Homework 3

Daniel Tshiani

2025-06-01

```r
library(ISLR)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

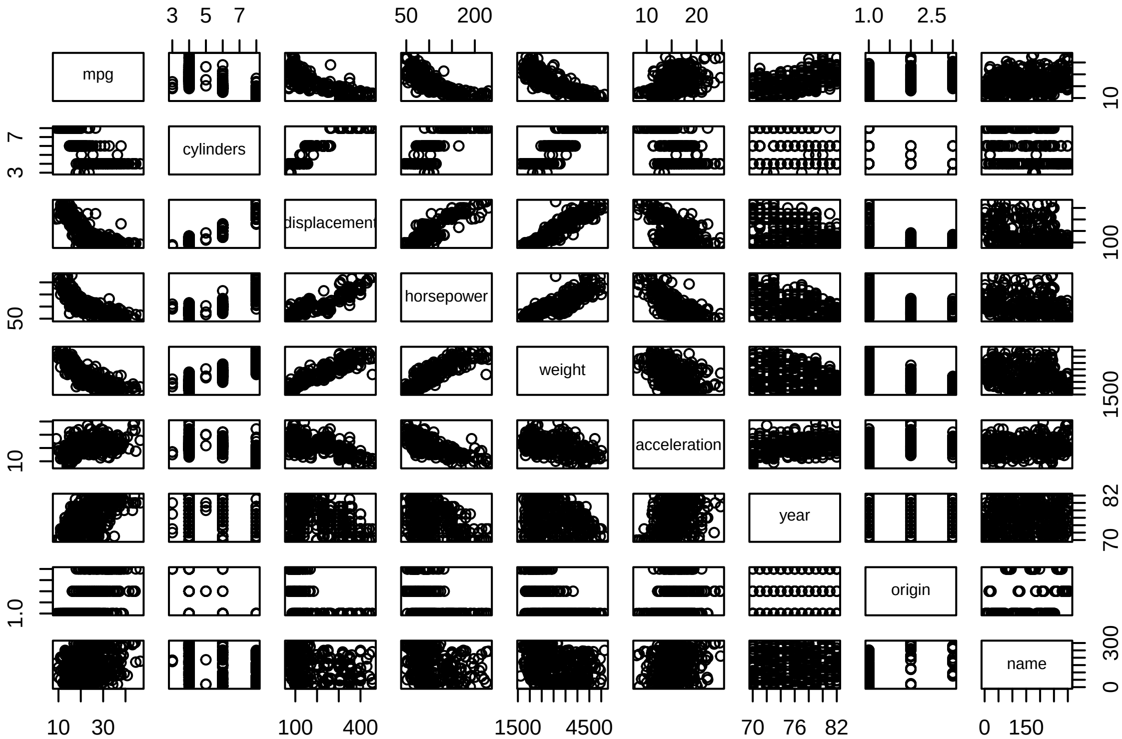
```r
library(ggplot2)
```

```r
load("../data/Auto-3.rda")
```

1

a

```r
pairs(Auto)
```



b

```r
Auto <- Auto%>%
  select(-name)

cor(Auto)
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

c

```r
lm <- lm(mpg ~ ., data = Auto)
summary(lm)
```

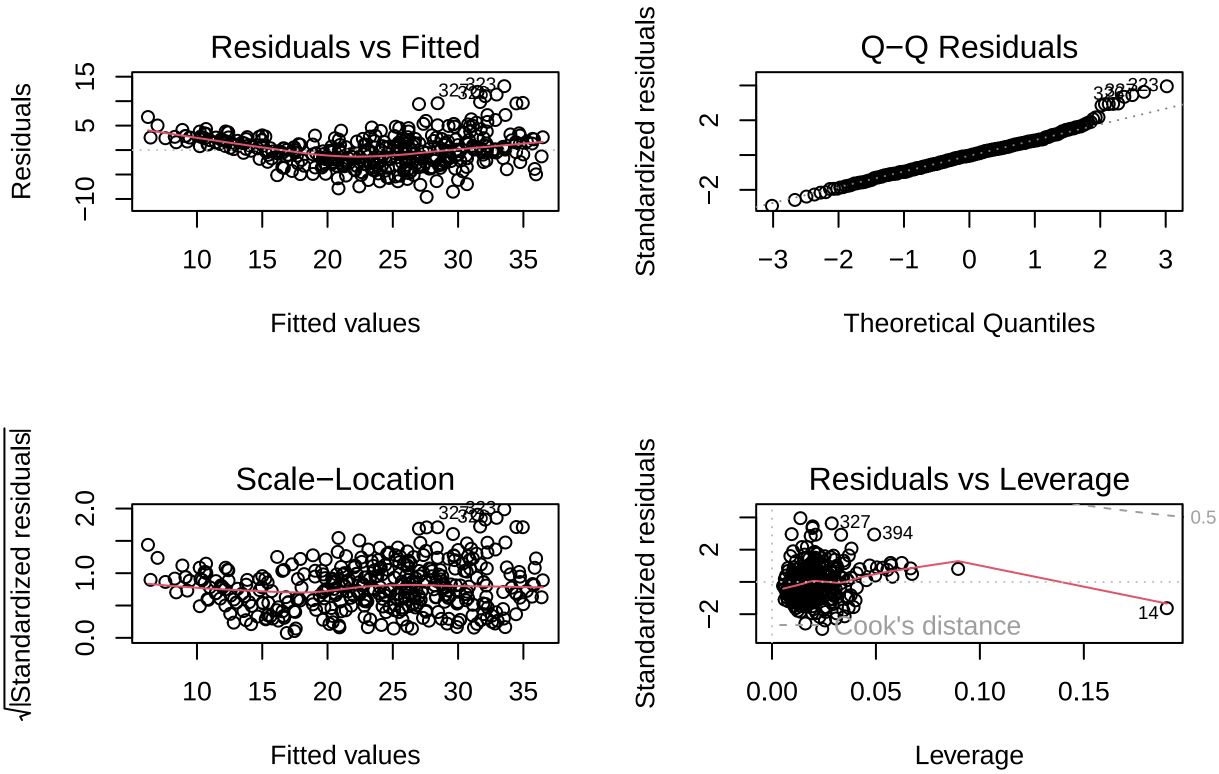
```
##
## Call:
```

```
## lm(formula = mpg ~ ., data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

displacement, year and origin all appear to have a positive significant relationships with mpg. weight appears to have a negative significant relationship with mpg. the other variables appear to have insignificant relationships with mpg.

d

```
par(mfrow=c(2,2))
plot(lm)
```



the residual plot suggest heteroscedaticity because the spread appears to increase. in the qq-plot towards the higher quartiles, the point deviate above the diagonal line which suggest a deviation from normality. it could possible be due to outliers or skewness.

in the residuals vs leverage plot, there is one point way out to the right at around 0.19 leverage. that point appears to have high influence on the model.

e

```
lm2 <- lm(mpg ~ . + weight * acceleration + cylinders * horsepower, data = Auto)
summary(lm2)
```

```
##
## Call:
## lm(formula = mpg ~ . + weight * acceleration + cylinders * horsepower,
##     data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9865 -1.6273 -0.0109  1.2923 11.9178
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.499e+00  8.175e+00   0.428 0.668893
## cylinders            -3.904e+00  5.587e-01  -6.988 1.25e-11 ***
## displacement         -3.033e-03  7.028e-03  -0.431 0.666354
## horsepower           -2.940e-01  3.514e-02  -8.365 1.14e-15 ***
## weight               -1.899e-03  1.708e-03  -1.112 0.266703
## acceleration          1.807e-01  2.939e-01   0.615 0.539018
## year                  7.470e-01  4.526e-02  16.506  < 2e-16 ***
```

```
## origin                8.668e-01  2.512e-01   3.451 0.000621 ***
## weight:acceleration  -1.235e-04  9.846e-05  -1.255 0.210320
## cylinders:horsepower  3.659e-02  4.754e-03   7.696 1.21e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.926 on 382 degrees of freedom
## Multiple R-squared:  0.8627, Adjusted R-squared:  0.8594
## F-statistic: 266.6 on 9 and 382 DF,  p-value: < 2.2e-16
```

the interaction between weight and acceleration does not appear to be statistically significant, however the interaction between cylinders and horsepower appears to be statistically significant.

f

```r
lm3 <- lm(mpg ~ . + log(horsepower), data = Auto)
summary(lm3)
```

```
##
## Call:
## lm(formula = mpg ~ . + log(horsepower), data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5777 -1.6623 -0.1213  1.4913 12.0230
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.674e+01  1.106e+01   7.839 4.54e-14 ***
## cylinders       -5.530e-02  2.907e-01  -0.190 0.849230
## displacement    -4.607e-03  7.108e-03  -0.648 0.517291
## horsepower       1.764e-01  2.269e-02   7.775 7.05e-14 ***
## weight          -3.366e-03  6.561e-04  -5.130 4.62e-07 ***
## acceleration    -3.277e-01  9.670e-02  -3.388 0.000776 ***
## year             7.421e-01  4.534e-02  16.368  < 2e-16 ***
## origin           8.976e-01  2.528e-01   3.551 0.000432 ***
## log(horsepower) -2.685e+01  2.652e+00 -10.127  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.959 on 383 degrees of freedom
## Multiple R-squared:  0.8592, Adjusted R-squared:  0.8562
## F-statistic: 292.1 on 8 and 383 DF,  p-value: < 2.2e-16
```

just focusing on horsepower: - when i took the log transformation of horsepower, it became statisticaly significant and the effect of horsepower on mpg became positive.

```r
lm4 <- lm(mpg ~ . + sqrt(horsepower), data = Auto)
summary(lm4)
```

```
##
## Call:
## lm(formula = mpg ~ . + sqrt(horsepower), data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -8.5402 -1.6717 -0.0778  1.4861 11.9754
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.299e+01  7.251e+00   5.929 6.82e-09 ***
## cylinders         6.037e-02  2.928e-01   0.206 0.836748
## displacement     -5.870e-03  7.156e-03  -0.820 0.412560
## horsepower        4.239e-01  4.532e-02   9.353  < 2e-16 ***
## weight           -3.285e-03  6.604e-04  -4.975 9.87e-07 ***
## acceleration     -3.342e-01  9.705e-02  -3.443 0.000638 ***
## year              7.398e-01  4.536e-02  16.308  < 2e-16 ***
## origin            9.159e-01  2.526e-01   3.626 0.000326 ***
## sqrt(horsepower) -1.050e+01  1.039e+00 -10.104  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.961 on 383 degrees of freedom
## Multiple R-squared:  0.8591, Adjusted R-squared:  0.8561
## F-statistic: 291.8 on 8 and 383 DF,  p-value: < 2.2e-16
```

just focusing on horsepower: - when i took the log transformation of horsepower, it became statisticalys significant and the effect of horsepower on mpg became positive. however, one unit change in horsepower increases the model 4 by 4.2 units compared to 1.7 units in model 3.

```r
lm5 <- lm(mpg ~ . + I(horsepower^2), data = Auto)
summary(lm5)
```

```
##
## Call:
## lm(formula = mpg ~ . + I(horsepower^2), data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5497 -1.7311 -0.2236  1.5877 11.9955
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.3236564  4.6247696   0.286 0.774872
## cylinders        0.3489063  0.3048310   1.145 0.253094
## displacement    -0.0075649  0.0073733  -1.026 0.305550
## horsepower      -0.3194633  0.0343447  -9.302  < 2e-16 ***
## weight          -0.0032712  0.0006787  -4.820 2.07e-06 ***
## acceleration    -0.3305981  0.0991849  -3.333 0.000942 ***
## year             0.7353414  0.0459918  15.989  < 2e-16 ***
## origin           1.0144130  0.2545545   3.985 8.08e-05 ***
## I(horsepower^2)  0.0010060  0.0001065   9.449  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.001 on 383 degrees of freedom
## Multiple R-squared:  0.8552, Adjusted R-squared:  0.8522
## F-statistic: 282.8 on 8 and 383 DF,  p-value: < 2.2e-16
```

interestingly, making horsepower quadratic also makes it statistically significant. however, the influence on horsepower on mpg has become negative.

2

```r
rm(list = ls())
load("../data/Carseats.rda")
```

a

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

```
##
## Call:
## glm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## 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
##
## (Dispersion parameter for gaussian family taken to be 6.113219)
##
##     Null deviance: 3182.3  on 399  degrees of freedom
## Residual deviance: 2420.8  on 396  degrees of freedom
## AIC: 1865.3
##
## Number of Fisher Scoring iterations: 2
```

b

Sales = 13.043 - 0.054 * Price - 0.022 * UrbanYes + 1.201 * USYes

Where: UrbanYes = 1 if Urban == "Yes", else 0 USYes = 1 if US == "Yes", else 0

Urban = "No", US = "No" Sales = 13.043 - 0.054 * Price

Urban = "Yes", US = "No" Sales = (13.043 - 0.022) - 0.054 * Price Sales = 13.021 - 0.054 * Price

Urban = "No", US = "Yes" Sales = (13.043 + 1.201) - 0.054 * Price Sales = 14.244 - 0.054 * Price

Urban = "Yes", US = "Yes" Sales = (13.043 - 0.022 + 1.201) - 0.054 * Price Sales = 14.222 - 0.054 * Price

c

looking at the p value, we would reject the hypothesis of beta_j = 0 for price and USYes. the p-value is less that 0.01

d

```r
reduced_model <- glm(Sales ~ Price + US, data=Carseats)
anova(reduced_model, lm)
```

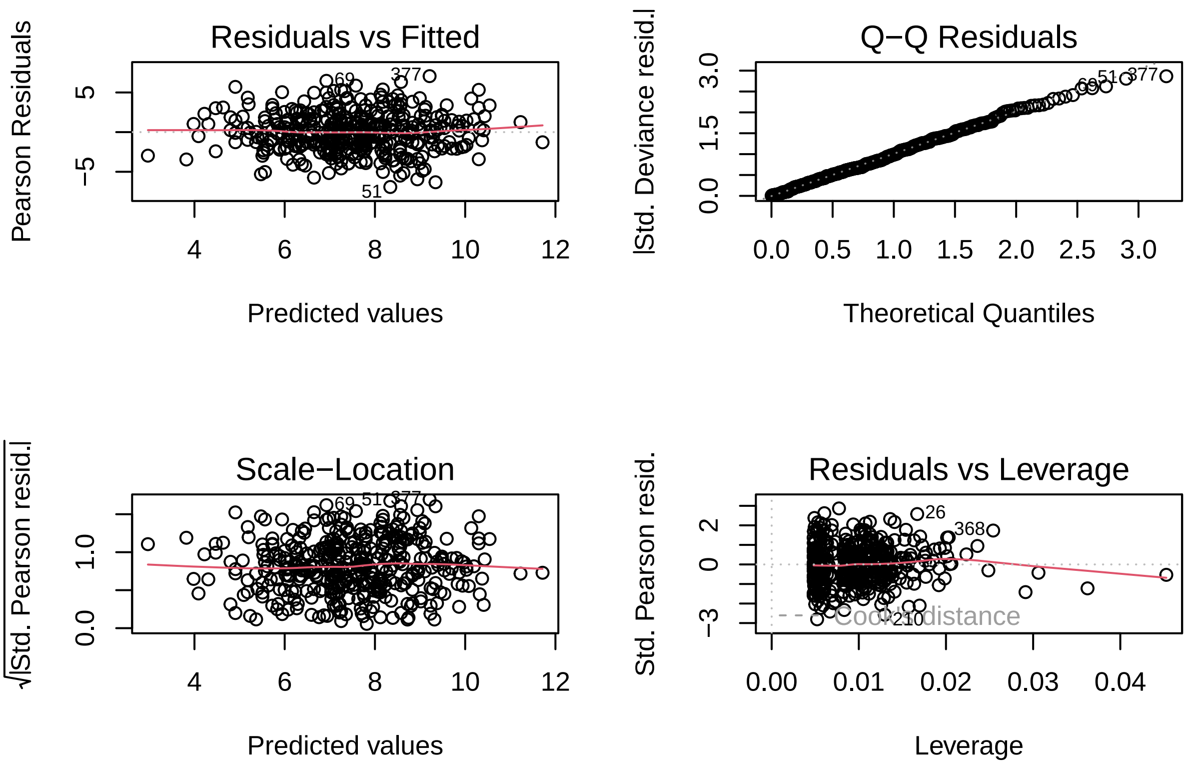
```
## Analysis of Deviance Table
```

```
##
## Model 1: Sales ~ Price + US
## Model 2: Sales ~ Price + Urban + US
##   Resid. Df Resid. Dev Df Deviance      F Pr(>F)
## 1       397     2420.9
## 2       396     2420.8  1  0.03979 0.0065 0.9357
```

the f-stat is 0.0065 which means that this partial f test agrees that urban does not sifnificantly improve the model.

e

```r
par(mfrow = c(2, 2))
plot(lm)
```




```r
outliers <- rstandard(lm)
outliers_df <- data.frame(outliers)

outliers_df <- outliers_df %>%
  filter(abs(outliers) > 3)

head(outliers_df)
```

```
## [1] outliers
## <0 rows> (or 0-length row.names)
```

to be honest, i am not finding anything out the ordinary with the data from the plots and there arent any outliers when using 3 as a threshold.

f

```r
Carseats$USyes <- 1 * (Carseats$US == "Yes")
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
vif(lm)
```

```
##    Price    Urban       US
## 1.005342 1.004203 1.005349
```

the vifs are under 2 which means that there is low multicolinearity.

g

```r
stud_resid <- rstudent(lm)
n <- nrow(Carseats)
p <- length(coef(lm))
alpha <- 0.05

# Bonferroni-adjusted critical t-value
t_crit <- qt(1 - alpha / (2 * n), df = n - p)

outliers <- which(abs(stud_resid) > t_crit)

Carseats[outliers, ]
```

```
##  [1] Sales       CompPrice   Income      Advertising Population  Price
##  [7] ShelveLoc   Age         Education   Urban       US          USyes
## <0 rows> (or 0-length row.names)
```

again, i do not get any outliers that are significant.

h

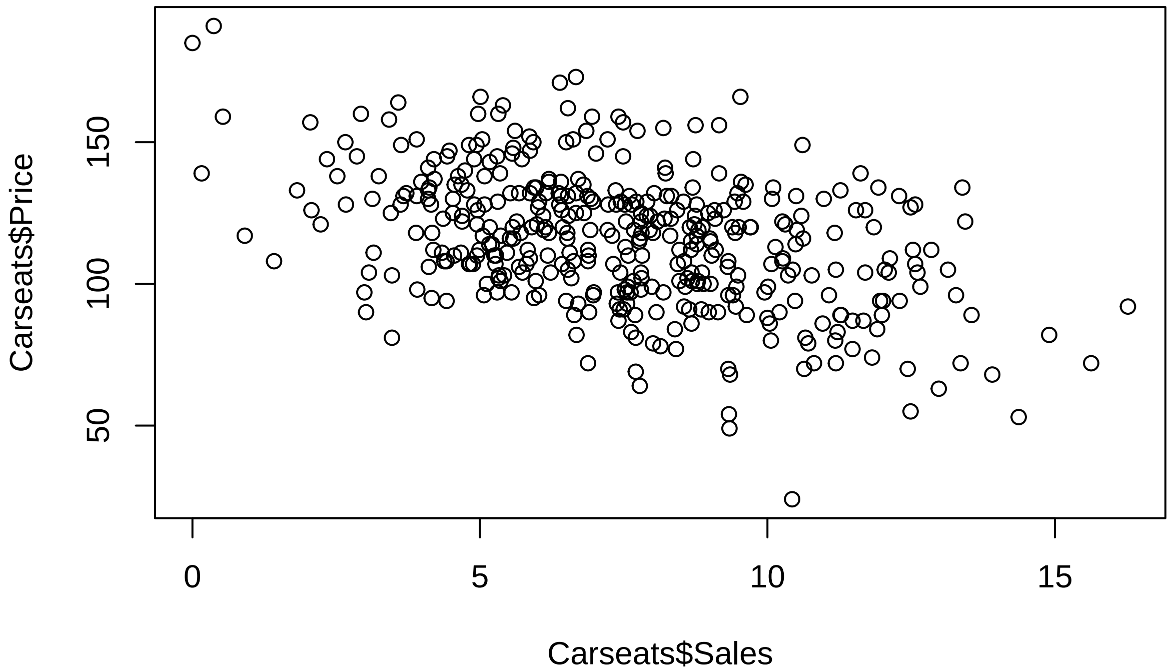
```r
residuals <- resid(lm)
shapiro.test(residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.99798, p-value = 0.9184
```

p value is greater than 0.05 which means the residuals are likely normally distributed.

i

```r
plot(Carseats$Sales, Carseats$Price)
```




```r
model_linear <- lm(Sales ~ Price, data = Carseats)
model_full <- lm(Sales ~ Price + I(Price^2), data = Carseats)
anova(model_linear, model_full)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Price
## Model 2: Sales ~ Price + I(Price^2)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    398 2552.2
## 2    397 2551.5  1   0.76682 0.1193   0.73
```

there is a linear relationship because in my test the f value is 0.1193 and the p value is0.73. this means the linear models performs just was well as the quadratic model.