Math 133 - Group Work 5

Pranav Jayakumar

February 10, 2025

Abstract

In this assignment we will fit and analyze a linear model to the Carseats data set from the ISLR2 library to predict sales of car seats based on various factors.

1 Data Analysis

1.1 Fitting the model

We will start by fitting a linear model of the following form:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n + \epsilon$$

where:

- $\hat{y}$ is the target (Sales)
- β_0 is the intercept coefficient
- β_n is the predictor coefficient
- x_n is the predictor variable
- ϵ is the residual error

```
        data <- Carseats
        sales_lm <- lm(Sales˜. data=data)
        summary(sales_lm)
```

1.1.1 Residuals

Min	1Q	Median	3Q	Max
-2.8692	-0.6908	0.0211	0.6636	3.4115

Table 1: Residuals

1.1.2 Coefficients

	Estimate	Std. Error	t value	Pr(> $\|t\|$)
(Intercept)	5.6606231	0.6034487	9.380	$< 2 \times 10^{-16}$
CompPrice	0.0928153	0.0041477	22.378	$< 2 \times 10^{-16}$
Income	0.0158028	0.0018451	8.565	2.58×10^{-16}
Advertising	0.1230951	0.0111237	11.066	$< 2 \times 10^{-16}$
Population	0.0002079	0.0003705	0.561	0.575
Price	-0.0953579	0.0026711	-35.700	$< 2 \times 10^{-16}$
ShelveLocGood	4.8501827	0.1531100	31.678	$< 2 \times 10^{-16}$
ShelveLocMedium	1.9567148	0.1261056	15.516	$< 2 \times 10^{-16}$
Age	-0.0460452	0.0031817	-14.472	$< 2 \times 10^{-16}$
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

Table 2: Coefficients

1.1.3 Model Summary

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	$< 2.2 \times 10^{-16}$

Table 3: Model Summary

1.2 Feature Engineering

We will now drop the insignificant terms and refit the multiple regression model with the new feature space. Our model will still be of the form:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n + \epsilon$$

We observe in our previous analysis that the terms Population, Education, Urban, and US are not significant ($p > 0.1$). These terms will be dropped from the feature space.

```
    sales_lmUpdated <- update(sales_lm, .~. , -Population-Education-Urban-US)
    summary(sales_lmUpdated)
```

1.2.1 Residuals

Min	1Q	Median	3Q	Max
-2.7728	-0.6954	0.0282	0.6732	3.3292

Table 4: Residuals Summary

1.2.2 Coefficients

	Estimate	Std. Error	t value	Pr(> $\|t\|$)
(Intercept)	5.475226	0.505005	10.84	2×10^{-16}
CompPrice	0.092571	0.004123	22.45	2×10^{-16}
Income	0.015785	0.001838	8.59	2×10^{-16}
Advertising	0.115903	0.007724	15.01	2×10^{-16}
Price	-0.095319	0.002670	-35.70	2×10^{-16}
ShelveLocGood	4.835675	0.152499	31.71	2×10^{-16}
ShelveLocMedium	1.951993	0.125375	15.57	2×10^{-16}
Age	-0.046128	0.003177	-14.52	2×10^{-16}

Table 5: Regression Coefficients

1.2.3 Model Summary

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	2×10^{-16}

Table 6: Model Summary

We will now conduct an Analysis of Variance (ANOVA) test to compare the reduced model with the full model.

```
        anova(sales_lmUpdated, sales_lm)
```

Model	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

Table 7: Analysis of Variance (ANOVA) Table

1.3 Interpreting Results

We observe the effect of qualitative variable ShelveLoc is both large and significant. ShelveLoc represents the quality of the location at which the car seat shelf is placed in a store.

The coefficients for ShelveLocGood and ShelveLocMedium are observed to be approximately 4.8357 and 1.952, respectively. This indicates good shelf locations and medium shelf locations yield approximately 4,835.7 and 1,952 more sales than bad shelf locations, respectively.