MIS-64036 Assignment 2

Steven Spence

11/7/2018

Business Analytics

Assignment II

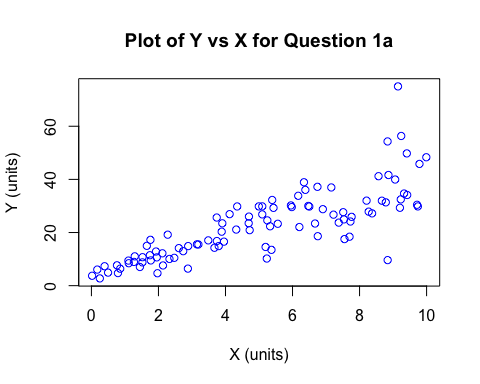
Instructions: Please answer all questions from 1 to 3. You should use R to solve the questions and include the screen shots in your submission. The Golden questions are optional and carries additional marks. This means that you will not lose marks if you do not answer that question. Your solution should include the calculation steps and your conclusion that is clearly expressed. Please use the link provided on the Blackboard, under the assessment section, to upload your submissions. Late submissions, up to two days, are subject to 30% penalty. Submissions made more than two days after the deadline will not be graded.

1. **Run the following code in R-studio to create two variables X and Y.**

# Creates X and Y variable - per instruction in Assignment 2  
  
set.seed(2017)  
X=runif(100)\*10  
Y=X\*4+3.45  
Y=rnorm(100)\*0.29\*Y+Y

1. **Plot Y against X. Include a screenshot of the plot in your submission. Using the File menu you can save the graph as a picture on your computer. Based on the plot do you think we can fit a linear model to explain Y based on X? (5 Marks)**

# Plot Y vs X on a graph with title, x-axis title, y-axis title, and blue colored points  
  
plot(X,Y, xlab = "X (units)", ylab = "Y (units)", main = "Plot of Y vs X for Question 1a", col = "blue")



From the graph shown above you can see a positive linear trend between X and Y; therefore, you can conclude that there is a linear model that can be create to explain this trend.

1. **Construct a simple linear model of Y based on X. Write the equation that explains Y based on X. What is the accuracy of this model? (5 Marks)**

# Creates a model for the created X and Y variables and displays the results  
  
Model = lm(Y~X)  
Model$coefficients

## (Intercept) X   
## 4.465490 3.610759

summary(Model)

##   
## Call:  
## lm(formula = Y ~ X)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26.755 -3.846 -0.387 4.318 37.503   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.4655 1.5537 2.874 0.00497 \*\*   
## X 3.6108 0.2666 13.542 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.756 on 98 degrees of freedom  
## Multiple R-squared: 0.6517, Adjusted R-squared: 0.6482   
## F-statistic: 183.4 on 1 and 98 DF, p-value: < 2.2e-16

The formula to explain Y based on X from our model is:

Y = 3.6108\*X + 4.4655

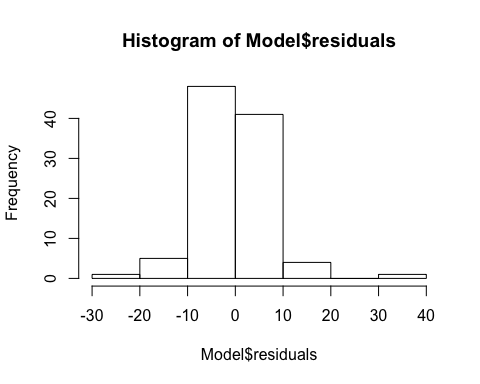
The R-squared value for this model is 0.6517 (adjusted R-squared is 0.6482), which means that the model explains 65.17% of the variability of the target response variable (Y in this case). This model is fairly robust and accurate for only having a single variable that accounts for approximately 65% of the variability.

1. **How the Coefficient of Determination, R2, of the model above is related to the correlation coefficient of X and Y? (5 marks)**

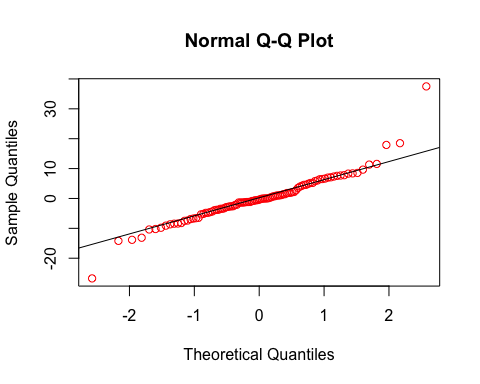
The coefficient of determination is the correlation coefficient squared, which is the proportion of variability of the dependent varibale explained by the independent variable. In this case, it states that 65.17% of the variability of Y is accounted for by the X variable.

1. **Investigate the appropriateness of using linear regression for this case (10 Marks). You may also find the story here relevant. More useful hints: #residual analysis, #pattern of residuals, #normality of residuals.**

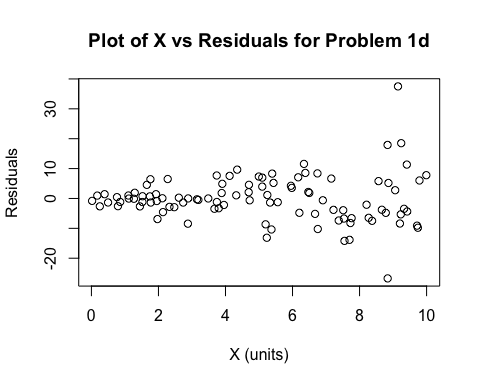
# Create a histogram of the residuals to determine if they are normal.  
  
hist(Model$residuals)



# Create a probability plot to also check for normality of the residuals.  
  
qqnorm(Model$residuals, col = "red")  
qqline(Model$residuals)



# Plots the residuals vs X to search for patterns  
  
plot(X,Model$residuals, xlab = "X (units)", ylab = "Residuals", main = "Plot of X vs Residuals for Problem 1d")



There are four assumptions that must be made for a regression model:

1. The model is linear

* Upon review of the graphs above, we can see that the residuals have a fairly linear distribution so it passes this assumption.

1. The error terms have constant variance

* Upon review of the residual vs X graph, we can see that the variances of the residuals appear to get larger as the X variable gets larger. So this is something that will have to be further reviewed to assess the validity of this model.

1. The error terms are independent

* Upon review of the residual plot vs X graph, we can see that the values appear to be independent of each other.

1. The error terms are normal

* Upon review of the histogram and normality plot, we can see the residuals are normal in this case.

Taking all of this information into account, we can state that the model passes three of the four assumptions; however, I would conclude that the team needs to review the model for additional improvement since we can see that there is nonconstant error variance as the value of X increases.

1. **We will use the ‘mtcars’ dataset for this question. The dataset is already included in your R distribution. The dataset shows some of the characteristics of different cars. The following shows few samples (i.e. the first 6 rows) of the dataset. The description of the dataset can be found here.**

# Displays the first six rows of the data set  
  
head(mtcars)

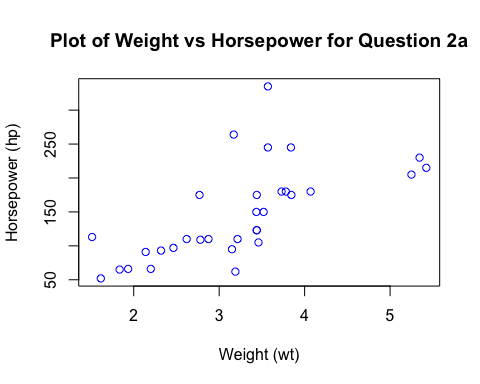
## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

1. **James wants to buy a car. He and his friend, Chris, have different opinions about the Horse Power (hp) of cars. James think the weight of a car (wt) can be used to estimate the Horse Power of the car while Chris thinks the fuel consumption expressed in Mile Per Gallon (mpg), is a better estimator of the (hp). Who do you think is right? Construct simple linear models using mtcars data to answer the question. (10 marks)**

# Creates a linear model for weight vs horsepower and displays a plot of the points  
  
Model2 = lm(hp~wt, data = mtcars)  
summary(Model2)

##   
## Call:  
## lm(formula = hp ~ wt, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -83.430 -33.596 -13.587 7.913 172.030   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.821 32.325 -0.056 0.955   
## wt 46.160 9.625 4.796 4.15e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 52.44 on 30 degrees of freedom  
## Multiple R-squared: 0.4339, Adjusted R-squared: 0.4151   
## F-statistic: 23 on 1 and 30 DF, p-value: 4.146e-05

plot(mtcars$wt,mtcars$hp, xlab = "Weight (wt)", ylab = "Horsepower (hp)", main = "Plot of Weight vs Horsepower for Question 2a", col = "blue")

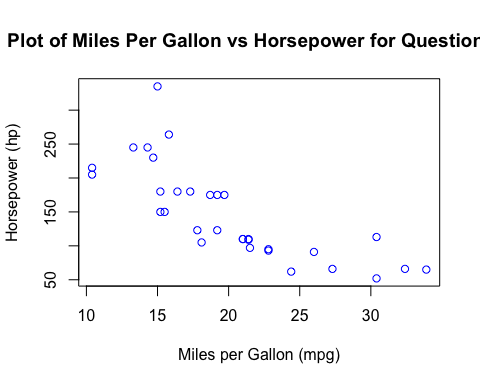


From this linear model we can see that weight results in a model that accounts for 43.39% of the variation in horsepower.

# Creates a linear model for mpg vs horsepower and displays a plot of the points  
  
Model3 = lm(hp~mpg, data = mtcars)  
summary(Model3)

##   
## Call:  
## lm(formula = hp ~ mpg, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -59.26 -28.93 -13.45 25.65 143.36   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 324.08 27.43 11.813 8.25e-13 \*\*\*  
## mpg -8.83 1.31 -6.742 1.79e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 43.95 on 30 degrees of freedom  
## Multiple R-squared: 0.6024, Adjusted R-squared: 0.5892   
## F-statistic: 45.46 on 1 and 30 DF, p-value: 1.788e-07

plot(mtcars$mpg,mtcars$hp, xlab = "Miles per Gallon (mpg)", ylab = "Horsepower (hp)", main = "Plot of Miles Per Gallon vs Horsepower for Question 2a", col = "blue")



From this linear model we can see that fuel efficiency results in a model that accounts for 60.24% of the variation in horsepower.

# Creates a linear model for weight and miles per gallon vs horsepower  
  
Model4 = lm(hp~wt+mpg, data = mtcars)  
summary(Model4)

##   
## Call:  
## lm(formula = hp ~ wt + mpg, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -59.42 -30.75 -12.07 24.82 141.84   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 349.287 103.509 3.374 0.00212 \*\*  
## wt -4.168 16.485 -0.253 0.80217   
## mpg -9.417 2.676 -3.519 0.00145 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 44.65 on 29 degrees of freedom  
## Multiple R-squared: 0.6033, Adjusted R-squared: 0.576   
## F-statistic: 22.05 on 2 and 29 DF, p-value: 1.505e-06

When we create a model that contains both weight and fuel efficiency, we see that the total model accounts for 60.33% of the variability in horsepower. Further analysis shows that fuel efficiency (mpg) is considered statistically significant in this model, while weight (wt) is not. Therefore, I would side with Chris and use fuel efficiency (mpg) as the estimator of horsepower.

1. **Build a model that uses the number of cylinders (cyl) and the mile per gallon (mpg) values of a car to predict the car Horse Power (hp).**

# Shows which variables are factor or numeric  
  
str(mtcars)

## 'data.frame': 32 obs. of 11 variables:  
## $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
## $ cyl : num 6 6 4 6 8 6 8 4 4 6 ...  
## $ disp: num 160 160 108 258 360 ...  
## $ hp : num 110 110 93 110 175 105 245 62 95 123 ...  
## $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
## $ wt : num 2.62 2.88 2.32 3.21 3.44 ...  
## $ qsec: num 16.5 17 18.6 19.4 17 ...  
## $ vs : num 0 0 1 1 0 1 0 1 1 1 ...  
## $ am : num 1 1 1 0 0 0 0 0 0 0 ...  
## $ gear: num 4 4 4 3 3 3 3 4 4 4 ...  
## $ carb: num 4 4 1 1 2 1 4 2 2 4 ...

Cyclinders is considered a numeric in this data frame; however, we need it to be considered a factor to get a more accurate model. As it currently stands, the model would consider the distance between the first and second category is the same as the distance between the second and third.

# Convert cylinder into a factor, rather than numeric, and verify it is indeed a factor now.  
  
mtcars$cyl = as.factor(mtcars$cyl)  
str(mtcars)

## 'data.frame': 32 obs. of 11 variables:  
## $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
## $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...  
## $ disp: num 160 160 108 258 360 ...  
## $ hp : num 110 110 93 110 175 105 245 62 95 123 ...  
## $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
## $ wt : num 2.62 2.88 2.32 3.21 3.44 ...  
## $ qsec: num 16.5 17 18.6 19.4 17 ...  
## $ vs : num 0 0 1 1 0 1 0 1 1 1 ...  
## $ am : num 1 1 1 0 0 0 0 0 0 0 ...  
## $ gear: num 4 4 4 3 3 3 3 4 4 4 ...  
## $ carb: num 4 4 1 1 2 1 4 2 2 4 ...

We can see from the code above that cylinders is now treated as a factor with three levels, rather than a numeric value.

# A multiple regression model is utilized in this case to build a model that represents horsepower as a result of cylinders and miles per gallon.  
  
Model5 = lm(hp~cyl+mpg, data = mtcars)  
summary(Model5)

##   
## Call:  
## lm(formula = hp ~ cyl + mpg, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -58.882 -20.904 -6.261 7.043 125.453   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 171.349 57.946 2.957 0.00625 \*\*  
## cyl6 16.623 23.197 0.717 0.47955   
## cyl8 88.105 28.819 3.057 0.00487 \*\*  
## mpg -3.327 2.133 -1.560 0.12995   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 37.01 on 28 degrees of freedom  
## Multiple R-squared: 0.7368, Adjusted R-squared: 0.7086   
## F-statistic: 26.12 on 3 and 28 DF, p-value: 2.888e-08

Model above treats 4 cylinder as the default value and shows the resulting coefficient values for 6 and 8 cylinders.

1. **Using this model, what is the estimated Horse Power of a car with 4 calendar and mpg of 22? (5 mark)**

# Predict the estimated horse power of a car with 4 cylinders and 22 mpg  
  
predict(Model5, data.frame(mpg = c(22), cyl = c("4")))

## 1   
## 98.15275

The prediction from this model is 98.15 horsepower for a car that has 4 cylinders and a fuel efficiency of 22 mpg.

1. **Construct an 85% confidence interval of your answer in the above question. Hint: use the predict function (5 mark)**

# Predict the estimated horse power of a car with 4 cylinders and 22 mpg at a 85% confidence interval  
  
predict(Model5, data.frame(mpg = c(22), cyl = c("4")), interval = "prediction", level = 0.85)

## fit lwr upr  
## 1 98.15275 39.06716 157.2383

From the model we can see that the lower limit of the 85% confidence interval is 39.07 horsepower and the upper limit is 157.24 horsepower.

1. **For this question, we are going to use BostonHousing dataset. The dataset is in ‘mlbench’ package, so we first need to instal the package, call the library and the load the dataset using the following commands:**

# Call and load the library for the BostonHousing dataset  
  
install.packages('mlbench')  
library(mlbench)  
data(BostonHousing)

You should have a dataframe with the name of BostonHousing in your Global environment now.

The dataset contains information about houses in different parts of Boston. Details of the dataset is explained here. Note the dataset is old, hence low house prices!

1. **Build a model to estimate the median value of owner-occupied homes (medv)based on the following variables: crime crate (crim), proportion of residential land zoned for lots over 25,000 sq.ft (zn), the local pupil-teacher ratio (ptratio) and weather the whether the tract bounds Chas River(chas). Is this an accurate model? (Hint check R2 ) (5 marks)**

library(mlbench)  
data(BostonHousing)  
  
# Check the data set to see which variable are considered numeric and which are factors.  
  
str(BostonHousing)

## 'data.frame': 506 obs. of 14 variables:  
## $ crim : num 0.00632 0.02731 0.02729 0.03237 0.06905 ...  
## $ zn : num 18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...  
## $ indus : num 2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...  
## $ chas : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ nox : num 0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...  
## $ rm : num 6.58 6.42 7.18 7 7.15 ...  
## $ age : num 65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...  
## $ dis : num 4.09 4.97 4.97 6.06 6.06 ...  
## $ rad : num 1 2 2 3 3 3 5 5 5 5 ...  
## $ tax : num 296 242 242 222 222 222 311 311 311 311 ...  
## $ ptratio: num 15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...  
## $ b : num 397 397 393 395 397 ...  
## $ lstat : num 4.98 9.14 4.03 2.94 5.33 ...  
## $ medv : num 24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...

# Create a linear model for median value based on crim, zn, ptratio, and chas.  
  
Model6 = lm(medv~crim+zn+ptratio+chas, data = BostonHousing)  
summary(Model6)

##   
## Call:  
## lm(formula = medv ~ crim + zn + ptratio + chas, data = BostonHousing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.282 -4.505 -0.986 2.650 32.656   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 49.91868 3.23497 15.431 < 2e-16 \*\*\*  
## crim -0.26018 0.04015 -6.480 2.20e-10 \*\*\*  
## zn 0.07073 0.01548 4.570 6.14e-06 \*\*\*  
## ptratio -1.49367 0.17144 -8.712 < 2e-16 \*\*\*  
## chas1 4.58393 1.31108 3.496 0.000514 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.388 on 501 degrees of freedom  
## Multiple R-squared: 0.3599, Adjusted R-squared: 0.3547   
## F-statistic: 70.41 on 4 and 501 DF, p-value: < 2.2e-16

Reviewing the R-squared values, we can see that the variables captured in this model (crim, zn, ptratio, and chas) capture 35.99% of the variability in median home value. This is a rather weak model in terms of accuracy and can be improved by adding more variables into the model.

1. **Use the estimated coefficient to answer these questions?**
2. **Imagine two houses that are identical in all aspects but one bounds the Chas River and the other does not. Which one is more expensive and by how much? (5 marks)**

Based on the coefficients, the resulting formula from our model is:

medv = 49.91868 - 0.26018*crim + 0.07073*zn - 1.49367*ptratio + 4.58393*chas1

Therefore, if the only difference between two houses is that one borders the Chas River, then we would only focus on the chas variable coefficient. The house that borders the river would be $4,583.93 more than the one that does not.

4.58393 (coeff of chas) \* 1 (value of chas) \* 1000 (medv in $1,000 units) = $4,583.93

1. **Imagine two houses that are identical in all aspects but in the neighborhood of one of them the pupil-teacher ratio is 15 and in the other one is 18. Which one is more expensive and by how much? (Golden Question: 10 extra marks if you answer)**

Based on the coefficients, the resulting formula from our model is:

medv = 49.91868 - 0.26018*crim + 0.07073*zn - 1.49367*ptratio + 4.58393*chas1

Therefore, if the only difference between two houses is the pupil-teacher ratio, then we would only focus on the ptratio variable coefficient. As a result, the house with the smaller pupil-teacher ratio value would be more expensive, because the coefficient is found to be negative in our model. The difference in values between the houses would be:

-1.49367 (coeff of ptratio) \* 0.03 (difference between ptratio values) \* 1000 (medv in $1,000 units) = $44.81

Therefore, the house with the lower pupil-teacher ratio would be $44.81 more expensive based on our model.

1. **Which of the variables are statistically important (i.e. related to the house price)? Hint: use the p-values of the coefficients to answer.(5 mark)**

Based on the model constructed from these variables, all of the variables (crim, zn, ptratio, and chas) were found to be statistically significant. This is true because all of the p-values calculated from our model at below the 0.05 threshold value for significance.

1. **Use the anova analysis and determine the order of importance of these four variables.(5 marks)**

# Returns the ANOVA results for the model used in this problem  
  
anova(Model6)

## Analysis of Variance Table  
##   
## Response: medv  
## Df Sum Sq Mean Sq F value Pr(>F)   
## crim 1 6440.8 6440.8 118.007 < 2.2e-16 \*\*\*  
## zn 1 3554.3 3554.3 65.122 5.253e-15 \*\*\*  
## ptratio 1 4709.5 4709.5 86.287 < 2.2e-16 \*\*\*  
## chas 1 667.2 667.2 12.224 0.0005137 \*\*\*  
## Residuals 501 27344.5 54.6   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Based on the ANOVA values returned, the order of importance of these variables are:

1. “crim” - accounts for 15.08% of variability in the model
2. “ptratio” - accounts for 11.02% of variability in the model
3. “zn” - accounts for 8.32% of variability in the model
4. “chas” - accounts for 1.56% of variability in the model

Additionally, the residuals in this model still account for 64.01% of variability in the model, so there is still a lot of room for improvement in the accuracy of this model.

