ASU W. P. Carey School of Business
Arizona State University

Pilgrim Bank Case, Part II

Team 12

We performed a hypothesis test as part of the first Pilgrim Bank case which was as follows:

$$H_0: \bar{\mu}_{online\_customer\_proftability} = \bar{\mu}_{offline\_customer\_proftability}$$

$$H_a: \bar{\mu}_{online\_customer\_proftability} > \bar{\mu}_{offline\_customer\_proftability}$$

We failed to reject the null hypothesis that the two means are equal as the z-score we found was lower than our z_{α} We performed a simple regression to determine whether the outcome would be the same. We determined bank profit was the dependent variable and online/offline customers was the independent variable. Using XLSTAT we performed a simple regression categorizing the independent variable as qualitative and using a confidence interval of 99%. Based on the output, if all assumptions regarding linear regression held true, the R^2 value was zero which means that 0% of the variance in profit is explained by the variance in online vs. offline customers. In addition, the variable was not statistically significant as the p-value was 21%. However, these results do not hold as the normality test shows the residuals do not follow a Normal distribution.

As the simple regression did not give us meaningful results, we decided to perform a multiple regression analysis considering the other demographic data given to Alan Green. To ensure our results are convincing and the regression is completed responsibly, we performed the following ten step process:

1. *Think about a model*

As the objective for Green is to determine whether online customers are more profitable for Pilgrim Bank, a multiple regression model appears appropriate as we have multiple variables at play including age, income, tenure, district, online vs. offline, and bill pay.

2. *Think about causality*

We believe it is possible an online customer brings more profit to Pilgrim Bank than an offline customer as there are lower costs associated with servicing the online customer's account. However, we realize there are other demographic factors at play. As such, we realize that changing the age or income or tenure of a customer as well as their status as an online or offline customer will affect the profitability of that specific customer. A customer who has been a bank customer longer (higher tenure in years) will result in more profit for the Bank. In addition, a customer who is older and has a higher income will also result in higher profit for the bank. As a result, we have determined most of the demographic variables could result in a positive, statistically significant regression coefficient that is worth investigating.

3. *Obtain, organize, and clean data*

The data provided to Green came from a reliable source at Pilgrim Bank and has over 30,000 sample points. When reviewing the data, we noted there were missing values in both the "Age" and "Income" columns. As these are independent variables for our analysis, we are not able to just throw out the rows

with missing values without first performing further analysis. Our first step for each of these variables was to determine how many data points were missing for each variable. By filtering on "Blanks" for each column of data there were 8,261 missing data points for Income and 8,289 missing data points for Age. As there are a total of only 31,634 data points, there was too much missing data to toss them. Our next step was to determine if the profitability for rows with missing Age or Income was different than the profitability for rows with the data. If so, we could increase the biasedness of our data if we simply deleted the rows with missing data. As a result, we ran a hypothesis test for both Age and Income to determine if the average profitability was different.

For the Age independent variable, our hypothesis test was as follows:

$$H_0: \bar{\mu}_{age\_blank\_proftability} = \bar{\mu}_{age\_value\_proftability}$$

$$H_a: \bar{\mu}_{age\_blank\_proftability} \neq \bar{\mu}_{age\_value\_proftability}$$

We calculated the mean of the age value profitability using the *AVERAGEIF* function when *9age* had a value ($125.19) and the mean of the age blank profitability when *9age* did not have a value ($72.96) and took the difference of the averages equaling $52.22. We noted the sample size was 31,634 by using the *COUNT* function. We calculated the standard deviation of each data group (i.e. age blank and age value) and squared the standard deviations to give us the variance as we are unable to simply add standard deviations together like we can with means. We then took the square root to find the standard deviation of the sample of $371.80. We next calculated the standard error of the sample of 2.09. We used a 99% confidence interval resulting in a z-score of 2.575. Our right and left confidence intervals were 5.38 and -5.38, respectively. As our calculated average difference of $52.22 does not fall within this confidence interval, we reject the null hypothesis and note the missing data for the Age independent variable is different and we must keep these data points.

For the Income independent variable, our hypothesis test was as follows:

$$H_0: \bar{\mu}_{income\_blank\_proftability} = \bar{\mu}_{income\_value\_proftability}$$

$$H_a: \bar{\mu}_{income\_blank\_proftability} \neq \bar{\mu}_{income\_value\_proftability}$$

We calculated the mean of the income value profitability when *9income* had a value ($125.69) and the mean of the income blank profitability when *9income* did not have a value ($71.36) and took the difference of the averages equaling $54.32. We noted the sample size was 31,634 by using the *COUNT* function. We calculated the standard deviation of each data group (i.e. income blank and income value) and squared the standard deviations to find the variance. We then took the square root to find the standard deviation of the sample of $370.84. We calculated the standard error of the sample of 2.09. We used a 99% confidence interval resulting in a z-score of 2.575. Our right and left confidence intervals were 5.37 and -5.37, respectively. As our calculated average difference of $54.32 does not fall within this

confidence interval, we reject the null hypothesis and note the missing data for the Income independent variable is different and we must keep these data points.

Therefore, due to the differences in profitability between rows with missing data points and rows without, for each of these independent variables we decided to estimate the missing data using the "Mean or mode" option for estimating missing data points in XLSTAT when we ran the multiple regression.

4. *Match the data to the theory*

Alan Green was given several different types of demographic data including age, income, tenure, district, and bill pay. We noted that district and bill pay were not variables that would be considered value-add to the multiple regression as the end goal is to determine whether online customers result in higher profitability for Pilgrim Bank. The district (1100, 1200, or 1300) and the bill pay data are not useful variables for our end goal, especially because the supplemental data dictionary attached to the case didn't include an explanation of what the bill pay variable represents, so we can't develop a theory around it. However, age, income, and tenure are useful metrics that give us more insight into the type of customer using online vs. offline banking. As such, we decided not to include district or bill pay in our multiple regression.

5. *Simple data analysis and understanding*

We used the Correlation function within XLSTAT to derive summary statistics, scatter plots, and the correlation matrix for the quantitative variables. We included tenure and profit as quantitative variables in this test as age, income, and online vs. offline were considered qualitative (see Step 7 for further detail). Based on the correlation test, the p-value for tenure and profit were statistically significant. Reviewing the scatter plots however, there seems to be a small correlation between profit and tenure which aligns with the R^2 of 3.7%. This means 3.7% of the variance in profit is explained by the variance in the tenure of the customer.

6. *Start simple, then use a richer model*

We performed a hypothesis test during our completion of Pilgrim Bank Case I. After performing a simple regression (see above) we noted that the residuals weren't normally distributed so we couldn't use our analysis, but even if we could, the coefficient on online customers wasn't statistically significant. However, we determined a multiple regression was needed to derive any statistically significant results as the data was not normally distributed for the simple regression, and there may be helpful information gleaned on the profitability of online customers when other factors are controlled for.

7. *Is the model and are the variables appropriate?*

When reviewing the data, we determined there were three qualitative variables given to Alan Green: online usage, age bucket, and income bucket. Online usage is considered a dummy variable where there is either a 1 or 0 value based on whether the customer is using online banking or does not use online banking. The aging and income variables were both bucket variables and could have been treated as

quantitative variables by taking the average of the bucket group. However, as the bucketing for each group was not equal (eg some buckets for the income variable spanned $15,000 while others spanned $10,000, $5,000, or $25,000, etc), we decided to treat these as qualitative variables to not bias the data.

8. *Check the assumptions about ε*

We ran the multiple regression analysis with XLSTAT and the following parameters: dependent variable of profit, independent quantitative variable of tenure, independent qualitative variables of age, income, and online usage, using a 99% confidence interval, and estimating the missing data with the mean or mode. When reviewing the test assumptions related to the normality of the residuals, one should reject the null hypothesis that the residuals follow a Normal distribution. As such, we are not able to derive any meaningful results from this data.

9. *Are results robust?*

As the residuals don't have constant variance and fail the normality test, the results are not robust.

10. *Interpret results*

If we were to assume the results were robust and all of the assumptions regarding residuals held, the adjusted R^2 would be interpreted as 6.6% of the variance in profit is explained by the variance in online usage, age, income, and tenure. Further, the p-value for the F-test is considered statistically significant with it being less than 0.0001. Each variable was also determined to be statistically significant as the p-value was 0.001 or less on the coefficients. When interpreting the model parameters, the base case for the multiple regression is online, oldest age group (65 years and older), and highest income bracket ($125,000 and more). As such, on average for a customer with a tenure of 0 years, online usage, age 65 years and older, and income $125,000 and more the profit for Pilgrim Bank equals $128.30 (i.e. intercept) holding all else equal. When a customer does not use online banking, on average the profit will decrease by $15.39 holding all else equal. In addition, when a customer changes to a different age bracket on average the profit will decrease. For example, if a customer is in the Age 2 bucket, on average profit will decrease by $95.52 holding all else equal. When a customer decreases their income bucket, on average profit will also decrease. For example, if a customer is in the Income 6 bucket, on average profit will decrease by $121.91 holding all else equal. When considering tenure, on average for every year a customer banks at Pilgrim Bank, profit will increase by $4.79 holding all else equal.