Exercise 4

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

Part (a)

Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

We would expect the training RSS for the cubic regression to be lower than the training RSS for the linear regression. This is because the cubic regression is able to more accurately fit the training data points better (even though the overall realtionship is linear). Because it has more coefficients to vary, the cubic regression will overfit to the data causing its RSS to be lower.

Part (b)

Answer (a) using test rather than training RSS.

Because the true relationship is linear, we would expect the linear model to preform better than the cubic model on data points that was not included in the training dataset. This is because the linear model accurately reflects the true behavior of the data, so while the training RSS might be higher due to error terms, the test RSS will be lower because the cubic model will not predict new daa points as well since it overfitted to the training data. Thus the test RSS for the linear regression will be lower.

Part (c)

Suppose that 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 regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

We would expect the cubic regression to have a lower training RSS, because it can fit to the training data better due to the extra coefficients in the model. The linear model is restricted to only one coefficient and one intercept, meaning it can't fit to an unknown relationship as well as a cubic (in general).

Part (d)

Answer (c) using test rather than training RSS.

Because we don't know the true relationship of X and Y, we can't say which one will perform better on data it has never seen before. It all depends on the actual relationship and how it comapres to the linear and cubic graphs!

Exercise 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 $\bar{x} = \bar{y} = 0$.

Because $\bar{x} = \bar{y} = 0$, we can say that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 0$, so our regression line is of the form $\hat{y} = \hat{\beta}_1 x$. Additionally, our expressions for $\hat{\beta}_1$ and TSS are simplifed to the following:

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$TSS = \sum y_i^2.$$

Note that when $\sum$ is used without bounds, it is implied that it means $\sum_{i=1}^{n}$.

Now we can use our simplified $\hat{y}$ formula to find RSS:

$$
\begin{aligned}
RSS &= \sum (y_i - \hat{y}_i)^2 \\
&= \sum \left(y_i - \frac{\sum x_j y_j}{\sum x_j^2} x_i \right)^2 \\
&= \sum \left(y_i^2 - 2 \frac{\sum x_j y_j}{\sum x_j^2} x_i y_i + \left(\frac{\sum x_j y_j}{\sum x_j^2} x_i \right)^2 \right) \\
&= \left(\sum y_i^2 \right) - 2 \frac{\sum x_j y_j}{\sum x_j^2} \left(\sum x_i y_i \right) + \left(\frac{\sum x_j y_j}{\sum x_j^2} \right)^2 \sum x_i^2 \\
&= TSS - 2 \frac{\left(\sum x_j y_j \right)^2}{\sum x_j^2} + \frac{\left(\sum x_j y_j \right)^2}{\sum x_j^2} \\
&= TSS - \frac{\left(\sum x_j y_j \right)^2}{\sum x_i^2}.
\end{aligned}
$$

Plugging this into the formula for R^2 gives:

$$
\begin{aligned}
R^2 &= \frac{TSS - RSS}{TSS} \\
&= \frac{TSS - \left(TSS - \frac{\left(\sum x_j y_j \right)^2}{\sum x_i^2} \right)}{\sum y_i^2} \\
&= \frac{\left(\sum x_j y_j \right)^2}{\sum x_i^2 \sum y_i^2}.
\end{aligned}
$$

Now let's look at the correlation coefficient r:

$$
r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}.
$$

Then we square the expression to get:

$$
r^2 = \frac{\left(\sum x_i y_i \right)^2}{\sum x_i^2 \sum y_i^2}
$$

This exactly the expression we found for R^2!! Therefore we have probed that $\boxed{R^2 = r^2}$.

Exercise 10

In [2]:
```
library(ISLR)
head(Carseats)

attach(Carseats)
```

A data.frame: 6 × 11

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Educatio
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<fct>	<dbl>	<dbl>
1	9.50	138	73	11	276	120	Bad	42	1
2	11.22	111	48	16	260	83	Good	65	1
3	10.06	113	35	10	269	80	Medium	59	1
4	7.40	117	100	4	466	97	Medium	55	1
5	4.15	141	64	3	340	128	Bad	38	1
6	10.81	124	113	13	501	72	Bad	78	1

Part(a)

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

In [4]:
```
model = lm(Sales ~ Price + Urban + US)
summary(model)
```

```
Call:
lm(formula = Sales ~ Price + Urban + US)

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
```

Part (b)

Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

In [12]:
```
print(contrasts(Urban))
print(contrasts(US))
```

```
    Yes
No    0
Yes   1
    Yes
No    0
Yes   1
```

The coefficient for Price is negative, meaning as price increases sales decrease (which makes sense). The coefficient for UrbanYes is -0.02, so if the store is urban then its sales decrease. The coefficient for USYes however is positive, meaning that if a store is in the US then it has more sales.

Part (c)

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

$$\text{Sales} \approx -0.05 \cdot \text{Price} - 0.02 \cdot \text{UrbanYes} + 1.20 \cdot \text{USYes}$$

where UrbanYes is 1 if it is urban and 0 otherwise, and USYes is 1 if in the US and 0 otherwise.

Part (d)

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

The p-value for everything except UrbanYes is extremely low, so we can reject the null hypothesis for Price and UrbanYes

Part (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 association with the outcome.

In [15]:
```
model1 = lm(Sales~Price+US)
summary(model1)
```

```
Call:
lm(formula = Sales ~ Price + US)

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
```

Part (f)

How well do the models in (a) and (e) fit the data?

The R^2 term for the models in both (a) and are 0.23, which means that only 23% of the variation in Sales is explained by the predictors. The exclusion of the Urban variable doesn't change the fit significantly.

Part (g)

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

In [16]:
```
print(confint(model1, level=0.95))
```

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

Part (h)

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

In [35]:
```
st_res = rstandard(model1)
st_res[which.min(st_res)]
st_res[which.max(st_res)]

lev = hatvalues(model1)
lev[which.max(lev)]
print(2*(2+1)/nrow(Carseats))
```

51: -2.81102167244627

377: 2.86508213825242

43: 0.0433376570371785

```
[1] 0.015
```

We see that the lowest and highest studentized residual is -2.8 and 2.9, both of which are within -3 and 3, so no significant outliers there. We see that twice the average leverage is 0.015, but the max leverage is 0.04, meaning there are some points that have a significantly high leverage!

In [34]:
```
lev[lev > 2*3/nrow(Carseats)]
```

43: 0.0433376570371785 **126:** 0.0259661351134241 **156:** 0.016106161622167 **157:** 0.0153555754048395 **160:** 0.0157073653087183 **166:** 0.0285666078977165 **172:** 0.0210140057511345 **175:** 0.0296867177031533 **192:** 0.0180391028373988 **204:** 0.0153555754048395 **209:** 0.0182347179910024 **270:** 0.0191949376316542 **273:** 0.0186873442250878 **314:** 0.0231647047366309 **316:** 0.0170488131217078 **357:** 0.0182789444516194 **366:** 0.0173988378853422 **368:** 0.0237070475051107 **384:** 0.0165139295890552 **387:** 0.0165546179328228

The values in bold above are the row index of the store, and the corresponding decimal is the leverage associated with it. These are all the stores with a leverage greater than twice the average (high leverage).