SML HW2

Christopher Marais

ISLR Chapter 3:

Import libraries

```
library("ISLR2")
```

Question 4

4.a With the cubic regression the RSS should be lower. This is because the bonus features in the cubic regression allows it to describe the training data with more precision due to its higher complexity.

4.b Because the true relationship is linear we expect the linear regression to perform better than the cubic regression. We expect the linear regression to have a lower test RSS than the cubic regression. The linear model should be able to generalize better than the cubic model. The cubic regression might more easily over fit on the training data.

4.c We still expect the cubic regression to have a lower RSS on the training data. The linear regression model will have a higher RSS than in (a.) because the true relationship is non-linear.

4.d On the spectrum on linearity if the true relationship is closer to what is described by the cubic regression, the cubic model will have a lower RSS. However, if the model is closer to the linear regression model the linear regression model might have a lower RSS. in thee case where the true relationship is closer to the cubic model in non-linearity it will result in the linear regression model will under fit, whereas if the true relationship is closer to the linear model then it will result in the cubic model to over fit to the training data.

Question 10

10.a

```
model1 = lm(Sales~Price+Urban+US,data=Carseats)
summary(model1)
```

```
##
## 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
```

10.b The (intercept) is the number of car seats that would be sold according to the model if all other features were ignored. The Price coefficient indicates that car seats sales will decrease by ~50 for every dollar the price goes increases. The UrbanYes coefficient is not at all significant with a very high p value of 0.936. It is therefore, uninformative. The USYes coefficient indicates that on average ~20% more car seats are sold if the shop is located in the US than if it were outside of the US.

10.c

```
attach(Carseats)
contrasts(US)
```

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

```
contrasts(Urban)
```

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

$$Sales = 13.04 - 0.05Price - 0.02UrbanYes + 1.20USYes$$

10.d

```
model2 = lm(Sales~.,data=Carseats)
summary(model2)
```

```
##
## Call:
## lm(formula = Sales ~ ., data = Carseats)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -2.8692 -0.6908  0.0211  0.6636  3.4115 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***
## CompPrice        0.0928153  0.0041477  22.378  < 2e-16 ***
## Income           0.0158028  0.0018451   8.565 2.58e-16 ***
## Advertising      0.1230951  0.0111237  11.066  < 2e-16 ***
## Population       0.0002079  0.0003705   0.561    0.575    
## Price           -0.0953579  0.0026711 -35.700  < 2e-16 ***
## ShelveLocGood    4.8501827  0.1531100  31.678  < 2e-16 ***
## ShelveLocMedium  1.9567148  0.1261056  15.516  < 2e-16 ***
## Age             -0.0460452  0.0031817 -14.472  < 2e-16 ***
## Education       -0.0211018  0.0197205  -1.070    0.285    
## UrbanYes         0.1228864  0.1129761   1.088    0.277    
## USYes           -0.1840928  0.1498423  -1.229    0.220    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 1.019 on 388 degrees of freedom
## Multiple R-squared:  0.8734, Adjusted R-squared:  0.8698 
## F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```

The null hypothesis can be rejected for all the predictors excluding Population, Education, UrbanYes, USYes.

10.e

```
model3 = lm(Sales~.-Education-Urban-US-Population,data=Carseats)
summary(model3)
```

```
## 
## Call:
## lm(formula = Sales ~ . - Education - Urban - US - Population, 
##     data = Carseats)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.7728 -0.6954  0.0282  0.6732  3.3292 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      5.475226   0.505005   10.84   <2e-16 ***
## CompPrice        0.092571   0.004123   22.45   <2e-16 ***
## Income           0.015785   0.001838    8.59   <2e-16 ***
## Advertising      0.115903   0.007724   15.01   <2e-16 ***
## Price           -0.095319   0.002670  -35.70   <2e-16 ***
## ShelveLocGood    4.835675   0.152499   31.71   <2e-16 ***
## ShelveLocMedium  1.951993   0.125375   15.57   <2e-16 ***
## Age              -0.046128   0.003177  -14.52   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.019 on 392 degrees of freedom
## Multiple R-squared:  0.872,  Adjusted R-squared:  0.8697
## F-statistic: 381.4 on 7 and 392 DF,  p-value: < 2.2e-16
```

10.f - The RSE goes down from 2.47 **model (a)** to 1.02 **model (e)**. The R2 statistic goes up from 0.24 **(a)** to 0.872 **(e)** and the F-statistic goes up from 41.52 to 381.4. - The statistical evidence clearly shows that **(e)** is a much better fit.

10.g

```
confint(model3)
```

```
##                       2.5 %      97.5 %
## (Intercept)      4.48236820  6.46808427
## CompPrice        0.08446498  0.10067795
## Income           0.01217210  0.01939784
## Advertising      0.10071856  0.13108825
## Price           -0.10056844 -0.09006946
## ShelveLocGood    4.53585700  5.13549250
## ShelveLocMedium  1.70550103  2.19848429
## Age             -0.05237301 -0.03988204
```

Typed Problem 1

Let $Y_1, ..., Y_n$ be iid rvs with $E(Y_i) = a$ and $E(Y_i^2) = b$, so that $Var(Y_i) = b - a^2$.

Define $T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, where $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$ is the sample mean.

1.1. (Optional) Use the properties/calculus of expectations to find $E(T)$. If you are not able to find $E(T)$, you can use use $E(T) = (n-1)Var(Y_i)$ in subsequent subproblems.

1.2. Suppose we estimate the population variance $Var(Y_i)$ by cT for some constant $c > 0$. What value of c results in an unbiased estimator of the population variance? (The answer you should get is $c = 1/(n-1)$.) Let $T_1 = cT$ be this unbiased estimator.

1.3. Let $Y_1, ..., Y_n$ be iid $Normal(\mu, \sigma^2)$, where μ and σ^2 are the population mean and variance, respectively. One can show that $T_2 = T/n$ is the MLE for σ^2; you can take this fact for granted.

Use R to examine the small-sample properties of T_1 and T_2 as follows:

(a) Generate the data as follows:

m=1000; n=4; # n is the sample size; m is the # of replications set.seed(0); M = matrix(rnorm(m*n),nrow=m); # default parameters in rnorm are mean=0, sd=1; M is an m-by-n matrix with replications of the experiment stored in rows

(b) For each row of M, evaluate and store values of T_1 and T_2, in separate vectors. (Optional): you can do this without loops using apply() function

(c) Plot histograms of T_1 and T_2.

(d) “Monte Carlo integration” is estimation of population moments of a rv X by the corresponding sample moments whenever one can simulate iid variates $X_1, X_2, \ldots$ from the sampling distribution of X. I.e., using the law of large numbers (and another result known as the continuous mapping theorem) $\bar{X}_n \to E(X)$ and $S_n^2 \to Var(X)$ as $n \to \infty$, where $\bar{X}_n$ and S_n^2 are the sample mean and the sample variance, respectively. Use “Monte Carlo integration” to estimate bias, variance and MSE of the two estimators. Specifically, you can estimate $E(T_1)$ and $E(T_2)$ using the respective sample means, and (population) variances of T_1 and T_2 using the sample variances of T_1 and T_2.

Briefly discuss your findings in (c) and (d).

1.4. Suppose we are now interested in the population standard deviation, i.e., $\sigma = \sqrt{\sigma^2}$. Explain/argue whether $\sqrt{T_1}$ is unbiased for estimation of σ, and why. Feel free to extend the simulation study in 1.3 to reinforce your answer.