

Homework 3

Jiao Qu A20386614, Yuan-An Liu A20375099, Zhenyu Zhang A20287371

1. First, indicate some summary statistics about the Auto Data Set

a) How many rows are there?

Answer: 392 rows # check it with: `nrow(Auto)`

b) What are the average weight, cylinders, mpg, displacement?

Answer: average weight:2978, cylinders:5.472, mpg:23.45, displacement: 194.4

c) How many unique cars are there in the data set? Assume a car is defined by its make and model (i.e., its "name"). You can ignore "year."

Answer: 371 unique cars

2. Let's assume you want to predict the fuel efficiency (mpg) of the car. If this is the case, state your null hypothesis.

Answer: There is no relationship among mpg, cylinders, displacement, horsepower, weight, acceleration, year. It means the coefficient of all the predictors are all zero.

3. Which features in the data set are numeric / quantitative and which are not?

Numeric/Quantitative: cylinders, horsepower, weight, acceleration, displacement, mpg, year, origin

Not: name

4. Look at the data set. If you can only pick three features for predicting mpg, which would you pick? (There's no right answer right now, but I want you to practice "guessing", which you would need sometimes when building models.)

Answer: Horsepower, weight, cylinders

5. Compute the correlations coefficients of all the numerical features. Which three feature pairs are most correlated and what are their correlation coefficients?

Answer: Weight-displacement 0.9329944

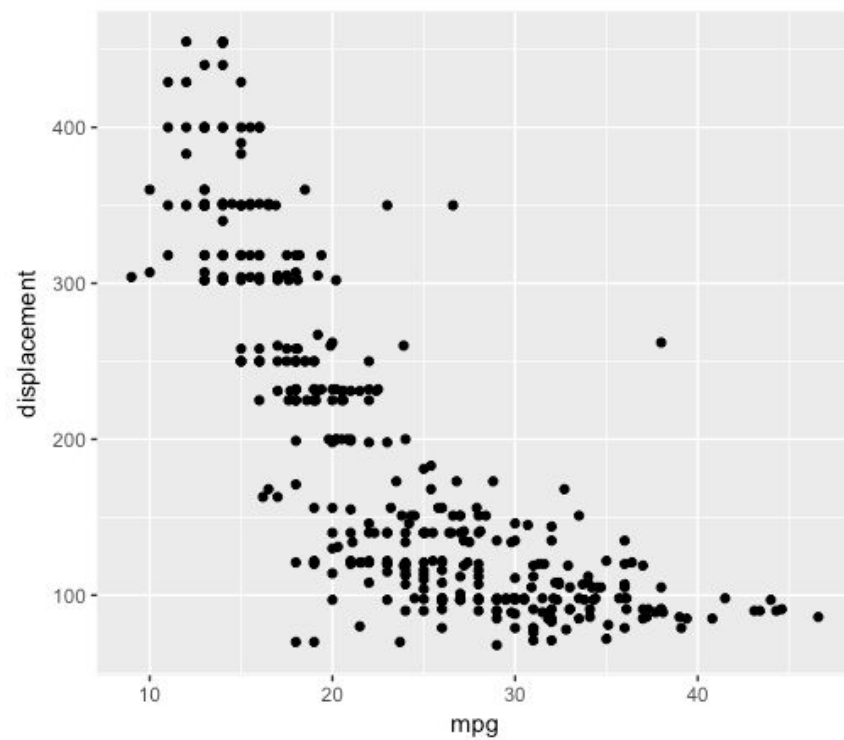
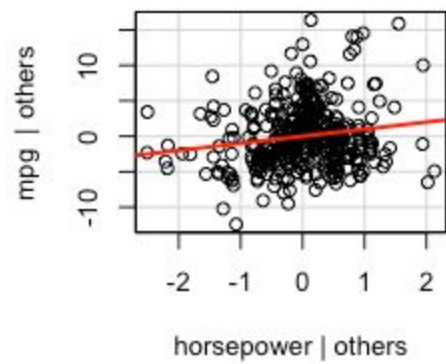
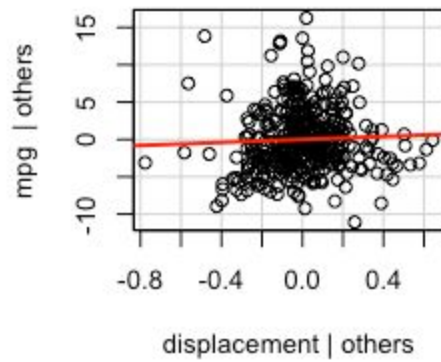
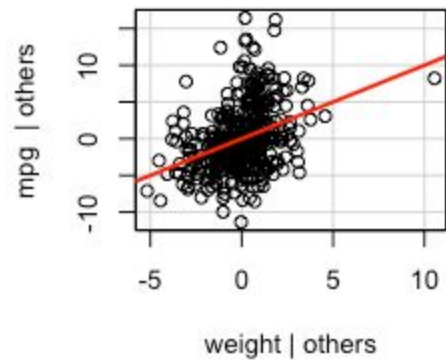
Horsepower-displacement 0.8972570

Horsepower-weight 0.8645377

6. Compute the "leverage statistic" for weight, displacement, and horsepower, assuming the model $\text{mpg} \sim \text{weight} + \text{displacement} + \text{horsepower}$. Plot the displacement values. Identify some possible high leverage points. What are the weight, displacement, horsepower, and leverage values for those points?

Answer: There are some values in "rstudent" > 3

Leverage Plots



7. Compute the variance inflation factor (VIF) for weight, displacement, and horsepower, assuming the model $\text{mpg} \sim \text{weight} + \text{displacement} + \text{horsepower}$. What are they?

Answer: VIF for Weight: 7.957383

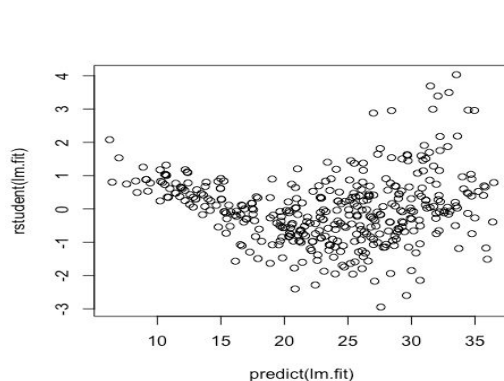
VIF for Displacement: 10.310539

VIF for Horsepower: 5.287295

8. Which numeric features have outliers? If any, indicate the number of outliers. Define an outlier as a value being beyond 3 standard deviations from the mean.

Model fitting

Answer: In p-values, it indicates that displacement, weight, year, and origin have a statistically significant relationship. Cylinders, horsepower, and acceleration have not. They have outliers.



There are outliers and leverage points, data values greater than 3, and point 14.

9. What is the best model you can fit, measured by adjusted R², with a single numeric feature for predicting mpg using only numeric features? Indicate your features and your adjusted R².

Best: $\text{MPG} \sim \text{Weight}$. Adjusted R-squared: 0.691

10. What is the best model you can fit, measured by adjusted R², with multiple regression (no polynomial regression, no interactions) for predicting mpg, without using the feature “name?” Indicate your features and your adjusted R². Must be greater than 0.8063.

Answer: $\text{mpg} \sim \text{horsepower} + \text{displacement} + \text{cylinders} + \text{acceleration} + \text{year}$

Multiple R-squared: 0.8858, Adjusted R-squared: 0.8487

11. Find the feature x, where the model $\text{mpg} \sim B_0 + B_1x + B_2x^2$ yields the highest percentage improvement in adjusted R². What is the model and what are the adjusted R² values with and without the x² feature and what is the percentage improvement?

Answer: Horsepower is the feature. The model with or without x² yields the same adjusted R² values. The adjusted R² values are 0.6049. So there is no improvement associated with x².

12. Find two features x and y where their interaction, $\text{mpg} \sim B_0 + B_1x + B_2y + B_3xy$, yields statistically significant B₁, B₂, B₃ coefficients. Do not use the feature, “name.” (Should be at least 0.70.)

Answer: Displacement and horsepower ; The Adjusted R-squared: 0.7446

13. Find a model that outperforms all the models above in terms of adjusted R² using any combination of transformed features and interactions. Do not use the feature, “name.” (Should be at least 0.8465.)

Answer: $\text{mpg} \sim \text{horsepower} + \text{horsepower} * \text{displacement} + \text{horsepower} * \text{cylinders} + \text{acceleration} + \text{horsepower} * \text{year}$
Multiple R-squared: 0.8598, Adjusted R-squared: 0.8569

14. Assume you have a model, $\text{mpg} \sim B_0 + B_1 * \text{cylinders} + B_2 * \text{displacement}$. You see that the coefficient for cylinders is not statistically significant. You remove cylinders from your model, but then your adjusted R2 decreases, so what should you do? Show a way you can do to the model, keeping both features, that makes both features statistically significant. Tell the adjusted R2 of the new Model.

Answer : For the model which is “ $\text{mpg} \sim B_0 + B_1 * \text{cylinders} + B_2 * \text{displacement}$ ”, the adjusted R2 is 0.6467. If you want to keep the adjusted R2 is nondecreasing under removing cylinders, you should add another feature to this model. So the new model is “ $\text{mpg} \sim B_0 + B_1 * \text{horsepower} + B_2 * \text{displacement}$ ”, and Multiple R-squared: 0.6643, Adjusted R-squared: 0.6626

ML engineering. You may want to use code from <https://bitbucket.org/waigen/ml>. Use the “advertising” data set for this question.

15. You built your own solver and now you have to tune it... Use gradient descent with 1000 iterations to solve $\text{Sales} \sim \text{TV} + \text{Radio} + \text{Newspaper}$. Try at least 5 different alpha values such that the best performing is not at the “end points” of alpha values. For example, if you try alphas of 0.001, 0.002, 0.004, 0.008, 0.016, and your best performing alpha is at 0.016, try larger alpha values until you find one where the error increases. Note that I do not want any “infinite error” solutions. Which of your alpha values yields the lowest adjusted R2? Show a plot where the X axis is the alpha value and the y axis is the adjusted R2. What is the percentage difference between your adjusted R2 value and what you get from lm?

```
Advertising=read.csv ("Advertising.csv", header =T,na.strings ="?")
```

```
# update of the Y-intercept, beta0
updateB0 = function(B0, B1, alpha, mydata) {
  n = nrow(mydata);
  newB0 = B0 - alpha * (1/n) * sum( B0 + B1 * mydata$x1 - mydata$y);
  return(newB0);
}
```

```
# update of the predictor coefficient, beta1
updateB1 = function(B0, B1, alpha, mydata) {
  n = nrow(mydata);
  newB1 = B1 - alpha * (1/n) * sum( (B0 + B1 * mydata$x1 - mydata$y) * mydata$x1);
  return(newB1);
}
```

```
set.seed(1);
mydata = Advertising;
mydata$Sales = mydata$TV + rnorm(nrow(mydata), 0, 1);
```

```
alpha = 0.16;
B0 = 0;
B1 = 0;
```

```

nIter = 1000;
errorList = NULL;

par(mfrow=c(1, 1));
plot(mydata, xlim = c(-1, 6), ylim = c(-1, 6))
abline(0, 1, lty = 2); # y = x
abline(a = B0, b = B1, lty = 2); # y = 0

for(i in 1:nIter) {
  error = computeCost(B0, B1, mydata);
  errorList = c(errorList, error);
  abline(a = B0, b = B1);
  print(paste("Error", round(error, 2), "B0", round(B0, 2), "B1", round(B1, 2)));
  newB0 = updateB0(B0, B1, alpha, mydata);
  newB1 = updateB1(B0, B1, alpha, mydata);
  B0 = newB0;
  B1 = newB1;
}

mylm = lm(y ~ x1, data = mydata);
summary(mylm);
abline(mylm, lty = 2, lwd = 2)

Advertising=read.csv("Advertising.csv", header = T, na.strings = "?")

# update of the Y-intercept, beta0
updateB0 = function(B0, B1, alpha, mydata) {
  n = nrow(mydata);
  newB0 = B0 - alpha * (1/n) * sum( B0 + B1 * mydata$x1 - mydata$y);
  return(newB0);
}

# update of the predictor coefficient, beta1
updateB1 = function(B0, B1, alpha, mydata) {
  n = nrow(mydata);
  newB1 = B1 - alpha * (1/n) * sum( (B0 + B1 * mydata$x1 - mydata$y) * mydata$x1);
  return(newB1);
}

set.seed(1);
mydata = Advertising;
mydata$Sales = mydata$TV + rnorm(nrow(mydata), 0, 1);

alpha = 0.16;
B0 = 0;
B1 = 0;
nIter = 1000;
errorList = NULL;

```

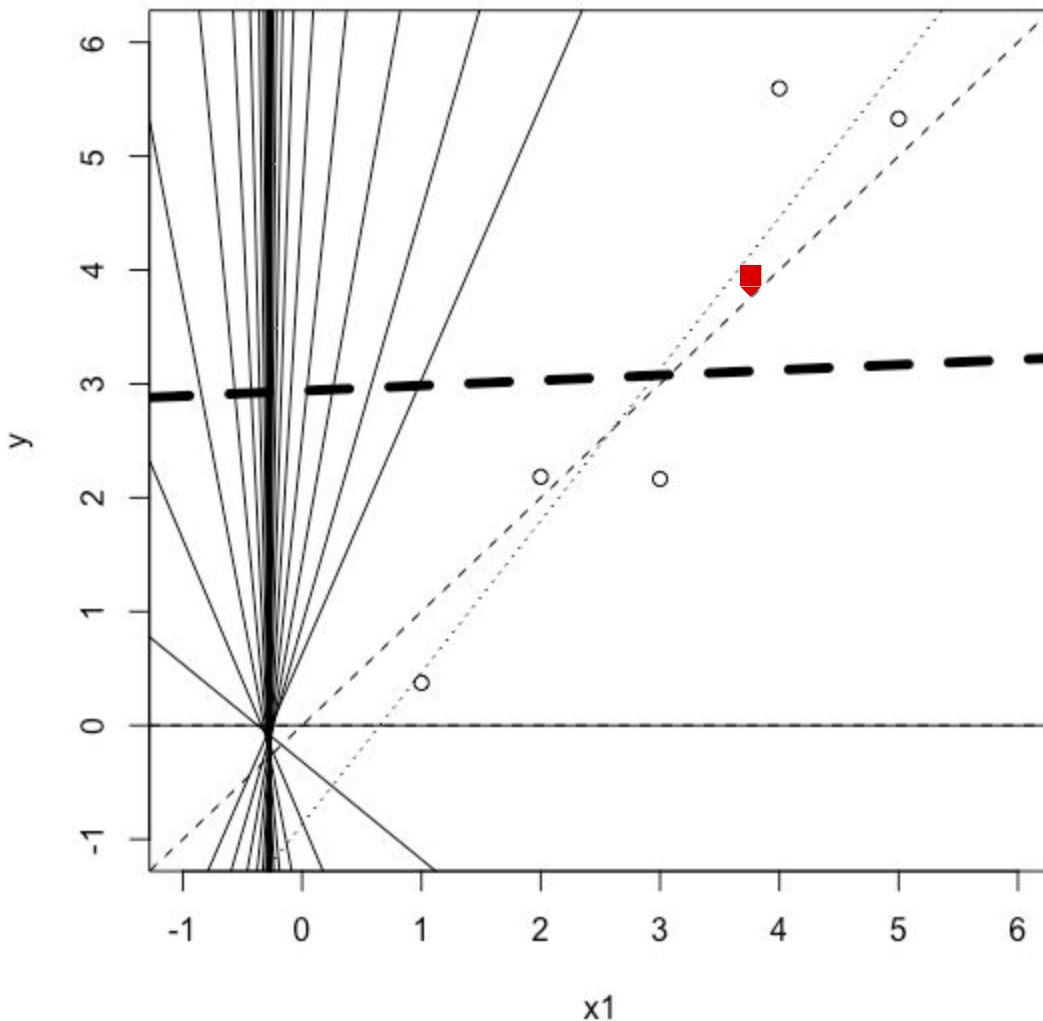
```

par(mfrow=c(1, 1));
plot(mydata, xlim = c(-1, 6), ylim = c(-1, 6))
abline(0, 1, lty = 2); # y = x
abline(a = B0, b = B1, lty = 2); # y = 0

for(i in 1:nIter) {
  error = computeCost(B0, B1, mydata);
  errorList = c(errorList, error);
  abline(a = B0, b = B1);
  print(paste("Error", round(error, 2), "B0", round(B0, 2), "B1", round(B1, 2)));
  newB0 = updateB0(B0, B1, alpha, mydata);
  newB1 = updateB1(B0, B1, alpha, mydata);
  B0 = newB0;
  B1 = newB1;
}


mylm = lm(y ~ x1, data = mydata);
summary(mylm);
abline(mylm, lty = 2, lwd = 2)

```



Multiple R-squared: 0.8972,

Adjusted R-squared: 0.8956

16. You want to see how well your home-made solution performs compared with more “out of the box” solutions. (In practice, if it is slower, you have to have a reason to use it.) Use `lm`, gradient descent as defined above (1000 iterations and the alpha of your choosing), and the normal equations to solve $\text{Sales} \sim \text{TV} + \text{Radio} + \text{Newspaper}$. Run 10 trials of each and time each trial. Report the average run times for each. Which solving technique is fastest? 

Answer: The average run time for each is 0.6562591 secs. The alpha choosing 0.20 is the fastest. The runtime is 0.5274041 secs

Critiquing real work. For the questions below, refer to the Chicago crime data from <http://www.chicagotribune.com/news/data/ct-crime-heat-analysis-htlmstory.html>.

17. Chicago crime data. Based on the plots, which crimes seem most correlated with temperature?

Indicate at least two.

Answer: 1. Shootings, other battery
2. Theft

18. Chicago crime data. Can you think of a confounding variable that might also explain the apparent increase in crime with temperature?

Answer: In our opinion, higher temperature can also indicate that people will be more active. Basically, it means there will be more pedestrians on the street. Thus, for people who want to commit crimes, they will tend to do these when they have more targets (victims). Therefore, in this case, we would say that numbers of pedestrian can be the confounding variable for the increase in crime with temperature.

19. Chicago crime data. What statistic(s) might one use to be more certain of the suggested crime trend?

Answer: We should go for using "Shootings, other battery" and "Theft." According to graphs, both of them have obviously increasing slopes. Positive slope coefficient indicates that it has stronger relationship.