STATS 415 HW 3

Jiahao Cheng

Feburary 2025

Problem 1

(a) Conclusion:
If the true relationship is linear, for the training RSS, the cubic regression's RSS would be lower than the linear regression's RSS.

Justification:
Since the cubic regression contains more predictors than the linear regression, the cubic model would always fit better to the train data, and explain more by predictors. Thus the RSS would be lower.

(b) Conclusion:
If the true relationship is linear, for the testing RSS, the linear regression's RSS would be lower than the cubic regression's RSS.

Justification:
Since the true relationship is linear, the cubic model would overfit on the train data, but perform badly on the test data. Meanwhile the linear model represents the true relationship and generalizes well on the test data. Thus the linear regression's RSS would be lower than the cubic regression's RSS.

(c) Conclusion:
If the true relationship is non-linear, for the training RSS, the cubic regression's RSS would be lower than the linear regression's RSS.

Justification:
Since the cubic regression contains more predictors than the linear regression, the cubic model would always fit better to the train data, and explain more by predictors to reduce the training RSS.

(d) Conclusion:
If the true relationship is non-linear, for the testing RSS, there is not enough information to tell which RSS is lower.

Justification:
Since the true relationship is non-linear, we don't know which models fit better on the test data, which might be the cubic one or the linear one. Thus the cubic RSS could be higher, even, or lower than the linear RSS.

Problem 2

(a) Plot 1 shows the regression model with all predictors and the values of coefficients. Based on the Multiple R-squared value 0.8734, we can tell that this model fits the data quite well since it explained 87.34% of the variance in Sales.

(b) From Plot 1, we can tell that Intercept, ComPrice, Income, Advertising, ShelveLoc(Good&Medium&Bad), and Age have significant p-values.

For variable Urban,
the null Hypothesis(H_0) is $\beta_{UrbanYes} = 0$, that Urban doesn't contribute well to this model.
the alternative Hypothesis(H_a) is $\beta_{UrbanYes} \neq 0$, that Urban contributes well to this model.
Since the p-value for UrbanYes is $0.277 (\geq 0.05)$, we fail to reject the H_0 .

```
> help(Carseats)
> model <- lm(Sales ~ ., data=Carseats)
> summary(model)

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
```

Figure 1: Full Model

(c) Plot 2 shows the reduced model.
Based on the Multiple R-squared, the model could explain 87.2% variance in Sales, which is quite close to the full model's 87.34%. So the reduced variables don't contribute much to this model. And the reduced model is better since it's simpler.

```
> reduce <- lm(Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + Age, data=Carseats)
> summary(reduce)

Call:
lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
    ShelveLoc + Age, 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
```

Figure 2: Reduced Model

(d) Plot 3 shows the result of anova().
The F test's p-value is $0.358 \geq 0.05$, so we fail to reject the H_0, and conclude that the reduced variables don't improve the model.
Based on the R^2 comparison in (c), we can conclude now that the reduced model is better, since it's simpler and explains as much as the full model.

```
> anova_result <- anova(reduce, model)
> anova_result
Analysis of Variance Table

Model 1: Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
    Age
Model 2: Sales ~ CompPrice + Income + Advertising + Population + Price +
    ShelveLoc + Age + Education + Urban + US
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    392 407.39
2    388 402.83  4    4.5533 1.0964  0.358
```

Figure 3: Anova Result

(e) Formula:

ShelveLoc=Good, $\hat{Sales} = 10.311 + 0.092 * CompPrice + 0.015 * Income + 0.116 * Advertising - 0.095 * Price - 0.046 * Age$

ShelveLoc=Medium, $\hat{Sales} = 7.427 + 0.092 * CompPrice + 0.015 * Income + 0.116 * Advertising - 0.095 * Price - 0.046 * Age$

ShelveLoc=Bad, $\hat{Sales} = 5.475 + 0.092 * CompPrice + 0.015 * Income + 0.116 * Advertising - 0.095 * Price - 0.046 * Age$

Interpretation:
$\beta_{Intercept} = 5.475$: The coefficient is just the proper vertical placement $E(y_i|x_i)$.
$\beta_{CompPrice} = 0.092$: other conditions are the same, two individuals who differ in variable $Price$ by 1 unit are expected to differ in $Sales$ by $\beta_{CompPrice}$ units.
$\beta_{Income} = 0.015$: other conditions are the same, two individuals who differ in variable $Income$ by 1 unit are expected to differ in $Sales$ by β_{Income} units.
$\beta_{Advertising} = 0.116$: other conditions are the same, two individuals who differ in variable $Advertising$ by 1 unit are expected to differ in $Sales$ by $\beta_{Advertising}$ units.
$\beta_{Price} = -0.095$: other conditions are the same, two individuals who differ in variable $Price$ by 1 unit are expected to differ in $Sales$ by β_{Price} units.
$\beta_{ShelveLocGood} = 4.836$: an $Good - ShelveLoc$ individual tends to have higher $Sale$ than a $Bad - ShelveLoc$ individual.
$\beta_{ShelveLocMedium} = 1.952$: an $Medium - ShelveLoc$ individual tends to have higher $Sale$ than a $Bad - ShelveLoc$ individual.
$\beta_{Age} = -0.046$: other conditions are the same, two individuals who differ in

variable Age by 1 unit are expected to differ in $Sales$ by β_{Age} units.

(f) Plot 4 shows the interactive model and its coefficients.
Interpretation:
$\beta_{Price*ShelveLocGood} = 0.006$: other conditions are the same, two $Good-ShelveLoc$ individuals who differ in variable $Price$ by 1 unit are expected to differ in $Sales$ by $\beta_{Price} + \beta_{ShelveLocGood}$
$\beta_{Price*ShelveMedium} = 0.004$: other conditions are the same, two $Medium-ShelveLoc$ individuals who differ in variable $Price$ by 1 unit are expected to differ in $Sales$ by $\beta_{Price} + \beta_{ShelveLocMedium}$
Other conditions are the same, two $Bad - ShelveLoc$ individuals who differ in variable $Price$ by 1 unit are expected to differ in $Sales$ by β_{Price} units.

Since the p-values are all greater than 0.05, we fail to reject H_0, so the interaction terms are not necessary.

```
> interact <- lm(Sales ~ CompPrice + Income + Advertising + Price * ShelveLoc + Age, data=Carseats)
> summary(interact)

Call:
lm(formula = Sales ~ CompPrice + Income + Advertising + Price *
    ShelveLoc + Age, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7984 -0.6896  0.0144  0.6743  3.3391

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            5.866758   0.696460   8.424 7.08e-16 ***
CompPrice              0.092592   0.004159  22.262  < 2e-16 ***
Income                 0.015766   0.001849   8.528 3.32e-16 ***
Advertising            0.116003   0.007746  14.975  < 2e-16 ***
Price                 -0.098594   0.004677 -21.082  < 2e-16 ***
ShelveLocGood          4.185088   0.747377   5.600 4.06e-08 ***
ShelveLocMedium        1.535031   0.628915   2.441   0.0151 *
Age                   -0.046494   0.003209 -14.490  < 2e-16 ***
Price:ShelveLocGood    0.005619   0.006300   0.892   0.3730
Price:ShelveLocMedium  0.003650   0.005386   0.678   0.4984
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.021 on 390 degrees of freedom
Multiple R-squared:  0.8723,    Adjusted R-squared:  0.8693
F-statistic: 295.9 on 9 and 390 DF,  p-value: < 2.2e-16
```

Figure 4: Interacted Model

(g) Plot 5 shows the anova result.
The p-value is $0.6593(\geq 0.05)$, so we fail to reject the H_0, including the interaction terms doesn't improve the model.
Thus the interaction terms aren't needed.

```
> anova(reduce, interact)
Analysis of Variance Table

Model 1: Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
    Age
Model 2: Sales ~ CompPrice + Income + Advertising + Price * ShelveLoc +
    Age
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    392 407.39
2    390 406.52  2   0.86946 0.4171 0.6593
```

Figure 5: Anova Results