**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).

**Answer Key**

**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**:

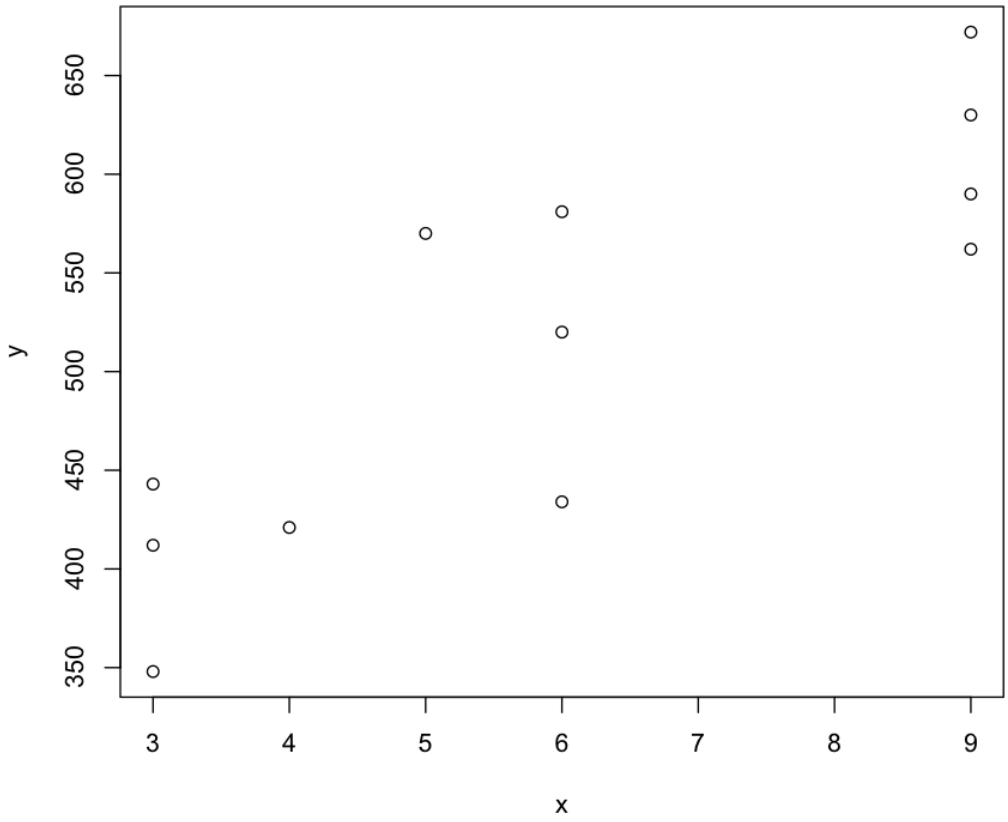
(1) All questions and sub-parts (Q1-Q3) of this h/w require hand calculations using the formulas that you learned from the course materials. You can use any calculator.
(2) For all project/programming questions, you need implementations in R and submit codes and plots (as requested). Please provide comments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

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




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

| **x** | **y** | **xy** | **$x^2$** |
|---|---|---|---|
| 6 | 520 | 3120 | 36 |
| 4 | 421 | 1684 | 16 |
| 6 | 581 | 3486 | 36 |
| 9 | 630 | 5670 | 81 |
| 3 | 412 | 1236 | 9 |
| 9 | 562 | 5058 | 81 |
| 6 | 434 | 2604 | 36 |
| 3 | 443 | 1329 | 9 |
| 9 | 590 | 5310 | 81 |
| 5 | 570 | 2850 | 25 |

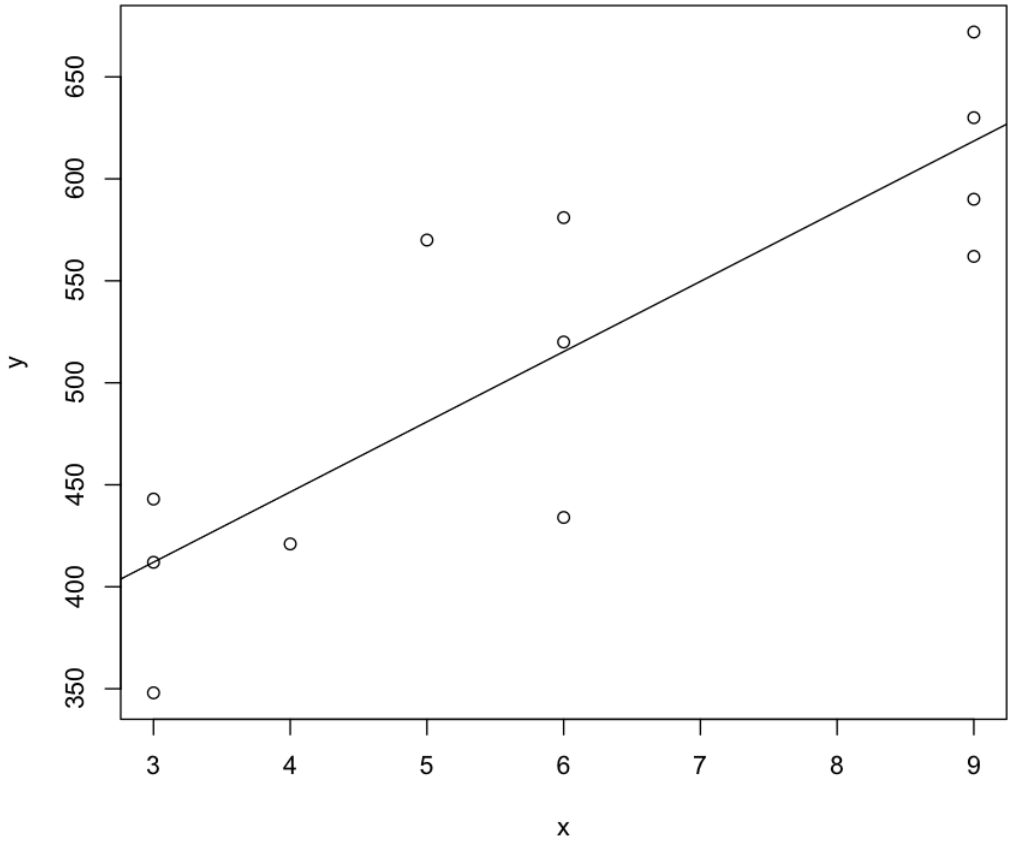
| | | | |
|---|---|---|---|
| 3 | 348 | 1044 | 9 |
| 9 | 672 | 6048 | 81 |

$\Sigma x = 72$
$\Sigma y = 6183$
$\Sigma xy = 39439$
$\Sigma x^2 = 500$

$a = [(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)]/[n(\Sigma x^2) - (\Sigma x)^2]$
$= (6183 * 500 - 72 * 39439)/(12 * 500 - 72^2)$
$= 308.69$

$b = [n(\Sigma xy) - (\Sigma x)(\Sigma y)]/[n(\Sigma x^2) - (\Sigma x)^2]$
$= (12 * 39439 - 72 * 6183)/(12 * 500 - 72^2)$
$= 34.43$

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




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

| x | y | y' | Residual | y' – mean (explained) | y – mean (total) |
|---|---|---|---|---|---|
| 6 | 520 | 515.25 | 4.75 | 0.00 | 4.75 |
| 4 | 421 | 446.40 | -25.40 | -68.86 | -94.24 |
| 6 | 581 | 515.25 | 65.75 | 0.00 | 65.75 |
| 9 | 630 | 618.53 | 11.48 | 103.28 | 114.75 |
| 3 | 412 | 411.97 | 0.03 | -103.28 | -103.25 |
| 9 | 562 | 618.53 | -56.53 | 103.28 | 46.75 |
| 6 | 434 | 515.25 | -81.25 | 0.00 | -81.25 |
| 3 | 443 | 411.97 | 31.03 | -103.28 | -72.25 |
| 9 | 590 | 618.53 | -28.53 | 103.28 | 74.75 |
| 5 | 570 | 480.82 | 89.18 | -34.43 | 54.75 |
| 3 | 348 | 411.97 | -63.97 | -103.28 | -167.25 |
| 9 | 672 | 618.53 | 53.48 | 103.28 | 156.75 |

sum of residuals = $0.02 \approx 0$

(e) How much of variation in y is explained by x? (5 points)
Mean of y is 515.25, see table in (d) for values.
$\Sigma(y' - \text{mean})^2 = 80590.16$
$\Sigma(y - \text{mean})^2 = 112192.25$
$r^2 = 80590.16 / 112192.25 = 0.7183$

(f) Compute the standard error of the estimate (5 points)
$s_{est} = \text{sqrt}(\Sigma(y - y')^2/(n - 2)) = \text{sqrt}(31599.88/(12 - 2)) = \text{sqrt}(3159.9) = 56.21$

(g) What are the predicted values for x = 2, 4, 6, 7, 10. (5 points)
$y'_i = x_i * 34.43 + 308.69$

| x | y |
|---|---|
| 2 | 377.54 |
| 4 | 446.39 |
| 6 | 515.25 |

| | |
|---|---|
| 7 | 549.67 |
| 10 | 652.95 |

(h) Test for significance of "r" (linear relationship) at $\alpha = 0.05$. (10 points)
$H_0$: rho = 0
$H_1$: rho $\neq$ 0
d.f. = 12 – 2 = 10
t = r * sqrt((n – 2) / (1 – $r^2$)) = sqrt(0.7183) * sqrt(10 / (1 - 0.7183)) = 5.050
CV for $\alpha = 0.05$ and 10 degrees of freedom: ±2.228
5.050 > 2.228, so reject – we claim that the correlation between x and y is significant.

(i) Compute the prediction interval for same $\alpha$ (5 points)

$t_{0.025}s_{est}$ sqrt(1 + 1/n + n(x – $x_{mean}$)$^2$/[n $\Sigma x^2 - (\Sigma x)^2$])
= 2.228 * 56.21 * sqrt(1 + 1/12 + 12(x – 6)$^2$/(12 * 500 – $72^2$))
= 125.24sqrt(1.083 + 0.0147(x – 6)$^2$)

Prediction interval: y' ± 125.24.05sqrt(1.083 + 0.0147(x – 6)$^2$)
= (34.43x + 308.69) ± 125.24sqrt(1.083 + 0.0147(x – 6)$^2$)

**Q2**. (20 points)

(a) 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).

**Simple:**
- Distribution of y for a given x is normal
- Std. dev. of that distribution is same for every x
- Mean is a linear function of x
- $y_i$'s are conditionally independent given the $x_i$'s (that is, residuals are independent of each other)

**Multiple:**
Same as simple, plus no (strong) multicollinearity ($x_i$'s are independent of each other)

**Logistic:**
- Errors are independent
- Log odds are a linear function of $x_i$'s
- No (strong) multicollinearity

(b) 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).

Simple (linearity assumption): Linear regression can be used on nonlinear data, but the fit will not be very good. This can be improved by (i) transforming the data, (ii) subset the data in such a way that for each subset we can fit a separate linear regression, or (iii) using higher-order polynomial terms (nonlinear regression).

Multiple (dependent variable is binary): If predictor variables are perfectly correlated, no unique solution will exist.

Logistic (assumes y to be ): If y is converted to binary (say from ordinal), then loss of information happens (means less powerful).

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

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

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

Since response variable is binary (0, 1), each error term can only take two values:

When $Y_i = 1$: $\epsilon_i = 1 - \beta_0 - \beta_1 X_i$

When $Y_i = 0$: $\epsilon_i = -\beta_0 - \beta_1 X_i$

Therefore, $\epsilon_i$ is not normally distributed

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

A similar problem with error terms $\epsilon_i$, is that they do not have equal variances when response variable is binary. We have

$$\sigma^2\{Y_i\} = E\{(Y_i - E\{Y_i\})^2\} = (1 - \pi_i)^2 \pi_i + (0 - \pi_i)^2(1 - \pi_i)$$

or

$$\sigma^2\{Y_i\} = \pi_i(1 - \pi_i) = (E\{Y_i\})(1 - E\{Y_i\})$$

The variance of $\epsilon_i$ is the same as that of $Y_i$ because $\epsilon_i = Y_i - \pi_i$ is a constant.

$$\sigma^2\{\epsilon_i\} = \pi_i(1 - \pi_i) = (E\{Y_i\})(1 - E\{Y_i\})$$

or

$$\sigma^2\{\epsilon_i\} = \pi_i(1-\pi_i) = (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i)$$

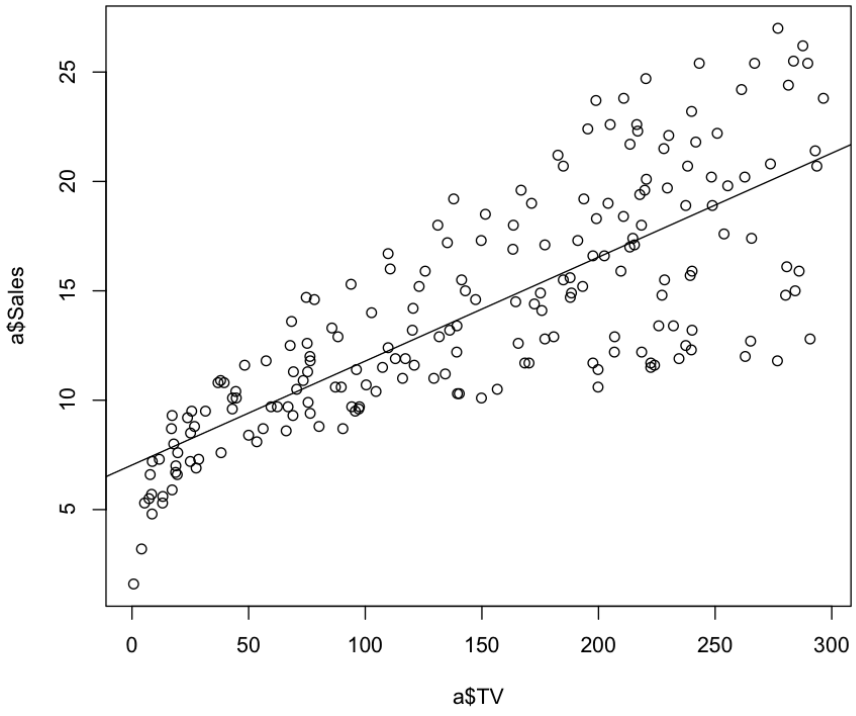
From the above equation, we can see that $\sigma^2\{\epsilon_i\}$ depends on $X_i$. Hence the error variances will differ at different values of X, and ordinary least squares will no longer be optimal.

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

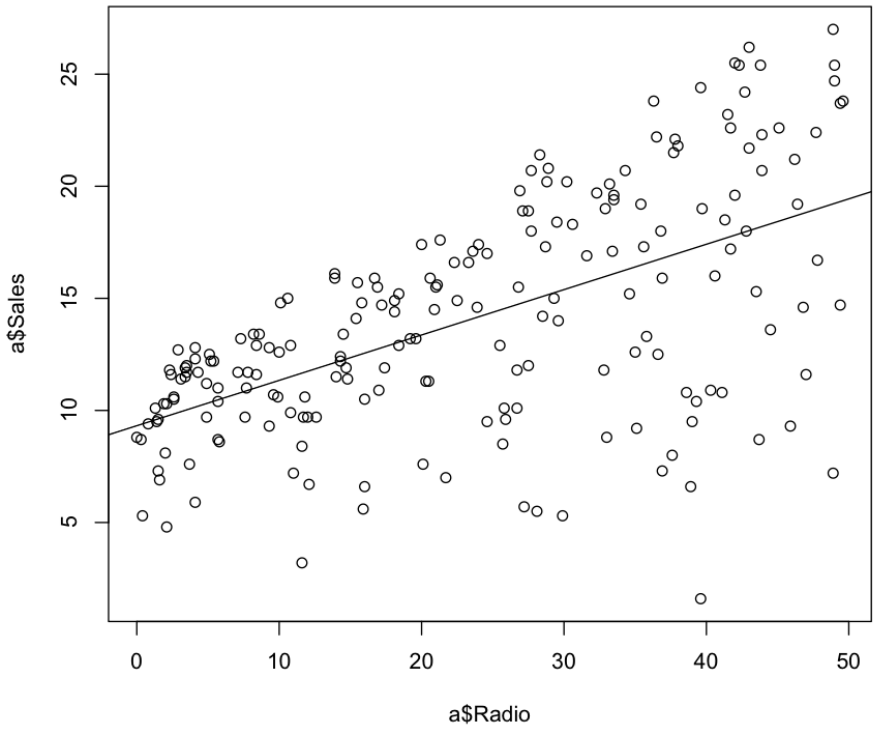
**Note:** You should provide answers (along with inline R code) in the pdf file. In addition submit R scripts.

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

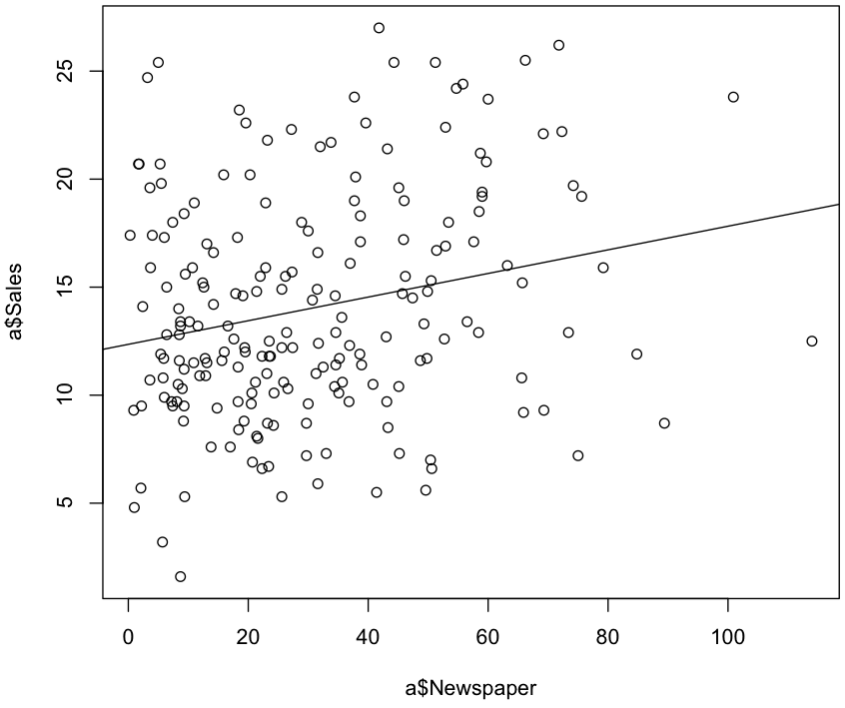
**(a)** Fit simple linear regression (separately) for each covariate. Provide scatter plots with fitted regression line. Which covariate provides best prediction?




residual std. err.: 3.259




residual std. err.: 4.275




residual std. err.: 5.092

Best prediction: TV

```
a <- read.csv("Advertising.csv");

plot(a$TV, a$Sales);
abline(lm(a$Sales ~ a$TV));
summary(lm(a$Sales ~ a$TV));

plot(a$Radio, a$Sales);
abline(lm(a$Sales ~ a$Radio));
summary(lm(a$Sales ~ a$Radio));

plot(a$Newspaper, a$Sales);
abline(lm(a$Sales ~ a$Newspaper));
summary(lm(a$Sales ~ a$Newspaper));
```

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

Sales = 2.939 + 0.046 * TV + 0.189 * Radio – 0.001 * Newspaper

```
summary(lm(a$Sales ~ a$TV + a$Radio + a$Newspaper));
```

**(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 $\beta_0 + \beta_1 X$ with $\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n$)

$P(Y \mid X_1, X_2, X_3) = s(-10.134 + 0.331X_1 + 0.181X_2 + 5.088X_3)$
where $s(x) = 1/(1 + \exp(-x))$

```
data <- read.csv('Downloads/hw3-q4c.txt', sep='\t');
summary(glm(data$Y ~ data$X1 + data$X2 + data$X3, family='binomial'));
```

**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**)

Possibilities include:

- Feature selection
- Data exploration (searching for outliers, etc.)