**ISQA 8080 Assignment 1 Due: By Tuesday, Sep. 24 2019, 5:30 PM (see Canvas for potential changes of the due date)**

**Name: \_\_\_Eric Troudt\_\_\_**

**NOTES:**

1. Use R for the calculations and implementation.
2. You need to submit this answer sheet and your R code.
3. Submit all documents in a zip file and upload it to Canvas. Name your Zip Folder with your name, A1, and the course # (Example: LastName-A1-ISQA 8080).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Regression Analysis (60 points)**

This question should be answered using the Carseats data set that you have available in Canvas. You can find a detailed description of the data variables by looking up the help function about the data set in the package ISLR: (library(ISLR) ?Carseats)

1. Fit a multiple regression model to predict Sales, using Price, Urban, and US as predictor variables. Show the R output for the lm model here.

Call:

lm(formula = .outcome ~ ., data = dat)

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

1. Write out the model in equation form (note that you have qualitative predictors in the model).

* Sales = 13.043469 + -0.054459\_Price + -0.021916\_Urban + 1.200573\_US

\*UrbanYes = 1, UrbanNo = 0

\*USYes = 1, USNo = 0

1. Provide an interpretation of each coefficient in the model. What does it tell you about the relationship between the target and each of the predictor variables?

* Price coefficient – With all other predictor variables held constant, a one unit increase in price (Dollars) of car seats leads to a decrease of $55.46 in car seat sales.
* Urban coefficient – stores in urban locations average $21.92 less in sales than non-urban stores.
* US coefficient – stores located in US average $1,200.57 more in sales than stores outside US.

1. For which of the predictors can you reject the null hypothesis H0: βj = 0?

- Price and US coefficients each have p-values close to 0, and allow us to reject null hypothesis of B == 0 for each of them.

1. Based on 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 (significance).

Call:

lm(formula = .outcome ~ ., data = dat)

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

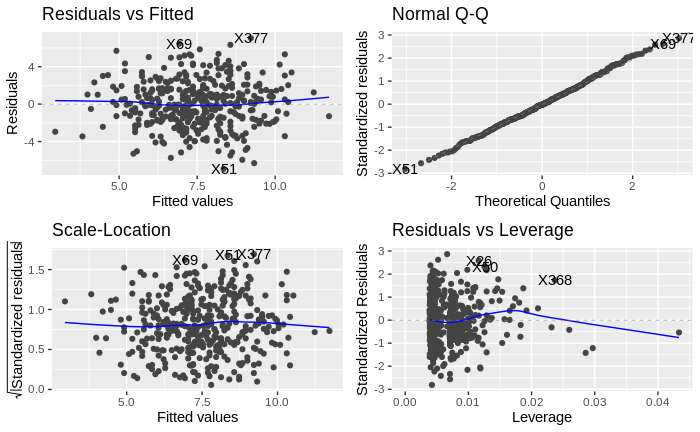
F-statistic: 62.43 on 2 and 397 DF, p-value: < 2.2e-16

1. How well do the models in (a) and (e) fit the data? Use both R2 and MSE/RMSE for this.

-The RMSE for both models is 2.46 (~ $2,460), which indicates that the average error of each sales prediction will be ~ 33% of the average sales value (~ $7,500).

-The R^2 is 0.2393 for each model (adj R^2 is negligibly smaller), indicating that both models only account for ~24% of the variability in the observed data.

1. Test the assumptions of the linear regression model, i.e., use the plot() function (or autoplot()) to test the constant variance and normality assumptions. Do they seem to be reasonable?



-Based on the smooth fit of the residual plots above, there is no discernible pattern that would suggest a non-linear relationship. Additionally there is no funnel shape to the residual plots, which indicates a valid assumption of constant variance.

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



2.5 % 97.5 %

(Intercept) 11.79032020 14.27126531

Price -0.06475984 -0.04419543

USYes 0.69151957 1.70776632

1. Using the model from (e), predict the Sales for following observation:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CompPrice | Income | Advertising | Population | Price | ShelveLoc | Age | Education | Urban | US |
| 138 | 73 | 11 | 276 | 120 | Bad | 42 | 17 | Yes | Yes |

-for Price=120, UrbanYes=1, USYes=1,

**sales=** **7.69312 ($7693.12)**

1. **Regression Analysis – Collinearity (40 points)**

This problem focuses on the *collinearity* problem and its potential effects on linear regression models.

1. Perform the following commands in R:

> set .seed (1)

> x1 <- runif (100)

> x2 <- 0.5\* x1 + rnorm (100) /10

> y <- 2 + 2\* x1 +0.3\* x2+rnorm (100)

The last line corresponds to creating a linear model in which y is

a function of x1 and x2. Write out the parametrized form of the linear model.

What are the regression (beta) coefficients?

Y = B0 + B1X1 + B2X2

Y = 2 + 2X1 + 0.3X2

B0 (Y intercept) = 2

B1 = 2

B2 = 0.3

1. What is the correlation coefficient between x1 and x2? Create a scatterplot displaying the relationship between the variables.

-Correlation coefficient = 0.835

1. Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are , , and ? How do these relate to the true *, ,* and ? Can you reject the null hypothesis *H*0 : = 0? How about the null hypothesis *H*0: = 0?

Call:

lm(formula = .outcome ~ ., data = dat)

Residuals:

Min 1Q Median 3Q Max

-2.8311 -0.7273 -0.0537 0.6338 2.3359

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.1305 0.2319 9.188 7.61e-15 \*\*\*

x1 1.4396 0.7212 1.996 0.0487 \*

x2 1.0097 1.1337 0.891 0.3754

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 on 97 degrees of freedom

Multiple R-squared: 0.2088, Adjusted R-squared: 0.1925

F-statistic: 12.8 on 2 and 97 DF, p-value: 1.164e-05

-Linear model: Y = B0 + B1X1 + B2X2

-Estimated linear model: B0 = 2.1305, B1 = 1.4396, B2 = 1.0097

-Compared with the true regression coefficients, B0 is relatively close, while B1 is underestimated by ~ 0.6 and B2 is overestimated by ~ 0.7.

-The p-value for B1 is 0.0487, which is sufficient statistical confidence to reject the null, however it’s right on the edge of the 0.05 level (at which point the null would not be rejected), so this must be kept in perspective.

-The p-value for B2 is 0.3754, which is far above the 0.05 threshold, and therefore the null cannot be rejected.

1. Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis *H*0: = 0?

Call:

lm(formula = .outcome ~ ., data = dat)

Residuals:

Min 1Q Median 3Q Max

-2.89495 -0.66874 -0.07785 0.59221 2.45560

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.1124 0.2307 9.155 8.27e-15 \*\*\*

x1 1.9759 0.3963 4.986 2.66e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.055 on 98 degrees of freedom

Multiple R-squared: 0.2024, Adjusted R-squared: 0.1942

F-statistic: 24.86 on 1 and 98 DF, p-value: 2.661e-06

-With the p-value for B1 = ~ 0, we can reject the null that B1 == 0 with high confidence.

1. Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis *H*0: = 0?

Call:

lm(formula = .outcome ~ ., data = dat)

Residuals:

Min 1Q Median 3Q Max

-2.62687 -0.75156 -0.03598 0.72383 2.44890

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.3899 0.1949 12.26 < 2e-16 \*\*\*

x2 2.8996 0.6330 4.58 1.37e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.072 on 98 degrees of freedom

Multiple R-squared: 0.1763, Adjusted R-squared: 0.1679

F-statistic: 20.98 on 1 and 98 DF, p-value: 1.366e-05

With the p-value for B1 = ~ 0, we can reject the null that B1 == 0 with high confidence.

1. Do the results obtained in (c)–(e) contradict each other? If yes, explain why this happens in this example.

- The results in c – e demonstrate the collinear effect of x1 and x2 (verified manually by calculating the *r* to be .835). When both variables correlate strongly, a linear model cannot effectively discern the individual effects of each predictor variable on the response. This in turn leads to a larger variability in coefficient estimates, which then decreases the t-statistic due to larger coefficient standard errors. With decreased t-statistics come larger p-values, and hence less powerful hypothesis tests for the linear model (textbook pg. 99, section 3.3.3)

1. Now suppose we obtain one additional observation, which was unfortunately incorrectly measured.

> x1 <- c(x1 , 0.1)

> x2 <- c(x2 , 0.8)

> y <- c(y,6)

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models?

In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

-After adding the new observation to the data-set, the modified regression model with both x1 and x2 included now exhibits strong statistical significance (p-value ~ 0.006) for x2, while the null hypothesis for x1 can no longer be rejected.

-Each of the two modified models consisting of only x1 or x2 both displayed strong significance, which is similar to the same models generated with the original data set (collinearity).

-For the two modified models in which x2 was included, an examination of the diagnostic plots revealed excessively high leverage from the new observation, and therefore may explain why x2 becomes significant in the modified model that contains both x1 and x2.