STAT 516 hw 3

Solutions

Chp 8 Ex 2

a)

We have $p = 10$ and $n = 500$ with $R^2 = 0.07$. The relationship between R^2 and the test statistic for the overall F-test gives

$$F_{\text{stat}} = \frac{n-(p+1)}{p}\frac{R^2}{1-R^2} = \frac{500-(10+1)}{10}\frac{0.07}{1-0.07} = 3.6806452.$$

We reject H_0: $\beta_j = 0$ for $j = 1, \ldots, 10$ (that is, the hypothesis that all the regression coefficients are equal to zero) at significance level 0.05 if $F_{\text{stat}} > F_{10,500-(10+1),0.05} = 1.8500646$.

Since $3.6806452 > 1.8500646$, we reject H_0. So there *is* some statistically significant relationship between the covariates and the response, even though the value of R^2 is small.

b)

Yes, the sociologist, having rejected H_0: $\beta_j = 0$ for $j = 1, \ldots, 10$, can accept the alternative hypothesis, which is that at least one of the covariates has a nonzero slope coefficient.

c)

The small value of R^2 indicates that the model explains only a small proportion of the total variability in the responses. Even though there is a statistically significant relationship between the response and at least one of the predictors, the small value of R^2 suggests that predictions based on this model will not be very accurate.

d)

Under $n = 50$, we would have

$$F_{\text{stat}} = \frac{n-(p+1)}{p}\frac{R^2}{1-R^2} = \frac{50-(10+1)}{10}\frac{0.07}{1-0.07} = 0.2935484,$$

which we would compare to the critical value $F_{10,50-11,0.05} = 2.083869$. We would fail to reject the null hypothesis that all the slope coefficients are equal to zero. So under the smaller sample size, the same value of R^2 would *not* indicate a statistically significant relationship between the covariates and the response.

Chp 8 Ex 8

```
goalmade <- read.table(file = "Data Tables 4th edition/Chapter 8/datatab_8_29.prn",
                       header= TRUE)
head(goalmade)
```

```
  obs weight height dash100 goalmade
1   1    130     71   11.50       15
2   2    149     74   12.23       19
3   3    170     70   12.26       11
4   4    177     71   12.65       15
5   5    188     69   10.26       12
6   6    210     73   12.76       17
```

a)

Use the `lm()` function to perform the regression:

```
lm_out <- lm(goalmade ~ weight + height + dash100, data = goalmade)
summary(lm_out)
```

```

Call:
lm(formula = goalmade ~ weight + height + dash100, data = goalmade)

Residuals:
    Min      1Q  Median      3Q     Max
```

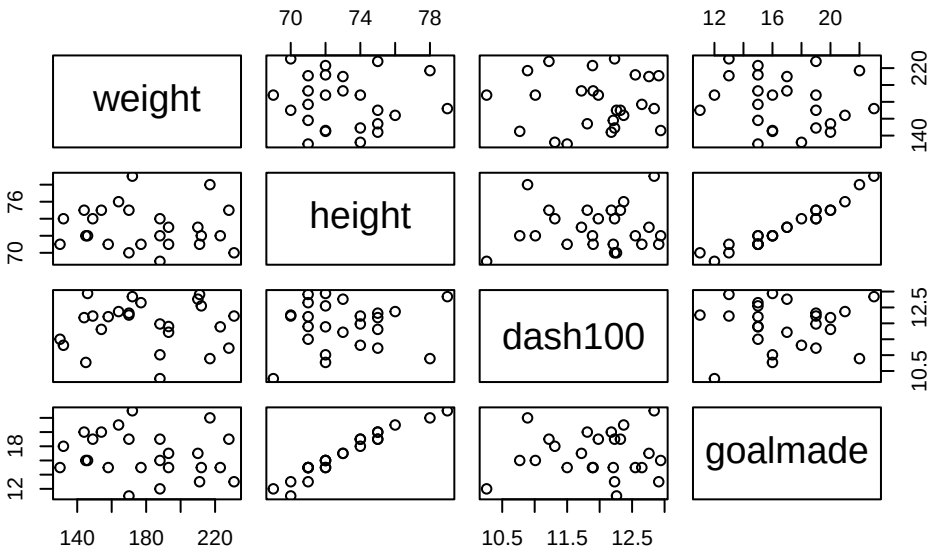
```
-2.2904 -0.2108  0.1617  0.4455  1.0493

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -67.364828   5.219781 -12.906 1.88e-11 ***
weight       -0.010570   0.005052  -2.092   0.0487 *
height        1.202711   0.062687  19.186 8.60e-15 ***
dash100      -0.141724   0.216264  -0.655   0.5194
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7583 on 21 degrees of freedom
Multiple R-squared:  0.9478,    Adjusted R-squared:  0.9403
F-statistic: 127.1 on 3 and 21 DF,  p-value: 1.267e-13
```

The overall F-test rejects the null hypothesis that none of the regressors are related to the response. The weight and height variables are significant at the $\alpha = 0.05$ significance level, though the p-value for the weight variable is just slightly less than 0.05. While height is positively related to goals made, weight appears to be negatively related.

b)

A good way to visualize multicollinearity is to make pairwise scatter plots of all the variables:

```
plot(goalmade[,-1]) # remove first column "obs"
```

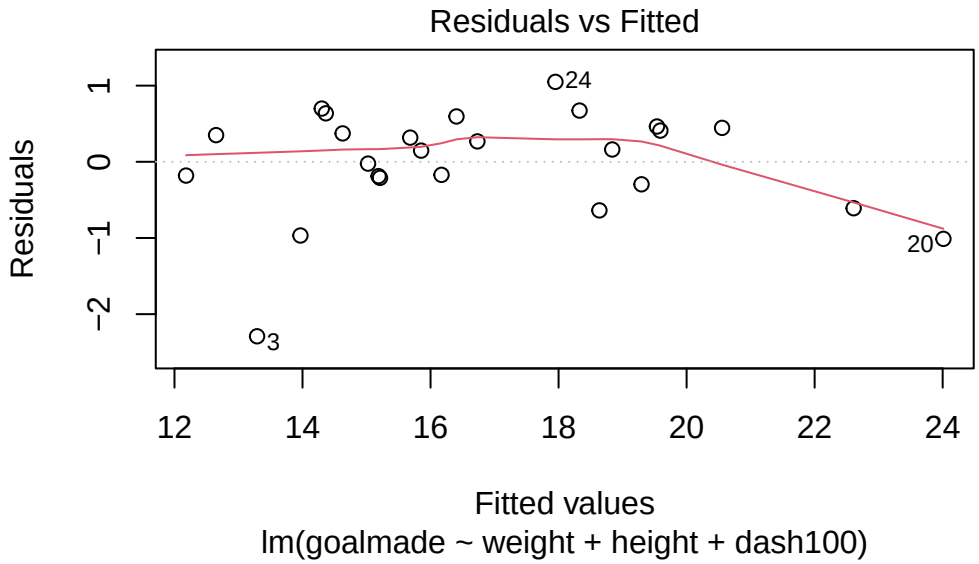


We see that height is positively correlated with goals made, but no other strong correlations are apparent in this plot. Multicollinearity refers only to correlations among the predictors, so from this plot, it appears that there is very little multicollinearity in these data.

c)

To check for outliers we can look at the residuals versus fitted values plot:

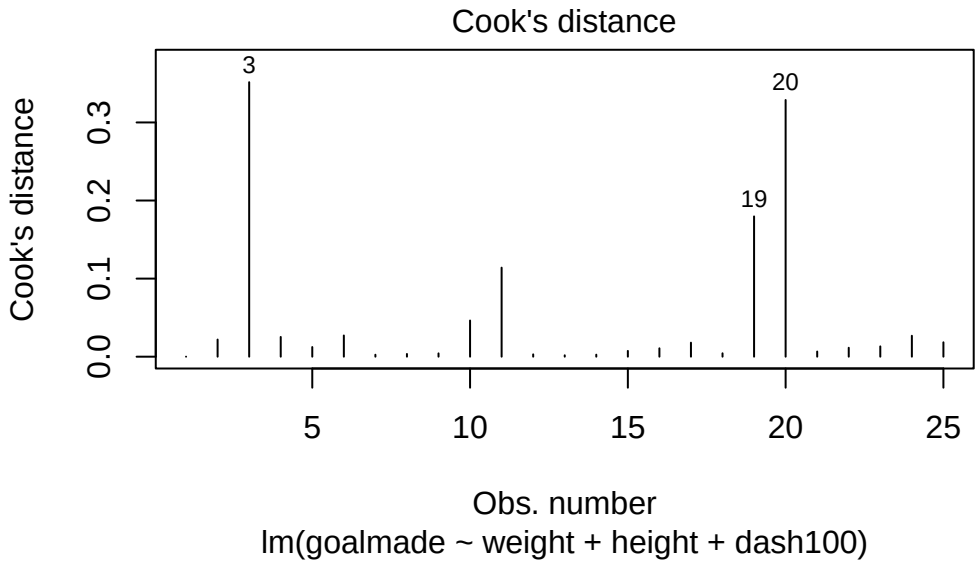
```
plot(lm_out,which = 1)
```



We see that there are a couple of large residuals, which may correspond to outliers.

Now make a plot of Cook's D values:

```
plot(lm_out, which = 4)
```



There are a few points which have Cook's D values rather larger than the other points. The most extreme observations are 3 and 20. Both of these players scored below the predicted value based on their covariate values (had negative residuals).

d)

One could try fitting a model after removing these outlying observations.

Suppose we remove observations 3 and 20.

Then we obtain:

```
goalmade2 <- goalmade[-c(3,20),]
lm2_out <- lm(goalmade ~ weight + height + dash100, data = goalmade2)
summary(lm2_out)
```

```
Call:
lm(formula = goalmade ~ weight + height + dash100, data = goalmade2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2299 -0.2430  0.1706  0.3108  0.9080

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.584438   4.263754 -16.085 1.60e-12 ***
weight       -0.012149   0.003431  -3.541  0.00218 **
height        1.199446   0.050818  23.603 1.54e-15 ***
dash100       0.016816   0.152003   0.111  0.91307
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5128 on 19 degrees of freedom
Multiple R-squared:  0.9686,    Adjusted R-squared:  0.9637
F-statistic: 195.6 on 3 and 19 DF,  p-value: 1.855e-14
```

We see that the resulting model is not very different. Height and weight are still significant predictors (weight appears now to be more significant) and each still influences the response in the same direction.

This suggests that the presence of the outliers in the data set does not change the fitted model by much.

Chp 8 Ex 14

```r
# use na.strings = "." to properly encode the missing values
hp <- read.table("Data Tables 4th edition/Chapter 1/datatab_1_2.prn",
                 header = FALSE,
                 na.strings = ".")
colnames(hp) <- c("obs","zip","age","bed",
                  "bath","size","lot","exter",
                  "garage","fp","price")
head(hp)
```

```
  obs zip age bed bath size    lot exter garage fp price
1   1   3  21   3    3  951  64904 other      0  0 30000
2   2   3  21   3    2 1036 217800 frame      0  0 39900
3   3   4   7   1    1  676  54450 other      2  0 46500
4   4   3   6   3    2 1456  51836 other      0  1 48600
5   5   1  51   3    1 1186  10857 other      1  0 51500
6   6   2  19   3    2 1456  40075 frame      0  0 56990
```

a)

i)

Select a model using backward selection with AIC.

```
lm_all <- lm(price ~ age + bed + bath + size + lot,data = hp)
step(lm_all,direction="backward",scope= formula(lm_all),trace = 0)
```

```
Call:
lm(formula = price ~ age + size + lot, data = hp)

Coefficients:
(Intercept)          age         size          lot
 -4.247e+04   -1.053e+03    9.062e+01    3.028e-01
```

The chosen model uses only the age, size, and lot variables.

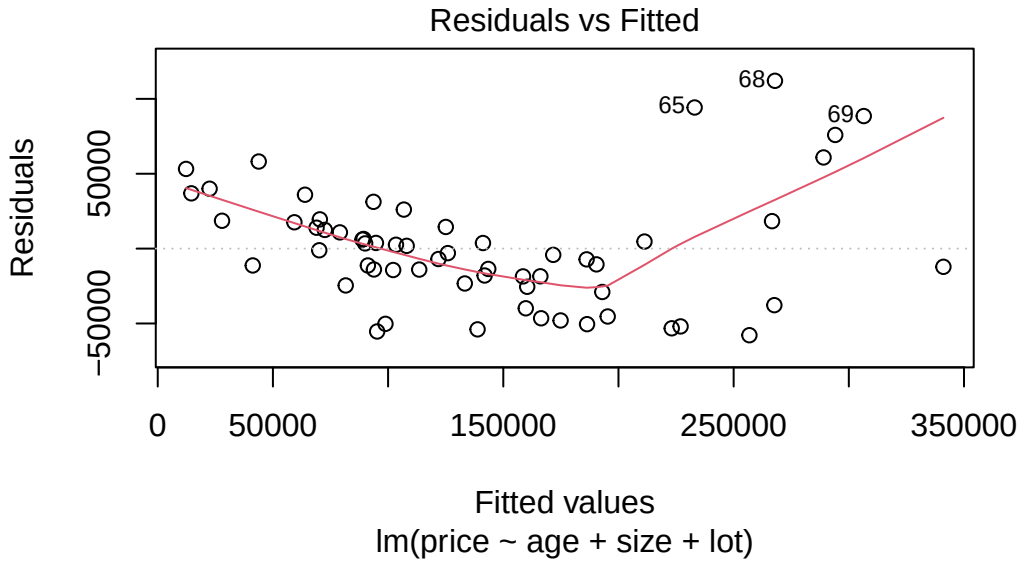
```
lm_chosen <- lm(price ~ age + size + lot,data = hp)
summary(lm_chosen)
```

```
Call:
lm(formula = price ~ age + size + lot, data = hp)

Residuals:
   Min     1Q Median     3Q    Max
-57842 -23624  -3599  17768 112067

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.247e+04  1.826e+04  -2.326 0.023667 *
age         -1.053e+03  4.134e+02  -2.547 0.013639 *
size         9.062e+01  6.503e+00  13.934  < 2e-16 ***
lot          3.028e-01  8.465e-02   3.577 0.000726 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39550 on 56 degrees of freedom
  (9 observations deleted due to missingness)
```

```
Multiple R-squared:  0.8063,    Adjusted R-squared:  0.796
F-statistic: 77.72 on 3 and 56 DF,  p-value: < 2.2e-16
```

Now look at the residuals versus fitted values plot.

```r
plot(lm_chosen,which = 1)
```



The residuals versus fitted values plot shows that the linear regression model is a poor fit to these data. There appears to be an "elbow" in the residuals versus fitted values plot around the fitted value 200000.

ii)

```r
lt200k <- which(hp$price <= 200000)
hp_lt200k <- hp[lt200k,]
lm_lt200k <-  lm(price ~ age + size + lot,data = hp_lt200k)
summary(lm_lt200k)
```

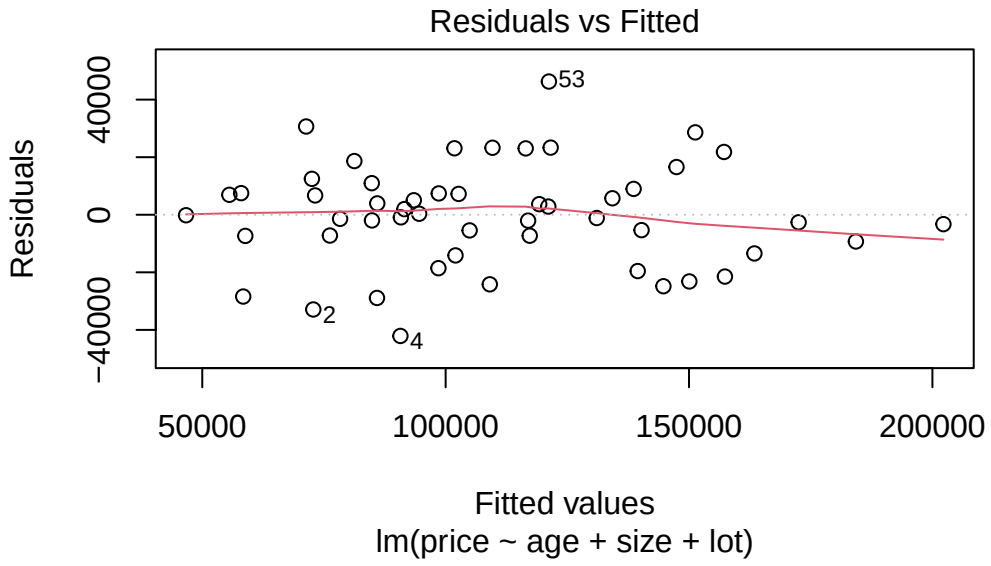
```
Call:
lm(formula = price ~ age + size + lot, data = hp_lt200k)

Residuals:
```

```
   Min     1Q Median     3Q    Max
-42107  -8313   -169   8247  46282

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 7401.34367 9910.82934   0.747    0.459
age         -313.26441  207.00640  -1.513    0.137
size          56.26836    4.24856  13.244   <2e-16 ***
lot            0.06286    0.05095   1.234    0.223
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18450 on 47 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.7986,    Adjusted R-squared:  0.7858
F-statistic: 62.14 on 3 and 47 DF,  p-value: < 2.2e-16
```

The fitted model has a fairly high value of R^2, suggesting that it might be good at making predictions. Each variable appears to be an important predictor of the price (though we must be cautious interpreting p-values after performing variable selection using the same data).

Now look at the residuals versus fitted values plot:

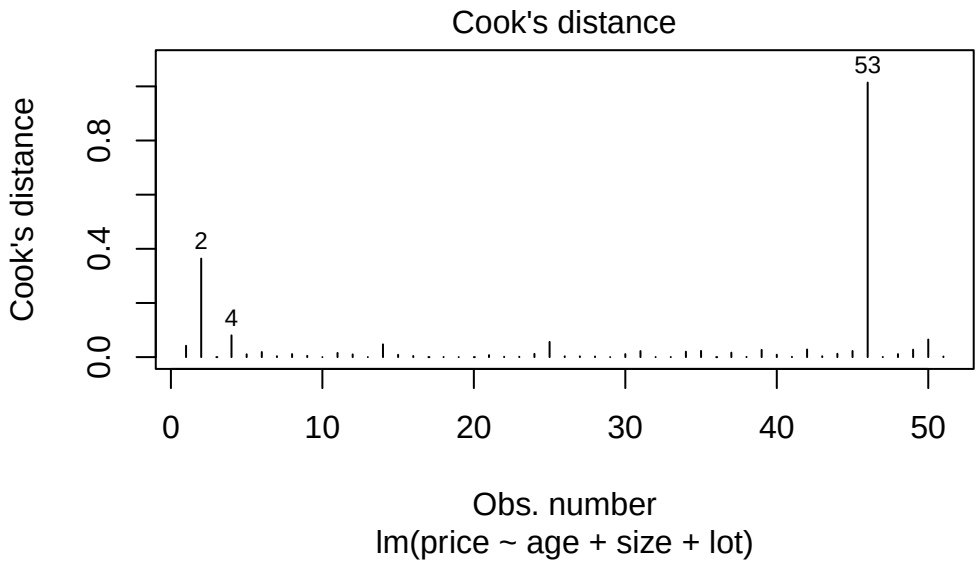
```
plot(lm_lt200k,which = 1)
```



The residuals versus fitted values plot is much improved. There do not appear to be any

extreme outliers.

```
plot(lm_lt200k,which = 4)
```



One observation has Cook's D above 1, but let's leave this observation in.

The regression based on the homes with prices less than 200k is a better fit to the data, so it will be more reliable for making predictions.

iii)

It could be useful for a home buyer to know its residual from one of these models. This is the difference in its actual price and its predicted price based on its characteristics. A positive residual may suggest that the home is over-priced, while a negative residual may mean that the home is a good bargain.

b)

Consider also the fp and garage variables while doing backward stepwise regression:

```
# manually remove missing values before using the step() function
hp_lt200k_noNA <- hp_lt200k[!is.na(hp_lt200k$lot),]
lm_all <- lm(price ~ age + bed + bath + size + lot + fp + garage,data = hp_lt200k_noNA)
step(lm_all,direction="backward",scope= formula(lm_all),trace = 0)
```

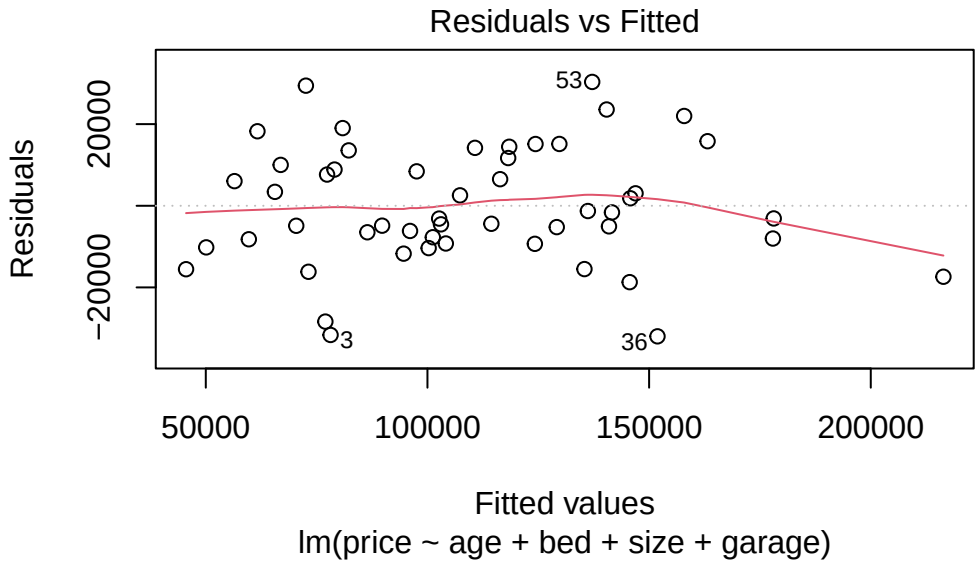
```
Call:
lm(formula = price ~ age + bed + size + garage, data = hp_lt200k_noNA)

Coefficients:
(Intercept)          age          bed         size       garage
   34485.53      -293.12    -11227.23        53.52     10378.72
```

The garage variable was included.

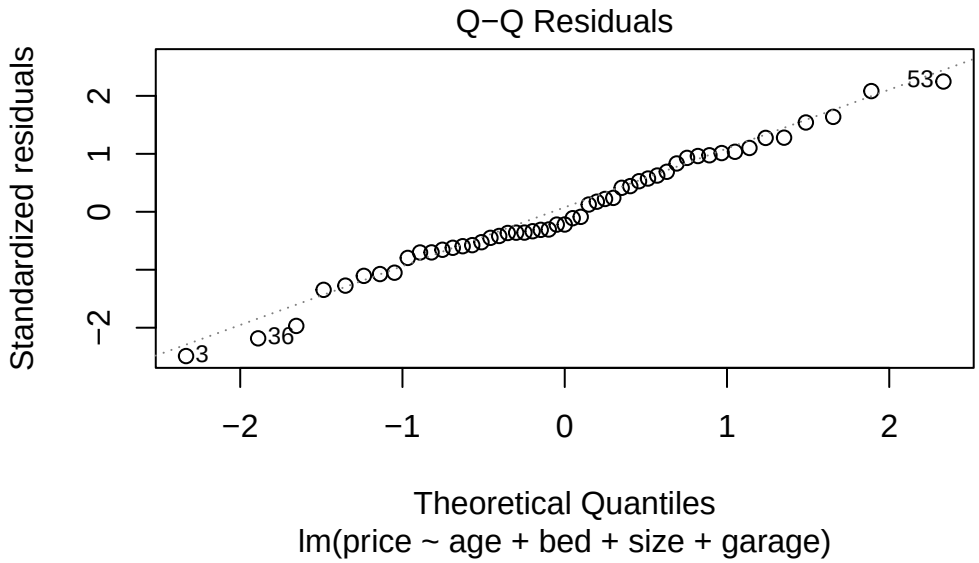
```
lm_garage <- lm(price ~ age + bed + size + garage,data = hp_lt200k_noNA)
summary(lm_garage)
```

```
Call:
lm(formula = price ~ age + bed + size + garage, data = hp_lt200k_noNA)

Residuals:
   Min     1Q Median     3Q    Max
-31987  -8724  -3106  10853  30337

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  34485.526  10833.530   3.183 0.002612 **
age           -293.120    170.069  -1.724 0.091508 .
bed         -11227.226   3813.627  -2.944 0.005066 **
size            53.520      4.521  11.839 1.45e-15 ***
garage       10378.720   2574.571   4.031 0.000207 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15150 on 46 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8671,    Adjusted R-squared:  0.8555
F-statistic: 75.03 on 4 and 46 DF,  p-value: < 2.2e-16
```

```
plot(lm_garage,which = 1)
```




```
plot(lm_garage,which = 2)
```



The model with the age, bed, size, and garage variables appears to be a pretty good model.

Chp 8 Ex 16

a)

The test statistic of the full-reduced model F test is equal to

```
n <- 50
p <- 3
s <- 1
SSEred <- 256
SSEfull <- 194
Fstat <- ((SSEred - SSEfull)/s) / (SSEfull / (n - (p + 1)))
```

The p-value is

```
pval <- 1 - pf(Fstat,s,n-(p+1))
pval
```

```
[1] 0.0003812353
```

So we reject H_0: $\beta_3 = 0$.

b)

The value of the F statistic is the square of $T_{\text{stat}} = \frac{\hat{\beta}_3}{\hat{\sigma}\sqrt{\Omega_{33}/n}}$. Since $\hat{\beta}_3 = 2.1$ is positive, T_{stat} is the positive square root of the F statistic value.

```
Tstat <- sqrt(Fstat)
Tstat
```

```
[1] 3.834192
```

c)

The estimated standard error of $\hat{\beta}_3$, which is given by $\text{se}\{\hat{\beta}_3\} = \hat{\sigma}\sqrt{\Omega_{33}/n}$, is equal to $\hat{\beta}_3/T_{\text{stat}}$.

```r
b3hat <- 2.1
se_b3hat <- b3hat / Tstat
se_b3hat
```

```
[1] 0.5477034
```

From here, a 95% confidence interval for β_3 can be constructed as

```r
lo <- b3hat - qt(.975,n - (p + 1)) * se_b3hat
up <- b3hat + qt(.975,n - (p + 1)) * se_b3hat
c(lo,up)
```

```
[1] 0.9975303 3.2024697
```

Chp 8 Ex 19

a)

For Model 1:

- $\text{SS}_{\text{Reg}} = \text{SS}_{\text{Total}} R^2 = 45.778 \cdot 0.07 = 3.20446$,
- $\text{SS}_{\text{Error}} = \text{SS}_{\text{Total}} - \text{SS}_{\text{Reg}} = 45.778 - 3.20446 = 42.57354$
- $F_{\text{stat}} = \frac{\text{SS}_{\text{Reg}}/p}{\text{SS}_{\text{Error}}/(n-(p+1))} = (3.20446/2)/(42.57354/(100-(2+1))) = 3.650538$
- p-value is $P(F > 3.650538)$, where $F \sim F_{2,97}$. This is `1 - pf(3.650538,2,97)` $= 0.0296089$.

So the overall F test rejects the null hypothesis.

For Model 2:

- $\text{SS}_{\text{Reg}} = \text{SS}_{\text{Total}} R^2 = 45.778 \cdot 0.19 = 8.69782$,
- $\text{SS}_{\text{Error}} = \text{SS}_{\text{Total}} - \text{SS}_{\text{Reg}} = 45.778 - 8.69782 = 37.08018$
- $F_{\text{stat}} = \frac{\text{SS}_{\text{Reg}}/p}{\text{SS}_{\text{Error}}/(n-(p+1))} = (8.69782/4)/(37.08018/(100-(4+1))) = 5.570988$
- p-value is $P(F > 5.570988)$, where $F \sim F_{4,95}$. This is `1 - pf(5.570988,4,95)` $= 4.5094394 \times 10^{-4}$.

So the overall F test rejects the null hypothesis for Model 2 also.

b)

The T statistic is $-0.56/0.17 = -3.294$, which is greater in absolute value than $t_{95,0.025} =$ `qt(.975,95)` $= 1.985251$, so we conclude that the type of work *is* associated with the response.

c)

A 95% confidence interval for the coefficient corresponding to type of work is given by

$$-0.56 \pm (1.985251)(0.17) = (-0.897, -0.223).$$

d)

The full-reduced model F test of whether type of work and employment length make a contribution to the model has test statistic equal to

$$\frac{(42.57354 - 37.08018)/2}{37.08018/(100 - (4+1))} = 7.037037$$

The critical value at significance level 0.05 is $F_{2,95,0.05} =$ `qf(.95,2,95)` $=3.0922174$.

So we reject the null hypothesis that type of work and employment length are uninformative to the response.