*

*#Simple linear Regression for Radio covariate*

*plot(adv\_data$Radio, adv\_data$Sales, xlab = "Sales", ylab = "Radio - Advertising",*

*main = "Scatter Plot of Radio - Sales", col = "firebrick",*

*type = "p", pch = 16)*

*model\_Radio = lm(adv\_data$Sales ~ adv\_data$Radio)*

*abline(model\_Radio, col = "steelblue", lw = 3)*

*print(model\_Radio$coefficients)*

*print(summary(model\_Radio))*

*#Simple linear Regression for Newspaper covariate*

*plot(adv\_data$Newspaper, adv\_data$Sales, xlab = "Sales", ylab = "Newspaper - Advertising",*

*main = "Scatter Plot of Radio - Newspaper", col = "firebrick",*

*type = "p", pch = 16)*

*model\_NewsPaper = lm(adv\_data$Sales ~ adv\_data$Newspaper)*

*abline(model\_NewsPaper, col = "steelblue", lw = 3)*

*print(model\_NewsPaper$coefficients)*

*print(summary(model\_NewsPaper))*

**Beta – values for each covariate**

**TV:** (Intercept) adv\_data$TV

7.03259355 0.04753664

**Radio:** (Intercept) adv\_data$Radio

9.3116381 0.2024958

**Newspaper** (Intercept) adv\_data$Newspaper

12.3514071 0.0546931

**Scatter Plots**







The ‘**TV**’ covariate best predicts the model, since it’s the most significant feature.

***(*b)** Fit multiple linear regression model for the data. Show resulting equation. How do you compare the β’s obtained with this model with corresponding β’s found in (a).

**Answer**

**R Code**

*# Clean the environment*

*rm(list = ls())*

*# Load data*

*adv\_data = read.csv("Advertising.csv")*

*adv\_data = adv\_data[,-1]*

*#Question 4(b)*

*# Multiple Linear Regression and the Beta-coefficient*

*model\_multivatiate = lm(adv\_data$Sales ~ adv\_data$TV +*

*adv\_data$Radio + adv\_data$Newspaper)*

*print(model\_multivatiate$coefficients)*

**Beta – Coefficients of multiple linear regression**

*(Intercept) adv\_data$TV adv\_data$Radio adv\_data$Newspaper*

*2.938889369 0.045764645 0.188530017 -0.001037493*

**Comparing the beta’s from the simple linear regression for each covariate.**

The beta’s obtained from linear and multiple linear regression for TV and Radio covariates is almost the same. While there is a difference between the beta’s for the covariate Newspaper. From linear regression it seems that there is a correlation between the Newspaper Covariate and the dependent variable sales. But from multiple linear regression we can see that beta value for Newspaper is slightly negative, which means that Newspaper advertisement does not affect the Sales. Since the value is only slightly negative we can say that its not adversely affecting the sales.

**(c)**. Fit logistic regression model for the dataset (hw3-q4c.txt). Note that this dataset contain 3 covariates, therefore you should use multiple logistic regression which is straight forward generalization of simple logistic regression (simply replace β0+β1X with β0+β1X1 + … + βnXn)

**Answer**

**R Code**

*#Question 4(c)*

*rm(list = ls())*

*# Load the new data*

*log\_data = read.table("hw3-q4c.txt", header = TRUE, sep = "\t")*

*log\_model = glm(log\_data$Y ~ log\_data$X1 + log\_data$X2 + log\_data$X3,*

*family = "binomial")*

*print(summary(log\_model))*

**Model Summary:**

*Deviance Residuals:*

*Min 1Q Median 3Q Max*

*-1.64148 -0.00008 0.00000 0.00135 1.41755*

*Coefficients:*

*Estimate Std. Error z value Pr(>|z|)*

*(Intercept) -10.1535 10.8398 -0.937 0.3489*

*log\_data$X1 0.3312 0.3007 1.101 0.2707*

*log\_data$X2 0.1809 0.1069 1.692 0.0907 .*

*log\_data$X3 5.0875 5.0820 1.001 0.3168*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*(Dispersion parameter for binomial family taken to be 1)*

*Null deviance: 91.4954 on 65 degrees of freedom*

*Residual deviance: 5.8129 on 62 degrees of freedom*

*AIC: 13.813*

*Number of Fisher Scoring iterations: 12*

**Q5**. Apply your data science skills to **improve** the model fitted in (4.c). In what sense your improved model is **better** than the model found in (4.c). [Note the term “improve”; that is you still have to use multiple logistic model only). Show your work. (Worth **2% of final grade points**)

**Answer**

**R Code**

*rm(list =ls())*

*# Applying log transform*

*trans\_data = read.table("hw3-q4c.txt", header = TRUE, sep = "\t")*

*trans\_data$X3 = log(trans\_data$X3)*

*trans\_model = glm(trans\_data$Y ~ trans\_data$X1 + trans\_data$X2 + trans\_data$X3,*

*family = "binomial")*

*print(summary(trans\_model))*

**Model Summary**

*Deviance Residuals:*

*Min 1Q Median 3Q Max*

*-1.57236 -0.00004 0.00000 0.00093 1.48872*

*Coefficients:*

*Estimate Std. Error z value Pr(>|z|)*

*(Intercept) -5.5860 7.3058 -0.765 0.4445*

*trans\_data$X1 0.3493 0.3636 0.960 0.3368*

*trans\_data$X2 0.1827 0.1084 1.685 0.0919 .*

*trans\_data$X3 8.3599 8.9293 0.936 0.3492*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*(Dispersion parameter for binomial family taken to be 1)*

*Null deviance: 91.4954 on 65 degrees of freedom*

*Residual deviance: 5.7925 on 62 degrees of freedom*

*AIC: 13.793*

*Number of Fisher Scoring iterations: 13*

As seen in the above model summary we have some improvements on the in terms of the Residual deviance and AIC as compared to the original model.

The log transformation is applied to the X3 covariate, since applying log transforms the space into a more concise space. We can only apply log transform to X3 since other covariates have negative values and hence log cannot be applied.

**References:**

1. <http://www.statisticssolutions.com/assumptions-of-logistic-regression/>
2. <https://en.wikipedia.org/wiki/Linear_regression>
3. <http://www.public.iastate.edu/~stat415/stephenson/stat415_chapter3.pdf>