HW4_repo

Yihao Zhang yzhng127 hwid:17

September 25, 2016

HW 4 Due Tuesday Sept 27, 2016. Upload R file to Moodle with name: HW2_490IDS_YOUR_CLASSID.R

Notice we are using the new system with your unique class ID. You should have received an email with

your unique class ID. Please make sure that ID is the only information on your hw that identifies you.

Do not remove any of the comments. These are marked by

Part 1: Linear Regression Concepts

These questions do not require coding but will explore some important concepts

from lecture 5.

"Regression" refers to the simple linear regression equation:

y = B0 + B1*x ## This homework will not discuss any multivariate regression.

1. (1 pt)

What is the interpretation of the coefficient B1?

(What meaning does it represent?)

Your answer

```
print("This represents the change of y corresponding to x")
```

```
## [1] "This represents the change of y corresponding to x"
```

2. (1 pt)

If the residual sum of squares (RSS) of my regression is exactly 0, what does

that mean about my model?

Your answer

```
print("This means your model fits the data perfectly. There's no error from your prediction and the observation")
```

```
## [1] "This means your model fits the data perfectly. There's no error from your prediction and the observation"
```

3. (2 pt)

Outliers are problems for many statistical methods, but are particularly problematic

for linear regression. Why is that? It may help to define what outlier means in this case.

(Hint: Think of how residuals are calculated)

Your answer

```
print("For linear regression, outliers are points that fall horizontally away from the center of the cloud (leverage points). Because distance to line is squared, models will change greated when containing these outliers.")
```

```
## [1] "For linear regression, outliers are points that fall horizontally away from the center of the cloud (leverage points). Because distance to line is squared, models will change greated when containing these outliers."
```

Part 2: Sampling and Point Estimation

The following problems will use the ggplot2movies data set and explore

the average movie length of films in the year 2000.

Load the data by running the following code

```
##install.packages("ggplot2movies")
library(ggplot2movies)
data(movies)
```

4. (2 pts)

Subset the data frame to ONLY include movies released in 2000.

Use the sample function to generate a vector of 1s and 2s that is the same

length as the subsetted data frame. Use this vector to split

the 'length' variable into two vectors, length1 and length2.

IMPORTANT: Make sure to run the following seed function before you run your sample

function. Run them back to back each time you want to run the sample function.

Check: If you did this properly, you will have 1035 elements in length1 and 1013 elements

in length2.

```
new_movies = subset(movies, movies$year == 2000)
set.seed(1848)
v1 = rep(1,1035)
v2 = rep(2,1013)
vec= c(v1,v2)
sample(vec)

##    [1] 1 1 2 1 2 2 1 1 2 2 2 2 1 1 2 1 1 1 2 1 1 2 2 2 1 2 1 2 1 2 2 1 1 2
##   [35] 1 1 1 2 1 2 1 2 2 2 1 1 2 2 2 1 2 1 2 2 1 2 1 1 2 2 1 1 2 2 1 1 2 1
##   [69] 2 2 1 2 1 1 1 1 1 1 1 1 2 2 1 2 2 2 2 1 2 1 2 1 2 2 2 1 1 1 2 1 2 1
##  [103] 2 2 2 2 1 2 2 2 1 2 2 2 1 1 2 2 1 2 1 1 2 2 1 2 1 2 1 1 2 2 2 2 1 2
##  [137] 1 1 1 1 1 2 1 1 1 1 1 2 2 2 1 2 1 2 1 1 1 2 1 1 2 1 1 2 2 1 2 2 1 2
##  [171] 1 2 1 2 2 1 2 1 1 1 2 2 1 1 2 2 1 2 1 2 2 2 1 1 2 1 1 1 1 2 2 1 2 1
##  [205] 1 2 2 1 2 1 2 1 1 1 2 1 1 1 1 1 1 2 2 2 1 1 1 1 1 2 1 1 1 1 1 2 1 1
##  [239] 1 2 1 1 2 2 2 2 2 2 2 1 2 2 2 1 2 1 1 1 2 1 1 2 1 2 2 2 2 1 1 1 2 2
##  [273] 2 2 2 2 1 1 2 2 2 2 2 1 2 2 1 1 2 1 1 1 2 2 1 1 1 1 1 1 2 2 1 2 1 1
##  [307] 2 1 2 1 1 1 2 2 2 2 1 1 1 2 2 1 1 1 2 2 1 1 1 1 1 1 1 2 1 1 1 2 1 1
##  [341] 2 2 2 1 1 1 2 2 2 1 2 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 2 2
##  [375] 2 2 2 1 2 1 1 1 2 2 1 1 1 2 1 1 2 1 2 1 2 2 1 2 2 1 2 2 1 2 2 1 1 1
##  [409] 2 1 1 1 2 1 1 2 1 1 1 1 1 1 2 1 1 1 1 2 1 2 1 2 1 2 1 1 2 1 1 1 2 1
##  [443] 1 1 1 2 2 2 1 1 1 1 2 2 1 2 1 1 1 1 2 1 1 2 2 2 2 1 2 2 1 1 1 2 2 1
##  [477] 1 2 1 1 1 1 1 2 2 2 2 2 1 2 2 1 2 2 2 1 1 2 2 1 2 1 2 2 2 1 1 1 1 2
##  [511] 2 2 1 1 1 1 1 1 2 2 2 1 2 1 1 2 2 2 1 1 1 1 1 2 2 2 1 2 2 1 1 1 2 2
##  [545] 1 2 1 1 2 2 2 1 2 1 1 2 2 1 1 2 2 2 1 2 2 2 1 1 1 1 2 1 2 2 1 1 1 1
##  [579] 2 1 2 2 1 2 2 1 1 2 2 1 1 2 2 1 2 1 2 2 1 2 2 1 1 2 1 2 2 2 2 1 1 2
##  [613] 1 1 2 1 1 2 1 1 1 1 1 2 1 2 1 1 2 2 1 2 1 2 2 2 1 2 1 1 2 2 1 2 1 1
##  [647] 2 2 1 1 2 2 1 2 1 1 2 1 1 1 1 2 2 2 1 2 2 1 1 2 2 1 1 2 2 2 1 2 1 1
##  [681] 2 1 2 2 1 2 2 1 1 2 1 1 1 2 2 2 1 1 1 2 2 1 1 2 1 1 1 2 2 1 1 2 1 2
##  [715] 1 1 2 1 1 1 1 2 2 2 2 2 1 2 1 1 2 1 2 1 2 1 1 2 2 1 1 2 1 2 2 1 2 1
##  [749] 1 2 1 2 2 2 1 1 2 2 2 1 2 1 1 2 1 1 1 2 1 2 2 1 2 1 2 2 2 2 1 1 2 1
##  [783] 2 2 1 1 1 1 2 1 2 2 2 2 1 1 2 2 2 1 2 2 1 1 1 2 1 1 2 2 2 2 2 1 1 1
##  [817] 2 2 2 1 2 1 2 1 1 1 1 1 1 2 2 2 1 1 2 2 1 1 2 1 1 1 1 2 1 1 1 1 2 2
##  [851] 1 2 2 2 2 1 2 2 2 2 1 1 1 1 2 1 2 1 2 2 2 2 2 2 2 1 1 1 1 2 2 2 2 2
##  [885] 2 1 1 1 2 2 1 2 1 1 2 1 2 1 1 1 1 1 1 2 1 1 2 2 2 1 2 2 1 2 2 2 1 1
##  [919] 1 2 2 2 1 2 1 1 1 1 1 2 1 1 1 1 1 2 2 1 2 2 2 2 1 1 2 1 2 2 2 1 2 1
##  [953] 1 2 2 1 2 2 2 1 1 2 1 1 2 2 1 1 1 1 1 2 2 1 2 1 2 1 1 2 2 2 2 2 2 2
##  [987] 2 2 2 1 1 2 2 2 2 2 1 2 1 1 1 2 2 2 1 2 1 1 2 2 1 1 1 1 2 2 1 2 2 2
## [1021] 1 2 1 2 2 2 1 2 2 2 1 2 2 2 2 1 1 1 1 2 1 2 2 2 2 1 1 1 2 1 2 1 2 1
## [1055] 1 1 2 1 1 1 2 1 2 1 2 2 1 2 2 1 2 1 1 1 1 2 1 2 1 2 2 2 1 2 2 2 2 2
## [1089] 2 1 1 2 1 1 1 1 2 2 2 2 1 1 2 2 1 1 2 1 2 2 2 1 1 2 2 1 1 1 2 1 1 2
## [1123] 1 2 2 2 1 2 1 2 1 1 2 1 1 1 1 2 1 1 1 1 1 1 2 1 2 1 1 1 1 2 1 2 1 2
## [1157] 1 2 2 1 2 1 2 2 1 1 1 1 1 2 1 1 2 2 2 2 1 1 1 1 1 2 2 1 2 2 1 2 2 1
## [1191] 2 2 2 1 2 1 2 2 1 2 1 2 2 2 2 2 2 1 2 2 2 1 2 1 2 1 2 2 1 1 2 2 2 1
## [1225] 2 2 1 2 1 1 1 2 1 1 2 2 2 1 2 1 2 2 1 1 2 2 1 1 1 2 2 2 2 2 2 2 2 2
## [1259] 2 1 2 2 2 2 1 1 2 2 1 2 2 2 1 2 1 1 1 2 2 2 2 2 1 1 2 2 2 1 1 2 2 1
## [1293] 2 1 1 1 1 2 1 1 2 2 2 1 1 2 1 1 1 1 2 2 2 1 2 1 1 1 2 1 1 1 1 2 1 1
## [1327] 1 1 1 2 1 1 2 1 2 1 2 1 1 2 2 2 1 2 1 2 2 2 2 1 1 1 1 1 1 1 1 1 2 2
## [1361] 2 1 1 2 1 2 1 1 2 1 1 2 1 1 2 1 2 2 2 2 1 1 1 2 1 1 1 1 1 1 2 1 2 1
## [1395] 2 1 1 2 1 2 2 2 2 1 1 1 1 2 1 2 1 1 1 2 1 1 1 2 1 2 1 2 2 1 2 1 1 2
## [1429] 1 1 2 1 2 1 2 2 2 2 1 2 1 1 1 2 2 1 1 2 2 1 1 1 2 2 2 2 2 2 2 2 1 2
## [1463] 1 2 1 1 2 1 2 1 1 1 2 1 2 1 2 1 1 1 2 1 2 2 2 1 1 1 1 2 2 2 2 1 1 2
## [1497] 1 2 1 2 1 2 1 1 1 1 2 2 2 1 1 2 1 2 1 2 1 2 1 1 2 2 2 1 1 2 2 2 2 2
## [1531] 1 2 1 2 2 2 1 2 1 1 2 2 2 1 2 2 2 1 1 1 1 2 1 1 1 2 1 2 2 1 2 2 2 2
## [1565] 1 2 1 1 1 2 2 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 1 2 2 1 1 1 2 1 2 1 2 1
## [1599] 2 1 1 1 1 2 1 2 1 2 2 2 2 1 2 1 1 2 1 1 2 2 2 1 2 1 2 1 1 2 2 1 1 2
## [1633] 2 1 1 2 1 2 2 1 1 2 2 1 2 2 1 2 2 2 1 2 1 2 1 2 2 1 2 2 1 1 2 2 2 2
## [1667] 1 2 1 2 2 2 1 1 1 1 1 1 2 1 2 2 2 1 1 2 2 2 1 2 2 2 1 2 1 2 1 1 1 2
## [1701] 1 1 1 2 1 1 1 2 1 2 2 2 2 2 1 2 2 2 2 1 2 2 1 1 2 2 1 1 2 1 2 1 2 1
## [1735] 2 2 1 2 2 2 2 1 1 1 2 2 2 1 1 2 2 1 1 1 2 2 2 1 2 2 2 2 1 2 2 1 2 1
## [1769] 2 1 1 2 1 2 1 1 1 2 1 2 2 2 2 2 1 2 1 2 1 1 2 2 1 1 2 1 2 2 1 1 1 2
## [1803] 2 1 2 2 2 1 2 2 2 2 2 1 1 2 2 2 1 1 2 1 2 2 2 2 2 1 2 1 2 2 2 2 2 2
## [1837] 1 2 2 1 2 2 2 1 1 1 2 1 2 2 1 1 2 2 1 2 1 1 2 1 1 1 2 1 1 1 1 2 1 1
## [1871] 1 1 2 1 2 2 2 1 1 1 2 2 2 2 2 2 2 1 1 1 1 1 2 1 2 2 2 2 1 1 2 1 2 1
## [1905] 1 1 2 1 1 2 2 1 1 1 1 1 2 1 2 1 2 2 1 2 2 1 2 1 1 1 2 1 1 1 1 2 1 2
## [1939] 1 2 1 2 1 2 2 1 2 1 2 1 2 2 2 1 1 2 2 1 1 2 1 1 1 2 2 1 2 1 2 2 1 1
## [1973] 1 1 1 1 2 2 2 1 2 2 1 1 2 2 1 2 2 1 2 1 2 2 1 1 1 2 1 1 2 2 1 2 2 2
## [2007] 1 2 2 1 1 1 1 2 2 2 1 1 1 2 1 1 1 1 1 2 2 2 2 2 1 1 2 1 1 1 1 1 1 2
## [2041] 2 2 1 2 2 2 2 2
```

```
length1 = as.numeric(new_movies$length[vec == 1])
length2 = as.numeric(new_movies$length[vec == 2])
# sample(...)
```

5. (3 pts)

Calculate the mean and the standard deviation for each of the two vectors, length1 and length2. Use this information to create a 95% confidence interval for your sample means. Compare the confidence intervals -- do they seem to agree or disagree?

Your answer here

```
m1 = mean(length1)
m2 = mean(length2)
std1 = sd(length1)
std2 = sd(length2)
err1 = qt(0.95,df = length(length1)-1)*std1/sqrt(length(length1))
err2 = qt(0.95,df = length(length2)-1)*std2/sqrt(length(length2))
ci1 = c(m1-err1,m1+err1)
ci2 = c(m2-err2,m2+err2)
print("yes they agree. The left and right of confidence intervals are very
close")
```

```
## [1] "yes they agree. The left and right of confidence intervals are very
close"
```

6. (4 pts)

Draw 100 observations from a standard normal distribution. Calculate the sample mean.

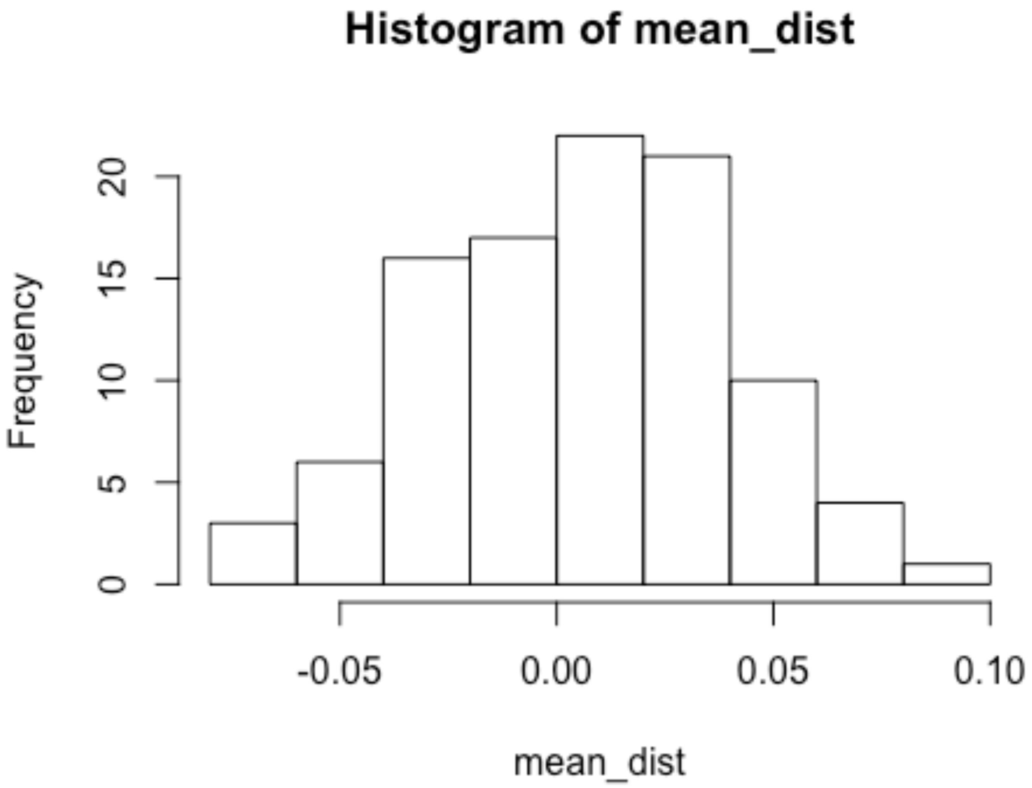
Repeat this 100 times, storing each sample mean in a vector called mean_dist.

Plot a histogram of mean_dist to display the sampling distribution.

How closely does your histogram resemble the standard normal? Explain why it does or does not.

Your answer here

```r
mean_dist = c()
for (i in 1:100){
  new = as.numeric(mean(rnorm(1000,0,1)))
  mean_dist[i] = new
}
hist(mean_dist)
```



```
print("This graph represents standard normal distribution pretty well.
Because of central limit theorem. The mean of a standard normal distribution
will be roughly a normal distribution. And the expected mean is 0 and
expected std is 1. So it's roughly a standard normal distribution")
```

```
## [1] "This graph represents standard normal distribution pretty well.
Because of central limit theorem. The mean of a standard normal distribution
will be roughly a normal distribution. And the expected mean is 0 and
expected std is 1. So it's roughly a standard normal distribution"
```

7. (3 pts)

Write a function that implements Q6.

Your answer here

```
HW.Bootstrap=function(distn,n,reps){
  set.seed(1848)

  #more lines here
  mean_dist = c()
  for (i in 1:reps){
    new = as.numeric(mean(rnorm(n,0,1)))
    mean_dist[i] = new
  }
}
```

Part 3: Linear Regression

This problem will use the Boston Housing data set.

Before starting this problem, we will declare a null hypthosesis that the

crime rate has no effect on the housing value for Boston suburbs.

That is: H0: B1 = 0

HA: B1 =/= 0

We will attempt to reject this hypothesis by using a linear regression

Load the data

```
housing <- read.table(url("https://archive.ics.uci.edu/ml/machine-learning-
databases/housing/housing.data"),sep="")
names(housing) <-
c("CRIM","ZN","INDUS","CHAS","NOX","RM","AGE","DIS","RAD","TAX","PTRATIO","B"
,"LSTAT","MEDV")
```

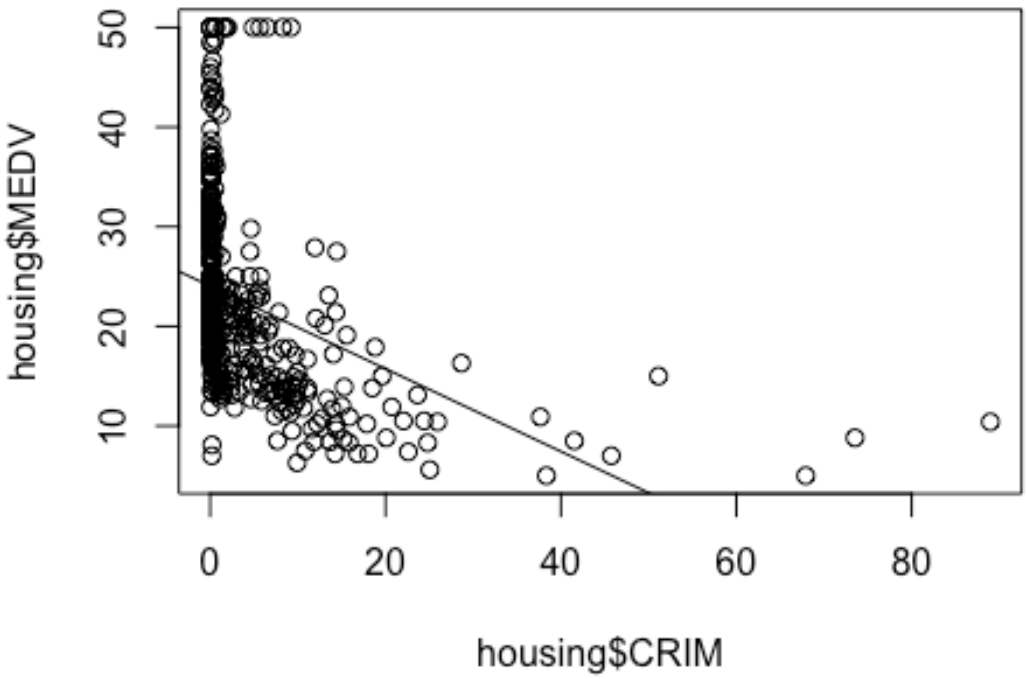
7. (2 pt)

Fit a linear regression using the housing data using CRIM (crime rate) to predict

MEDV (median home value). Examine the model diagnostics using plot(). Would you consider this a good

model or not? Explain.

```
line = lm(housing$MEDV ~ housing$CRIM)
plot(housing$CRIM,housing$MEDV,type = "p")
abline(line)
```



```
print("no,because lots of datas are outliers, especially when CRIM is 0.")

## [1] "no,because lots of datas are outliers, especially when CRIM is 0."
```

8. (2 pts)

Using the information from summary() on your model, create a 95% confidence interval

for the CRIM coefficient

```
summary(line)

##
## Call:
## lm(formula = housing$MEDV ~ housing$CRIM)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.957  -5.449  -2.007   2.512  29.800
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.03311    0.40914   58.74   <2e-16 ***
## housing$CRIM -0.41519    0.04389   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.484 on 504 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491
## F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
err_crim = qt(0.95,df = length(housing$CRIM)-
1)*sd(housing$CRIM)/sqrt(length(housing$CRIM))
ci_crim = c(mean(housing$CRIM)-err_crim, mean(housing$CRIM)+err_crim)
```

9. (2 pts)

Based on the result from question 8, would you reject the null hypothesis or not?

(Assume a significance level of 0.05). Explain.

Your answer

```
print("Yes I would reject the null hypothesis. Because firstly B1 is not in
the confidance interval. Secondly the p-value is really small, which means we
can reject the null hypothesis")
```

```
## [1] "Yes I would reject the null hypothesis. Because firstly B1 is not in
the confidance interval. Secondly the p-value is really small, which means we
can reject the null hypothesis"
```

10. (1 pt)

Pretend that the null hypothesis is true. Based on your decision in the previous

question, would you be committing a decision error? If so, which one?

Your answer

```
print("I would then associate housing value with other parameters and ignore
crime rate. Then I will make a mistake in predicting housing value because I
ignored one important aspect")
```

```
## [1] "I would then associate housing value with other parameters and ignore
crime rate. Then I will make a mistake in predicting housing value because I
ignored one important aspect"
```

11. (1 pt)

Use the variable definitions from this site:

https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names

Discuss what your regression results mean in the context of the data (using appropriate units)

(Hint: Think back to Question 1)

Your answer

```
print("The result means that housing value has a negative relationship with
the crime rate. Because the coefficient b1 is negative")
```

```
## [1] "The result means that housing value has a negative relationship with
the crime rate. Because the coefficient b1 is negative"
```

12. (2 pt)

Describe the LifeCycle of Data for Part 3 of this homework.

```
print("Identify the problem: The relationship between housing value and other
conditions. Design data requirement: Use online housing data. Pre-processing
data: choose only crime rate. Analysis: Do linear regression. Visualize:
plot")
```

```
## [1] "Identify the problem: The relationship between housing value and
other conditions. Design data requirement: Use online housing data. Pre-
processing data: choose only crime rate. Analysis: Do linear regression.
Visualize: plot"
```