Count Model Regression

Case Study

Najib Mozahem

May 1, 2019

The Dataset

- id: unique student identifier
- gpa: overall GPA of the student
- total_fail: the total number of courses in which the student has failed (this is the dependent variable)
- college: whether the student is in the engineering school or the business school (one means business, two means engineering)
- gender: whether the student is a male or a female (one means female, two means male)
- english: the average grade on all English courses taken by the student (data is taken from a non-English speaking country where the language of instruction in university is English)
- total_courses: the total number of courses taken by the student so far in the university

Univariable Tests

- The first thing that we should do when conducting regression analysis is to perform univariate analysis, where we try and uncover whether there is a relationship between the dependent variable and each independent variable separately.
- Once we have a good idea about the nature of these individual relationships, we can start building the model.
- In the case of count data, it is always a good idea to look at the histogram of the dependent variable in order to get an idea of the variable that we are dealing with.

Histogram of Dependent Variable

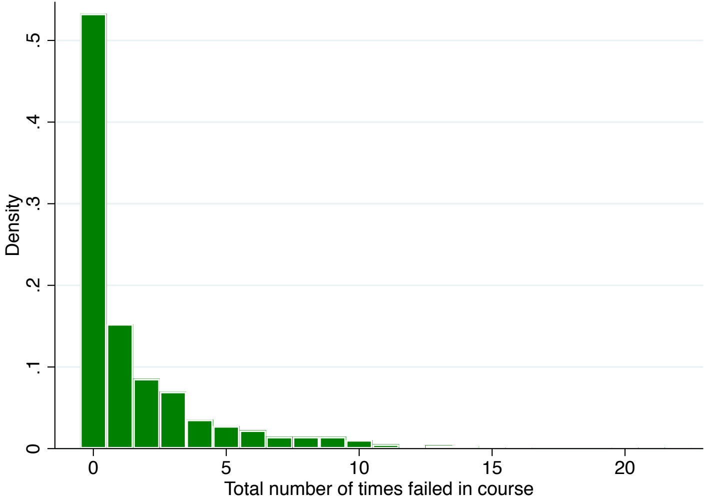

Figure: Histogram of the dependent variable.

Continuous Variables

In linear regression, when we have a continuous independent variable, we start our analysis by plotting a scatter plot. Graphs are also useful as a starting step in count models, but their shape is different from what we are used to due to the nature of the dependent variable.

Continuous Variables - GPA (Scatter plot

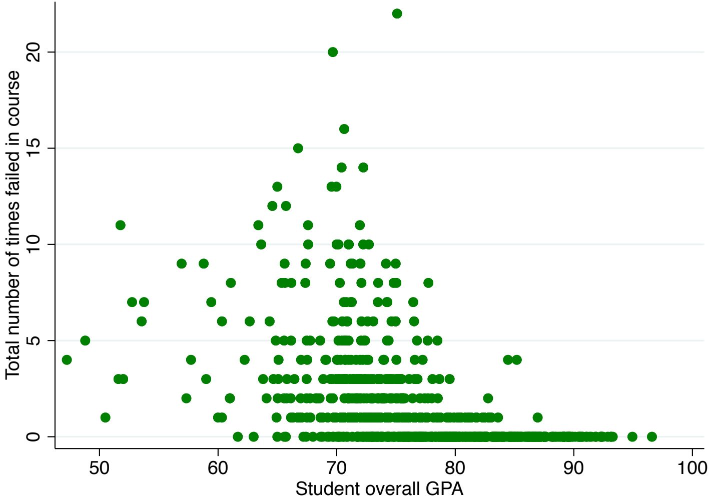

Figure: Scatterplot of total_fail and GPA.

Continuous Variables - GPA (Smoothed scatter plot)

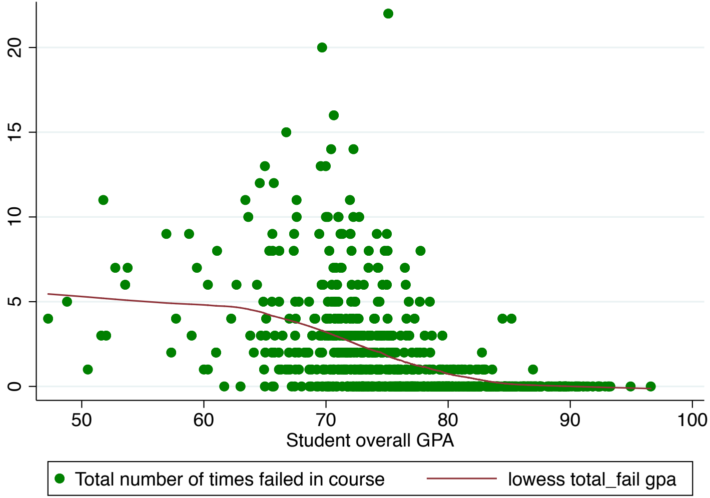

Figure: Smoothed scatterplot.

Continuous Variables - GPA (Poisson Regression)

```
Iteration 0:   log likelihood = -1495.7527
Iteration 1:   log likelihood = -1495.6504
Iteration 2:   log likelihood = -1495.6504

Poisson regression                                Number of obs   =        760
                                                  LR chi2(1)      =     817.86
                                                  Prob > chi2     =     0.0000
Log likelihood = -1495.6504                       Pseudo R2       =     0.2147

------------------------------------------------------------------------------
  total_fail |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |  -.0952663   .0030879   -30.85   0.000    -.1013185   -.0892141
       _cons |   7.528266   .2187085    34.42   0.000     7.099605    7.956927
------------------------------------------------------------------------------
```

Continuous Variables - GPA (Poisson Regression - IRR)

```
Iteration 0:   log likelihood = -1495.7527
Iteration 1:   log likelihood = -1495.6504
Iteration 2:   log likelihood = -1495.6504

Poisson regression                              Number of obs     =        760
                                                LR chi2(1)        =     817.86
                                                Prob > chi2       =     0.0000
Log likelihood = -1495.6504                     Pseudo R2         =     0.2147
```

total_fail	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	.9091308	.0028073	-30.85	0.000	.9036452	.9146498
_cons	1859.877	406.7711	34.42	0.000	1211.488	2855.284

Note: _cons estimates baseline incidence rate.

Exposure

We have however, disregarded one important factor so far, and it is the exposure time. As you recall from the theory part of the course, we need to account for the fact that different subjects were exposed to the probability of the event occurring for different period of time. In our dataset, the variable total_courses contains the total number of courses taken by the student. Statistical packages allow us to iclude an exposure variable. The output after dincluding total_courses as ab exposure variable is shown below.

Exposure

```
Iteration 0:   log likelihood =  -1146.808
Iteration 1:   log likelihood = -1146.7986
Iteration 2:   log likelihood = -1146.7986

Poisson regression                              Number of obs     =        760
                                                LR chi2(1)        =    1345.83
                                                Prob > chi2       =     0.0000
Log likelihood = -1146.7986                     Pseudo R2         =     0.3698

------------------------------------------------------------------------------
  total_fail |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |    .868657   .0029586   -41.34   0.000     .8628775    .8744752
       _cons |   1694.273   408.1211    30.87   0.000     1056.681    2716.584
ln(total_~s) |          1  (exposure)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Exposure

- As we can see, the value of IRR for the variable gpa is now slightly different.
- We also see that there is a new row in the regression table that contains the elements ln(total_courses).
- From this point forward, we will be including the variable total_courses as an exposure in all the output.

Continuous Variables - English (Smoothed scatter plot)

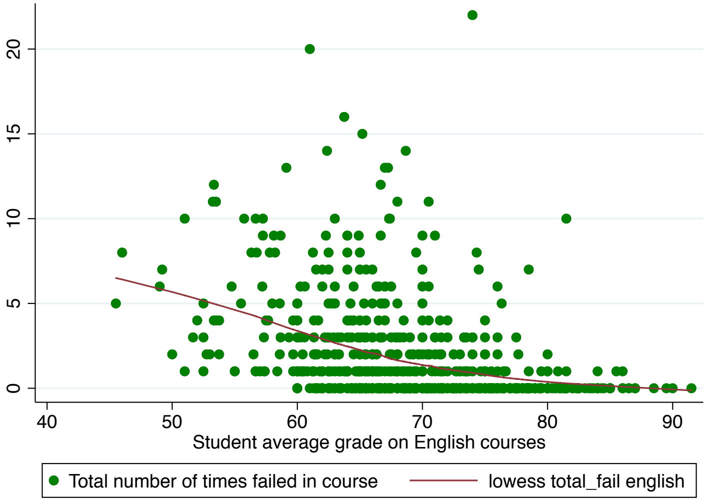

Figure: Smoothed scatterplot of total_fail and english.

Continuous Variables - English (Poisson Regression - IRR)

```
Iteration 0:   log likelihood = -1406.8576
Iteration 1:   log likelihood = -1406.8565
Iteration 2:   log likelihood = -1406.8565

Poisson regression                              Number of obs     =        755
                                                LR chi2(1)        =     801.08
                                                Prob > chi2       =     0.0000
Log likelihood = -1406.8565                     Pseudo R2         =     0.2216
```

total_fail	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
english	.8893731	.0036963	-28.21	0.000	.882158	.8966472
_cons	131.3423	35.34142	18.13	0.000	77.51119	222.5589
ln(total_~s)	1	(exposure)				

Note: _cons estimates baseline incidence rate.

Binary Variables

Now that we have seen how to analyze the relationship between the binary dependent variable and a continuous independent variable, we move onto other types of variables. Looking at our dataset, we notice that the variables gender and college are binary. Both take on two values.

Binary Variables - College (Poisson Regression)

```
Iteration 0:   log likelihood = -1813.5521
Iteration 1:   log likelihood = -1813.5521

Poisson regression                              Number of obs     =        760
                                                LR chi2(1)        =      12.33
                                                Prob > chi2       =     0.0004
Log likelihood = -1813.5521                     Pseudo R2         =     0.0034

------------------------------------------------------------------------------
  total_fail |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     college |
 Engineering |   .8180037   .0465096    -3.53   0.000     .7317423    .9144341
       _cons |   .0571091   .0024783   -65.97   0.000     .0524525     .062179
ln(total_~s) |          1  (exposure)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Binary Variables - Gender (Poisson Regression)

```
Iteration 0:   log likelihood = -1712.8214
Iteration 1:   log likelihood = -1712.8145
Iteration 2:   log likelihood = -1712.8145

Poisson regression                              Number of obs     =        760
                                                LR chi2(1)        =     213.80
                                                Prob > chi2       =     0.0000
Log likelihood = -1712.8145                     Pseudo R2         =     0.0587

------------------------------------------------------------------------------
  total_fail |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |
       male  |   2.845421   .2288983    13.00   0.000     2.430369    3.331356
       _cons |   .0224076   .0016702   -50.96   0.000      .019362    .0259322
ln(total_~s) |          1  (exposure)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Multivariate Analysis

```
Iteration 0:   log likelihood = -1099.3037
Iteration 1:   log likelihood = -1099.2339
Iteration 2:   log likelihood = -1099.2339

Poisson regression                              Number of obs     =        755
                                                LR chi2(4)        =    1416.32
                                                Prob > chi2       =     0.0000
Log likelihood = -1099.2339                     Pseudo R2         =     0.3918

------------------------------------------------------------------------------
  total_fail |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   .8876014   .0041204   -25.68   0.000     .8795622     .895714
     english |   .9663382      .0051    -6.49   0.000      .956394    .9763859
             |
     college |
 Engineering |   1.113613   .0669225     1.79   0.073     .9898769    1.252815
             |
      gender |
        male |   1.413999   .1209597     4.05   0.000     1.195731    1.672109
       _cons |   2418.674   780.7668    24.14   0.000     1284.703    4553.568
ln(total_~s) |          1  (exposure)
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Negative Binomial Regression

Now it is time to see whether the data displays overdispersion. As you recall, when there is evidence that overdispersion exists, we will need to fit a negative binomial model that will estimate the new parameter alpha.

Negative Binomial Regression

```
Negative binomial regression                    Number of obs     =        755
                                                LR chi2(4)        =     540.91
Dispersion     = mean                           Prob > chi2       =     0.0000
Log likelihood = -987.46614                     Pseudo R2         =     0.2150
```

total_fail	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	.8562923	.0082241	-16.15	0.000	.8403241	.8725639
english	.9626258	.0078959	-4.64	0.000	.9472738	.9782266
college						
Engineering	1.221273	.1176047	2.08	0.038	1.011218	1.474961
gender						
male	1.327607	.1531766	2.46	0.014	1.058912	1.664484
_cons	39032.16	26223.31	15.74	0.000	10460.5	145644
ln(total_~s)	1	(exposure)				
/lnalpha	-.6676535	.1361952			-.9345912	-.4007157
alpha	.5129107	.069856			.3927464	.6698404

```
Note: Estimates are transformed only in the first equation.
Note: _cons estimates baseline incidence rate.
LR test of alpha=0: chibar2(01) = 223.54               Prob >= chibar2 = 0.000
```

Zero-Inflated Models

- When we plotted the histogram of the dependent variables, we noted that the number of zeros seems to be too high.
- This means that perhaps a zero-inflated model would be of better use.
- Since we have found that there is overdispersion in the data, it would make sense for us to fit the zero-inflated negative binomial regression model.

Zero-Inflated Model

```
Zero-inflated negative binomial regression      Number of obs     =        755
                                                Nonzero obs       =        351
                                                Zero obs          =        404

Inflation model = logit                         LR chi2(4)        =     253.69
Log likelihood  = -957.9385                     Prob > chi2       =     0.0000
```

total_fail	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
total_fail						
gpa	-.1057381	.0090785	-11.65	0.000	-.1235317	-.0879445
english	-.0304713	.0076049	-4.01	0.000	-.0453766	-.015566
college						
Engineering	.1139091	.0943508	1.21	0.227	-.071015	.2988332
gender						
male	.2515483	.1220756	2.06	0.039	.0122845	.4908122
_cons	6.801989	.66255	10.27	0.000	5.503415	8.100564
ln(total_~s)	1	(exposure)				
inflate						
gpa	.3518453	.0541493	6.50	0.000	.2457145	.4579761
english	.0159122	.0350812	0.45	0.650	-.0528457	.0846701
college						
Engineering	-.3981901	.4403973	-0.90	0.366	-1.261353	.4649727
gender						
male	-.0967932	.4380667	-0.22	0.825	-.9553881	.7618018
_cons	-28.72857	3.672004	-7.82	0.000	-35.92556	-21.53157
/lnalpha	-1.320756	.1988137	-6.64	0.000	-1.710424	-.9310886
alpha	.2669333	.05307			.1807891	.3941244

Zero-Inflated Model

In some statistical packages, it is possible to request that the exponentiated coefficients be diplayed.

Zero-Inflated Model - Exponentiated Coefficients

```
zinb (N=755): Factor change in expected count
  Observed SD:  2.8663
Count equation: Factor change in expected count for those not always 0
```

	b	z	P>|z|	e^b	e^bStdX	SDofX
gpa	-0.1057	-11.647	0.000	0.900	0.450	7.555
english	-0.0305	-4.007	0.000	0.970	0.800	7.338
college						
Engineering	0.1139	1.207	0.227	1.121	1.058	0.494
gender						
male	0.2515	2.061	0.039	1.286	1.125	0.467
constant	6.8020	10.266	0.000	.	.	.
alpha						
lnalpha	-1.3208	.	.	.	.	.
alpha	0.2669	.	.	.	.	.

```
Binary equation: factor change in odds of always 0
```

	b	z	P>|z|	e^b	e^bStdX	SDofX
gpa	0.3518	6.498	0.000	1.422	14.269	7.555
english	0.0159	0.454	0.650	1.016	1.124	7.338
college						
Engineering	-0.3982	-0.904	0.366	0.672	0.821	0.494
gender						
male	-0.0968	-0.221	0.825	0.908	0.956	0.467
constant	-28.7286	-7.824	0.000	.	.	.

Zero-Inflated Model

We are interested in the p-values of the inflate part since we originally included all the independent variables in order to see which ones might be inflating the number of zeros. We had previously seen that that in the “inflate” part, only the variable gpa is significant. This would indicate that we might be better off fitting a model that only included gpa in the ”inflate” part:

Zero-Inflated Model - Updated Model

```
Zero-inflated negative binomial regression      Number of obs     =        755
                                                Nonzero obs       =        351
                                                Zero obs          =        404

Inflation model = logit                         LR chi2(4)        =     259.07
Log likelihood  = -958.4367                     Prob > chi2       =     0.0000
```

total_fail	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
total_fail						
gpa	-.1059945	.0090624	-11.70	0.000	-.1237565	-.0882324
english	-.0316688	.0073566	-4.30	0.000	-.0460876	-.0172501
college						
Engineering	.1447259	.0880362	1.64	0.100	-.0278219	.3172737
gender						
male	.2656364	.1119332	2.37	0.018	.0462513	.4850214
_cons	6.86581	.6550236	10.48	0.000	5.581988	8.149633
ln(total_~s)	1	(exposure)				
inflate						
gpa	.360155	.0459039	7.85	0.000	.270185	.4501251
_cons	-28.61654	3.626404	-7.89	0.000	-35.72416	-21.50892
/lnalpha	-1.312343	.1968799	-6.67	0.000	-1.69822	-.9264651
alpha	.2691887	.0529978			.183009	.3959509

Comparing Count Models

So which is it? Is it the negative binomial model or the zero-inflated binomial model? This is where we need to compare the four models: Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial.

Comparing Count Models - Graphing the Erros of each Model

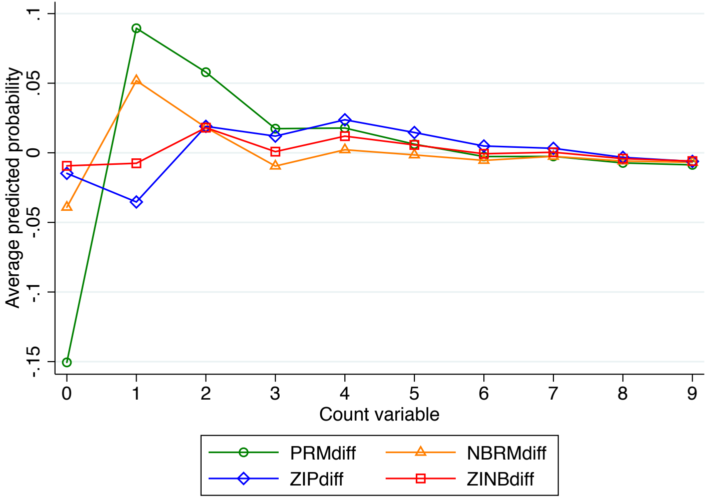

Figure: Difference between observed and predicted values in each of the four models.

Comparing Count Models - AIC and BIC Statistics

Variable	PRM	NBRM	ZIP	ZINB
total_fail				
Student overall GPA	0.928	0.859	0.966	0.941
	-17.66	-11.68	-6.74	-4.75
Student average grade on Eng~h	0.967	0.971	0.985	0.975
	-6.92	-2.84	-2.95	-2.65
College				
Engineering	1.218	1.450	1.177	1.291
	3.28	3.04	2.53	2.25
Gender				
male	1.553	1.426	1.519	1.459
	5.11	2.56	4.54	2.77
Constant	2409.199	5.01e+05	66.849	689.760
	25.64	14.77	11.20	7.17
lnalpha				
Constant		1.222		0.731
		2.05		-2.26
inflate				
Student overall GPA			1.318	1.485
			11.59	7.91
Constant			0.000	0.000
			-11.61	-7.83
Statistics				
alpha		1.222		
N	755	755	755	755
ll	-1427.307	-1098.999	-1195.208	-1064.134
bic	2887.747	2237.759	2436.802	2181.282
aic	2864.614	2209.999	2404.415	2144.268

legend: b/t

Comparing Count Models - Best-fit Model

Therefore, it is safe to deduce that the zero-inflated negative binomial model is the best suited model to be used with this dataset.

Visualizing the Results

There are two ways to look at the results:

- Predict the number of outcomes
- Predict the probability that the event will happen a certain number of times

Visualizing the Results - Predicting the Number of Outcomes

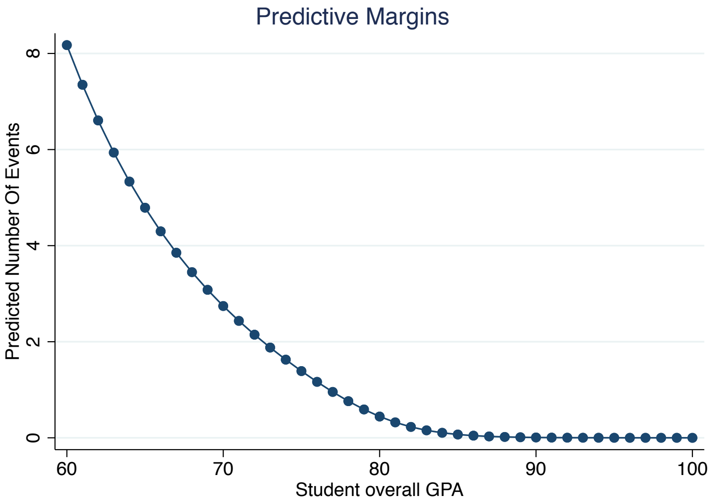

Figure: Using marginsplot to visualize the relationship between the expected number of events and GPA.

Visualizing the Results - Predicting the Number of Outcomes

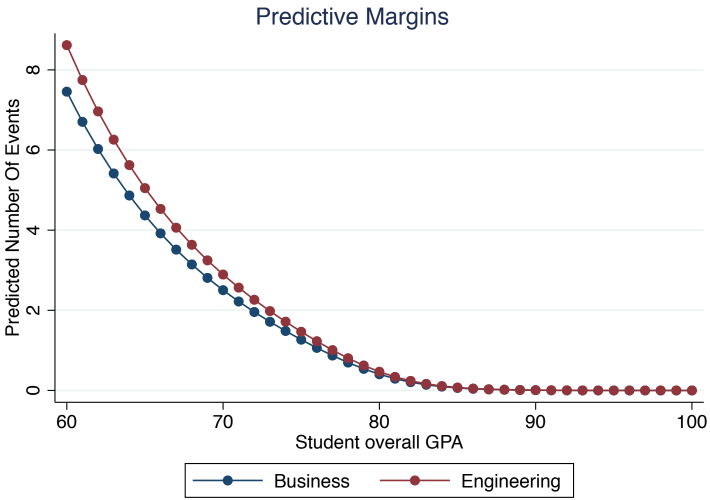

Figure: Visualizing the effect that two variables have on the expected number of outcomes.

Visualizing the Results - Predicting the Probability

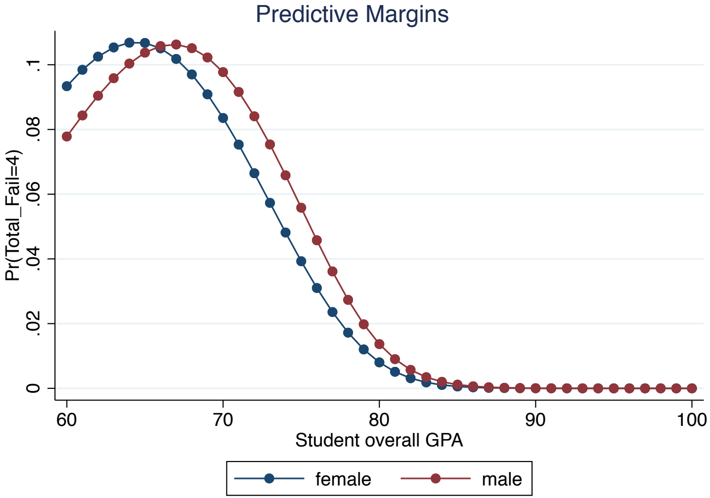

Figure: Plotting the probability that the event will occur exactly four times.