**Pilgrim Bank**

**Part 1: Heterogeneity in profitability**

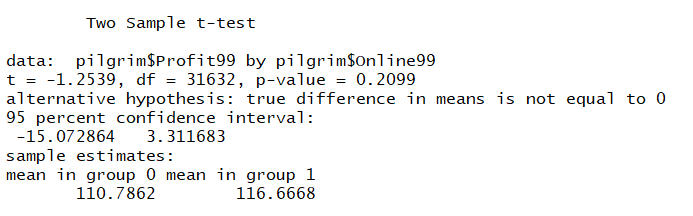
1. **21.5%,** is the minimum percentage required to realize 100% of their current profit.
2. **46.8%** of customers either return no profitability or return a negative profitability. If Pilgrim Bank only targeted profitable customers they could realize **122%** of total profit.
3. In the banking industry, variability across customers in terms of profitability may be attributed to the amount of services a customer uses (e.g. going to a branch), fees paid, and balance held.

**Part 2: Effect of online/offline – introductory analysis**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Mean** | **Standard Error** | **95% Confidence Interval** |
| **Online = 1** | $116.68 | 1.59 | 113.47 – 119.85 |
| **Offline = 0** | $110.78 | 1.52 | 107.74 – 113.84 |

The standard error represents a standardized distance from the mean, so 64% of the sample resides between Mean – SE and the Mean + SE; the confidence interval reflects where the mean lies between two values with 95% confidence. From the visual, it appears that there is a significant difference in the means between online and offline customers.

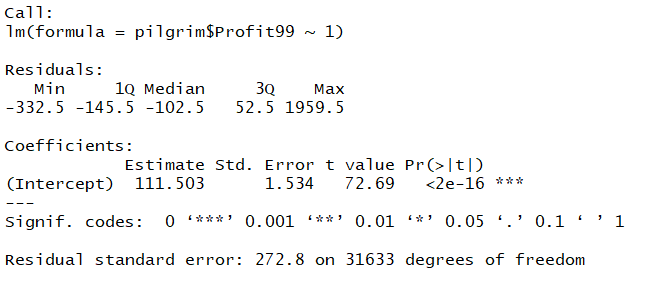
|  |  |  |
| --- | --- | --- |
| **T-Value** | **Degrees of Freedom** | **p-Value** |
| -1.2539 | 31632 | 0.2099 |



The T-Test is used to determine if two sets of data are significantly different from each other. In this case, we are testing to see if the profitability of online customers is significantly different than the profitability of offline customers, in this case, there is no significant difference between the two segments, t(31632) = -1.253, p = 0.211.

**Part 3: Simple Regression**

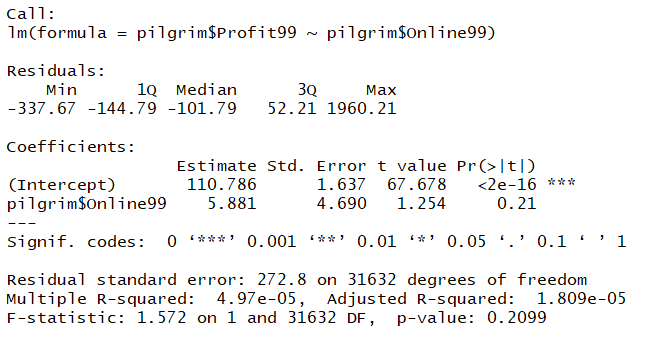
**Model 1**



The average profit of all customers is $111.50, without any other variables the average is the best guess of each customer’s contributed profit to the bank.

**Model 2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **T Value** | **P Value** |
| **Intercept** | 110.786 | 1.637 | 67.678 | <2e -16\*\*\* |
| **Pilgrim$online99** | 5.881 | 4.690 | 1.254 | 0.21 |

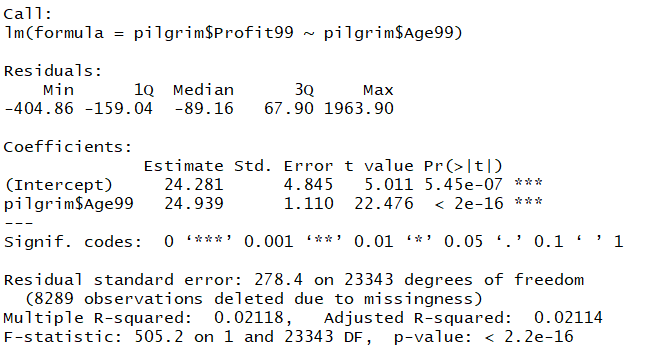


This model produced the same results as the T-Test, as in online customers have an average profitability of $116 ($ 110.79 + $ 5.88). While the profitability of offline customers average s $ 110.79. It demonstrates that online customers contribute an effect of $ 5.88 more than their offline counterparts, however, like the T-Test, the regression suggests that the effect of Online on Profitability is not significant.

The intercept is an estimation of the profitability of offline customers (Profitability = 110.78 + 5.88(0) = 110.78). The coefficient for online99 represents the profitability of an online customer (Profitability = 110.78 + 5.88(1) = 116.68).

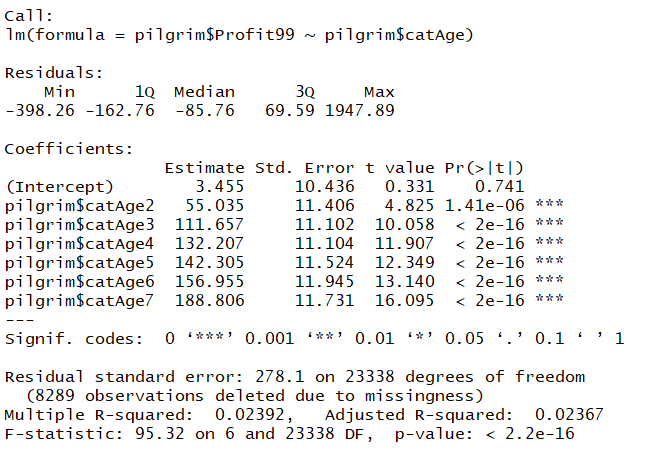
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **T Value** | **P Value** |
| **Intercept** | 24.281 | 4.845 | 5.011 | 5.45e-07\*\*\* |
| **Pilgrim$Age99** | 24.939 | 1.110 | 22.476 | < 2e-16\*\*\* |

**Model 3**

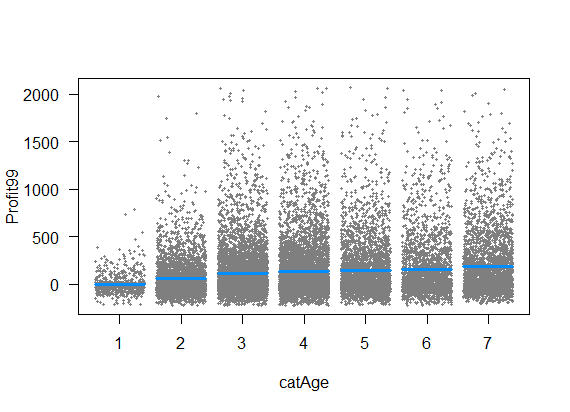
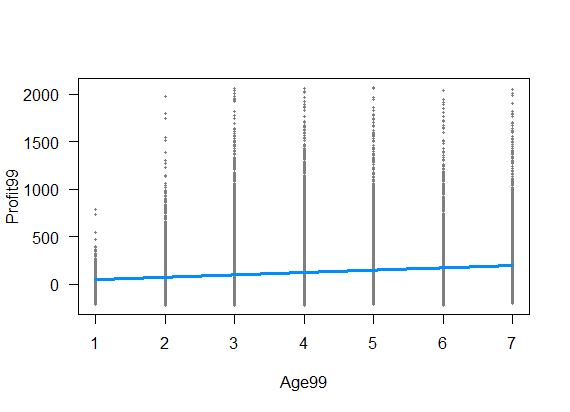


The coefficient for age represents an additional $24.94 a person contributes for every year they age on a scale from 1 to 7. The intercept represents the average profit for a customer age group 0, the best way to interpret it, everyone on average contributes at least $24.28. As no age group 0 exists, there is not much practical interpretation of the intercept alone. The models show that older customers are more profitable for the bank than younger ones.

**Model 4**



The intercept represents the prediction in profitability for customers residing in Age Group 1. Each parameter represents the predicted profit contributed depending on which age category a customer falls into. Each categorical age is compared against Age Group 1, as Age Group 1 is built into the intercept.

**Charts**

In terms of their predictions the continuous model enforces a strict linearity while the categorical model does not, making the categorical model more flexible in its predictions.

Age as a continuous predictor offers a simpler model, one value needs to substitute into the model as opposed to multiple for the dummy coded categorical age model.

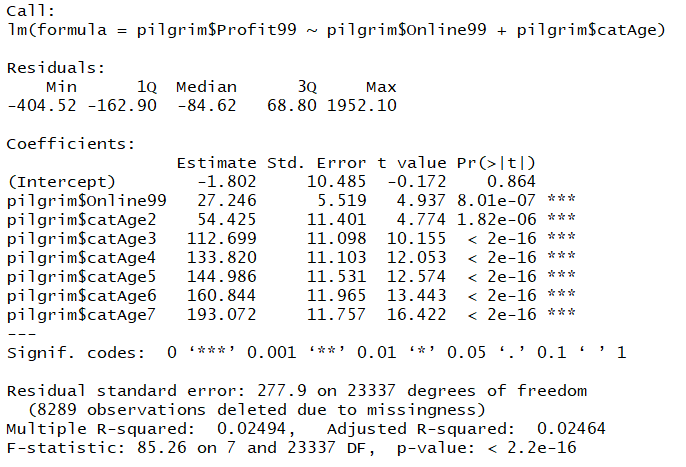
The categorical age model explains slightly more of the variance than the continuous age model (R2 = 0.02392 vs 0.02118).

**Part 4: Multiple Regression**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Online** | | | **Offline** | | |
| **Mean Age** | 3.34 | | | 4.15 | | |
| **Mean Income** | 5.95 | | | 5.38 | | |
| **Mean Tenure** | 8.65 | | | 10.37 | | |
| **Geographic Distribution** | **1100** | **1200** | **1300** | **1100** | **1200** | **1300** |
| 6% | 83% | 10% | 10% | 76% | 13% |

There are apparent differences between the groups, the offline customers are older, have a longer tenure, and are distributed geographically in a different way than the online customers. This raises the issue that we are not comparing profitability of online and offline customers directly as there is a host of confounds.

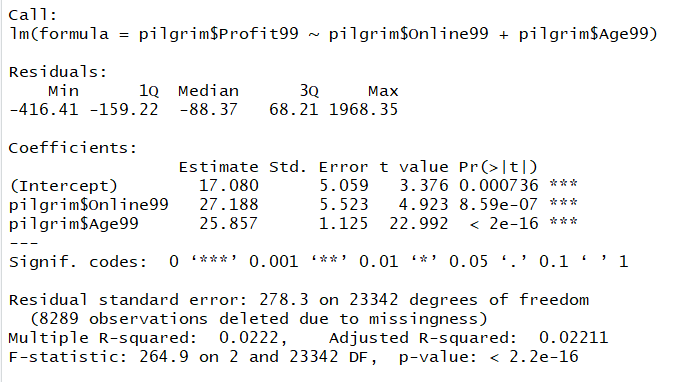
**Model 5**



In the model 5 (categorical age), the coefficients are interpreted in comparison to the intercept (as Age Group 1 is built into the intercept) at each level, no age category has influence on another age category. As age is categorical, the variables are dummy coded, so those who fall into an age category only receive a value of 1 there, while the other age categories receive a 0, when making prediction.

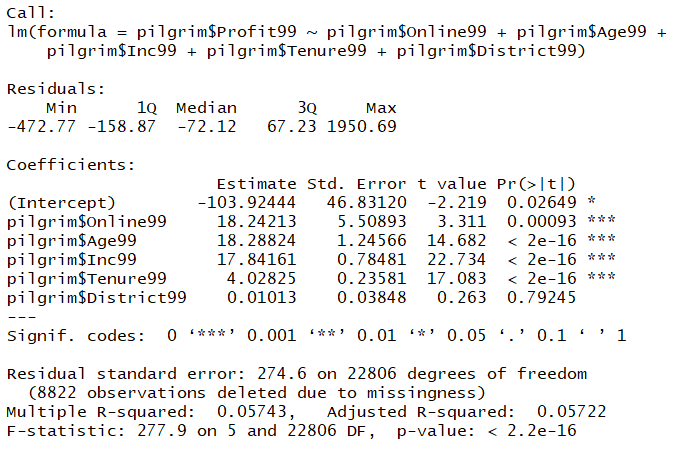
For example, those in age category 4 who are online customers are predicted to have: in profitability. This model, as compared to Model 6 has more flexibility in its interpretations.

**Model 6**



In the model with continuous age, the coefficient is interpreted as 26.857 \* age category (ranging from 1 to 7). For example, an online customer in age 4 would be predicted to have in profitability. This model is less flexible than model 5, as it enforces a rigid linearity. I think model 5 is superior, because of the flexibility in predictions it can make.

**Model 7**



The model shows that being online, older, having more income, and having a longer tenure have significant main effects on profitability, i.e. those who are higher on each parameter contribute more to profitability.

This model demonstrates that being online is more profitable for the bank, when the other demographic variables are controlled for, while model 2 suggested that online was not associated with profitability.

Model 7 predicts different values because it uses more predictors and ensures that the groups are controlled for, meaning, we are making an apples to apples comparison on what contributes to profitability, thus reducing the effect of confounds.

Model 7 explains 5% of the variability in the data, this is a lot, given that life has many random variables that contribute to one’s banking that this model cannot account for.

A note, this model treats district as a continuous variable, different predictions may arise if it is treated as categorical.

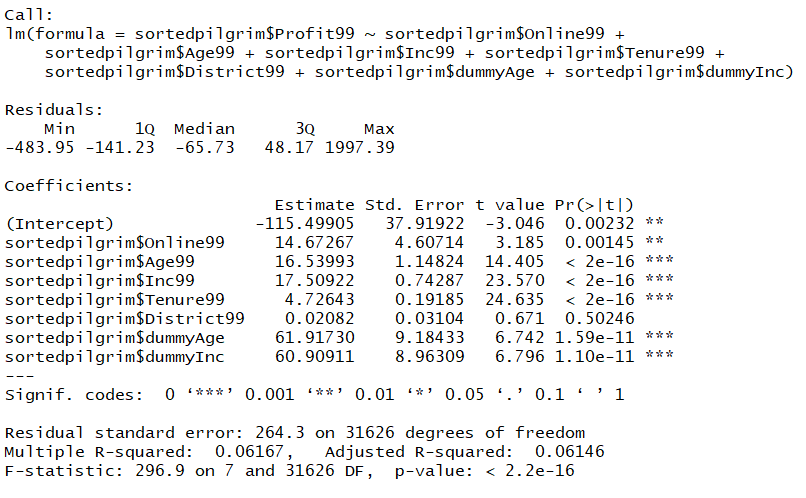
Customer 1: Online, Age Group 6, Income Group 8, Tenure 5 years, District 1200

Customer 2: Offline, Age Group 2, Income Group 3, Tenure 20 years, District 1300

**Part 5: Missing Data**

It is an issue to ignore customers with missing data, it may be the case that we are systematically biasing our estimates because the missing data is missing for a systematic reason. Older customers may have more demographic information because they have left a ‘paper trail’ over the years whereas young customers are just now starting their lives.

**Model 8**



Model 8 shows that age, income, tenure have a significant main effect on profitability, the higher each is, the more profit they contribute to the bank. Those who are missing age or income contribute $61 and $60 to profitability:

Missing both age and income:

The age99 and Inc99 coefficients are multiplied by 0 in the above case.

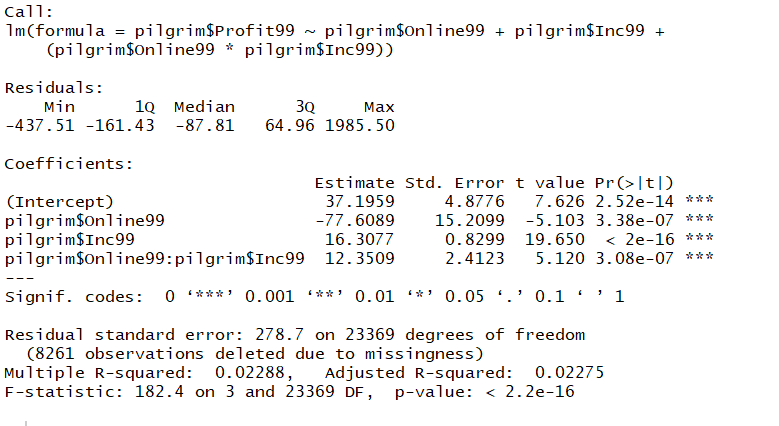
The results are different than Model 7, indicating that the missing data is systematic. I have more faith in model 8, because it offers a prediction for those that have missing data.

Customer: Online, Age and Income are Missing, Tenure 5 years, District 1200

+ 17.51 \* (0) + 4.73 \* (5) + 0.02 \* (1200) + 61.92 \* (1) + 60.91 \* (1) = $69.65

**Part 6: Interaction Modeling**

**Model 9**



There is a significant effect of income and being online on profitability. Those online contribute less profit and those with more income contribute more profit. There is a significant interaction of online on profitability that is dependent on income. Those who are online are less profitable until they reach income level 7, then they are more profitable than those offline.

This pattern may exist because those who are online may be younger and if they have less income they may require more services from which results in costs to the bank and they contribute less money into their accounts; so, there is less opportunity for the bank to profit from the customer. However, as people earn more money, and are online they require fewer of the bank’s resources (such as tellers) and cost the bank less than those who must go inside of a branch.

**Part 7: Reflection and Recommendations**

Online customers are not necessarily more profitable than offline customers. Once they reach a certain level of income they become more profitable. Based on the analysis, management should convert higher income offline customers into online customers, to cut the costs of servicing those clients. As the internet continues to take shape (because we reside in 2001 for this case) and it will be imperative for the bank to have a digital presence in which it can acquire customers smoothly and seamlessly on the desktop, on mobile, and with an omnichannel save & resume feature. These will work to cut costs for all customers and eventually assist in cutting costs in servicing customers, making the bank more profitable.