# MGEC11 Fall 2018 *Introduction to Regression Analysis* Victor Yu

## ASSIGNMENT

Last Name Zhang

Fist Name Zhengyi (Ethan)

Student number 1002926569

Student Email [ethanzy.zhang@mail.utoronto.ca](mailto:ethanzy.zhang@mail.utoronto.ca)

## Question 1

* The full model of the regression:

*voteA = 32.117 + 0.342prtystrA + 0.038expendA – 0.031expendB*

*– 6.63 x10-6 expendA × expendB + 𝛆*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 4 | 27660.9659 | 6915.241476 | 55.86385737 | 6.76732E-30 |
| Residual | 168 | 20796.28265 | 123.7873967 |  |  |
| Total | 172 | 48457.24855 |  |  |  |

* The restricted model of regression of *expendB*, holding *prtystrA* and *expendA* fixed:

*voteA = 22.067 + 0.448prtystrA + 0.020expendA + 𝛆*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 2 | 10808.58595 | 5404.292976 | 24.4027209 | 4.8219E-10 |
| Residual | 170 | 37648.6626 | 221.4627212 |  |  |
| Total | 172 | 48457.24855 |  |  |  |

Number of explanatory variables dropped q = 2;

F = [(37648.6626 - 20796.28265)/2] / [20796.28265/168] = 68.070

* The restricted model of regression of expendA, holding prtystrB and expendB fixed:

*voteA = 28.677 + 0.579prtystrA – 0.023expendB + 𝛆*

Number of explanatory variables dropped q = 2;

F = [(34188.781 - 20796.28265)/2] / [20796.28265/168] = 54.095

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA | |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 2 | 14268.468 | 7134.234 | 35.474 | 0.000 |
| Residual | 170 | 34188.781 | 201.110 |  |  |
| Total | 172 | 48457.249 |  |  |  |

* The expected sign for *β4* is not obvious. The T-stat for variable *expendA × expendB* is -0.923 > -1.974 (critical value for df = 168) and P-value = 0.358 which is very large. We cannot reject *β4*is insignificant. i.e. *β4* is not significant.

## Question 2

Estimating the regression model, the equation form:

***voteA = (32.117) + 0.342(prtystrA) + 0.038(expendA)***

***(4.591) (0.088) (0.005)***

***– 0.031(expendB) – 6.63 x10-6 (expendA×expendB) + 𝛆***

***(0.005) (0.000)***

***n = 173 R2 = 0.570832367 adj.R2 = 0.56061409***

The interaction term *expendA×expendB* is statistically insignificant with t-stat = -0.923. While the coefficient on prtystrA, expendA and expendB are significant.

## Question 3

The average of *expendA* is 310.6111.

Partial derivative of *voteA* on *expendB:*

d(*voteA*) = -0.00000663(*expendA*) - 0.031

Set *expendA* = 300,

Δ(*voteA*) = 100 (-0.00000663(300) - 0.031) = -3.30

An increase of $100,000 spent by Candidate B would lead to a decrease of 3.30 votes on *voteA*. It is not a large effect.

## Question 4

Partial derivative of *voteA* on *expendA:*

d(*voteA*) = -0.00000663(*expendB*) + 0.038

Set *expendB* = 100,

Δ(*voteA*) = 100 (-0.00000663(100) + 0.038) = 3.734

The effect of spending $100,000 more by Candidate A would increase *voteA* by 3.734.

## Question 5

***voteA = (18.196) + 0.157(prtystrA) - 0.007 (expendA)***

***(2.568) (0.050) (0.003)***

***+ 0.004(expendB) + 0.494 (shareA) + 𝛆***

***(0.003) (0.025)***

***n = 173 R2 = 0.86816 adj.R2 = 0.*** ***865016***

It would not make sense changing shareA while holding expendA and expendB. Because *shareA* = *expendA* / (*expendA* + *expendB*). If shareA is change, at least either expendA or expendB must be changed simultaneously.

## Question 6

***voteA = (37.661) + 0.252(prtystrA) + 3.793(democA)***

***(4.736) (0.071) (1.407)***

***+ 5.779(lexpendA) – 6.238(lexpendB) + 𝛆***

***(0.392) (0.397)***

***n = 173 R2 = 0.80116 adj.R2 = 0.79643***

Why R2 = 0:

Since E(Σ𝛆 𝔦 )= 0, then avg(𝛆 𝔦) → 0;

Regress on 𝛆 𝔦 , TTS = Σ(𝛆 𝔦 – 0)2 = Σ 𝛆 𝔦2 ;

Since Σ 𝛆 𝔦2 is the minimal sum of residual for this set of independent variables after regression on *voteA*, if we minimize the sum of residual when regressing 𝛆 𝔦 , we would have the same result -- Σ 𝛆 𝔦2. i.e. SSR = Σ 𝛆 𝔦2.

R2 = 1-SSR/TTS = 1 - Σ 𝛆 𝔦2/ Σ 𝛆 𝔦2 = 1 – 1 = 0.

## Question 7

## Question 8

I expect *β1 and β3* to be negative, since the smaller the schools are or higher percentile it ranks, higher GPA its students tends to achieve; I expect *β4* to be positive, since higher SAT the students get, higher GPA they tend to achieve; I am not sure about *β5* and *β6* because it is not clear to me whether being female or athlete would have a significant impact on their GPA.

## Question 9

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 6 | 524.819 | 87.470 | 284.589 | 0 |
| Residual | 4130 | 1269.376 | 0.307 |  |  |
| Total | 4136 | 1794.196 |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 1.241 | 0.079 | 15.616 | 0.000 |
| hsize | -0.057 | 0.016 | -3.477 | 0.001 |
| hsizeq | 0.005 | 0.002 | 2.079 | 0.038 |
| hsperc | -0.013 | 0.001 | -23.068 | 0.000 |
| sat | 0.002 | 0.000 | 24.640 | 0.000 |
| female | 0.155 | 0.018 | 8.602 | 0.000 |
| athlete | 0.169 | 0.042 | 3.998 | 0.000 |

***colgpa = (1.241) – 0.057 (hsize) + 0.005 (hsizeq)***

***(0.079) (0.016) (0.002)***

***– 0.013 (hsperc) + 0.002 (sat)***

***(0.001) (0.00007)***

***+ 0.155 (female) + 0.169 (athlete) + 𝛆***

***(0.018) (0.042)***

***n = 4137 R2 = 0.292509512 adj.R2 = 0.291481681***

The estimated GPA differential between athletes and non-athletes is 0.169. It is very statically significant. The coefficient for *athlete* has t-stat = 3.998, which is greater than the critical value = 1.96.

## Question 10

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 5 | 338.217 | 67.643 | 191.922 | 0.000 |
| Residual | 4131 | 1455.979 | 0.352 |  |  |
| Total | 4136 | 1794.196 |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 3.048 | 0.033 | 92.594 | 0.000 |
| hsize | -0.053 | 0.018 | -3.050 | 0.002 |
| hsizeq | 0.005 | 0.002 | 2.210 | 0.027 |
| hsperc | -0.017 | 0.001 | -29.086 | 0.000 |
| female | 0.058 | 0.019 | 3.089 | 0.002 |
| athlete | 0.005 | 0.045 | 0.122 | 0.903 |

***colgpa = (3.048) – 0.053 (hsize) + 0.005 (hsizeq)***

***(0.033) (0.018) (0.002)***

***– 0.017 (hsperc) + 0.058 (female)***

***(0.001) (0.019)***

***+ 0.05 (athlete) + 𝛆***

***(0.045)***

***n = 4137 R2 = 0.188507 adj.R2 = 0.187524***

Now, the estimated GPA differential between athletes and non-athletes dropped from 0.169 to 0.005. The reason for that is

1) In the new model, coefficient of variable athlete is no longer statically significant. Its t-stat has dropped to 0.122 which is way smaller than critical value = 1.96.

2) Since we do not control SAT, in the new model intercept has gone up more than twice to 3.048, which would likely make the boosting impact from a single variable smaller than before.

## Question 11

To test the difference between women athletes and women non-athletes, we chose women non-athletes as base group:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 7 | 524.821 | 74.974 | 243.876 | 0 |
| Residual | 4129 | 1269.374 | 0.307 |  |  |
| Total | 4136 | 1794.196 |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 1.396 | 0.076 | 18.478 | 0.000 |
| hsize | -0.057 | 0.016 | -3.470 | 0.001 |
| hsizesq | 0.005 | 0.002 | 2.075 | 0.038 |
| hsperc | -0.013 | 0.001 | -23.056 | 0.000 |
| sat | 0.002 | 0.000 | 24.618 | 0.000 |
| feath | 0.175 | 0.084 | 2.084 | 0.037 |
| malath | 0.013 | 0.049 | 0.263 | 0.793 |
| malnoath | -0.155 | 0.018 | -8.443 | 0.000 |

From the coefficient of feath (female athlete), we estimate that female athlete achieves 0.175 more GPA than female non-athlete. It is also very statically significant since t-stat = -8.443 < -1.96

## Question 12

To test if gender effect sat on colgpa, we create a new variable femalesat = female \* sat

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 7 | 524.868 | 74.981 | 243.906 | 0.000 |
| Residual | 4129 | 1269.328 | 0.307 |  |  |
| Total | 4136 | 1794.196 |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 1.264 | 0.097 | 12.962 | 0.000 |
| hsize | -0.057 | 0.016 | -3.480 | 0.001 |
| hsizesq | 0.005 | 0.002 | 2.083 | 0.037 |
| hsperc | -0.013 | 0.001 | -23.053 | 0.000 |
| sat | 0.002 | 0.000 | 19.089 | 0.000 |
| female | 0.102 | 0.134 | 0.765 | 0.445 |
| athlete | 0.168 | 0.043 | 3.944 | 0.000 |
| femalesat | 0.0001 | 0.0001 | 0.3965 | 0.6917 |

From the result, femalesat is practically and statically no impact on colgpa. i.e. There is no significant effect of sat on colgpa differ by gender.