SL HW 6

Joshua Ingram

11/1/2019

Problem 1

```r
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
##   lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
## 4  2.94 33.4
## 5  5.33 36.2
## 6  5.21 28.7
```

a.

```r
mu.hat <- mean(Boston$medv)
mu.hat
```

```
## [1] 22.53281
```

$\hat{\mu} = 22.53281$

b.

```r
se <- sd(Boston$medv)/sqrt(length(Boston$medv))
se
```

```
## [1] 0.4088611
```

The standard error for $\hat{\mu}$ is 0.409, which tells us the typical sampled value of $\hat{\mu}$ will fall within 0.409 units away from the population value.

c.

```r
set.seed(1)
x.mean <- function(x,i) { mean(x[i,14]) }
boot.sim1 <- boot(data = Boston, statistic = x.mean, R = 1000)
sd.boot <- sd(boot.sim1$t)
sd.boot
```

```
## [1] 0.4106622
```

This standard error estimate is very close to the se found in the thereotical approach, being .002 units greater.

d.

```r
t.test(Boston$medv)
```

```
##
##  One Sample t-test
##
## data:  Boston$medv
## t = 55.111, df = 505, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  21.72953 23.33608
## sample estimates:
## mean of x
##  22.53281
```

```r
mu.hat + (c(-1, 1) * 1.96 * sd.boot)
```

```
## [1] 21.72791 23.33770
```

The bootstrap confidence interval, (21.73, 23.34), is basically indentical to the interval found using t.test, (21.73, 23.34).

e.

```r
median(Boston$medv)
```

```
## [1] 21.2
```

$\hat{\mu}_{med} = 21.2$

f.

```r
set.seed(1)
x.median <- function(x,i) { median(x[i,14]) }
boot.sim2 <- boot(data = Boston, statistic = x.median, R = 1000)
sd.boot2 <- sd(boot.sim2$t)
sd.boot2
```

```
## [1] 0.3778075
```

standard error estimate using bootstrap simulation: 0.379

g.

```r
quantile(Boston$medv, .90)
```

```
##  90%
## 34.8
```

$\hat{\mu}_{0.1} = 34.8$

h.

```r
set.seed(1)
x.quantile <- function(x,i) { quantile(x[i,14], .90) }
boot.sim3 <- boot(data = Boston, statistic = x.quantile, R = 1000)
sd.boot3 <- sd(boot.sim3$t)
sd.boot3
```

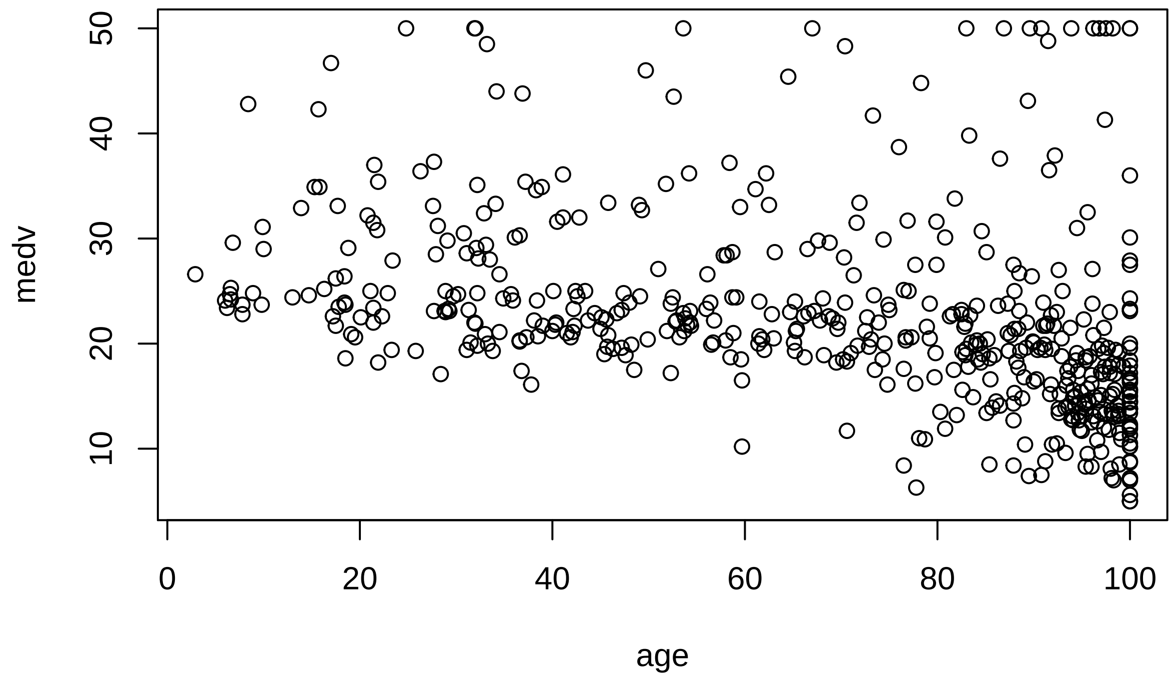
```
## [1] 1.14822
```

standard error estimate of $\hat{\mu}_{0.1} = 1.15$

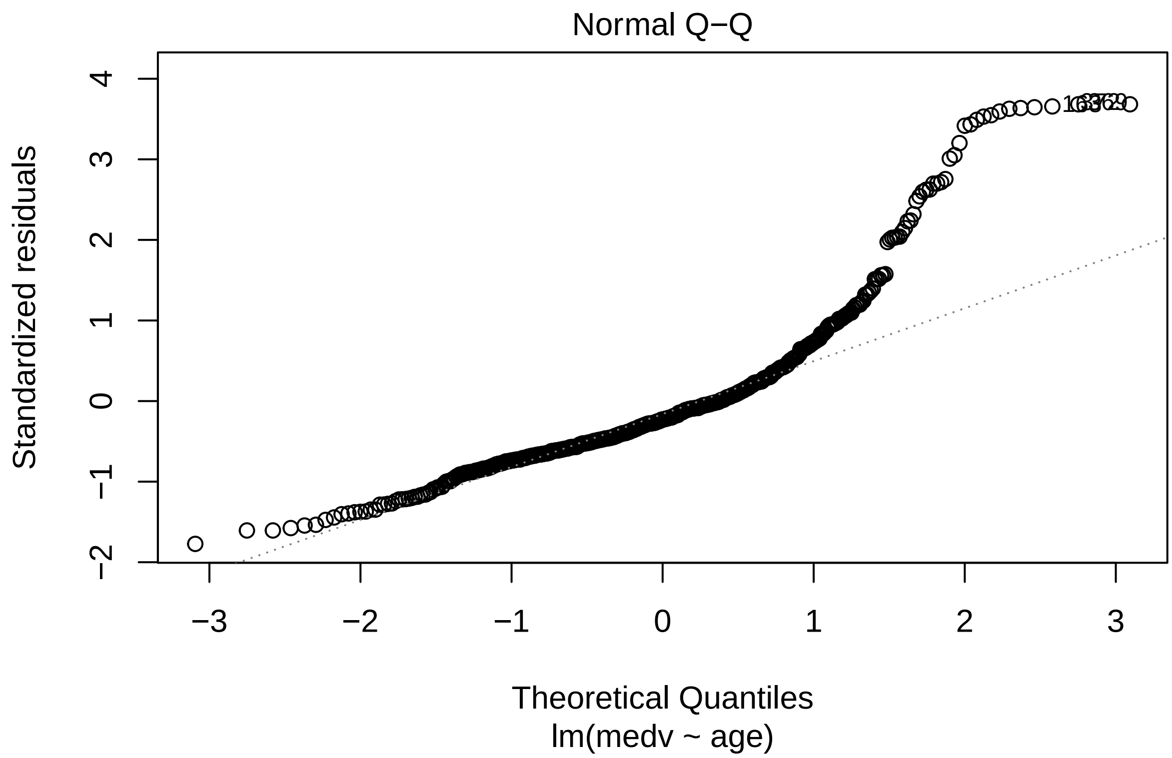
Problem 2

a.

```r
plot(medv ~ age, data = Boston)
```




```
lm.obj <- lm(medv~age, data = Boston)
plot(lm.obj, which = 2)
```



Based on the plot of the model, there does not seem to be constant variance throughout, Focusing on the QQ-plot, the standardized residuals do not follow the "expected" values that should be along the dotted line. It is very clear that the normality assumption is not satisfied.

b.

```
sum.lm <- summary(lm.obj)
sum.lm
```

```
##
## Call:
## lm(formula = medv ~ age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.097  -5.138  -1.958   2.397  31.338
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.97868    0.99911  31.006   <2e-16 ***
## age         -0.12316    0.01348  -9.137   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.527 on 504 degrees of freedom
```

```
## Multiple R-squared:  0.1421, Adjusted R-squared:  0.1404
## F-statistic: 83.48 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
CI <- -0.12316 + (c(-1, 1) *(1.96 * 0.01348))
CI
```

```
## [1] -0.1495808 -0.0967392
```

95% confidence interval: -0.1496, -0.0967

c.

```
set.seed(1)
x.beta <- function(x,i) { coef(lm(medv~age, data=Boston, subset = i))[2] }
boot.sim4 <- boot(data = Boston, statistic = x.beta, R = 1000)
x.betafull <- coef(lm(medv~age, data = Boston,))[2]
bias <- mean(boot.sim4$t) - x.betafull
unbiased.est <- boot.sim4$t - bias
c(quantile(unbiased.est, 0.025), quantile(unbiased.est, 0.975))
```

```
##        2.5%       97.5%
## -0.14923421 -0.09822158
```

This confidence interval is nearly the same as the "classical" approach using the standard error given in the summary.

d.

It would generally be best to trust the bootstrap confidence interval more, especially when we have a lack of normality in our residuals (as seen here). Under the classical appraoch, like a Walk Confidence Interval, we assume normality.

Problem 3

1.

a.

π_A = probability of receiving an A $\pi_A = \frac{e^{-6+0.05(hoursstudied)+(undergradGPA)}}{1+e^{-6+0.05(hoursstudied)+(undergradGPA)}}$

b.

```
pi.1 = exp(-6 + (.05*40) + 3.5)/(1 + exp(-6 + (.05*40) + 3.5))
pi.1
```

```
## [1] 0.3775407
```

c.

derived algebraiclly by hand by solving for X_1 in the logistic regression model

$X_1 = 50$ hours

2.

a.

Solving for π in the odds ratio $\pi_{default} = .27$

b.

$\frac{.19}{1-.19}$ = odds of defaulting: .235

Problem 4

a.

```
Auto$mpg01 <- ifelse(Auto$mpg> median(Auto$mpg), 1, 0)
Auto$mpg <- NULL
log.reg <- glm(mpg01 ~ .-name, family = binomial, data = Auto)
summary(log.reg)
```

```
##
## Call:
## glm(formula = mpg01 ~ . - name, family = binomial, data = Auto)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4277  -0.1061   0.0080   0.2123   3.1631
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -17.154875   5.763805  -2.976 0.002917 **
## cylinders     -0.162589   0.423195  -0.384 0.700835
## displacement   0.002095   0.012034   0.174 0.861789
## horsepower    -0.041019   0.023872  -1.718 0.085750 .
## weight        -0.004315   0.001140  -3.784 0.000154 ***
## acceleration   0.016065   0.141462   0.114 0.909582
## year           0.429459   0.075225   5.709 1.14e-08 ***
## origin         0.477339   0.362014   1.319 0.187314
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 543.43  on 391  degrees of freedom
## Residual deviance: 157.54  on 384  degrees of freedom
```

```
## AIC: 173.54
##
## Number of Fisher Scoring iterations: 8
```

b.

year and weight are the most significant predictors.

odds interpretation:

year - for every one year increase, we expect the odds of having high gas mileage will increase by a factor of $e^{0.429}$

weight - for every one pound increase, we expect the odds of having high gas mileage will increase by a factor of $e^{-0.004315}$

"simple" interpretation:

year - for every one year increase, we expect the probability of having high gas mileage will increase

weight - for every one pound increase in weight, we expect the probability of having high gas mileage to decrease