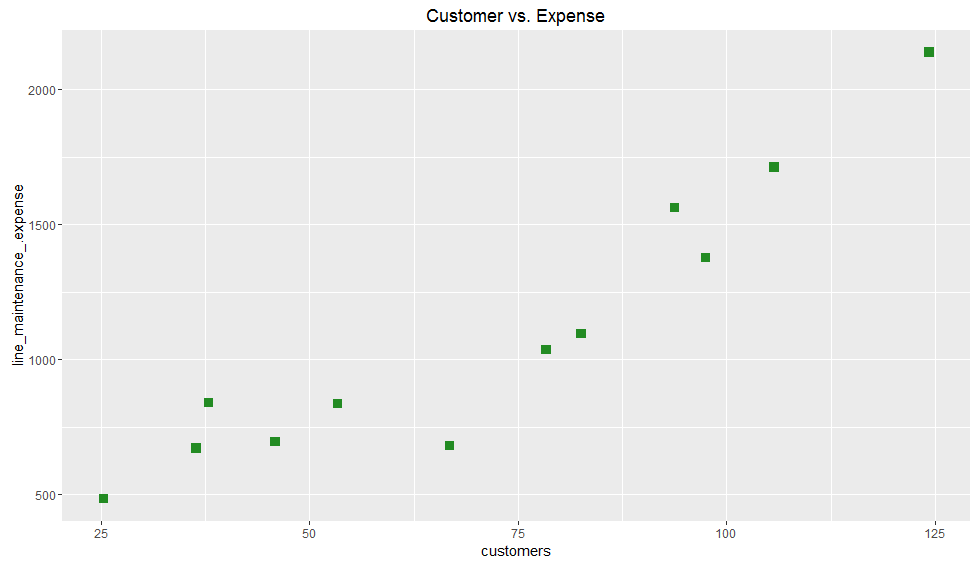
**Leonardo Ji**

**A. Enter data in spreadsheet.**

> phone\_service\_data <- read.csv("C:/Users/lj015625/Desktop/ML Class/HW3/phone\_service\_data.csv")

> attach(phone\_service\_data)

**B. Create a scatter plot of these data.**



**C. What is the estimated regression equation?**

Y = 33.321 + 15.016 \* Customers

> reg1 <- lm(line\_maintenance\_.expense~customers, data=phone\_service\_data)

Call:

lm(formula = line\_maintenance\_.expense ~ customers, data = phone\_service\_data)

Residuals:

Min 1Q Median 3Q Max

-354.49 -133.00 36.30 99.62 238.79

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 33.321 138.704 0.240 0.815

customers 15.016 1.807 8.311 0.00000842 \*\*\*

---

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

Residual standard error: 187.7 on 10 degrees of freedom

Multiple R-squared: 0.8735, Adjusted R-squared: 0.8609

F-statistic: 69.07 on 1 and 10 DF, p-value: 0.00000842

> anova(reg1)

Analysis of Variance Table

Response: line\_maintenance\_.expense

Df Sum Sq Mean Sq F value Pr(>F)

customers 1 2433767 2433767 69.07 0.00000842 \*\*\*

Residuals 10 352364 35236

---

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

Analysis of Variance Table

Response: line\_maintenance\_.expense

Df Sum Sq Mean Sq F value Pr(>F)

customers 1 2433767 2433767 69.07 0.00000842 \*\*\*

Residuals 10 352364 35236

---

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

**D. Interpret the value for R-squared obtained using the equation from question 3.**

R-squared = Regression Sum of Squared / total sum of squared = 0.8735

Which means 87.35% of the expenses can be explained by the linear equation model.

The residual is only 12.65%. The anova function gives value of sum of regression and sum of residual. Or we can calculate the number.

> totalSumOfSquares = sum((phone\_service\_data$line\_maintenance\_.expense - mean(phone\_service\_data$line\_maintenance\_.expense))^2)

> residualSumOfSquares = sum(reg1$residuals^2)

> regressionSumOfSquares = sum((reg1$fitted.values - mean(phone\_service\_data$line\_maintenance\_.expense))^2)

> R2 = regressionSumOfSquares/totalSumOfSquares

> R2

[1] 0.8735291

**E. Using equation from C what level of line maintenance expense would be expected for a phone company with 75,000 customers? Show how you arrive at this value.**

Y = 33.321 + 15.016 \* Customers

Y = 33.321 + 15.016 \* 75

= 1159.521

**F. Supposed a phone company with 75,000 customers reports a line maintenance expense of 1,500,000. Based on the results of the linear model, should Nolan view this amount as reasonable or excessive.**

SE = 187.7

95% confidence interval is: 1159.521± (1.96) \*(187.7) = (791.629, 1527.413)

The value is reasonable since it is inside the 95% confidence interval.

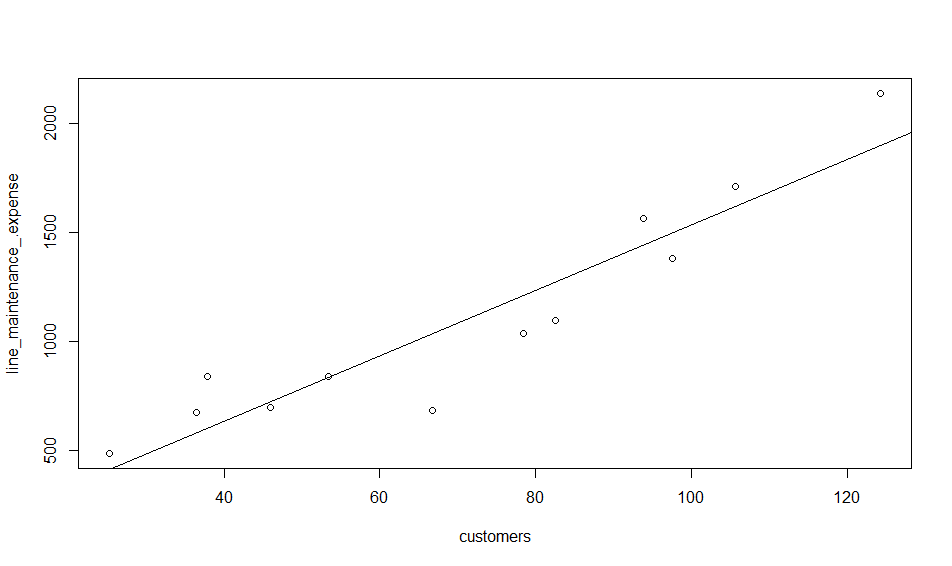
**G. Predict the value using model and draw fitted line.**

Yes, the predicted values are shown using predict function. The linear regression line appears appropriate.

> predict(reg1)

1 2 3 4 5 6 7 8 9 10 11 12

413.2249 579.9023 602.4263 722.5542 835.1740 1036.3882 1210.5736 1273.6407 1441.8197 1497.3789 1620.5099 1899.8072



**H. Use regression to estimate the parameters for the quadratic equation Y = b0 + b1\*X + b2\*X2**

Y = 707.47474 + (-7.39221) \*X + 0.15430 \* X2

> customerSquareTerm <- customers^2

> reg2 <- lm(line\_maintenance\_.expense~customers+customerSquareTerm, data=phone\_service\_data)

> summary(reg2)

Call:

lm(formula = line\_maintenance\_.expense ~ customers + customerSquareTerm,

data = phone\_service\_data)

Residuals:

Min 1Q Median 3Q Max

-220.32 -59.47 -16.24 67.12 191.37

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 707.47474 230.49267 3.069 0.0134 \*

customers -7.39221 7.03366 -1.051 0.3207

customerSquareTerm 0.15430 0.04761 3.241 0.0101 \*

---

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

Residual standard error: 134.4 on 9 degrees of freedom

Multiple R-squared: 0.9416, Adjusted R-squared: 0.9287

F-statistic: 72.61 on 2 and 9 DF, p-value: 0.000002801

F-statistic: 69.07 on 1 and 10 DF, p-value: 0.00000842

**I. Interpret the value of R2?**

R-squared = 0.9416

The R2 value is much better than single term linear regression model. Therefore adding a squared term to the model produced a much better model for the data.

**J. What is the adjusted R2?**

Adjusted R-squared: 0.9287

Adjusted R-squared taken into account the number of terms in the model. Since there are two terms in the model (customer, and customer squared) the adjusted R-squared is slightly lower than R-squared.

**K. What level of line maintenance expense would be expected for a phone company with 75,000 customers according to this new estimated regression function? Show how you arrive at this value.**

Y = 707.47474 + (-7.39221) \*X + 0.15430 \*X2

= 707.47474 - 7.39221 \* 75 + 0.15430 \* 75^2

= 1020.996

**L. Calculate the estimated line maintenance expense that would be predicted by the quadratic regression function for each company in the sample.**

> predict(reg2)

1 2 3 4 5 6 7 8 9 10 11 12

619.2208 642.8461 648.9551 693.2634 752.7404 902.2206 1076.3698 1149.6635 1371.7278 1453.5952 1650.0881 2172.7093

**M. Suppose that a phone company with 75,000 customers reports a line maintenance expense of $1,500,000. Based on the results of the quadratic model, should Nolan view this amount as reasonable or excessive?**

SE = 134.4

95% confidence interval is: 1020.996 ± (1.96) \*(134.4) = (757.572, 1284.42)

The value is excessive since the it is outside the 95% confidence interval.

**N. Which of the two regression functions would you suggest that Nolan use for prediction purpose?**

Model two is better because higher R-squared value, and lower residual standard error. Model two P value is smaller; the 95% confidence interval is narrower (residual error is correlated). The 75,000 customers phone company with expense of $1,500,000 was not rejected based on model one. Model two was able to reject it.

R code:

# scientific notation

options(scipen=999)

library(psych)

library(ggplot2)

phone\_service\_data <- read.csv("C:/Users/lj015625/Desktop/ML Class/HW3/phone\_service\_data.csv")

attach(phone\_service\_data)

summary(phone\_service\_data)

describe(phone\_service\_data)

ggplot(phone\_service\_data, aes(x = customers, y = line\_maintenance\_.expense)) +

geom\_point(color = "forestgreen", shape=15, size = 3) +

ggtitle("Customer vs. Expense")

reg1 <- lm(line\_maintenance\_.expense~customers, data=phone\_service\_data)

summary(reg1)

anova(reg1)

totalSumOfSquares = sum((phone\_service\_data$line\_maintenance\_.expense - mean(phone\_service\_data$line\_maintenance\_.expense))^2)

residualSumOfSquares = sum(reg1$residuals^2)

regressionSumOfSquares = sum((reg1$fitted.values - mean(phone\_service\_data$line\_maintenance\_.expense))^2)

R2 = regressionSumOfSquares/totalSumOfSquares

R2

predict(reg1)

plot(customers, line\_maintenance\_.expense)

abline(reg1)

customerSquareTerm <- customers^2

reg2 <- lm(line\_maintenance\_.expense~customers+customerSquareTerm, data=phone\_service\_data)

summary(reg2)

anova(reg2)

predict(reg2)