Analysing the Correlation Between U-Pass and T-Card Repurchases

Overview

- ➢ Multiple Linear Regression: attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. The multiple linear regression method will be used in order to estimate the number of T-Cards and U-Passes misplaced or stolen in forthcoming academic years at UTM. Independent variables to be used in the regression analysis include, but are not limited to, the student population at UTM in the past 4 years, the participant's year of study, program of study, means of transportation during the academic period, and his/her registration status at UTM (full-time vs part-time).
- ➢ Bootstrapping: In the process of completing our analysis, we found out that bootstrapping did not fill in missing values, and we thus opted for the statistical technique of imputation.
- ➢ Imputation: Our missing NA data was to be filled with imputation. However the amount of missing data was relatively small compared to the size of the sample. Therefore we could have either omitted the samples containing missing data, or left them in. Due to the missing data samples being so small in number, and the NAs being so unevenly distributed among fields, we decided to keep the samples with NAs so that our Linear Regression model wouldn't be biased.

Key Statistics of Interest

/ Year of Study
Frequency Table:

1	105
2	114
3	74
4	70
5	37

Mean = 2.55

/ Total UPASS Re-Purchased in the Fall/Winter term: 114

/ Total UPASS Re-Purchased in the Summer term: 18

/ Proportion of Total UPASS Re-Purchased to Total T-Card Re-Purchased: 132/336 = 0.3929

/ Total Number of T Cards lost in 2016 (n=400): 79

/ Total Number of T Cards lost in 2015 (n=294): 123

/ Total Number of T Cards lost in 2014 (n=180): 99

/ Total Number of T Cards lost in 2013 (n=106):29

/ Total Number of T Cards lost in 2012 (n=36): 6

Table 1 represents the data of T-card lost per person per year. It was calculated by ∑(T-cards lost per students in each year of study in i^{th} year) divided by the sample size of 400 (i = 2016, 2015, 2014, 2013, 2012). Thus, the results are as follows:

2016: (1/400)(40+13+7+8+11) = 0.1975 = 19.75%

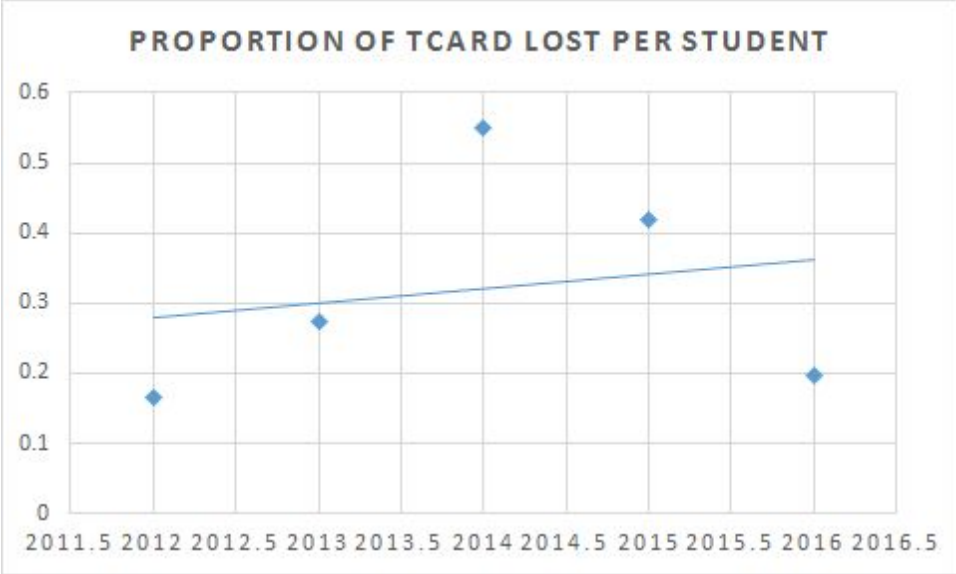
↓
2012: (1/36)(6) = 0.1667 = 16.67%

Table 1:

T-Card Lost	2016	2015	2014	2013	2012
1st year	40	N/A	N/A	N/A	N/A
2nd year	13	72	N/A	N/A	N/A
3rd year	7	20	48	N/A	N/A
4th year	8	17	31	24	N/A
5th year and above	11	14	20	5	6
total	79	123	99	29	6
T-card lost per person	***0.1975***	***0.41837***	***0.55***	***0.27358***	***0.16667***
sample size	400	294	180	106	36

After obtaining the values above, we have plotted scatter plot graph (Graph 1) representing the proportion of T-Cards lost per student per year and also included the fitted line to see the linear progression at which students in 2017 are likely to lose their T-cards.

Graph 1:



/ Total Number of U-Pass lost in 2016 (n=400): 65

/ Total Number of U-Pass lost in 2015 (n=294): 59

/ Total Number of U-Pass lost in 2014 (n=180): 65

/ Total Number of U-Pass lost in 2013 (n=106): 15

/ Total Number of U-Pass lost in 2012 (n=36): 5

Table 2 represents the data of U-Pass lost per person per year. It was calculated by ∑(U-Pass lost per students in each year of study in i^{th} year) divided by the sample size of 400 (i = 2016, 2015, 2014, 2013, 2012). Similarly, we created a scatter plot Graph 2, as per below. The calculation results are as follows:
2016: (1/400)(26+5+1+22) = 0.1625 = 16.25%

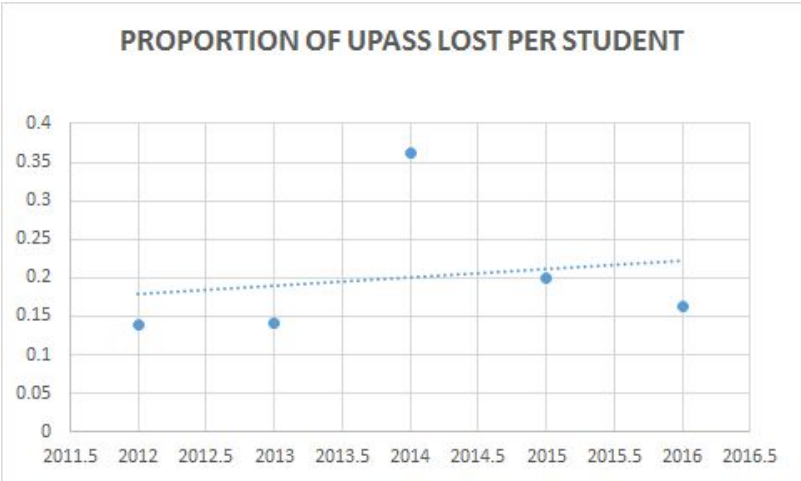
↓

2012: (1/36)(5) = 0.1389=13.89%

Table 2:

-PASS lost	2016	2015	2014	2013	2012
st year	26	N/A	N/A	N/A	N/A
nd year	5	32	N/A	N/A	N/A
rd year	1	12	29	N/A	N/A
th year	22	8	17	14	N/A
th and above	11	7	19	1	5
otal	65	59	65	15	5
-Passs lost per person	***0.1625***	***0.20068***	***0.36111***	***0.14151***	***0.13889***
ample size	400	294	180	106	36

Graph 2:



Linear Regression Report

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon_i$

Where Y_i = Total number of Tcard that the ith student has lost in his/her university career. For i=1,...,400

β_0: Intercept

X_1 : Year of study

X_2 : Total number of U-Pass that the student has lost in his/her university career

X_3 : whether the student carries the Tcard and U-Pass separately or not - if yes, 1, if no, 0

X_4 : whether the student lives off campus or on campus - if he/she lives off campus, 1, if he/she lives on campus, 0

X_5 : whether the student drive to school or not. if yes, 1, if no, 0

X_6 : Number of days in which the student comes to campus per week

X_7 : whether the student uses T card to purchase any commodity on campus or not. if yes, 1, if no, 0

$\mathcal{E}_i$: error

```
>modeltcard<-lm(Total_Number_Tcard_Loss ~ Year_Of_Study + Total_Number_Upass_Loss + Separate + Off_Campus + Drive + Days + Purchases)
> summary(modeltcard)

Call:
lm(formula = Total_Number_Tcard_Loss ~ Year_Of_Study + Total_Number_Upass_Loss +
    Separate + Off_Campus + Drive + Days + Purchases)

Residuals:
   Min      1Q  Median      3Q     Max
-1.8735 -0.5838 -0.2295  0.3148  4.2408

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)                -0.55473    0.37400  -1.483   0.1389
Year_Of_Study               0.24649    0.04452   5.537 6.04e-08 ***
Total_Number_Upass_Loss     0.74059    0.07862   9.421  < 2e-16 ***
Separate                    0.09148    0.13164   0.695   0.4876
Off_Campus                 -0.15554    0.16463  -0.945   0.3454
Drive                      -0.11350    0.13223  -0.858   0.3913
Days                        0.13865    0.06096   2.274   0.0235 *
Purchases                   0.28323    0.18066   1.568   0.1178
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9469 on 351 degrees of freedom
(41 observations deleted due to being missing)
Multiple R-squared: 0.3137,    Adjusted R-squared: 0.3
F-statistic: 22.92 on 7 and 351 DF,  p-value: < 2.2e-16
```

Thus, we can say that there is a very strong evidence to show that total number of U-Pass loss per student is positively correlated to total number of T Card loss per student based on the estimate value 0.74059 which is the slope of their relationship. Also there are strong evidence to show that number of days in which students come to campus affect the number of T card lost per student significantly, that is, the more often students come to school, the more T card will be gone missing. Additionally, there is a very strong evidence to show that the year of study is positively related to the total number of T-Card lost.

Moving forward, below is the model for U-Passes:
Since
> cor(Total_Number_Tcard_Loss, Total_Number_Upass_Loss)
[1] 0.49588
And from the previous model, we know that the total number of T-Card loss is significantly correlated, we will test for other dependent variables.

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon_i$

Where Y_i = Total number of U-Pass that the ith student has lost in his/her university career.

For i = 1,...,400

β_0: Intercept

X_1 : Year of study

X_2 : whether the student carries the Tcard and U-Pass separately or not - if yes, 1, if no, 0

X_3 :whether the student lives off campus or on campus - if he/she lives off campus, 1, if he/she lives on campus, 0

X_4 : whether the student drive to school or not. if yes, 1, if no, 0

X_5 : Number of days in which the student comes to campus per week

X_6 : whether the student uses T card to purchase any commodity on campus or not. if yes, 1, if no, 0

ε_i : error

```
> modelupass = lm(Total_Number_Upass_Loss ~ Year_Of_Study + Separate + Off_Campus + Drive + Days + Purchases)

> summary(modelupass)
Call:
lm(formula = Total_Number_Upass_Loss ~ Year_Of_Study + Separate +
    Off_Campus + Drive + Days + Purchases)

Residuals:
   Min      1Q  Median      3Q     Max
-0.6995 -0.3344 -0.1839  0.0550  3.6180

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.006456   0.253566  -0.025    0.980
Year_Of_Study  0.132557   0.029344   4.517 8.56e-06 ***
Separate          -0.071852   0.089165  -0.806    0.421
Off_Campus     -0.087935   0.111517  -0.789    0.431
Drive             -0.099961   0.089493  -1.117    0.265
Days               0.026224   0.041307   0.635    0.526
```

Purchases -0.088687 0.122397 -0.725 0.469

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.642 on 352 degrees of freedom
(again, 41 observations deleted due to being missing)
Multiple R-squared: 0.07093, Adjusted R-squared: 0.05509
F-statistic: 4.479 on 6 and 352 DF, p-value: 0.0002187

In this model, we set the Yi as the number of U-Passes lost per student. The model showed that there is a very strong evidence to support that the upper year students lost more U-Pass than lower year student. There is insufficient evidence to show the relationship between the total number of U-Passes lost and all of the independent variables except for the year of study, that were set up in the model.

Analysis Conclusion:

As per the regression analysis we can conclude that there's strong correlation between the number of T-cards lost and the number of U-Passes lost. This correlation was important to prove in order to proceed with our further assumptions. Since we have calculated the proportion of T-Cards and U-Passes lost per student per year we can assume with confidence that the future amount of T-Cards and U-Passes lost will be positively related and will correspond to the linear progression of the calculated proportions above. The below Graph 3 represents the expected UTM population over the period from 2012-2017.[1] It is predicted that in 2017 UTM population will consist of 14781 students. Due to the fitted line for both proportions (T-Cards and U-Passes) being positively sloped, as per the Graphs 1 & 2, the proportion will be growing at a steady rate, meaning the amount of T-Cards lost and U-Passes lost will be approaching infinity by every subsequent year.Therefore, for the purpose of this project we will be using the mean proportion of T-Cards and U-Passes lost per student to estimate our prediction.
Thus, assuming N=14781 (from Graph 3), the expected amount of future T-Cards to Upasses lost will be as follows:

$$p_{Tcard2017} = (0.1975+0.41837 + 0.55+0.27358 + 0.16667) /4 = 0.40153$$

$\therefore N \times p_{Tcard2017} = 14781 \times 0.40153 = 5935.01493 \cong 5936$ T-cards will be lost in total in 2017

$$p_{UPass2017} = (0.1625+0.20068+0.36111+0.14151+0.13889)/4 = 0.2512$$

$\therefore N \times p_{UPass2017} = 14781 \times 0.2512 = 3712.9872 \cong 3713$ U-Passes will be lost in total in 2017.

[1] Hammond, Cindy Ferencz. *Diane Crocker Registrar and Director of Enrolment Management Academic Affairs Committee - Oct. 25, 2020119 Enrolment Demographics*. https://www.utm.utoronto.ca/governance/sites/files/governance/public/shared/AAC_supporting_Docs_2011-12/October_25__2011/DC_s_AAC_Enrolment_presentation_-_Oct_25__2011.pdf

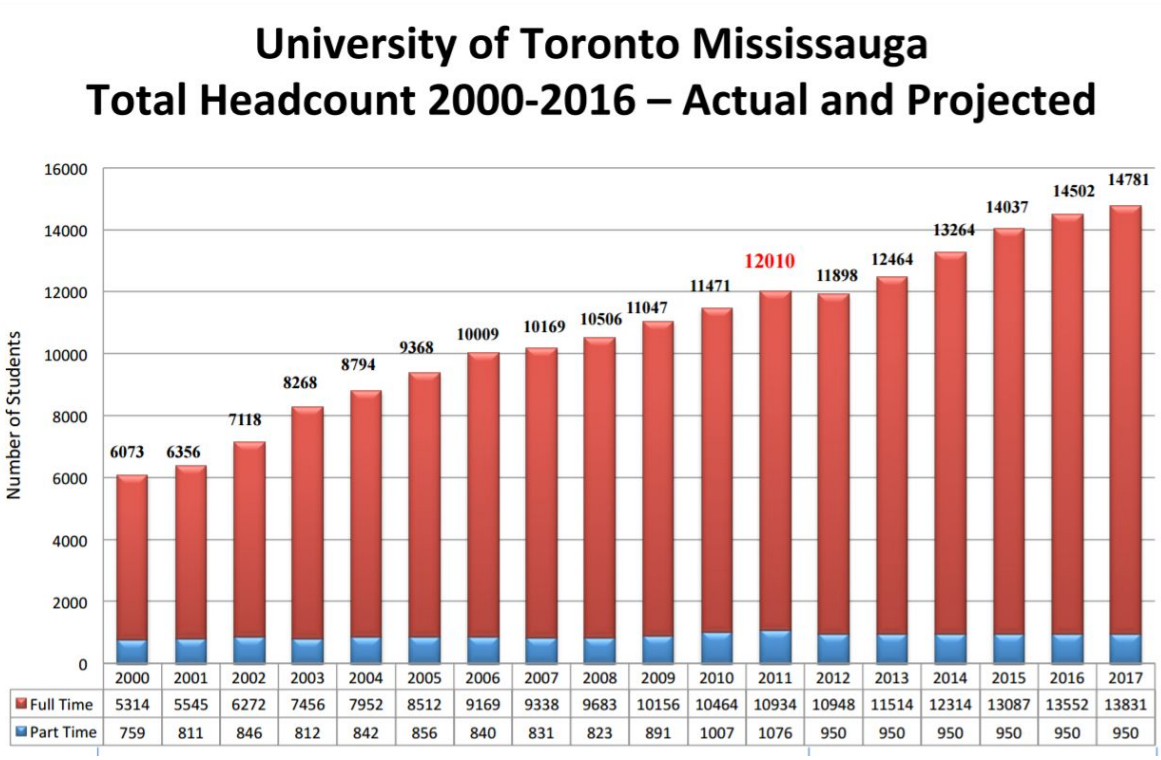
Assuming that the proportion of U-Passes lost:

- during Fall-Winter term will be 114/(114+18)= 0.8636
- During Summer term will be 18/(114+18) = 0.1364

Which leads to estimated U-Passes lost in Fall-Winter term being 0.8636× 3713=3206.68 ≅3207, and U-Passes lost in Summer term being 0.1364×3713 = 506.32≅506

Therefore, considering the cost of replacement for both items, the total amount of money spent by future UTM students on 2017 will be equal to $12(cost for T-Card replacement) × 5936 + $60(cost for U-Pass replacement during summer term)×506 + $100(cost for U-Pass replacement during Fall/Winter term) × 3207 = $422,292

Graph 3:



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Full Time	5314	5545	6272	7456	7952	8512	9169	9338	9683	10156	10464	10934	10948	11514	12314	13087	13552	13831
Part Time	759	811	846	812	842	856	840	831	823	891	1007	1076	950	950	950	950	950	950