STAT 632 Homework 3

Winnie Lu

2022-03-09

title: “Homework 3 STAT632” author: “Winnie Lu” date: “3/3/2022” output: pdf_document: default html_document: default —

Exercise 1

(a)

```
library(ISLR)
```

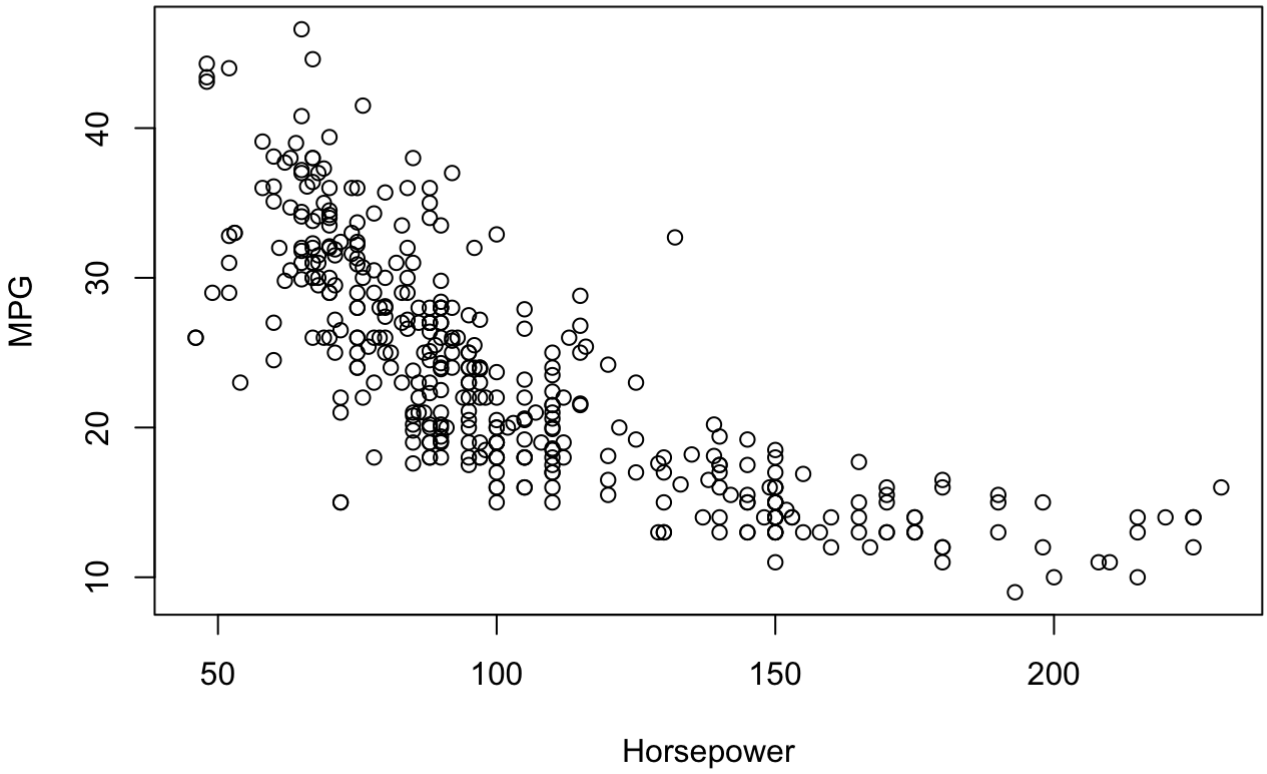
```
library(ggplot2)
```

```
head(Auto)
```

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                       name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

```
help(Auto)
```

```
cars <-plot(mpg ~ horsepower, data=Auto, xlab="Horsepower", ylab="MPG")
```



(b)

```
lm1 <-lm(mpg ~ horsepower + I(horsepower^2), data=Auto)
summary(lm1)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower + I(horsepower^2), data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7135  -2.5943  -0.0859   2.2868  15.8961
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     56.9000997  1.8004268   31.60   <2e-16 ***
## horsepower      -0.4661896  0.0311246  -14.98   <2e-16 ***
## I(horsepower^2)  0.0012305  0.0001221   10.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.374 on 389 degrees of freedom
## Multiple R-squared:  0.6876, Adjusted R-squared:  0.686
## F-statistic:   428 on 2 and 389 DF,  p-value: < 2.2e-16
```

(c)

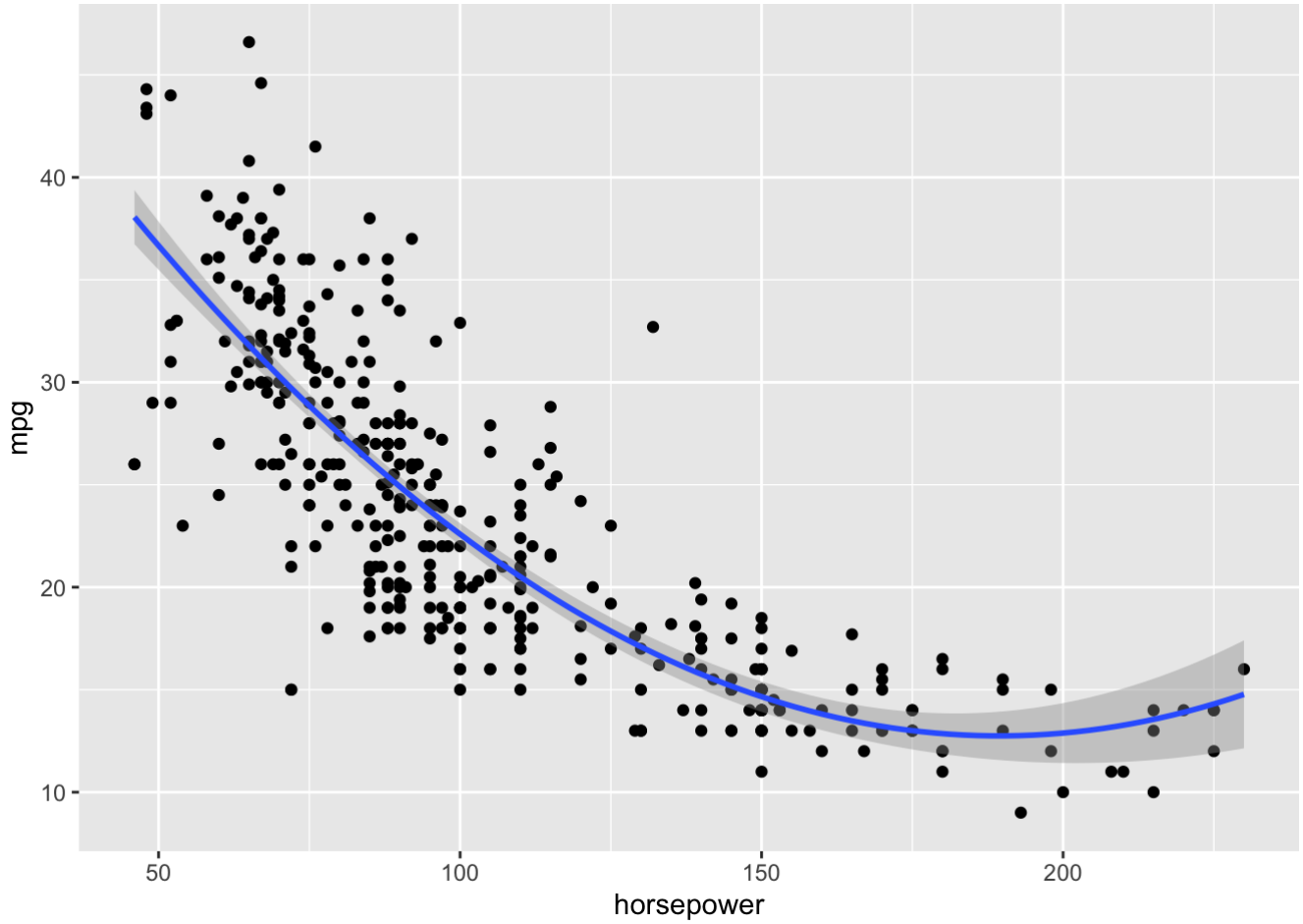
- The fitted regression model : $\hat{y} = 56.90 - 0.466x + 0.0012x^2$
- The 95% prediction interval for the MPG of a vehicle with 150 horsepower is (6.027, 23.29). The predicted MPG for an individual vehicle with 150 horsepower is 14.65 MPG.

```
lm1 <-lm(mpg ~ horsepower + I(horsepower^2), data=Auto)
new_x <-data.frame(horsepower=150)
predict(lm1, newdata=new_x, interval="prediction")
```

```
##        fit      lwr      upr
## 1 14.65872 6.027273 23.29016
```

(d)

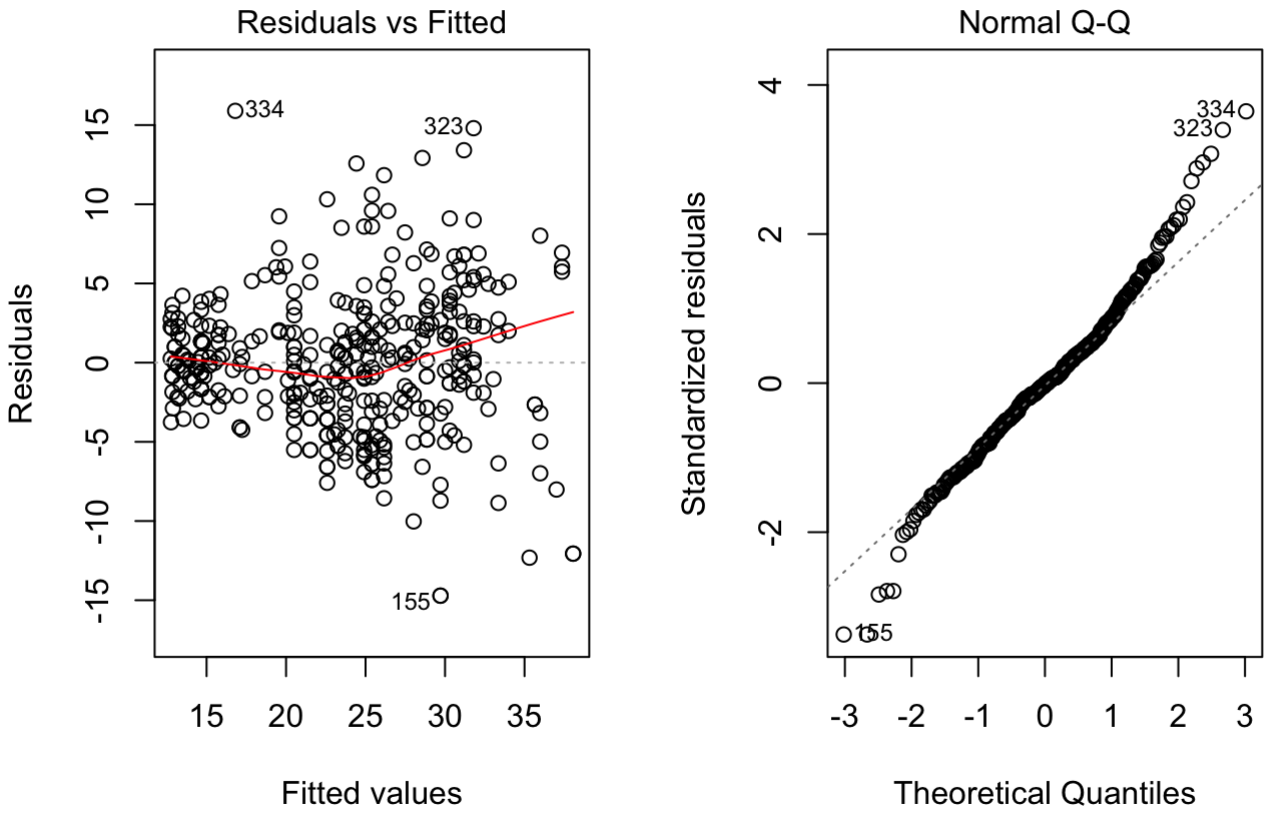
```
library(ggplot2)
ggplot(data=Auto, aes(horsepower, mpg)) +
    geom_point() +
    stat_smooth(method='lm', formula = y ~ poly(x,2))
```



(e)

- the LINE assumptions still hold for the polynomial regression model: (linearity, independent, normality and equal variance.).
- In the Residuals vs Fitted plot, we can see that the points are mostly randomly scattered; however there is a slight fanning pattern. This tells us that the data may deviate from having equal variance. Additionally, most of the points are scattered around 0. There are some possible outliers present.
- In the QQ plot, we assess normality; majority of the data points fall onto the line which indicates that the data, for the most part, follows a normal distribution. However, it's important to note that the tails of the graph do not fall onto the line, which means that these points contribute to the data set deviating from normal behavior.

```
par(mfrow=c(1,2))
plot(lm1, 1:2)
```



Exercise 2

(a)

```
library(ISLR)
head(Carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 5  4.15       141     64           3        340   128       Bad  38        13
## 6 10.81       124    113          13        501    72       Bad  78        16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

```
lm2 <-lm(Sales ~ Price + Urban + US, data=Carseats)
summary(lm2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b)

- A unit increase in the price (by one dollar), with the other predictors(US and Urban) held fixed, is associated with a decrease in Sales by $0.054.
- Our dummy variable, (categorical variable), US: the Sales of Carseats in the US is $1.20 higher than Carseats outside of the US, when all other predictors (Price and Urban) is held fixed.
- Our other dummy variable, Urban, to indicate if the store is in an urban or rural location: the Sales of the Carseats in an Urban location is $0.02 lower than the sales of Carseats in rural locations, when all other predictors (price and US) are held fixed.

(c)

- The predicted multiple linear regression model : $\hat{Sales} = 13.04 - 0.054Price - 0.02Urban + 1.20US$

(d)

- We can reject the $H_0 : \beta_j = 0$, for the predictors Price and US; the p-value for both are less than $\alpha = 0.05$. There is an effect on Carseats sales that is associated with price and location (US or not in US). In contrast, the predictor, Urban is insignificant, given that the p-value is > $\alpha = 0.05$.

(e)

- - We notice that the predictor **Urban**, is insignificant, given that the p-value is > $\alpha = .05$. As a result, we remove it from the model and run the MLR again without it. The final regression model is: $\hat{Sales} = 13.03 - 0.05Price + 1.19US$

```
lm3 <-lm(Sales ~ Price + US, data= Carseats)
summary(lm3)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

(f)

- After removing the Urban predictor, R^2 had decreased and $R^2{}_{adj}$ increased, both not by much. About 24% of the variability in Sales is explained by the predictors for both models in (a) and (e). This value is low, which indicates that the predictive power of our model is subpar; we have a lot of uncertainty, even though most of our predictors are significant.

```
s1 <-summary(lm2)
s2 <-summary(lm3)

s1$r.squared
```

```
## [1] 0.2392754
```

```
s2$r.squared
```

```
## [1] 0.2392629
```

```
s1$adj.r.squared
```

```
## [1] 0.2335123
```

```
s2$adj.r.squared
```

```
## [1] 0.2354305
```

(g)

- None of the 95% confidence intervals contain 0, so there is a relationship between the predictors Price, US and Sales of Carseats.
- With all other predictors held fixed, we are 95% confident that with an increase in price by a dollar, there is an expected decrease in Sales ranging from -0.06 to -0.04.
- With all other predictors held fixed, we are 95% confident that with an increase in one unit of Carseats produced in the US, there is an expected increase in Sales ranging from 0.69 to 1.71.

```
confint(lm3) #after we've removed Urban
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```