Chapter 3

Conceptual

1. Describe the null hypothesis to which the p-values given in Table 3.4 correspond. Explain what the conclusions you can draw based on these p-values. Your explanation should be phrased in terms of `sales`, `TV`, `radio`, and `newspaper`, rather than in terms of the coefficients of the linear model.

`Intercept`. H_0 : In the absence of advertising, sales are zero. ($\beta_0 = 0$)
H_A : In the absence of advertising, sales are not zero. ($\beta_0 \neq 0$)

`TV`. H_0 : Given a fixed level of `radio` and `newspaper` advertising, a change in `TV` advertising will not effect sales. ($\beta_1 = 0$)
H_A : Given a fixed level of `radio` and `newspaper` advertising, a change in `TV` advertising will effect sales. ($\beta_1 \neq 0$)

`radio`. H_0 : Given a fixed level of `TV` and `newspaper` advertising, a change in `radio` advertising will not effect sales. ($\beta_2 = 0$)
H_A : Given a fixed level of `TV` and `newspaper` advertising, a change in `radio` advertising will effect sales. ($\beta_2 \neq 0$)

`newspaper`. H_0 : Given a fixed level of `TV` and `radio` advertising, a change in `newspaper` advertising will not effect sales. ($\beta_3 = 0$)
H_A : Given a fixed level of `TV` and `radio` advertising, a change in `newspaper` advertising will effect sales. ($\beta_3 \neq 0$)

Since the first three null hypotheses are all < 0.0001, we can reject their null hypotheses in favor of their alternative at the 0.0001 significance level. Thus we are 99.99% confident that changes in `TV` and `radio` will effect `Sales`, however `Sales` will not be zero if no money is spent on advertising. Since the p-value for `newspaper` is high, we cannot conclude that `newspaper` has an effect on `Sales`.

2. Carefully explain the differences between the KNN classifier and the KNN regression methods.

The KNN classifier is used to predict values of a qualitative response variable. KNN regression uses a similar formula to predict values of a quantitative response variable.

3. Suppose we have a data set with five predictors, $X_1 =$ GPA, $X_2 =$ IQ, $X_3 =$ Gender (1 for female and 0 for male), $X_4 =$ Interaction between GPA and IQ and $X_5 =$ Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$ and $\hat{\beta}_5 = -10$.

   (a) Which answer is correct and why?

i. For a fixed value of IQ and GPA, males earn more on average than females.

ii. For a fixed value of IQ and GPA, females earn more on average than males.

iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

iv. For a fixed value of GPA and IQ, females earn more on average than males provided the GPA is high enough.

Since $\hat{\beta}_3 > 0$, females will have a larger starting salary than males if the interaction term is ignored. However, since $\hat{\beta}_5 < 0$, the interaction term will be negative and thus decrease the premium for females as GPA increases and eventually the interaction term will dominate the initial premium. Thus, iii. is correct.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

$$50 + 20(4.0) + 0.07(110) + 35(1) + 0.01(4.0)(110) - 10(4.0)(1) = 137.1 \quad (1)$$

(c) True or False: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

False. To decide whether the interaction term is significant we would need its p-value to be small. Since IQ is relatively large in scale we would actually expect any related coefficients to be small, even when they are highly significant.

4. I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fir a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

(a) Suppose the true relationship between X and Y is linear. Consider the training residual sum of squares for the linear regression, and also the cubic regression. Would we expect one to be lower than the other, would we expect them to be about the same, or is there not enough information to tell? Justify you answer.

The amount would depend on the data, but we would expect the cubic model to have a lower training RSS since the extra flexibility of the model will allow it to overfit the data.

(b) Answer (a) using test rather than training RSS.

The linear model should have a lower test RSS, since the cubic model will overfit the data, increasing its test RSS.

(c) Suppose the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear and cubic regressions. Would we expect them to be the same, or is there not enough information to tell? Justify your answer.

We would again expect the cubic model to have a smaller training RSS, since the model is more flexible and thus more capable of closely fitting the given data.

(d) Answer (c) using the test rather than the training RSS.

Here there is not enough information to tell, since the sizes of the training RSS would depend on the degree of nonlinearity in the data. (i.e. whether the extra flexibility of

the cubic model is causing overfitting or if that degree of nonlinearity is actually present in the data.)

5. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form

$$\hat{y}_i = x_i \hat{\beta}_i, \tag{2}$$

where

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{j=1}^{n} x_i j^2}. \tag{3}$$

Show that we can write

$$\hat{y}_i = \sum_{k=1}^{n} a_k y_k. \tag{4}$$

What is a_k?

Note that,

$$\begin{aligned} \hat{y}_i &= \frac{\sum_{k=1}^{n} x_k y_k}{\sum_{j=1}^{n} x_j^2} x_i \\ &= \sum_{k=1}^{n} \left(\frac{x_i}{\sum_{j=1}^{n} x_j^2} \right) x_k y_k. \end{aligned} \tag{5}$$

Thus, $\hat{y}_i = \sum_{k=1}^{n} a_k y_k$ where $a_k = \frac{x_i x_k}{\sum_{j=1}^{n} x_j^2}$.

6. Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$.

Note that the point (a, b) lies on the line $y = \hat{\beta}_1 x + \hat{\beta}_0$ is and only if $b = \hat{\beta}_1 a + \hat{\beta}_0$. Thus since, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, it follows that $\bar{y} = \hat{\beta}_1 \bar{x} + \hat{\beta}_0$ and hence $(\bar{x}, \bar{y})$ is on the least-squares line.

7. It is claimed in the text that in the case of simple linear regression of Y onto X, the R^2 statistic is equal to the square of the correlation between X and Y. Prove that this is the case. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.

Applied

8. This question involves the use of simple linear regression on the `Auto` data set.

   (a) Use the `lm()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Use the `summary()` function to print the results.

```
> attach(Auto)
> lm.fit=lm(mpg~horsepower)
> summary(lm.fit)
```

```
Call:
lm(formula = mpg ~ horsepower)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861  0.717499   55.66   <2e-16 ***
horsepower  -0.157845  0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .  0.1    1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

Comment on the output. For example:

i. Is there a relationship between the predictor and the response?
There is a significant relationship between mpg an horsepower. Testing the null-hypothesis that all coefficients are equal to zero we obtain a F-statisitic of 599.7 witha p-value of $2.2 \times 10^{-16} \approx 0$. Thus we may reject the null hypothesis, and conclude that there is a relationship between mpg and horsepower, at any reasonable significance level.

ii. How strong is the relationship between the predictor and the response?
The relationship is somewhat strong as the regression has an $R^2 = 0.6059$, which indicates that approximately 60% of the variability in mpg is explained by horsepower.

iii. Is the relationship between the predictor and the response positive or negative?
The relationship is negative, since the coefficient of horsepower is less than zero.

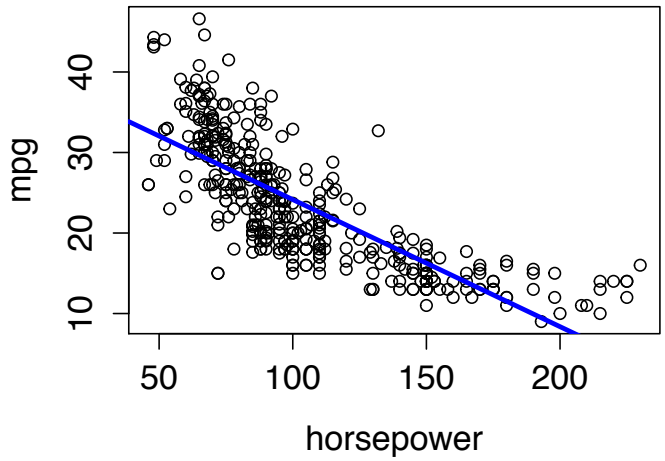
iv. What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95% confidence and prediction intervals?

```
> predict(lm.fit, data.frame(horsepower=c(98)), interval="confidence")
       fit      lwr      upr
1 24.46708 23.97308 24.96108
> predict(lm.fit, data.frame(horsepower=c(98)), interval="prediction")
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

(b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.

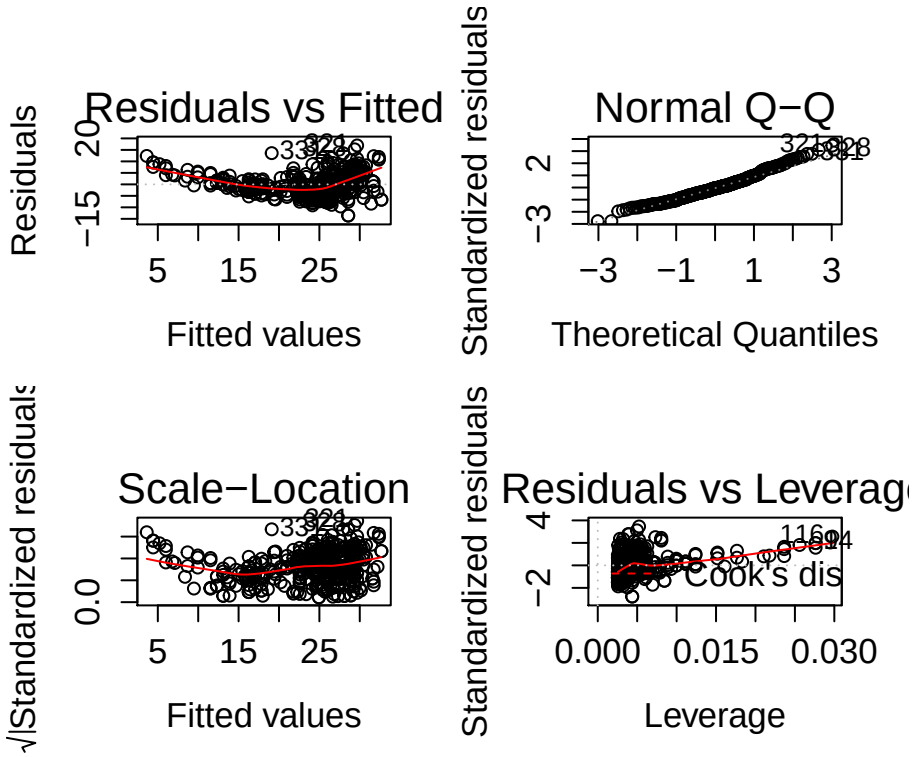
```
> plot(horsepower, mpg)
```

```
> abline(lm.fit, col="blue", lwd=3)
```



(c) Use the `plot()` function to produce diagnostic plots of the least squares fit. Comment on any problems you see with the fit.

```
> par(mfrow=c(2,2))
> plot(lm.fit)
```

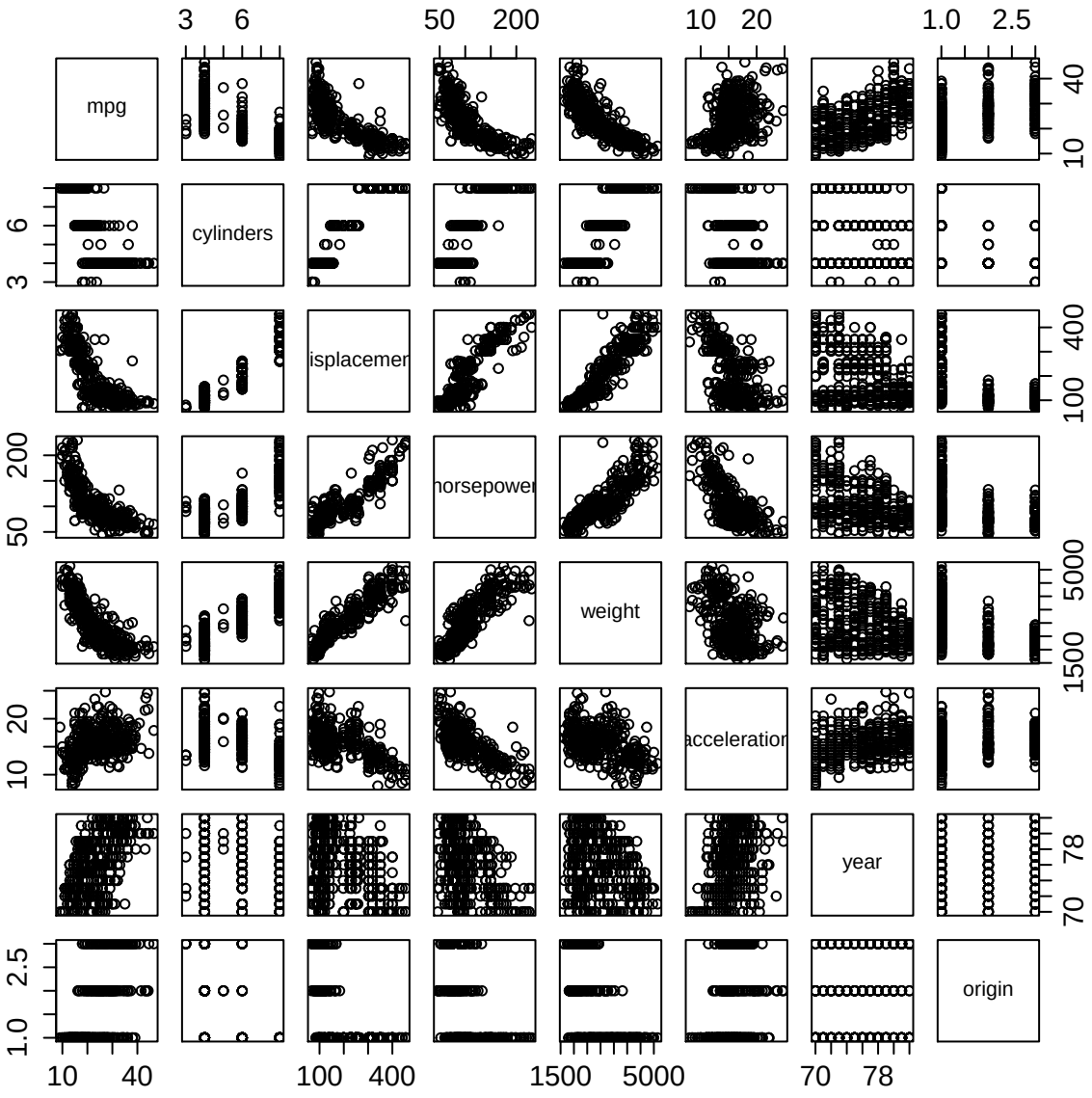


The residuals vs. fitted graph shows that small and large values of the predictor have substantially larger residuals. This indicates nonlinearity in the true relationship between mpg and horsepower.

9. This question involves the use of multiple linear regression on the `Auto` data set.

   (a) Produce a scatter plot matrix which includes all of the variables in the data set.

```
> plot(Auto[1:8])
```



   (b) Compute the matrix of correlations between the variables using the function `cor()`. You

will need to exclude the name variable, which is qualitative.

```
> cor(Auto[1:8])
                mpg  cylinders displacement horsepower weight acceleration
                    year    origin
mpg          1.0000000 -0.7776175 -0.8051269 -0.7784268 -0.8322442 0.4233285
    0.5805410 0.5652088
cylinders   -0.7776175 1.0000000   0.9508233 0.8429834 0.8975273  -0.5046834
    -0.3456474 -0.5689316
displacement -0.8051269 0.9508233 1.0000000 0.8972570 0.9329944  -0.5438005
    -0.3698552 -0.6145351
horsepower  -0.7784268 0.8429834   0.8972570 1.0000000 0.8645377  -0.6891955
    -0.4163615 -0.4551715
weight      -0.8322442 0.8975273   0.9329944 0.8645377 1.0000000  -0.4168392
    -0.3091199 -0.5850054
acceleration 0.4233285 -0.5046834 -0.5438005 -0.6891955 -0.4168392 1.0000000
    0.2903161 0.2127458
year         0.5805410 -0.3456474 -0.3698552 -0.4163615 -0.3091199 0.2903161
    1.0000000 0.1815277
origin       0.5652088 -0.5689316 -0.6145351 -0.4551715 -0.5850054 0.2127458
    0.1815277 1.0000000
```

(c) Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results.

```
> lm.fit2=lm(mpg~.-name, data=Auto)
> summary(lm.fit2)

Call:
lm(formula = mpg ~ . - name, 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
```

Comment on the output. For instance:

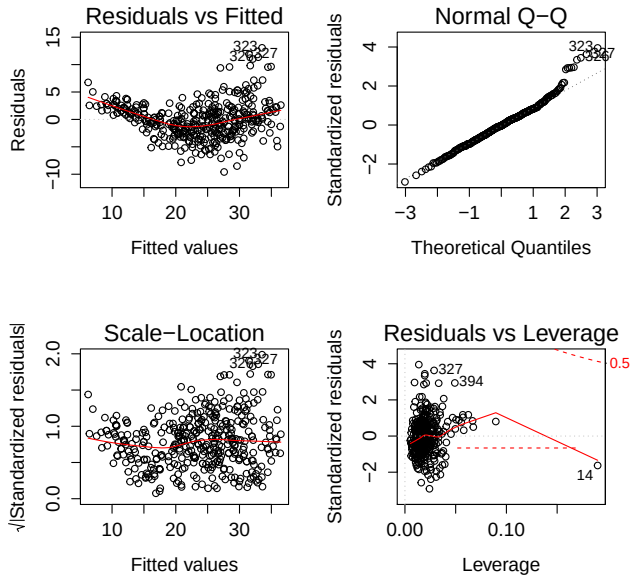
i. Is there a relationship between the predictors and the response?
There is a significant relationship between mpg and the remaining predictors. Testing the null-hypothesis that all coefficients are equal to zero we obtain a F-statisitic of 252.4 with a p-value of $2.2 \times 10^{-16} \approx 0$. Thus we may reject the null hypothesis, and conclude that there is a relationship between mpg and horsepower, at any reasonable significance level.

ii. Which predictors appear to have a statistically significant relationship to the response?
The displacement, weight, year and origin all have a statistically significant relationship with mpg. Weight, year and origin are all significant at the 0.001 level and displacement is significant at the 0.01 level.

iii. What does the coefficient for the `year` variable suggest?
The coefficient for year is $0.75 > 0$. This indicates that year and mpg are positively correlated. In particular cars have become more fuel efficient over time.

(d) Use the `plot()` function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

```
> plot(lm.fit2)
```



The u shape of the fitted vs. residuals plots suggests some nonlinearity of the data. The fitted values vs. studentized residuals all lie within the range ± 3, and hence seem reasonable. The point 14 in the leverage vs. standardized residuals plot has large leverage, but relatively modest residual.

(e) Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

```
> tAuto=Auto[1:8]
> lm.fitfull=lm(mpg~.+.*., tAuto)
> summary(lm.fitfull)

Call:
lm(formula = mpg ~ . + . * ., data = tAuto)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6303 -1.4481 0.0596 1.2739 11.1386

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            3.548e+01 5.314e+01  0.668  0.50475
cylinders              6.989e+00 8.248e+00  0.847  0.39738
displacement          -4.785e-01 1.894e-01 -2.527  0.01192 *
horsepower             5.034e-01 3.470e-01  1.451  0.14769
weight                 4.133e-03 1.759e-02  0.235  0.81442
acceleration          -5.859e+00 2.174e+00 -2.696  0.00735 **
year                   6.974e-01 6.097e-01  1.144  0.25340
origin                -2.090e+01 7.097e+00 -2.944  0.00345 **
cylinders:displacement -3.383e-03 6.455e-03 -0.524  0.60051
cylinders:horsepower   1.161e-02 2.420e-02  0.480  0.63157
cylinders:weight       3.575e-04 8.955e-04  0.399  0.69000
cylinders:acceleration 2.779e-01 1.664e-01  1.670  0.09584 .
cylinders:year        -1.741e-01 9.714e-02 -1.793  0.07389 .
cylinders:origin       4.022e-01 4.926e-01  0.816  0.41482
displacement:horsepower -8.491e-05 2.885e-04 -0.294  0.76867
displacement:weight    2.472e-05 1.470e-05  1.682  0.09342 .
displacement:acceleration -3.479e-03 3.342e-03 -1.041  0.29853
displacement:year      5.934e-03 2.391e-03  2.482  0.01352 *
displacement:origin    2.398e-02 1.947e-02  1.232  0.21875
horsepower:weight     -1.968e-05 2.924e-05 -0.673  0.50124
horsepower:acceleration -7.213e-03 3.719e-03 -1.939  0.05325 .
horsepower:year       -5.838e-03 3.938e-03 -1.482  0.13916
horsepower:origin      2.233e-03 2.930e-02  0.076  0.93931
weight:acceleration    2.346e-04 2.289e-04  1.025  0.30596
weight:year           -2.245e-04 2.127e-04 -1.056  0.29182
weight:origin         -5.789e-04 1.591e-03 -0.364  0.71623
acceleration:year      5.562e-02 2.558e-02  2.174  0.03033 *
acceleration:origin    4.583e-01 1.567e-01  2.926  0.00365 **
year:origin            1.393e-01 7.399e-02  1.882  0.06062 .
---
```

```
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 2.695 on 363 degrees of freedom
Multiple R-squared: 0.8893, Adjusted R-squared: 0.8808
F-statistic: 104.2 on 28 and 363 DF, p-value: < 2.2e-16
```

Fitting the model with all possible interaction terms we see that on the displacement:year, acceleration:year and acceleration:origin interaction terms have significance. We now fit a model containing only these interaction terms.

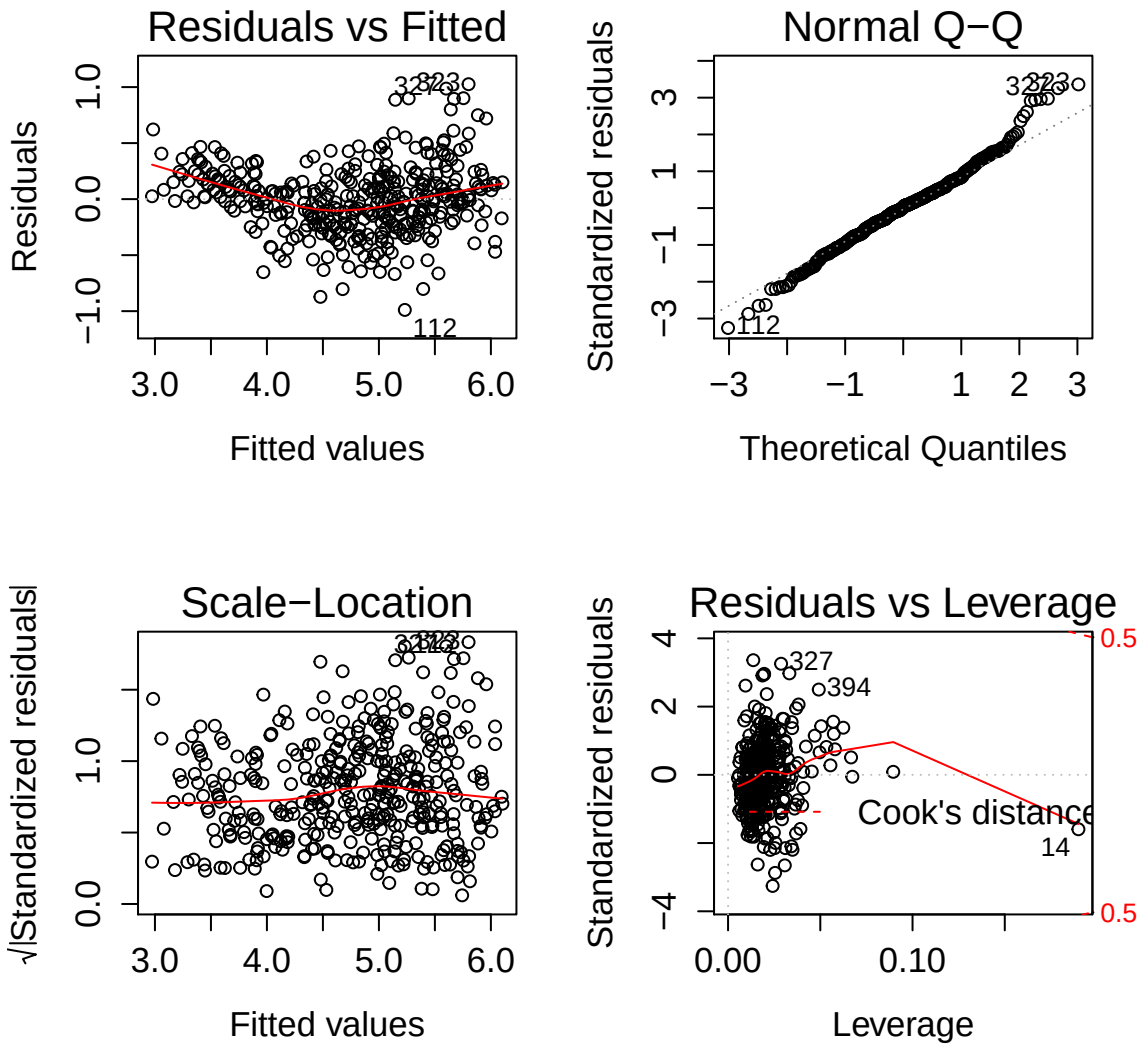
```
> lm.fit2=lm(mpg~.+displacement:year+acceleration:year+acceleration:origin,
    tAuto)
> summary(lm.fit2)

Call:
lm(formula = mpg ~ . + displacement:year + acceleration:year +
    acceleration:origin, data = tAuto)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7659 -1.8996 0.0241 1.4837 12.2739

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       11.8757343 27.4715000 0.432  0.6658
cylinders         -0.1269380 0.3026441 -0.419  0.6751
displacement       0.1708125 0.0430357  3.969 8.62e-05 ***
horsepower        -0.0370510 0.0126847 -2.921  0.0037 **
weight            -0.0048741 0.0006195 -7.868 3.76e-14 ***
acceleration      -3.6459893 1.4303639 -2.549  0.0112 *
year               0.5476463 0.3561807  1.538  0.1250
origin            -7.0940248 1.5825945 -4.483 9.77e-06 ***
displacement:year -0.0022809 0.0005711 -3.994 7.81e-05 ***
acceleration:year  0.0374758 0.0186441  2.010  0.0451 *
acceleration:origin 0.5078440 0.0957537 5.304 1.93e-07 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 3.012 on 381 degrees of freedom
Multiple R-squared: 0.8548, Adjusted R-squared: 0.851
F-statistic: 224.4 on 10 and 381 DF, p-value: < 2.2e-16
```

The displacement:year and acceleration:origin terms are both highly significant, however the acceleration:year term is only moderately significant. These terms make intuitive sense since the negative coefficient displacement:year indicates that as technology advances and combustion becomes more efficient, cars pay less of a mpg penalty for increased displacement. The positive coefficient of acceleration:origin indicates that European and Japanese (origin=2 and 3, resp) performance cars are relatively more fuel efficient than American (origin=1) performance cars. The interaction terms increased

R^2 by a little over 3% and hence are only moderately helpful.

(f) Try a few different transformations of the variables, such as $\log(X), \sqrt{X}, X^2$. Comment on your findings.

```
> lm.fit3=lm(log(mpg)~., data=tAuto)
> summary(lm.fit3)

Call:
lm(formula = log(mpg) ~ ., data = tAuto)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40955 -0.06533 0.00079 0.06785 0.33925

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.751e+00 1.662e-01 10.533 < 2e-16 ***
cylinders   -2.795e-02 1.157e-02 -2.415 0.01619 *
displacement 6.362e-04 2.690e-04  2.365  0.01852 *
horsepower  -1.475e-03 4.935e-04 -2.989 0.00298 **
weight      -2.551e-04 2.334e-05 -10.931 < 2e-16 ***
acceleration -1.348e-03 3.538e-03 -0.381 0.70339
year         2.958e-02 1.824e-03 16.211 < 2e-16 ***
origin       4.071e-02 9.955e-03  4.089 5.28e-05 ***
---
Signif. codes: 0 ***  0.001 **  0.01 *  0.05 .  0.1    1

Residual standard error: 0.1191 on 384 degrees of freedom
Multiple R-squared: 0.8795, Adjusted R-squared: 0.8773
F-statistic: 400.4 on 7 and 384 DF, p-value: < 2.2e-16

> plot(lm.fit4)
```



Applying the log to mpg substantially flattens the residuals vs. fitted graph, it also increases R^2 by over 5% and substantially increases the F statistic.

10. This question should be answered using the `Carseats` data set.

    (a) Fit a multiple regression model to predict `Sales` using `Price`, `Urban` and `US`.

```
> attach(Carseats)
> lm.fit=lm(Sales~Price+Urban+US, data=Carseats)
> summary(lm.fit)

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) Provide an interpretation of each coefficient in the model. Be careful - Some of the variables in the model are qualitative!

The coefficient of Price indicates that for each \$1 increase in price total sales will decrease by 0.0546 thousand dollars. The coefficient of UrbanYes indicates that sales are 0.0219 thousand dollars lower in urban areas. The coefficient of USYes indicates that sales are 1.2 thousand dollars higher in US markets.

(c) Write out the model in equation form, being careful to handle the qualitative variables properly.

$$\text{Sales} = 13.043 - 0.054 \times \text{Price} - 0.0212 \times \text{UrbanYes} + 1.201 \times \text{USYes} \tag{6}$$

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

We can reject the null hypothesis $H_0 : \beta_j = 0$, for Intercept, Price and USYes at any significance level. We cannot reject the hypothesis for UrbanYes (unless we can stomach only 6.4% confidence).

(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of an association with the outcome.

```
> lm.fit2=lm(Sales~Price+US, data=Carseats)
> summary(lm.fit2)

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:
```

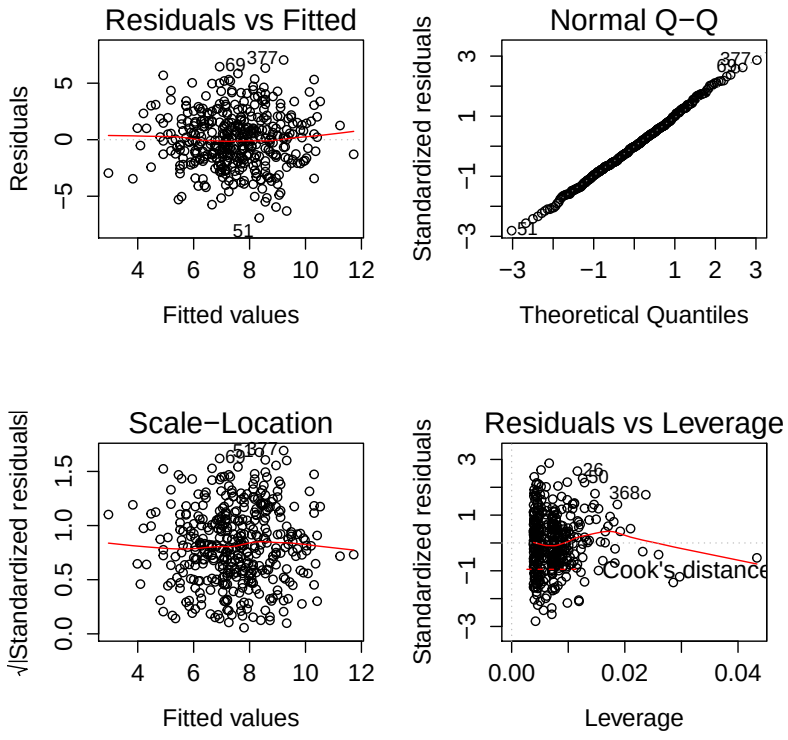
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.03079  0.63098 20.652 < 2e-16 ***
Price       -0.05448  0.00523 -10.416 < 2e-16 ***
USYes        1.19964  0.25846  4.641 4.71e-06 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

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
```

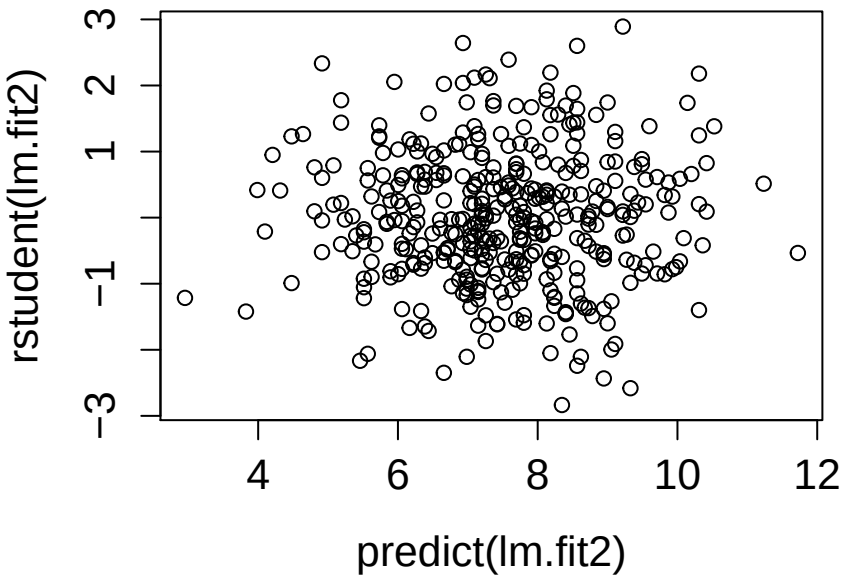
(f) How well do the models from (a) and (e) fit the data? Neither of the models are particularly good fits to the data. Both only explain about 24% of the variability in Sales.

(g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

```
> confint(lm.fit2)
                2.5 %     97.5 %
(Intercept) 11.79032020 14.27126531
Price       -0.06475984 -0.04419543
USYes        0.69151957 1.70776632
```

(h) Is there evidence of outliers or high leverage observations in the model from (e)?





There is no evidence of outliers since the studentized residuals are within the ± 3 range. There are some high leverage points, but none have a particularly large residual.

11. In this problem we will investigate the t-statistic for the null hypothesis $H_0 : \beta - 0$. in simple linear regression without an intercept. To begin, we generate a predictor x and a response y as follows.

```
> set.seed(1)
> x=rnorm(100)
> y=2*x+rnorm(100)
```

(a) Perform a simple linear regression of y onto x, *without* an intercept. Report the coefficient estimate $\hat{\beta}$, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results.

```
> summary(lm.fity)

Call:
lm(formula = y ~ x + 0)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9154 -0.6472 -0.1771 0.5056 2.3109

Coefficients:
```

```
  Estimate Std. Error t value Pr(>|t|)
x   1.9939     0.1065   18.73   <2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.9586 on 99 degrees of freedom
Multiple R-squared: 0.7798, Adjusted R-squared: 0.7776
F-statistic: 350.7 on 1 and 99 DF, p-value: < 2.2e-16
```

The estimate for $\hat{\beta}$ is 1.9939 with an associated standard error of 0.1065. The t-statistic and p-value associated with the null hypothesis $H_0 : \beta = 0$, are 18.73 and $< 2e - 16$, respectively. From this we can reject the null hypothesis at any significance level.

(b) Now perform a simple linear regression of x onto y without an intercept, and report the coefficient estimate, the standard error, and the corresponding t-statistic and p-values associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results.

```
> lm.fitx=lm(x~y+0)
> summary(lm.fitx)

Call:
lm(formula = x ~ y + 0)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8699 -0.2368 0.1030 0.2858 0.8938

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
y  0.39111    0.02089   18.73   <2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.4246 on 99 degrees of freedom
Multiple R-squared: 0.7798, Adjusted R-squared: 0.7776
F-statistic: 350.7 on 1 and 99 DF, p-value: < 2.2e-16
```

The estimate for $\hat{\beta}$ is 0.39111 with an associated standard error of 0.02089. The t-statistic and p-value associated with the null hypothesis $H_0 : \beta = 0$, are 18.73 and $< 2e - 16$, respectively. From this we can reject the null hypothesis at any significance level.

(c) What is the relationship between the results obtained in (a) and (b)?
While the two fitted lines are distinct but they have identical t-statistic and p-values associated to the null hypothesis $H_0 : \beta = 0$.

(d) For the regression of Y onto X without an intercept, the t-statistic for $H_0 : \beta = 0$ takes

the form $\frac{\hat{\beta}}{\text{SE}(\hat{\beta})}$ where $\hat{\beta}$ is given by (3.38) and where

$$\text{SE}(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}{(n-1)\sum_{j=1}^{n} x_j^2}} \tag{7}$$

Show algebraically and confirm numerically in R, that the t-statistic can be written as

$$t = \frac{\sqrt{n-1}\sum_{i=1}^{n} x_i y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} x_i y_i\right)^2}} \tag{8}$$

Note that

$$\begin{aligned} t =& \frac{\hat{\beta}}{\text{SE}(\hat{\beta})} \\ =& \frac{\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}{(n-1)\sum_{j=1}^{n} x_j^2}}} \\ =& \frac{\left(\sum_{i=1}^{n} x_i y_i\right)\sqrt{(n-1)\sum_{j=1}^{n} x_j^2}}{\left(\sum_{i=1}^{n} x_i^2\right)\sqrt{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}} \\ =& \frac{\sqrt{(n-1)}\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}}. \end{aligned} \tag{9}$$

Now

$$\begin{aligned} \sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2 =& \sum_{i=1}^{n}(y_i^2 - 2y_i x_i\hat{\beta} + x_i^2\hat{\beta}^2) \\ =& \sum_{i=1}^{n} y_i^2 - 2\hat{\beta}\sum_{i=1}^{n} y_i x_i + \hat{\beta}^2\sum_{i=1}^{n} x_i^2 \\ =& \sum_{i=1}^{n} y_i^2 - 2\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right)\sum_{i=1}^{n} y_i x_i + \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right)^2\sum_{i=1}^{n} x_i^2 \\ =& \sum_{i=1}^{n} y_i^2 - 2\frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2} + \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2} \\ =& \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2}. \end{aligned} \tag{10}$$

Thus

$$
\begin{aligned}
t &= \frac{\sqrt{(n-1)}\sum_{i=1}^{n} x_i y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2\right)}}. \\
&= \frac{\sqrt{(n-1)}\sum_{i=1}^{n} x_i y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2}\right)}}. \\
&= \frac{\sqrt{n-1}\sum_{i=1}^{n} x_i y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} x_i y_i\right)^2}}.
\end{aligned}
\tag{11}
$$

(e) Using the results from (d), argue that the t-statistic for the regression of y onto x is the same as the t-statistic for the regression of x onto y.

Note that the above expression (8) is symmetric in x and y, thus the t-statistic of the regression of x onto y is identical to the t-statistic of the regression of y onto x.

(f) In R show that when regression is performed *with* an intercept, the t-statistic for $H_0 : \beta = 0$ is the same for the regression of y onto x as it is for the regression of x onto y.

```
> lm.fit1=lm(y~x)
> summary(lm.fit1)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8768 -0.6138 -0.1395 0.5394 2.3462

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03769  0.09699 -0.389   0.698
x            1.99894  0.10773 18.556  <2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.9628 on 98 degrees of freedom
Multiple R-squared: 0.7784, Adjusted R-squared: 0.7762
F-statistic: 344.3 on 1 and 98 DF, p-value: < 2.2e-16

> lm.fit2=lm(x~y)
> summary(lm.fit2)

Call:
lm(formula = x ~ y)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-0.90848 -0.28101 0.06274 0.24570 0.85736

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.03880   0.04266   0.91    0.365
y           0.38942   0.02099  18.56   <2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1    1

Residual standard error: 0.4249 on 98 degrees of freedom
Multiple R-squared: 0.7784, Adjusted R-squared: 0.7762
F-statistic: 344.3 on 1 and 98 DF, p-value: < 2.2e-16
```

Note the t-statistics are identical for the two regressions.

12. This problem involves simple linear regression without an intercept.

    (a) Recall that the coefficient estimate $\hat{\beta}$ for the linear regression of Y onto X without an intercept is given by $\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{j=1}^{n} x_j^2}$. Under what circumstance is the coefficient estimate for the regression of X onto Y the same as the coefficient estimate for the regression of Y on to X?

    Since the numerator of the above expression is symmetric in x and y, the coefficients will be the same if and only if $\sum_{j=1}^{n} x_j^2 = \sum_{j=1}^{n} y_j^2$

    (b) Generate an example in R with $n = 100$ observations in which the coefficient estimate for the regression of X onto Y is different from the coefficient estimate for the regression of Y onto X.

```
> x=rnorm(100)
> y=4*x+.01*rnorm(100)
> lm.fit1=lm(y~x+0)
> lm.fit2=lm(x~y+0)
> summary(lm.fit1)

Call:
lm(formula = y ~ x + 0)

Residuals:
     Min       1Q   Median       3Q      Max
-0.020375 -0.006005 -0.000025 0.007450 0.036722

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x 4.0001813 0.0009484   4218   <2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1    1

Residual standard error: 0.01017 on 99 degrees of freedom
```

```
Multiple R-squared:    1,   Adjusted R-squared:    1
F-statistic: 1.779e+07 on 1 and 99 DF, p-value: < 2.2e-16

> summary(lm.fit2)

Call:
lm(formula = x ~ y + 0)

Residuals:
      Min        1Q    Median        3Q       Max
-0.0091901 -0.0018617 0.0000015 0.0014996 0.0050939

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
y 2.500e-01 5.927e-05   4218   <2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1    1

Residual standard error: 0.002543 on 99 degrees of freedom
Multiple R-squared:    1,   Adjusted R-squared:    1
F-statistic: 1.779e+07 on 1 and 99 DF, p-value: < 2.2e-16
```

(c) Generate an example in R with $n = 100$ observations in which the coefficient estimate for the regression of X onto Y is the same as the coefficient estimate for the regression of Y onto X.

```
> x=rnorm(100)
> y=sample(x)
> lm.fit1=lm(y~x+0)
> lm.fit2=lm(x~y+0)
> summary(lm.fit1)

Call:
lm(formula = y ~ x + 0)

Residuals:
     Min       1Q   Median       3Q      Max
-2.30940 -0.79090 0.06058 0.68602 2.38698

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x  0.09651    0.10003   0.965    0.337

Residual standard error: 1.025 on 99 degrees of freedom
Multiple R-squared: 0.009314, Adjusted R-squared: -0.0006931
F-statistic: 0.9307 on 1 and 99 DF, p-value: 0.337

> summary(lm.fit2)

Call:
```

```
lm(formula = x ~ y + 0)

Residuals:
     Min       1Q   Median       3Q      Max
-2.34313 -0.71072 -0.01516 0.74626 2.45342

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
y  0.09651    0.10003   0.965    0.337

Residual standard error: 1.025 on 99 degrees of freedom
Multiple R-squared: 0.009314, Adjusted R-squared: -0.0006931
F-statistic: 0.9307 on 1 and 99 DF, p-value: 0.337
```

13. In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use `set.seed(1)` prior to starting part (a) to ensure consistent results.

    (a) Using the `rnorm()` function, create a vector `x`, containing 100 observations drawn from a $N(0,1)$ distribution. This represents a feature, X.

    ```
    > set.seed(1)
    > x=rnorm(100)
    ```

    (b) Using the `rnorm()` function, create a vector `eps`, containing 100 observations drawn from a $N(0,0.25)$ distribution. This represents a feature, X.

    ```
    > eps=rnorm(100,0,0.5)
    ```

    (c) Using `x` and `eps`, generate a vector `y` according to the model

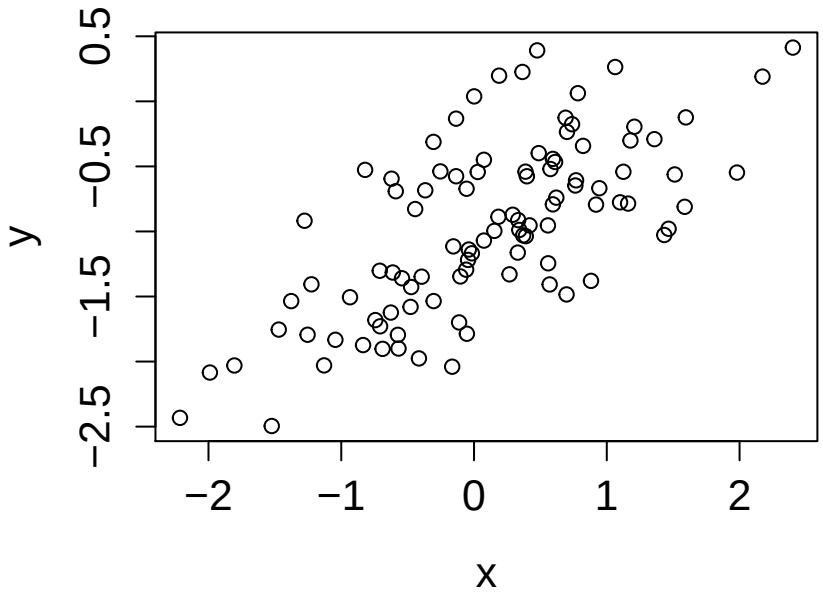
    $$Y = -1 + 0.5X + \epsilon \tag{12}$$

    What is the length of the vector `y`? What are the values of β_0 and β_1 in this linear model?

    ```
    > y=-1+0.5*x+eps
    ```

    The vector y has the same length as x, (i.e. 100). Also, $\beta_0 = -1$ and $\beta_1 = 0.5$.

    (d) Create a scatterplot displaying the relationship between `x` and `y`. Comment on what you observe.



The variables x and y are positively correlated, but there is a fair amount of noise.

(e) Fit a least squares model to predict y using x. Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1?

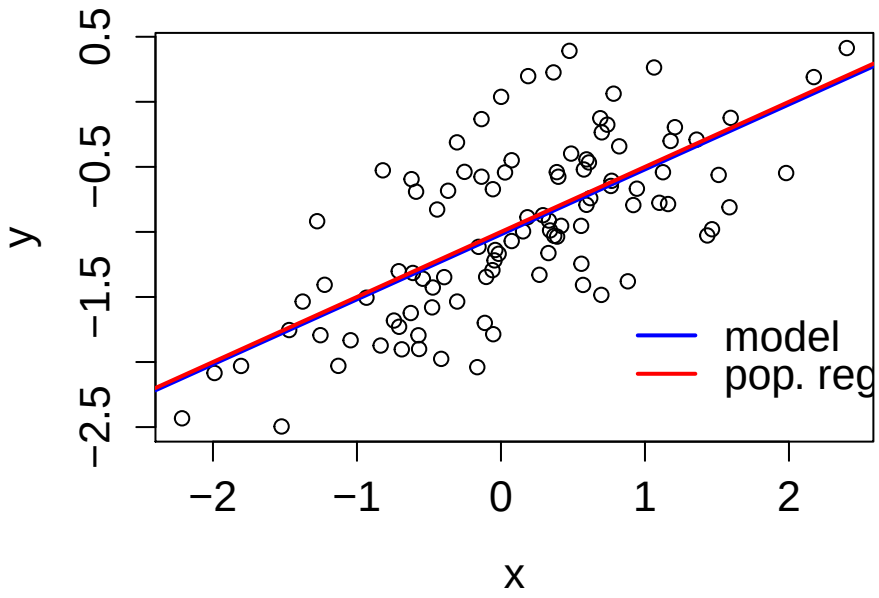
```
> lm.fit1=lm(y~x)
> summary(lm.fit1)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.93842 -0.30688 -0.06975 0.26970 1.17309

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.01885  0.04849 -21.010 < 2e-16 ***
x            0.49947   0.05386  9.273 4.58e-15 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1    1

Residual standard error: 0.4814 on 98 degrees of freedom
Multiple R-squared: 0.4674, Adjusted R-squared: 0.4619
F-statistic: 85.99 on 1 and 98 DF, p-value: 4.583e-15
```

The least squares regression is $y = 0.49947x - 1.01885$. This is very close to the popula-

tion regression line of $y = 0.5x - 1$. The estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are very good estimates of β_0 and β_1.

(f) Display the least squares line on the scatterplot obtained in (d). Draw the population regression line on the plot, in a different color. Use the `legend()` command to create an appropriate legend.

```
> plot(x,y)
> abline(lm.fit1,col="blue", lwd=2)
> abline(-1,0.5,col="red", lwd=2)
> legend(0.75,-1.5,legend=c("model", "pop. reg."), col=c("blue", "red"),
    lwd=2, bty="n")
```



(g) Now fit a polynomial regression model that predicts y using x and x^2. Is there evidence that the quadratic term improves the model fit? Explain your answer.

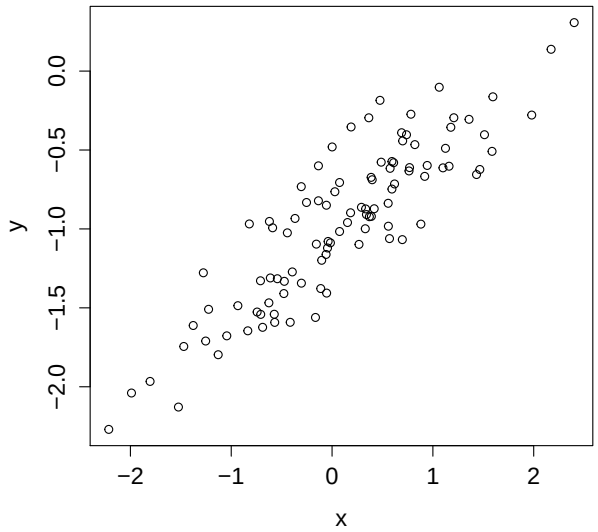
```
> lm.fit3=lm(y~x+I(x^2))
> summary(lm.fit3)

Call:
lm(formula = y ~ x + I(x^2))

Residuals:
     Min       1Q   Median       3Q      Max
-0.98252 -0.31270 -0.06441 0.29014 1.13500
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.97164  0.05883 -16.517 < 2e-16 ***
x            0.50858   0.05399  9.420 2.4e-15 ***
I(x^2)      -0.05946   0.04238 -1.403   0.164
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.479 on 97 degrees of freedom
Multiple R-squared: 0.4779, Adjusted R-squared: 0.4672
F-statistic: 44.4 on 2 and 97 DF, p-value: 2.038e-14
```

There is little evidence that the quadratic model improves the fit. The R^2 increases slightly, but the p-value of the t-statistic for the coefficient of the quadratic term is 0.164 and hence we can reject the hypothesis that the quadratic term is significant at the 0.1 significance level.

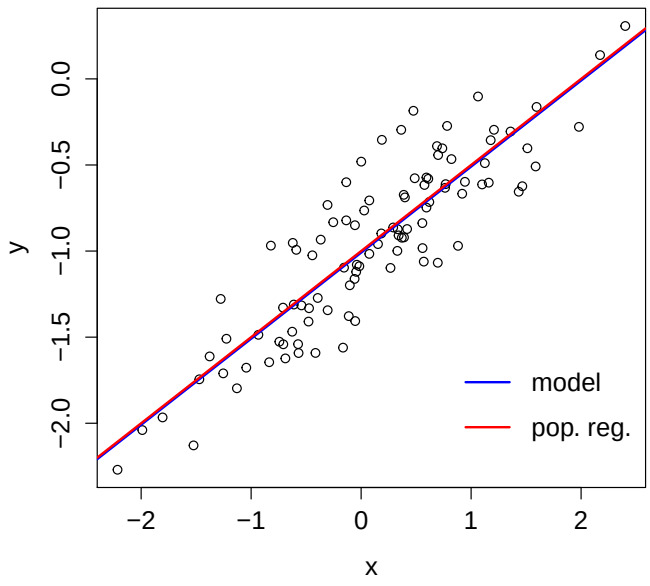
(h) Repeat (a)-(f) after modifying the data generation process in such a way that there is less noise in the data. The model (3.39) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term ϵ in (b). Describe your results.

```
> set.seed(1)
> x=rnorm(100)
> eps=rnorm(100, 0,0.25)
> y=-1+0.5*x+eps
> plot(x,y)
```



```
> lm.fit4=lm(y~x)
> summary(lm.fit4)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46921 -0.15344 -0.03487 0.13485 0.58654

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.00942  0.02425 -41.63  <2e-16 ***
x            0.49973   0.02693  18.56  <2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.2407 on 98 degrees of freedom
Multiple R-squared: 0.7784, Adjusted R-squared: 0.7762
F-statistic: 344.3 on 1 and 98 DF, p-value: < 2.2e-16
```

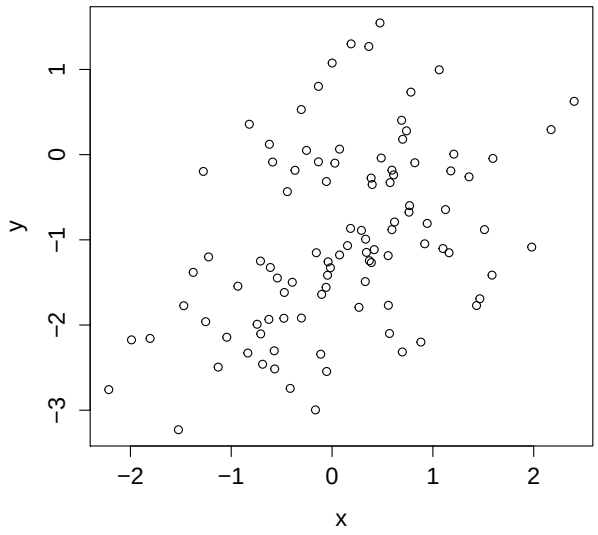
```
> plot(x,y)
> abline(lm.fit4, col="blue", lwd=2)
> abline(-1,0.5, col="red", lwd=2)
> legend(0.75,-1.5, legend=c("model", "pop. reg."), col=c("blue", "red"),
    lwd=2, bty="n")
```



This less noisy data has a stronger correlation between x and y. The R^2 of the model is substantially higher and the coefficient estimates are closer to the true population values.

(i) Repeat (a)-(f) after modifying the data generation process in such a way that there is more noise in the data. The model (3.39) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term ϵ in (b). Describe your results.

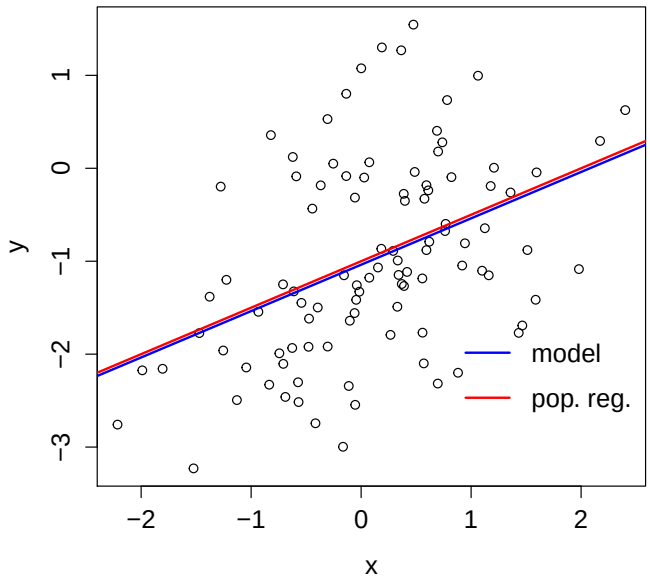
```
> set.seed(1)
> x=rnorm(100)
> eps=rnorm(100, 0,1)
> y=-1+0.5*x+eps
> plot(x,y)
```




```
> lm.fit5=lm(y~x)
> summary(lm.fit5)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8768 -0.6138 -0.1395 0.5394 2.3462

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.03769  0.09699 -10.699 < 2e-16 ***
x            0.49894  0.10773  4.632 1.12e-05 ***
```

```
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.9628 on 98 degrees of freedom
Multiple R-squared: 0.1796, Adjusted R-squared: 0.1712
F-statistic: 21.45 on 1 and 98 DF, p-value: 1.117e-05
```

```
> plot(x,y)
> abline(lm.fit5, col="blue", lwd=2)
> abline(-1,0.5, col="red", lwd=2)
> legend(0.75,-1.5, legend=c("model", "pop. reg."), col=c("blue", "red"),
    lwd=2, bty="n")
```



This more noisy data has a weaker correlation between x and y. The R^2 of the fitted model has decreased substantially and the coefficient estimates are further from the population values.

(j) What are the confidence intervals for β_0 and β_1 based on the original data set, the noisier data set and the less noisy data set? Comment on your results.

```
> confint(lm.fit1)
                2.5 %     97.5 %
(Intercept) -1.1150804 -0.9226122
x            0.3925794 0.6063602
> confint(lm.fit4)
                2.5 %     97.5 %
```

```
(Intercept) -1.0575402 -0.9613061
x           0.4462897 0.5531801
> confint(lm.fit5)
                2.5 %    97.5 %
(Intercept) -1.2301607 -0.8452245
x           0.2851588 0.7127204
```

The confidence intervals all contain the true population parameters. However, the intervals are longer for the noisier data and narrower for the less noisy data. This makes sense since less noise should translate into more confidence in our estimates.

14. This problem focuses on the collinearity problem.

(a) Perform the following commands in R:

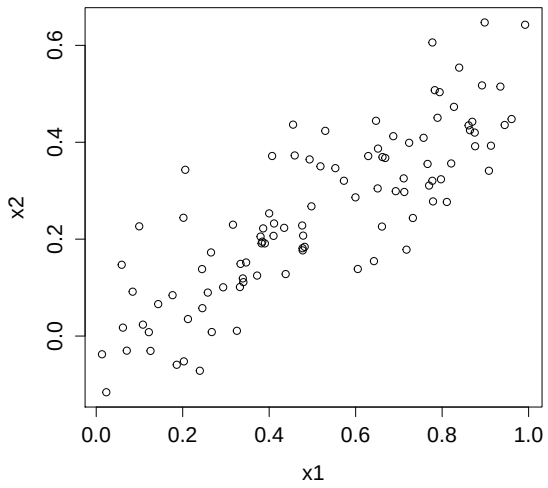
```
> set.seed(1)
> x1=runif(100)
> x2=0.5*x1+rnorm(100)/10
> y=2+2*x1+0.3*x2+rnorm(100)
```

The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

The regression coefficients are 2, 2 and 0.3 for the intercept, x1 and x2, resp.

(b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.

```
> cor(x1,x2)
[1] 0.8351212
```



(c) Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are $\hat{\beta}_0$,$\hat{\beta}_1$ and$\hat{\beta}_2$? How do these relate to the true β_0, β_1 and β_2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

```
> lm.fit=lm(y~x1+x2)
> summary(lm.fit)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8311 -0.7273 -0.0537 0.6338 2.3359

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.1305     0.2319   9.188 7.61e-15 ***
x1           1.4396     0.7212   1.996   0.0487 *
x2           1.0097     1.1337   0.891   0.3754
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.056 on 97 degrees of freedom
Multiple R-squared: 0.2088, Adjusted R-squared: 0.1925
F-statistic: 12.8 on 2 and 97 DF, p-value: 1.164e-05
```

For the fitted model we have $\hat{\beta}_0 = 2.1305$,$\hat{\beta}_1 = 1.4396$ and$\hat{\beta}_2 = 1.0097$. The estimate for $\hat{\beta}_0$ is reasonable, but the estimates for $\hat{\beta}_1$ and$\hat{\beta}_2$ differ substantially from the known values of β_1 and β_2. The hypothesis $H_0 : \beta_1 = 0$ can be rejected at the 0.05 significance level, but not at the 0.01 significance level. The hypothesis $H_0 : \beta_2 = 0$ cannot be rejected at any reasonable significance level.

(d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

```
> lm.fit2=lm(y~x1)
> summary(lm.fit2)

Call:
lm(formula = y ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.89495 -0.66874 -0.07785 0.59221 2.45560

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.1124     0.2307   9.155 8.27e-15 ***
x1           1.9759     0.3963   4.986 2.66e-06 ***
---
```

```
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.055 on 98 degrees of freedom
Multiple R-squared: 0.2024, Adjusted R-squared: 0.1942
F-statistic: 24.86 on 1 and 98 DF, p-value: 2.661e-06
```

This regression has a comparable fit to the regression using both variables as the RSE and R^2 are similar for the two models. However, with the one variable model we have much more confidence in the estimate for β_1. The p-value for the t-statistic corresponding to $\hat{\beta}_1$ is 2.66×10^{-6}. Hence we can reject $H_0 : \beta_1 = 0$ at the 0.001 significance level.

(e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

```
> lm.fit3=lm(y~x2)
> summary(lm.fit3)

Call:
lm(formula = y ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.62687 -0.75156 -0.03598 0.72383 2.44890

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.3899     0.1949  12.26  < 2e-16 ***
x2           2.8996     0.6330   4.58 1.37e-05 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.072 on 98 degrees of freedom
Multiple R-squared: 0.1763, Adjusted R-squared: 0.1679
F-statistic: 20.98 on 1 and 98 DF, p-value: 1.366e-05
```

The fit here is a bit worse than the previous two, with a slightly larger RSE and a substantially smaller R^2. The p-value for the t-statistic corresponding to $\hat{\beta}_1$ is 1.37×10^{-5}. Hence we can reject $H_0 : \beta_1 = 0$ at the 0.001 significance level.

(f) Do the results obtained in (c)-(e) contradict each other? Explain your answer.

These results do not contradict each other. We can be confident that there is a correlation between y and $x1$ and between y and $x2$ independently, but since $x1$ and $x2$ are themselves correlated, in the regression cannot decide how much of the value of y to attribute to $x1$ and how much to attribute to $x2$.

(g) Now suppose we obtain one additional observation, which was unfortunately mismeasured.

```
> x1=c(x1,0.1)
```

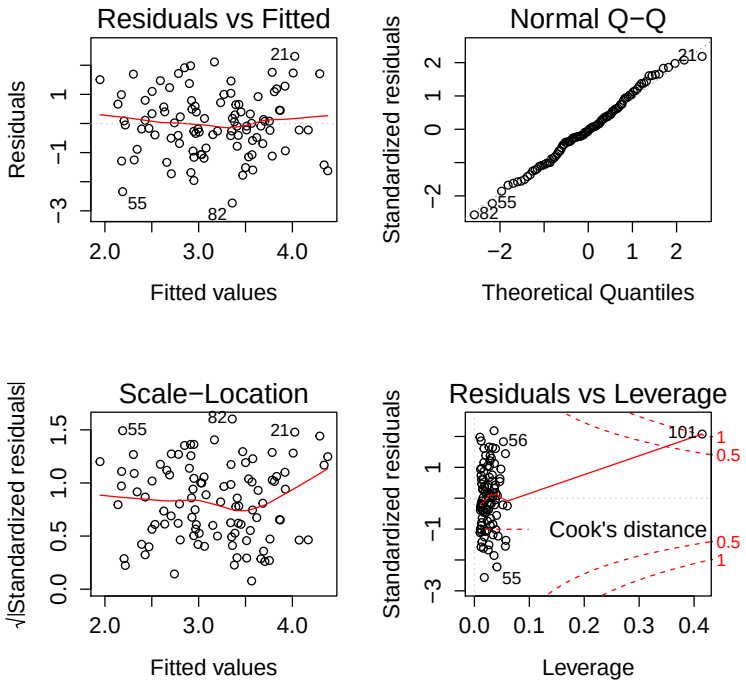
```
> x2=c(x2,0.8)
> y=c(y,6)
```

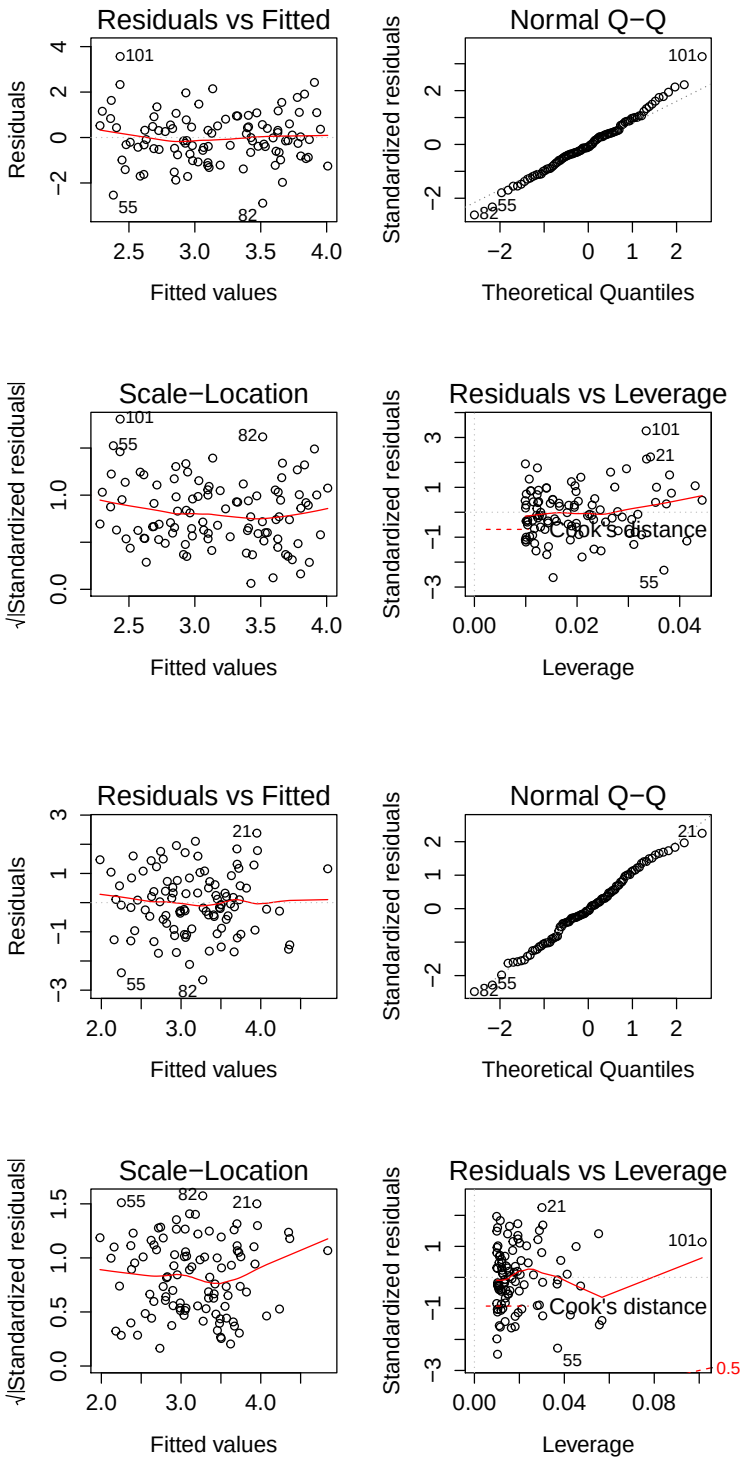
Re-fit the linear models from (c)-(e) using this new data. What effect does this new observation have on each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

```
> lm.fit1=lm(y~x1+x2)
> summary(lm.fit1)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.73348 -0.69318 -0.05263 0.66385 2.30619

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.2267     0.2314   9.624 7.91e-16 ***
x1           0.5394     0.5922   0.911  0.36458
x2           2.5146     0.8977   2.801  0.00614 **
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.075 on 98 degrees of freedom
Multiple R-squared: 0.2188, Adjusted R-squared: 0.2029
F-statistic: 13.72 on 2 and 98 DF, p-value: 5.564e-06

> lm.fit2=lm(y~x1)
> summary(lm.fit2)

Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8897 -0.6556 -0.0909 0.5682 3.5665

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.2569     0.2390   9.445 1.78e-15 ***
x1           1.7657     0.4124   4.282 4.29e-05 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.111 on 99 degrees of freedom
Multiple R-squared: 0.1562, Adjusted R-squared: 0.1477
F-statistic: 18.33 on 1 and 99 DF, p-value: 4.295e-05

> lm.fit3=lm(y~x2)
```

```
> summary(lm.fit3)

Call:
lm(formula = y ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.64729 -0.71021 -0.06899  0.72699  2.38074

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.3451     0.1912  12.264  < 2e-16 ***
x2           3.1190     0.6040   5.164 1.25e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.074 on 99 degrees of freedom
Multiple R-squared: 0.2122,  Adjusted R-squared: 0.2042
F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

This extra observation has little effect on the RSE or R^2 of the first model. It does however have a substantial effect on the fitted coefficients. In particular, now $\hat{\beta}_2$ is significant and $\hat{\beta}_1$ is not. The second model now substantially underestimates β_1 and has a worse RSE and R^2. The third model now over approximates β_1, though has a comparable RSE and R^2.





The new value is a high leverage point in the first and third models. It is an outlier in the second model, but not in the first or third.

15. This problem involve the Boston data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

    (a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

```
> library(MASS)
> Boston$chas<-factor(Boston$chas, labels=c("Y", "N"))
> attach(Boston)
> lm.zn=lm(crim~zn)
> summary(lm.zn)

Call:
lm(formula = crim ~ zn)

Residuals:
   Min     1Q Median     3Q    Max
-4.429 -4.222 -2.620  1.250 84.523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.45369    0.41722  10.675  < 2e-16 ***
zn          -0.07393    0.01609  -4.594 5.51e-06 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 8.435 on 504 degrees of freedom
Multiple R-squared: 0.04019, Adjusted R-squared: 0.03828
F-statistic: 21.1 on 1 and 504 DF, p-value: 5.506e-06

> lm.indus=lm(crim~indus)
> summary(lm.indus)

Call:
lm(formula = crim ~ indus)

Residuals:
    Min      1Q  Median      3Q     Max
-11.972  -2.698  -0.736   0.712  81.813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.06374    0.66723  -3.093  0.00209 **
indus        0.50978    0.05102   9.991  < 2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 7.866 on 504 degrees of freedom
Multiple R-squared: 0.1653, Adjusted R-squared: 0.1637
```

```
F-statistic: 99.82 on 1 and 504 DF, p-value: < 2.2e-16

> lm.chas=lm(crim~chas)
> summary(lm.chas)

Call:
lm(formula = crim ~ chas)

Residuals:
   Min     1Q Median     3Q    Max
-3.738 -3.661 -3.435  0.018 85.232

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.7444     0.3961   9.453   <2e-16 ***
chasN       -1.8928     1.5061  -1.257    0.209
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 8.597 on 504 degrees of freedom
Multiple R-squared: 0.003124, Adjusted R-squared: 0.001146
F-statistic: 1.579 on 1 and 504 DF, p-value: 0.2094

> lm.nox=lm(crim~nox)
> summary(lm.nox)

Call:
lm(formula = crim ~ nox)

Residuals:
    Min      1Q  Median      3Q     Max
-12.371  -2.738  -0.974   0.559  81.728

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.720     1.699  -8.073 5.08e-15 ***
nox          31.249     2.999  10.419  < 2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 7.81 on 504 degrees of freedom
Multiple R-squared: 0.1772, Adjusted R-squared: 0.1756
F-statistic: 108.6 on 1 and 504 DF, p-value: < 2.2e-16

> lm.rm=lm(crim~rm)
> summary(lm.rm)

Call:
lm(formula = crim ~ rm)

Residuals:
```

```
   Min     1Q Median     3Q    Max
-6.604 -3.952 -2.654 0.989 87.197

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.482     3.365   6.088 2.27e-09 ***
rm           -2.684     0.532  -5.045 6.35e-07 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 8.401 on 504 degrees of freedom
Multiple R-squared: 0.04807, Adjusted R-squared: 0.04618
F-statistic: 25.45 on 1 and 504 DF, p-value: 6.347e-07

> lm.age=lm(crim~age)
> summary(lm.age)

Call:
lm(formula = crim ~ age)

Residuals:
   Min     1Q Median     3Q    Max
-6.789 -4.257 -1.230 1.527 82.849

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.77791  0.94398 -4.002 7.22e-05 ***
age          0.10779  0.01274  8.463 2.85e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 8.057 on 504 degrees of freedom
Multiple R-squared: 0.1244, Adjusted R-squared: 0.1227
F-statistic: 71.62 on 1 and 504 DF, p-value: 2.855e-16

> lm.dis=lm(crim~dis)
> summary(lm.dis)

Call:
lm(formula = crim ~ dis)

Residuals:
   Min     1Q Median     3Q    Max
-6.708 -4.134 -1.527 1.516 81.674

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.4993     0.7304  13.006  <2e-16 ***
dis         -1.5509     0.1683  -9.213  <2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

```
Residual standard error: 7.965 on 504 degrees of freedom
Multiple R-squared: 0.1441, Adjusted R-squared: 0.1425
F-statistic: 84.89 on 1 and 504 DF, p-value: < 2.2e-16

> lm.rad=lm(crim~rad)
> summary(lm.rad)

Call:
lm(formula = crim ~ rad)

Residuals:
    Min      1Q  Median      3Q     Max
-10.164  -1.381  -0.141   0.660  76.433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.28716  0.44348  -5.157 3.61e-07 ***
rad          0.61791  0.03433  17.998  < 2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 6.718 on 504 degrees of freedom
Multiple R-squared: 0.3913, Adjusted R-squared: 0.39
F-statistic: 323.9 on 1 and 504 DF, p-value: < 2.2e-16

> lm.tax=lm(crim~tax)
> summary(lm.tax)

Call:
lm(formula = crim ~ tax)

Residuals:
    Min      1Q  Median      3Q     Max
-12.513  -2.738  -0.194   1.065  77.696

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.528369  0.815809  -10.45  <2e-16 ***
tax          0.029742  0.001847   16.10  <2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 6.997 on 504 degrees of freedom
Multiple R-squared: 0.3396, Adjusted R-squared: 0.3383
F-statistic: 259.2 on 1 and 504 DF, p-value: < 2.2e-16

> lm.ptratio=lm(crim~ptratio)
> summary(lm.ptratio)

Call:
```

```
lm(formula = crim ~ ptratio)

Residuals:
   Min     1Q Median     3Q    Max
-7.654 -3.985 -1.912  1.825 83.353

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.6469     3.1473  -5.607 3.40e-08 ***
ptratio       1.1520     0.1694   6.801 2.94e-11 ***
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .  0.1    1

Residual standard error: 8.24 on 504 degrees of freedom
Multiple R-squared:  0.08407,   Adjusted R-squared:  0.08225
F-statistic: 46.26 on 1 and 504 DF,  p-value: 2.943e-11

> lm.black=lm(crim~black)
> summary(lm.black)

Call:
lm(formula = crim ~ black)

Residuals:
    Min      1Q  Median      3Q     Max
-13.756  -2.299  -2.095  -1.296  86.822

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.553529   1.425903  11.609  <2e-16 ***
black       -0.036280   0.003873  -9.367  <2e-16 ***
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .  0.1    1

Residual standard error: 7.946 on 504 degrees of freedom
Multiple R-squared:  0.1483,    Adjusted R-squared:  0.1466
F-statistic: 87.74 on 1 and 504 DF,  p-value: < 2.2e-16

> lm.lstat=lm(crim~lstat)
> summary(lm.lstat)

Call:
lm(formula = crim ~ lstat)

Residuals:
    Min      1Q  Median      3Q     Max
-13.925  -2.822  -0.664   1.079  82.862

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.33054    0.69376  -4.801 2.09e-06 ***
```

```
lstat       0.54880   0.04776 11.491 < 2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 7.664 on 504 degrees of freedom
Multiple R-squared: 0.2076, Adjusted R-squared: 0.206
F-statistic: 132 on 1 and 504 DF, p-value: < 2.2e-16

> lm.medv=lm(crim~medv)
> summary(lm.medv)

Call:
lm(formula = crim ~ medv)

Residuals:
   Min     1Q Median     3Q    Max
-9.071 -4.022 -2.343 1.298 80.957

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.79654  0.93419  12.63  <2e-16 ***
medv        -0.36316  0.03839  -9.46  <2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 7.934 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
```

All of the predictors except chas have statistically significant associations at any practical level of significance. The predictors rad and tax have the highest R^2 values at 0.3913 and 0.3396 respectively. The predictor lstat has an R^2 of $0,2076$. The predictors medv, black, dis, age, noxand indus all have R^2 values between 0.1 and 0.2. The remaining predictors all have R^2 values less than 0.1.

(b) Fit a multiple regression model to predict the response using all of the predictors. Describe you results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

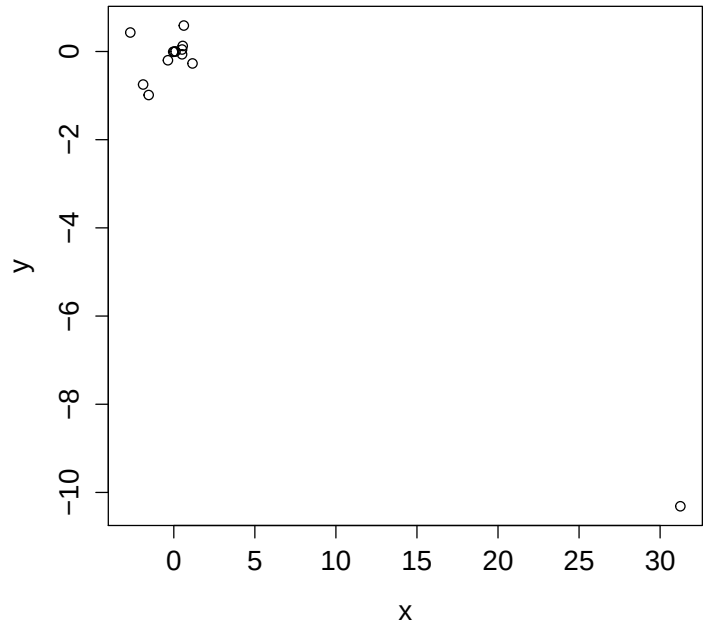
```
> lm.multi=lm(crim~., data=Boston)
> summary(lm.multi)

Call:
lm(formula = crim ~ ., data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-9.924 -2.120 -0.353 1.019 75.051

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 17.033228  7.234903  2.354 0.018949 *
zn           0.044855  0.018734  2.394 0.017025 *
indus       -0.063855  0.083407 -0.766 0.444294
chasN       -0.749134  1.180147 -0.635 0.525867
nox        -10.313535  5.275536 -1.955 0.051152 .
rm           0.430131  0.612830  0.702 0.483089
age          0.001452  0.017925  0.081 0.935488
dis         -0.987176  0.281817 -3.503 0.000502 ***
rad          0.588209  0.088049  6.680 6.46e-11 ***
tax         -0.003780  0.005156 -0.733 0.463793
ptratio     -0.271081  0.186450 -1.454 0.146611
black       -0.007538  0.003673 -2.052 0.040702 *
lstat        0.126211  0.075725  1.667 0.096208 .
medv        -0.198887  0.060516 -3.287 0.001087 **
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1    1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared: 0.454, Adjusted R-squared: 0.4396
F-statistic: 31.47 on 13 and 492 DF, p-value: < 2.2e-16
```

The multiple regression has an R^2 of 0.454 which is much better that each of the simple regressions. We can reject $H_0 : \beta_j = 0$ at any significance level for dis and rad. We can reject $H_0 : \beta_j = 0$ at the 0.01 significance level, but not the 0.001 level, for medv. We can reject $H_0 : \beta_j = 0$ at the 0.05 significance level, but not the 0.01 level, for zn and black. We can reject $H_0 : \beta_j = 0$ at the 0.1 significance level, but not the 0.05 level, for nox and lstat. The remainder cannot be rejected at any reasonable significance level.

(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed asa single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

```
> x = c(coefficients(lm.zn)[2], coefficients(lm.indus)[2],
    coefficients(lm.chas)[2], coefficients(lm.nox)[2], coefficients(lm.rm)[2],
    coefficients(lm.age)[2], coefficients(lm.dis)[2], coefficients(lm.rad)[2],
    coefficients(lm.tax)[2], coefficients(lm.ptratio)[2],
    coefficients(lm.black)[2], coefficients(lm.lstat)[2],
    coefficients(lm.medv)[2])
> y = coefficients(lm.all)[2:14]
> plot(x,y)
```



The coefficient of nox is radically different in the simple and multiple regressions. The remaining coefficients are comparable in the two models.

(d) Is there evidence of a nonlinear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon \tag{13}$$

```
> lm.pzn=lm(crim~poly(zn,3))
> summary(lm.pzn)

Call:
lm(formula = crim ~ poly(zn, 3))

Residuals:
   Min     1Q Median     3Q    Max
-4.821 -4.614 -1.294  0.473 84.130

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.6135     0.3722   9.709  < 2e-16 ***
poly(zn, 3)1 -38.7498   8.3722  -4.628  4.7e-06 ***
poly(zn, 3)2  23.9398   8.3722   2.859  0.00442 **
poly(zn, 3)3 -10.0719   8.3722  -1.203  0.22954
---
Signif. codes: 0 ***  0.001 **  0.01 *  0.05 .  0.1    1
```

```
Residual standard error: 8.372 on 502 degrees of freedom
Multiple R-squared: 0.05824, Adjusted R-squared: 0.05261
F-statistic: 10.35 on 3 and 502 DF, p-value: 1.281e-06

> lm.pindus=lm(crim~poly(indus,3))
> summary(lm.pindus)

Call:
lm(formula = crim ~ poly(indus, 3))

Residuals:
   Min     1Q Median     3Q    Max
-8.278 -2.514  0.054  0.764 79.713

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.614      0.330  10.950 < 2e-16 ***
poly(indus, 3)1 78.591      7.423  10.587 < 2e-16 ***
poly(indus, 3)2 -24.395     7.423  -3.286 0.00109 **
poly(indus, 3)3 -54.130     7.423  -7.292 1.2e-12 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 7.423 on 502 degrees of freedom
Multiple R-squared: 0.2597, Adjusted R-squared: 0.2552
F-statistic: 58.69 on 3 and 502 DF, p-value: < 2.2e-16

> lm.pnox=lm(crim~poly(nox,3))
> summary(lm.pnox)

Call:
lm(formula = crim ~ poly(nox, 3))

Residuals:
   Min     1Q Median     3Q    Max
-9.110 -2.068 -0.255  0.739 78.302

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.3216  11.237 < 2e-16 ***
poly(nox, 3)1 81.3720     7.2336  11.249 < 2e-16 ***
poly(nox, 3)2 -28.8286    7.2336  -3.985 7.74e-05 ***
poly(nox, 3)3 -60.3619    7.2336  -8.345 6.96e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 7.234 on 502 degrees of freedom
Multiple R-squared: 0.297, Adjusted R-squared: 0.2928
F-statistic: 70.69 on 3 and 502 DF, p-value: < 2.2e-16
```

```
> lm.prm=lm(crim~poly(rm,3))
> summary(lm.prm)

Call:
lm(formula = crim ~ poly(rm, 3))

Residuals:
    Min      1Q  Median      3Q     Max
-18.485  -3.468  -2.221  -0.015  87.219

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.6135     0.3703   9.758  < 2e-16 ***
poly(rm, 3)1 -42.3794   8.3297  -5.088 5.13e-07 ***
poly(rm, 3)2 26.5768    8.3297   3.191  0.00151 **
poly(rm, 3)3 -5.5103    8.3297  -0.662  0.50858
---
Signif. codes: 0 ***  0.001 **  0.01 *  0.05 .  0.1    1

Residual standard error: 8.33 on 502 degrees of freedom
Multiple R-squared: 0.06779, Adjusted R-squared: 0.06222
F-statistic: 12.17 on 3 and 502 DF, p-value: 1.067e-07

> lm.page=lm(crim~poly(age,3))
> summary(lm.page)

Call:
lm(formula = crim ~ poly(age, 3))

Residuals:
   Min     1Q Median    3Q    Max
-9.762 -2.673 -0.516 0.019 82.842

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.3485  10.368  < 2e-16 ***
poly(age, 3)1 68.1820    7.8397   8.697  < 2e-16 ***
poly(age, 3)2 37.4845    7.8397   4.781 2.29e-06 ***
poly(age, 3)3 21.3532    7.8397   2.724  0.00668 **
---
Signif. codes: 0 ***  0.001 **  0.01 *  0.05 .  0.1    1

Residual standard error: 7.84 on 502 degrees of freedom
Multiple R-squared: 0.1742, Adjusted R-squared: 0.1693
F-statistic: 35.31 on 3 and 502 DF, p-value: < 2.2e-16

> lm.pdis=lm(crim~poly(dis,3))
> summary(lm.pdis)

Call:
lm(formula = crim ~ poly(dis, 3))
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-10.757  -2.588   0.031   1.267  76.378

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.3259  11.087  < 2e-16 ***
poly(dis, 3)1 -73.3886     7.3315 -10.010  < 2e-16 ***
poly(dis, 3)2  56.3730     7.3315   7.689 7.87e-14 ***
poly(dis, 3)3 -42.6219     7.3315  -5.814 1.09e-08 ***
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .  0.1    1

Residual standard error: 7.331 on 502 degrees of freedom
Multiple R-squared:  0.2778,    Adjusted R-squared:  0.2735
F-statistic: 64.37 on 3 and 502 DF,  p-value: < 2.2e-16

> lm.prad=lm(crim~poly(rad,3))
> summary(lm.prad)

Call:
lm(formula = crim ~ poly(rad, 3))

Residuals:
    Min      1Q  Median      3Q     Max
-10.381  -0.412  -0.269   0.179  76.217

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.2971  12.164  < 2e-16 ***
poly(rad, 3)1 120.9074     6.6824  18.093  < 2e-16 ***
poly(rad, 3)2  17.4923     6.6824   2.618  0.00912 **
poly(rad, 3)3   4.6985     6.6824   0.703  0.48231
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .  0.1    1

Residual standard error: 6.682 on 502 degrees of freedom
Multiple R-squared:  0.4,    Adjusted R-squared:  0.3965
F-statistic: 111.6 on 3 and 502 DF,  p-value: < 2.2e-16

> lm.ptax=lm(crim~poly(tax,3))
> summary(lm.ptax)

Call:
lm(formula = crim ~ poly(tax, 3))

Residuals:
    Min      1Q  Median      3Q     Max
-13.273  -1.389   0.046   0.536  76.950
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.3047 11.860 < 2e-16 ***
poly(tax, 3)1 112.6458   6.8537 16.436 < 2e-16 ***
poly(tax, 3)2 32.0873    6.8537  4.682 3.67e-06 ***
poly(tax, 3)3 -7.9968    6.8537 -1.167   0.244
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 6.854 on 502 degrees of freedom
Multiple R-squared: 0.3689, Adjusted R-squared: 0.3651
F-statistic: 97.8 on 3 and 502 DF, p-value: < 2.2e-16

> lm.pptratio=lm(crim~poly(ptratio,3))
> summary(lm.pptratio)

Call:
lm(formula = crim ~ poly(ptratio, 3))

Residuals:
   Min     1Q Median    3Q    Max
-6.833 -4.146 -1.655 1.408 82.697

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.614      0.361  10.008 < 2e-16 ***
poly(ptratio, 3)1 56.045      8.122   6.901 1.57e-11 ***
poly(ptratio, 3)2 24.775      8.122   3.050 0.00241 **
poly(ptratio, 3)3 -22.280     8.122  -2.743 0.00630 **
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 8.122 on 502 degrees of freedom
Multiple R-squared: 0.1138, Adjusted R-squared: 0.1085
F-statistic: 21.48 on 3 and 502 DF, p-value: 4.171e-13

> lm.pblack=lm(crim~poly(black,3))
> summary(lm.pblack)

Call:
lm(formula = crim ~ poly(black, 3))

Residuals:
    Min      1Q  Median      3Q     Max
-13.096  -2.343  -2.128  -1.439  86.790

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.6135     0.3536  10.218  <2e-16 ***
poly(black, 3)1 -74.4312   7.9546  -9.357  <2e-16 ***
poly(black, 3)2 5.9264     7.9546   0.745   0.457
```

```
poly(black, 3)3 -4.8346   7.9546 -0.608   0.544
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 7.955 on 502 degrees of freedom
Multiple R-squared: 0.1498, Adjusted R-squared: 0.1448
F-statistic: 29.49 on 3 and 502 DF, p-value: < 2.2e-16

> lm.plstat=lm(crim~poly(lstat,3))
> summary(lm.plstat)

Call:
lm(formula = crim ~ poly(lstat, 3))

Residuals:
    Min      1Q  Median      3Q     Max
-15.234 -2.151 -0.486  0.066 83.353

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.6135     0.3392  10.654  <2e-16 ***
poly(lstat, 3)1 88.0697    7.6294  11.543  <2e-16 ***
poly(lstat, 3)2 15.8882    7.6294   2.082  0.0378 *
poly(lstat, 3)3 -11.5740   7.6294  -1.517  0.1299
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 7.629 on 502 degrees of freedom
Multiple R-squared: 0.2179, Adjusted R-squared: 0.2133
F-statistic: 46.63 on 3 and 502 DF, p-value: < 2.2e-16

> lm.pmedv=lm(crim~poly(medv,3))
> summary(lm.pmedv)

Call:
lm(formula = crim ~ poly(medv, 3))

Residuals:
    Min      1Q  Median      3Q     Max
-24.427 -1.976 -0.437  0.439 73.655

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.614      0.292  12.374 < 2e-16 ***
poly(medv, 3)1 -75.058     6.569 -11.426 < 2e-16 ***
poly(medv, 3)2  88.086     6.569  13.409 < 2e-16 ***
poly(medv, 3)3 -48.033     6.569  -7.312 1.05e-12 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 6.569 on 502 degrees of freedom
```

```
Multiple R-squared: 0.4202, Adjusted R-squared: 0.4167
F-statistic: 121.3 on 3 and 502 DF, p-value: < 2.2e-16
```

All the predictors, with the exception of black, show some evidence of a nonlinear relationship.