Jesse Goodman, Joyce Lu, Aditya Mittal, Jackson Sousa
STA 108, Spring 2022
Professor Emanuela Furfaro
June 7, 2022

STA 108 Final Project Report

I. Introduction

For our final project, we chose to analyze the CarSeats dataset to answer the question: *how do different predictor variables affect car seat sales?* To answer this question, we used a regression model to explain car seat sales, our response variable Y, given variables related to the product and different locations.

II. Summary Statistics

The CarSeat dataset contains 400 observations with 11 total variables. The dataset contains data of car seats sales at different stores located through various geographical regions across the world. More specifically, the data contains three categorical variables–ShelveLoc (3 levels: Bad, Medium Good), Urban (2 levels: Yes, No), and U.S. (2 levels: Yes, No)– and eight quantitative variables –Sales, CompPrice, Income, Advertising, Population, Price, Age, Education. Attached below is a table showing summary statistics (mean, median, variance, max, min) of each quantitative variable. For categorical variables, we count the number of times each factor level was observed.

Table 1: Summary Statistics for Qualitative Variables

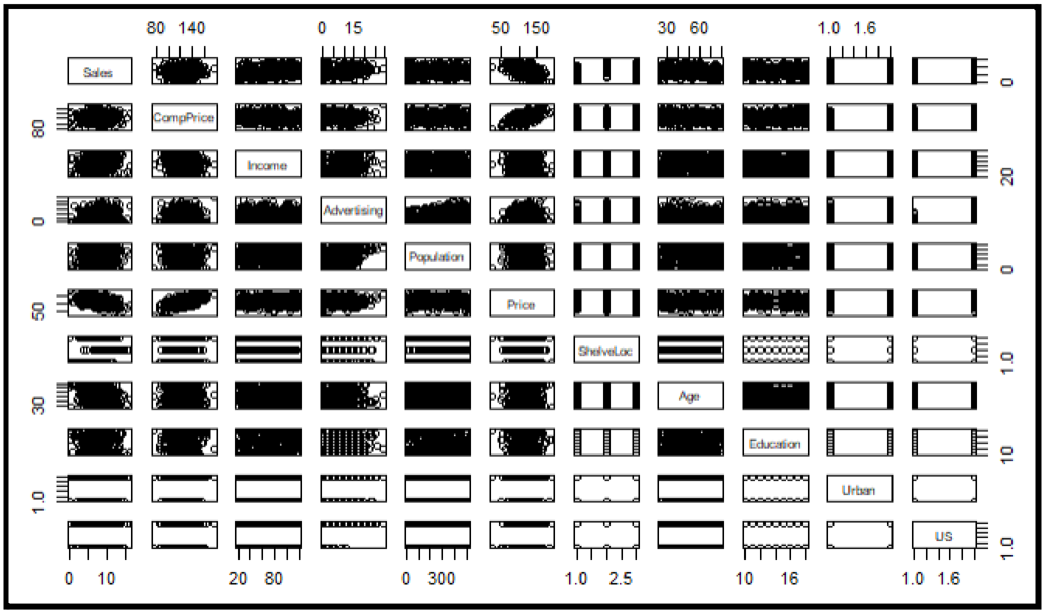
ShelveLoc	Urban	US
Bad: 96	Yes: 118	Yes: 142
Medium: 219	No: 282	No: 258
Good: 85		

Table 2: Summary Statistics for Quantitative variables.

	Sales	CompPrice	Income	Advertising	Population	Price	Age	Education
Mean	7.496	125	68.66	6.635	264.8	115.8	53.52	13.9
Median	7.490	125	69.00	5	272	117	54.50	14.0
Variance	7.9756	235.1472	783.2182	44.2273	21719.81	560.5844	262.4496	6.8671
Min	0.00	77	21.00	0.000	10.0	24	25.00	10.0
Max	16.27	175	120.00	29.000	509.0	191	80.00	18.0

```
                 Sales   CompPrice       Income  Advertising   Population       Price          Age    Education
Sales       1.00000000  0.06407873  0.151950979  0.269506781  0.050470984 -0.44495073 -0.231815440 -0.051955242
CompPrice   0.06407873  1.00000000 -0.080653423 -0.024198788 -0.094706516  0.58484777 -0.100238817  0.025197050
Income      0.15195098 -0.08065342  1.000000000  0.058994706 -0.007876994 -0.05669820 -0.004670094 -0.056855422
Advertising 0.26950678 -0.02419879  0.058994706  1.000000000  0.265652145  0.04453687 -0.004557497 -0.033594307
Population  0.05047098 -0.09470652 -0.007876994  0.265652145  1.000000000 -0.01214362 -0.042663355 -0.106378231
Price      -0.44495073  0.58484777 -0.056698202  0.044536874 -0.012143620  1.00000000 -0.102176839  0.011746599
Age        -0.23181544 -0.10023882 -0.004670094 -0.004557497 -0.042663355 -0.10217684  1.000000000  0.006488032
Education  -0.05195524  0.02519705 -0.056855422 -0.033594307 -0.106378231  0.01174660  0.006488032  1.000000000
```

Above and below is the correlation coefficient matrix (among quantitative variables) and the scatterplot matrix:



III. Interpretation

It was difficult to make solid inferences from the scatterplot matrix due to the many points being clustered together. There is a linear association between car seat sales (our response variable Y) and Price and between Price and CompPrice; however, there appears to be no other direct correlation between sales of car seats and any other quantitative variables. The relationship between Sales and Price is negative, suggesting that as price increases, sales decrease. There is also a positive relationship between CompPrice and Price, so as competitor prices increase, sales increase. Furthermore, the sales of car seats appears to vary among different factor levels of the variable "ShelveLoc," suggesting potential correlation between them. From the correlation coefficient matrix, we can see that there are not many strong correlations between the variables since the largest correlation coefficient is .585 between Price and CompPrice. In relation with our response variable Sales, the most correlated variables are: Income (0.1519), Advertising (0.2695), Price (-0.4449), Age (-0.2318). The low correlation values between our response variable and other potential predictors suggests there is not strong correlation between them. With this information, we can pick our predictor variable for the first order regression model.

IV. First Order Regression Model

For our regression model for car seat sales, we chose the following predictors: Income, Advertising, Population, Price, CompPrice, and ShelveLoc (Bad, Good, and Medium).

Income, Advertising, Population, Price, and CompPrice were our quantitative variables. As seen in the scatterplot matrix above, there is some correlation between combinations of these variables, however the correlation value isn't high. This means our model's explanatory power will likely be increased by the addition of these predictors without encountering much multicollinearity. This conviction is further supported by the correlation matrix which shows that none of the variables have a correlation coefficient higher than .585. Additionally, we chose ShelfLoc as our qualitative variable as there was a visible difference in car seat sales among different factor levels in the scatter plot matrix and to see how important the placement of car seats in stores were on sales. Upon initially looking at the data, we chose these predictors because we felt that they would have a statistically significant and a holistic impact on the sales of car seats. We chose variables that looked at different aspects of a person's life, so we could predict the relationships between the variables. We believe that income will affect car seat sales, because the more money a family has, the more likely they are to spend on car seats. Advertising, if done properly, should impact sales since the car seat exposure will increase and more customers will make purchases, reflected in the 0.269 correlation coefficient between Sales and Advertising. Population would be another important factor as more people would lead to a higher demand for car seats and thus more sales. Price is correlated because sales are generally higher for lower-costing products, an assumption that is supported by the -0.445 correlation coefficient and the strongest correlation between variables. Similar reasoning follows for CompPrice, as people would factor in other competitor prices before making a purchase.

V. Fitting First Order Regression Model

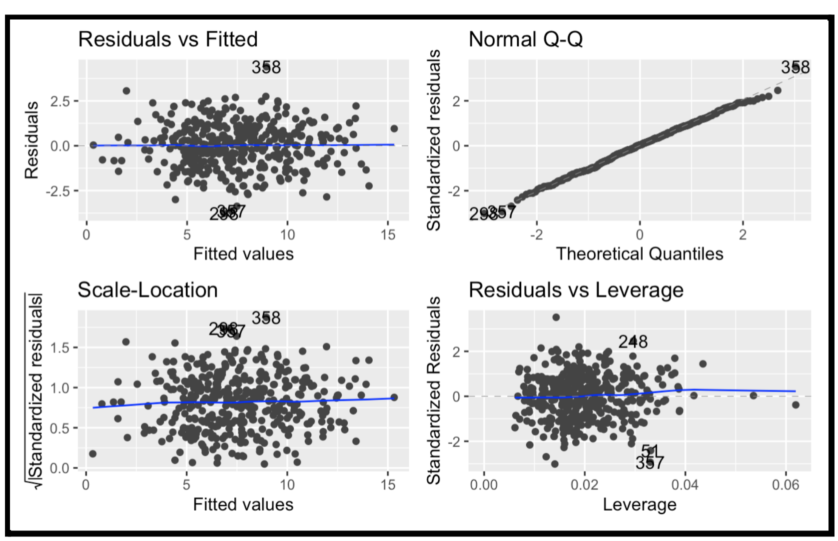
Below is an attached summary of our regression model:

```
Call:
lm(formula = Sales ~ Income + Advertising + Population + Price + 
    CompPrice + as.factor(ShelveLoc), data = carseat)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7734 -0.8482  0.0439  0.8796  4.4036 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               2.2325388  0.5907544   3.779 0.000182 ***
Income                    0.0161410  0.0022757   7.093 6.18e-12 ***
Advertising               0.1129511  0.0099123  11.395  < 2e-16 ***
Population                0.0005555  0.0004476   1.241 0.215308    
Price                    -0.0933876  0.0033021 -28.281  < 2e-16 ***
CompPrice                 0.0962863  0.0051201  18.806  < 2e-16 ***
as.factor(ShelveLoc)Good  4.8058555  0.1888318  25.450  < 2e-16 ***
as.factor(ShelveLoc)Medium  1.8597886  0.1551353  11.988  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.262 on 392 degrees of freedom
Multiple R-squared:  0.8039,	Adjusted R-squared:  0.8004 
F-statistic: 229.6 on 7 and 392 DF,  p-value: < 2.2e-16
```

This regression model assumes that the variables are independent from one another. We go into the relationships between the variables later on in the report. Using α = 0.05, we can assume that certain of these variables are statistically significant because their p-value is less than α. These variables include Income, Advertising, Price, CompPrice, and ShelveLoc. From an initial look, it appears the "Population" is not a statistically significant predictor due to its high p-value. With an adjusted R-Squared value of 0.8004, we can conclude that 80.04% of the variation of car seat sales can be explained by our model; thus, these variables are approx. accurate predictors. Later in the report, we remove variables that are not statistically significant.

VI. Model Diagnostics

We can use the residual plots to test the assumption of linear regression and if any deviations occur. The Residuals vs Fitted plot (containing a horizontal line without distinct patterns), is an indication of an approx. linear relationship. From this plot, it appears observations 298, 357, and 358



could be potential outliers. From the Normal QQ plot, the data is approximately normal as most residuals lie on/very close to the dashed QQ line. The Scale Location plot is used to check the homoscedasticity of variance of the residuals and also suggests that 298, 357, & 358 may be of possible concern. As there is no clear trend between the standardized residuals, we can conclude the condition of homoscedasticity is also met. The Residuals vs Leverage plot suggests that 51, 248, 357 are of concern. We see a few high leverage points as well, but because they lie along the line the assumption is that they are not heavily influential to our model. No points have both high leverage and high residuals, so we conclude there are no influential points. Based on the plots, it appears that point 298, 357, and 358 could be problematic because they are marked by R.

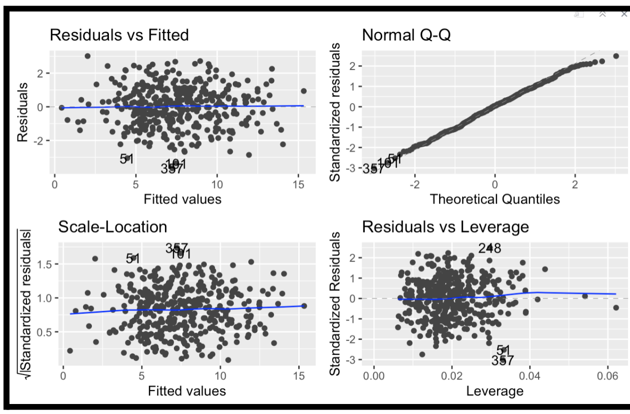
VII. Multicollinearity

```
                         GVIF Df GVIF^(1/(2*Df))
Income               1.016606  1        1.008269
Advertising          1.089114  1        1.043606
Population           1.090421  1        1.044232
Price                1.531984  1        1.237733
CompPrice            1.544998  1        1.242979
as.factor(ShelveLoc) 1.014327  2        1.003563
```

The rule of thumb when evaluating multicollinearity is usually a threshold of 3, 5, or 10. The GVIF is far below even the lowest threshold of 3, so we can conclude that none of our predictors are multicollinear. It is not necessary to remove any variables for remedial measures.

VIII. Remedial Measures

```
Call:
lm(formula = Sales ~ Income + Advertising + Population + Price +
    CompPrice + as.factor(ShelveLoc), data = newdata.new.no.outliers)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6561 -0.8538  0.0517  0.8822  3.0147

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 2.1603004  0.5772906   3.742  0.00021 ***
Income                      0.0161929  0.0022186   7.299 1.64e-12 ***
Advertising                 0.1154502  0.0096794  11.927  < 2e-16 ***
Population                  0.0005214  0.0004363   1.195  0.23287
Price                      -0.0928387  0.0032252 -28.785  < 2e-16 ***
CompPrice                   0.0966043  0.0049924  19.350  < 2e-16 ***
as.factor(ShelveLoc)Good    4.7606003  0.1845477  25.796  < 2e-16 ***
as.factor(ShelveLoc)Medium  1.7973881  0.1519042  11.832  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.23 on 390 degrees of freedom
Multiple R-squared:  0.8114,	Adjusted R-squared:  0.808
F-statistic: 239.7 on 7 and 390 DF,  p-value: < 2.2e-16
```



As there was no multicollinearity, we do not need to remove any predictors or conduct transformation of variables. Instead, we focus on removing outliers to potentially reduce the overall squared sum of errors. Using a cut-off of 3, we removed any points that had the absolute value of standardized residuals greater than the cutoff. With this, we can see our adjusted R^2 has slightly improved to 0.808, indicating more of the variation in car seat sales is now explained by our model. All residuals now lie between plus/minus 3 standard deviations, and the Residual vs Fitted plot suggests data is now completely linear. Based on the new residual plots, it appears that point 51 and 357 could still be of concern. However, we did not find these points to be outliers so we keep them in our dataset and move forward. Both assumptions of normality and homoscedasticity still hold.

IX. Brute Force Analysis

R2adj <int>	CP <int>	BIC <int>
6	6	6

From the brute force analysis, we should include the following predictors in our final model: Income, Advertising, Price, CompPrice, ShelveLoc (include both factor levels). All three criterions, R^2 adjusted, Mallow's CP, and BIC indicate we should select the same number of predictors so we can ultimately choose any criterion, although BIC is preferred. Essentially, we are removing the predictor "Population" as it was not recommended by the analysis. The results of this algorithm agree with our earlier statement that stated Population was a statistically insignificant predictor (p-value 0.2 > alpha). Thus, we can go ahead and remove this predictor.

```
Subset selection object
Call: regsubsets.formula(Sales ~ Income + Advertising + Population +
    Price + CompPrice + as.factor(ShelveLoc), data = newdata.new.no.outliers,
    nvmax = p2 - 1)
7 Variables  (and intercept)
                           Forced in Forced out
Income                         FALSE      FALSE
Advertising                    FALSE      FALSE
Population                     FALSE      FALSE
Price                          FALSE      FALSE
CompPrice                      FALSE      FALSE
as.factor(ShelveLoc)Good       FALSE      FALSE
as.factor(ShelveLoc)Medium     FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: exhaustive
         Income Advertising Population Price CompPrice
1  ( 1 ) " "    " "         " "        " "   " "
2  ( 1 ) " "    " "         " "        "*"   " "
3  ( 1 ) " "    " "         " "        "*"   "*"
4  ( 1 ) " "    "*"         " "        "*"   "*"
5  ( 1 ) " "    "*"         " "        "*"   "*"
6  ( 1 ) "*"    "*"         " "        "*"   "*"
         as.factor(ShelveLoc)Good as.factor(ShelveLoc)Medium
1  ( 1 ) "*"                      " "
2  ( 1 ) "*"                      " "
3  ( 1 ) "*"                      " "
4  ( 1 ) "*"                      " "
5  ( 1 ) "*"                      "*"
6  ( 1 ) "*"                      "*"
```

X. Stepwise Selection

We now consider all the variables in our dataset and use stepwise selection to narrow predictors. More specifically, we are using 3 methods: backward, forward, and stepwise selection. Using backward stepwise selection, we drop the following variables:

Step <S3: AsIs>	Df <dbl>	Deviance <dbl>	Resid. Df <dbl>	Resid. Dev <dbl>	AIC <dbl>
	NA	NA	386	382.2681	7.948715
- Population	1	0.1777065	387	382.4458	6.133692
- Urban	1	0.9485401	388	383.3943	5.119588

Thus, our final model with backward stepwise selection would include all 10 predictors variables except Population and Urban. Our model would include: ShelveLoc, Price, CompPrice, Advertising, Age, Income, US, and Education. Using Forward and Bidirectional analysis, we get the same set of predictors to be added to our final model.

Step <S3: AsIs>	Df <dbl>	Deviance <dbl>	Resid. Df <dbl>	Resid. Dev <dbl>	AIC <dbl>
	NA	NA	397	3128.2945	822.592910
+ ShelveLoc	-2	998.443091	395	2129.8514	673.587479
+ Price	-1	696.455269	394	1433.3961	517.977222
+ CompPrice	-1	499.708806	393	933.6873	349.370454
+ Advertising	-1	261.790569	392	671.8967	220.413756
+ Age	-1	206.581637	391	465.3151	76.192593
+ Income	-1	77.146582	390	388.1685	6.045047
+ US	-1	2.472411	389	385.6961	5.501907
+ Education	-1	2.301776	388	383.3943	5.119588

Thus, using these selection methods, our final model would include ShelveLoc, Price, CompPrice, Advertising, Age, Income, US, and Education. Ultimately, it appears that all backward, forward, and bidirectional stepwise selection all result in the same predictors being selected, so we can choose any method in this step to choose our predictors.

XI. Stepwise Selection vs. Brute Force

The brute force method only includes 5 predictor variables while stepwise selection selected eight variables (included Age, US, and Education in addition to the brute-force variables). Based on the principle of parsimony, the brute force method would be a better model since there are fewer variables.

XII. Comparing Brute Force vs. Full Model (all variables) using Anova

$$H_0 : \beta_{Population} = \beta_{Age} = \beta_{Education} = \beta_{Urban} = \beta_{US} = 0$$

H_A : At least one of the above $\neq 0$

```
Analysis of Variance Table

Model 1: Sales ~ as.factor(ShelveLoc) + Price + CompPrice + Advertising +
    Income
Model 2: Sales ~ CompPrice + Income + Advertising + Population + Price +
    ShelveLoc + Age + Education + Urban + US
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    391 592.13
2    386 382.27  5    209.87 42.383 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value < alpha = 0.05, we reject H_0 and conclude we cannot drop the variables from the model. In other words, we reject the reduced model. With this information, we should use the full model instead as the other predictors appear to have a statistically significant impact on the sales of car seats.

XIII. Final Comments and Conclusions

Considering the model listed in (IX), we can see that certain variables fit the model better than others. In the model listed above, it goes as follows from most important to least important, ShelveLoc - Good, Price, CompPrice, Advertising, ShelveLoc - Medium, Income, and lastly, Population (which is why we dropped it). The model knows this because of the exhaustive model that it ran, trying to find the most influential and accurate variables. The model finds the variable that is most closely related, and then continues until all of the variables are used.

XIV. DataSet Creation

	CompPrice	Income	Advertising	Price	ShelveLoc
1	89	32	9	47	Bad
2	98	22	22	200	Bad
3	114	94	40	156	Medium
4	50	37	2	103	Good
5	174	76	13	28	Medium

For our dataset creation, we decided to create a dataset ourselves. For this dataset, we decided on values that would logically make sense and fit in our data. We then attempt to run our model on this dataset to predict car seat sales.

XV. Point Predictions

```
       1         2        3        4         5
8.118562 -3.820499 6.875110 3.185705 21.012907
```

Point Prediction based on each observation using predict() function in R. These figures represent the CarSeat Sales from our own simulated dataset. Each specific number in the point prediction refers to the observation # in the data. The equation below is our expected model equation. The predicted value of the Car Seat Sales function, shown below, produces the similar outputs using the following equation:

$$\hat{y}_h = 2.347 + 4.752 * ShelveLocGood + 1.787 * ShelveLocMedium - .0927 * Price + .0960 * CompPrice + .1185 * Advertising + .0161 * Income$$