**BANA 200 Take Home Final Exam**

**Due Friday, September 10th on Canvas by 6PM Pacific Standard Time (1AM UTC Time)**

**100 Points**

# Q1) Training and Test Samples Regression (25 Points)

Starbucks is very interested in drivers that may affect a customer’s willingness to recommend Starbucks to others. In order to help management answer this question, do the following:

a. First divide the data into a **training and test sample**. Specifically, the first 5,000 observations should be the training sample, and the last 1,121 observations should be the test sample.

b. Run a **multiple regression** on the training sample using “recommend” as the dependent variable and X1 – X22 as the 22 independent variables. **Paste** the results of your regression analysis below (including all of the regression estimates and significance levels). **How many** of the 22 predictor variables are significant at the 5% level (have a p-value less than 0.05)? Reportthe **R2 value** on the training sample and comment.

c. Using your regression model estimated from part b) above, **calculate the out-of-sample R2 value** for the 1121 observations in the test sample and report it below. **Compare** the R2 value from the training sample to the R2 value you calculated in the test sample. **What can you conclude** about the model’s ability to predict “recommend” in the test sample? **How much of a difference** is there in the R2 values between the training and test samples?

# Answer:

**b)**

Call:

lm(formula = recommend ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 +

X9 + X10 + X11 + X12 + X13 + X14 + X15 + X16 + X17 + X18 +

X19 + X20 + X21 + X22, data = train)

Residuals:

Min 1Q Median 3Q Max

-7.1782 -1.3350 0.0671 1.4136 5.5228

Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | -3.468181 | 0.228723 | -15.163 | < 0.0000000000000002 \*\*\* |
| X1 | 0.636703 | 0.049315 | 12.911 | < 0.0000000000000002 \*\*\* |
| X2 | 0.204943 | 0.056005 | 3.659 | 0.000255 \*\*\* |
| X3 | 0.099213 | 0.049019 | 2.024 | 0.043027 \* |
| X4 | -0.106475 | 0.048375 | -2.201 | 0.027781 \* |
| X5 | 0.23097 | 0.045277 | 5.101 | 0.000000349850 \*\*\* |
| X6 | -0.047469 | 0.044719 | -1.061 | 0.28852 |
| X7 | 0.398114 | 0.047722 | 8.342 | < 0.0000000000000002 \*\*\* |
| X8 | 0.121266 | 0.043533 | 2.786 | 0.005363 \*\* |
| X9 | -0.125841 | 0.05058 | -2.488 | 0.012880 \* |
| X10 | 0.350641 | 0.054517 | 6.432 | 0.000000000138 \*\*\* |
| X11 | -0.080362 | 0.045805 | -1.754 | 0.079418 . |
| X12 | -0.162291 | 0.046437 | -3.495 | 0.000478 \*\*\* |
| X13 | 0.381262 | 0.036721 | 10.383 | < 0.0000000000000002 \*\*\* |
| X14 | -0.158025 | 0.049612 | -3.185 | 0.001455 \*\* |
| X15 | 0.095494 | 0.033382 | 2.861 | 0.004246 \*\* |
| X16 | 0.433071 | 0.048476 | 8.934 | < 0.0000000000000002 \*\*\* |
| X17 | 0.065832 | 0.050932 | 1.293 | 0.196225 |
| X18 | -0.003917 | 0.051929 | -0.075 | 0.939874 |
| X19 | -0.265705 | 0.047039 | -5.649 | 0.000000017076 \*\*\* |
| X20 | 0.305544 | 0.054046 | 5.653 | 0.000000016610 \*\*\* |
| X21 | 0.085914 | 0.044156 | 1.946 | 0.051747 . |
| X22 | 0.246201 | 0.049649 | 4.959 | 0.000000732915 \*\*\* |

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

Residual standard error: 1.929 on 4977 degrees of freedom

Multiple R-squared: 0.3547, Adjusted R-squared: 0.3519

F-statistic: 124.4 on 22 and 4977 DF, p-value: < 0.00000000000000022

**17 of the 22 predictor variables** are significant at the 5% level (have a p-value less than 0.05)

The **R2 value on the training sample is 0.3547** and the **adjusted R2 value is 0.3519**. The R2 value usually varies between 0 and 1. It is a statistical measure of how close the data is to the fitted regression line. It can be seen that the R2 value of the model is **quite low**. This show that the **model** **does not fit the training set well**.

**c)**

The out-of-sample R2 value for the 1121 observations in the test sample is **0.2943872**.

The R2 value for the test sample is **0.06032641 less** than the R2 value for the training sample. This shows that the model is only slightly better for the training set than the test set and this shows that the **model is** **not overfitting much**.

However, the R2 score obtained for the model in the test sample shows that the model’s ability to predict “recommend” in the test sample is **quite low** and thus the model is **not a good predictor** of recommend.

There is a difference of **0.06032641** between the R2 values between the training and test samples.

# Q2) Variable selection (25 Points)

Using only the training sample, perform a **forward variable selection** procedure by using “recommend” as the dependent variable and X1 – X22 as the 22 predictor variables.

**Paste** the results of your regression results based on the final variables selected below.

**Which variables** were dropped?

**What is the R2** of the forward selection model?

When you compare the R2 of the full model (with all 22 variables) and the R2 of the model using forward selection, **by how much did the R2 go down by**?

What can you **conclude** about how much those dropped variables really matter?

# Answer:

Call:

lm(formula = recommend ~ X1 + X16 + X7 + X13 + X5 + X20 + X10 +

X2 + X12 + X15 + X19 + X22 + X4 + X14 + X9 + X8 + X17 + X21 +

X3)

Residuals:

Min 1Q Median 3Q Max

-7.0004 -1.3590 0.0847 1.4309 5.5210

Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | -3.34234 | 0.20537 | -16.275 | < 0.0000000000000002 \*\*\* |
| X1 | 0.62487 | 0.04457 | 14.019 | < 0.0000000000000002 \*\*\* |
| X16 | 0.42453 | 0.04385 | 9.68 | < 0.0000000000000002 \*\*\* |
| X7 | 0.40688 | 0.0429 | 9.484 | < 0.0000000000000002 \*\*\* |
| X13 | 0.3848 | 0.03355 | 11.471 | < 0.0000000000000002 \*\*\* |
| X5 | 0.19762 | 0.04026 | 4.908 | 0.0000009421658 \*\*\* |
| X20 | 0.25822 | 0.04718 | 5.473 | 0.0000000461117 \*\*\* |
| X10 | 0.32749 | 0.04832 | 6.777 | 0.0000000000134 \*\*\* |
| X2 | 0.28549 | 0.05082 | 5.617 | 0.0000000202503 \*\*\* |
| X12 | -0.16328 | 0.04109 | -3.973 | 0.0000716669281 \*\*\* |
| X15 | 0.09492 | 0.03052 | 3.11 | 0.001881 \*\* |
| X19 | -0.2221 | 0.04203 | -5.284 | 0.0000001309084 \*\*\* |
| X22 | 0.17301 | 0.04503 | 3.842 | 0.000123 \*\*\* |
| X4 | -0.16294 | 0.04405 | -3.699 | 0.000219 \*\*\* |
| X14 | -0.14249 | 0.04517 | -3.154 | 0.001617 \*\* |
| X9 | -0.14782 | 0.0463 | -3.193 | 0.001417 \*\* |
| X8 | 0.10274 | 0.03974 | 2.586 | 0.009746 \*\* |
| X17 | 0.08568 | 0.04591 | 1.866 | 0.062068 . |
| X21 | 0.07454 | 0.03997 | 1.865 | 0.062197 . |
| X3 | 0.07138 | 0.04422 | 1.614 | 0.106524 |

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

Residual standard error: 1.949 on 6101 degrees of freedom

Multiple R-squared: 0.3443, Adjusted R-squared: 0.3422

F-statistic: 168.6 on 19 and 6101 DF, p-value: < 0.00000000000000022

3 variables were dropped. They are:- **X6,X11 and X18**

The R2 of the forward selection model is **0.3443**.

When you compare the R2 of the full model (with all 22 variables) and the R2 of the model using forward selection **the R2 value dropped by 0.0104**. **(0.3547 - 0.3443)**

From this it can be said that the **dropped variables do not matter at all** or matter only marginally.

# Q3 Cluster Analysis and Interpretation (25 Points)

a. Using all of the data (all 6121 observations), **create a data matrix called “X”** which includes the 22 predictor variables: X1, X2, …, X24, X25. Your data matrix X should have 6121 rows and 22 columns.

b. Once you have created your data matrix, **use the NbClust procedure** discussed in class to determine the optimal number of segments (clusters) for X. Use Euclidean distance as the distance measure, the minimum number of clusters to test = 2, and the maximum number of clusters to test = 10. Make sure to specify method = “kmeans”. Based on the analysis performed, **what are the optimal number of clusters for X**? Use the “majority rule” to determine the optimal number of clusters. **Paste the** **bar chart** you obtained from the analysis in R below and **report the optimal number of clusters** to use. Note: It might take several minutes for the analysis to run, as it is computationally intensive…

c. Using the optimal number of clusters you found in part Q3b above, run a **k-means cluster analysis** on the X matrix (perform a k-means cluster analysis on the X matrix using X1 – X22). Set “centers =” to the optimal number of clusters you found in step Q3b, and set the iter.max = 1000 and nstart=100. **Report** **below how many customers** are in Cluster 1 and how many customers are in Cluster 2.

d. Executive management has asked you to **identify the “most satisfied” segment** of customers. Examine the cluster centers from your k-means analysis and identify the segment of customers that seem the most satisfied. Hint: The most satisfied segment should be the one that generally has the highest average ratings (the highest cluster center values for X1 – X22). Once you have identified the most satisfied segment of customers, **flag this segment** and set them aside. **Report below the cluster center values for X1, X2, X3, X4, and X5** (rounded to two decimal places) for this most satisfied segment of customers.

e. Executive management wants to know by how much more the “most satisfied” segment you have identified in Step 3d above is willing to recommend Starbucks as compared to all other customers. In order to answer this question, do the following:

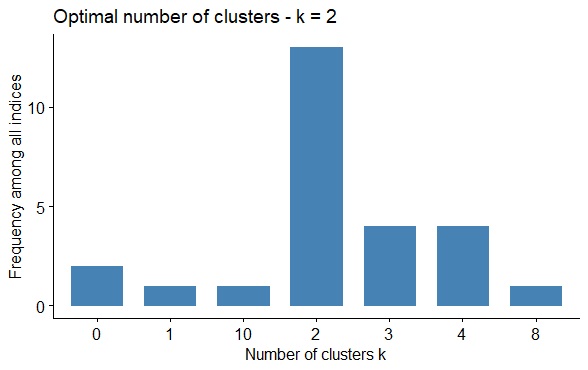
1. **Split your overall data sample** into two groups: “Most Satisfied” and “All Other”. The “most satisfied” group of customers should consist of the one segment that is most satisfied based on Step 3d above, and “All Other” customers will include all other customers that are not in the “most satisfied” segment.

2. Next, **run two separate regression analyses** for each group. Use “recommend” as the dependent variable for both regressions and X1 – X22 as the 22 predictor variables. Just to be clear: You are running two separate regressions for each one of the two groups: One regression for “Most Satisfied” and a separate regression for “All Other”.

3. For each one of the regressions, **report the average predicted values**. That is, **extract the two sets of predicted values** from the two lm objects by using the “fitted.values” function, and for each regression, take the **average of these fitted values** and report these two averages below. By how much more (in terms of average predicted “recommend”) is the “Most Satisfied” segment likely to recommend Starbucks? **Round all answers to two decimal places**.

# Answer:

The optimal number of clusters is **2.**

****

There are **3230 customers** in cluster 1 and **2891 customers** in cluster 2.

The **most satisfied segment** should be the one that generally has the highest average ratings and it is **cluster 1.**

The **cluster center values** for clusters 1 and 2 are given below with cluster 1 values for X1-X5 being highlighted:-

**1** 2

X1 **4.13** 3.28

X2 **4.16** 3.27

X3 **4.22** 3.49

X4 **4.17** 3.25

X5 **4.24** 3.46

The average of the fitted values for the most satisfied group is **7.33**.

The average of the fitted values for the all others group is **5.19**.

The difference is **2.14**.

# Q4 “What-If” Analysis (25 Points)

Management wants to figure out by how much more it can increase the “All Other” segment’s **willingness to recommend Starbucks to others**. It has conducted some market research and plans to invest in a series of advertisements. Based on the preliminary market research, management believes that it can increase each customer’s ratings in the “All Other” segment by one point for X1, X2, X7, X8, and X10. Starbucks has asked you to recalculate the average willingness to recommend for the “All Other” segment if each customer’s survey ratings increases by 1 points for X1, X2, X7, X8, and X10 in that segment.

This question is asking you to do the following:

a. For the “All Other” segment only, **increase X1, X2, X7, X8, and X10 by one point**. For example, if Customer 1 has X1 = 3, you should set X1 = 4 for this customer. However if Customer 1 has X1 = 5, you should NOT change his or her rating. Remember: the surveys are on a 5 point scale so you should not have any ratings above a “5”.

b. Once you have changed the ratings for X1, X2, X7, X8, and X10 for the “All Other” segment, only use your **existing regression model results from Q3 to recalculate the predicted “recommend”** for the “All Other” segment. So, recalculate the predicted “recommend” for the “All Other” segment but make sure to use the new values for X1, X2, X7, X8, and X10 as the basis for these predictions. Do NOT rerun your regression analysis: **Use the existing regression results to recalculate “recommend”.**

c. Once you have recalculated the predicted values for all customers in the “All Other” segment (based on the updated values for X1, X2, X7, X8, and X10), **take the average of these new predicted values and report this average below**, rounded to two decimal places. By how much is the willingness to recommend expected to increase by if Starbucks can get the “All Other” customer segment to be one point more satisfied for X1, X2, X7, X8, and X10? **Does this seem like a worthwhile thing** to do? **Comment** on whether the change seems significant or not.

# Answer:

The average of the new predicted values is **7.05.**

The willingness to recommend is **expected to increase by** **1.86** if Starbucks can get the “All Other” customer segment to be one point more satisfied for X1, X2, X7, X8, and X10.

The increase of 1.86 corresponds to an **increase of 35.74%** over the previous value. This seems to be a **extremely high increase,** considering that only 5 of the independent variables were increased by only one point. Therefore this seems to be a **worthwhile thing to do** by management if increasing the 5 independent variables can be done at less cost. However if **the cost of increasing the 5 independent variables by a point is high** then perhaps it is not a worthwhile thing to do.

The change seems to be **extremely significant** considering that the new average of the All Other customer segment, **7.328483** is now 35.74% higher than the earlier value of **5.192667.** However this **depends on the cost** of increasing the 5 independent variables by 1 point each.