Logistic Regression

Case Study

Najib Mozahem

May 1, 2019

The Dataset

- withdraw: this is the dependent variable which records whether the student withdrew from the course or finished the course (zero means continued, one means withdraw)
- college: whether the student is in the engineering school or the business school (zero means business, one means engineering)
- gender: whether the student is a male or a female (zero means female, one means male)
- gpa: overall GPA of the student
- semester: records whether the course was taken in the spring, fall, or summer semester (zero means fall, one means spring, two means summer)
- level: records whether the level of the course (zero means remedial, one means one-hundred level course, two means two-hundred level course, three means three-hundred level course, four means four-hundred level course, and five means five-hundred level course)

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 logistic regression, but their shape is different from what we are used to due to the nature of the dependent variable. For example, let us produce a scatter plot of the dependent variable withdraw and the continuous independent variable GPA.

Continuous Variables - GPA (Scatter plot)

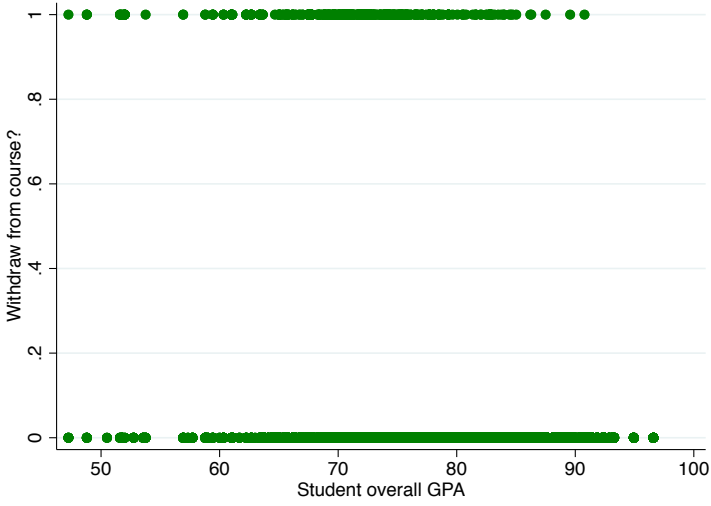

Figure: Scatter plot of withdraw and GPA.

Continuous Variables - GPA (Averages)

-> withdraw = No withdraw

Variable	Obs	Mean	Std. Dev.	Min	Max
gpa	24,656	77.38047	6.622922	47.25	96.59

-> withdraw = Withdraw

Variable	Obs	Mean	Std. Dev.	Min	Max
gpa	504	71.18786	6.447055	47.25	90.78

Continuous Variables - GPA (Regression)

```
Logistic regression                                 Number of obs     =     25,160
                                                    LR chi2(1)        =     425.42
                                                    Prob > chi2       =     0.0000
Log likelihood = -2257.0651                         Pseudo R2         =     0.0861
```

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	.8718886	.0057075	-20.94	0.000	.8607737	.8831471
_cons	550.52	259.6654	13.38	0.000	218.4159	1387.592

Note: _cons estimates baseline odds.

Continuous Variables - GPA (Regression showing coefficients)

Logistic regression

Log likelihood = -2257.0651

Number of obs	=	25,160
LR chi2(1)	=	425.42
Prob > chi2	=	0.0000
Pseudo R2	=	0.0861

withdraw	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	-.1370936	.0065461	-20.94	0.000	-.1499237	-.1242635
_cons	6.310863	.471673	13.38	0.000	5.386401	7.235325

Testing Linearity: the Box-Tidwell Test

Logistic regression — Number of obs = 25,160
LR chi2(2) = 442.63
Prob > chi2 = 0.0000
Log likelihood = -2248.4614 — Pseudo R2 = 0.0896

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	4.819065	2.114356	3.58	0.000	2.039387	11.38744
boxtid	.7209326	.0605389	-3.90	0.000	.6115284	.8499095
_cons	1.20e-07	6.88e-07	-2.78	0.005	1.59e-12	.0090799

Note: _cons estimates baseline odds.

Testing Linearity: Loess Curve

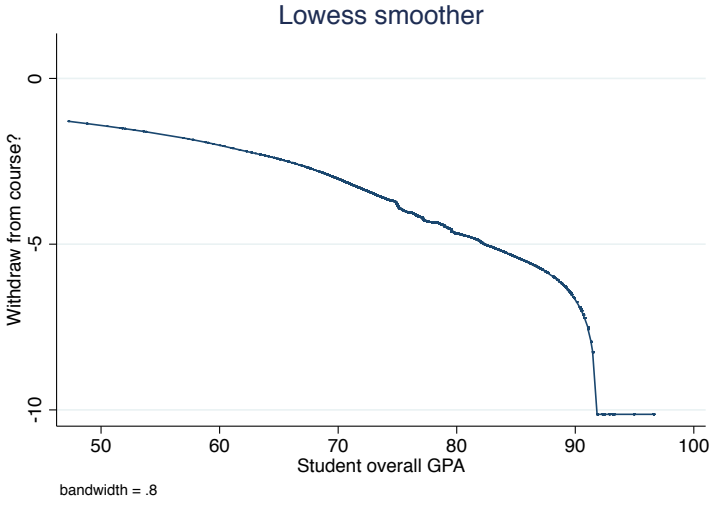

Figure: Loess curve of withdraw and GPA.

Testing Linearity: Linearity of Slopes Test

In order to perform the linearity of slopes test, we need to categorize the continuous variable GPA. In this case, we will generate a new variable, which we named gpacat, by cutting the variable GPA into groups: GPA between 40 and 50 in one group, 50 and 60 in a second group, 60 and 70 in a third group, 70 and 80 in a fourth group, 80 and 90 in a fifth group, and finally 90 and 100 in a sixth group.

Testing Linearity: Linearity of Slopes Test

gpacat	Freq.	Percent	Cum.
40	19	0.08	0.08
50	154	0.61	0.69
60	2,009	7.98	8.67
70	14,912	59.27	67.94
80	7,212	28.66	96.61
90	854	3.39	100.00
Total	25,160	100.00	

Testing Linearity: Linearity of Slopes Test

Logistic regression — Number of obs = 25,160; LR chi2(5) = 321.65; Prob > chi2 = 0.0000; Log likelihood = -2308.9538; Pseudo R2 = 0.0651

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpacat						
50	.5597015	.3423477	-0.95	0.343	.1687756	1.856109
60	.2509294	.1430813	-2.42	0.015	.0820714	.7672047
70	.0814491	.0460666	-4.43	0.000	.0268817	.2467833
80	.0188127	.0110433	-6.77	0.000	.0059537	.0594453
90	.0043962	.0050468	-4.73	0.000	.0004634	.0417097
_cons	.2666667	.1500617	-2.35	0.019	.0885056	.8034643

Note: _cons estimates baseline odds.

Testing Linearity: Linearity of Slopes Test

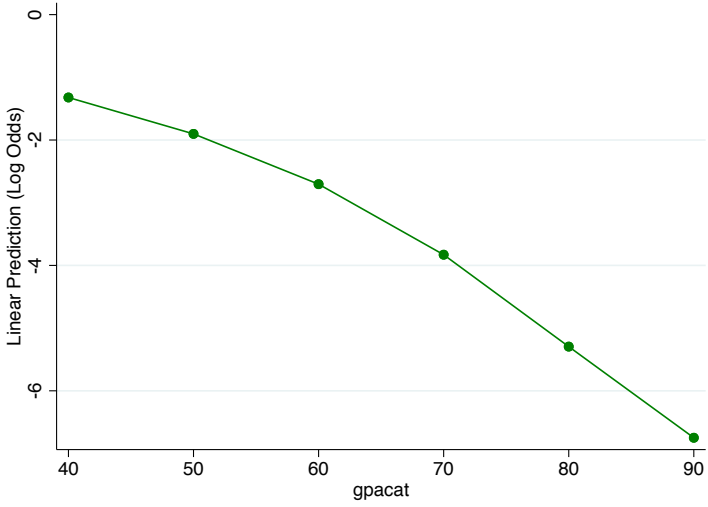

Figure: Testing the linearity of the slopes assumption.

Continuous Variables - Including a Quadratic Term

Logistic regression

Number of obs	= 25,160
LR chi2(2)	= 441.91
Prob > chi2	= 0.0000
Log likelihood = -2248.8207	Pseudo R2 = 0.0895

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	1.206675	.1030634	2.20	0.028	1.020677	1.426568
gpa2	.9976426	.0006178	-3.81	0.000	.9964324	.9988542
_cons	.0086151	.0253454	-1.62	0.106	.000027	2.750891

Note: _cons estimates baseline odds.

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. While graphs are used to investigate the relationship when a continuous variable is involved, contingency tables are used to investigate the relationship when the independent variable is binary.

Binary Variables - Gender (Contingency table)

Key
frequency *row percentage*

Gender	Withdraw from course? No withdr	Withdraw	Total
female	7,957 99.05	76 0.95	8,033 100.00
male	16,699 97.50	428 2.50	17,127 100.00
Total	24,656 98.00	504 2.00	25,160 100.00

Binary Variables - Gender (Logistic regression)

Logistic regression — Number of obs = 25,160
LR chi2(1) = 76.62
Prob > chi2 = 0.0000
Log likelihood = -2431.4659 — Pseudo R2 = 0.0155

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gender						
male	2.683423	.3360166	7.88	0.000	2.099433	3.429857
_cons	.0095513	.0011008	-40.35	0.000	.0076201	.0119721

Note: _cons estimates baseline odds.

Binary Variables - College (Contingency table)

Key
frequency *row percentage*

College	Withdraw from course? No withdr	Withdraw	Total
Business	9,079 97.64	219 2.36	9,298 100.00
Engineering	15,577 98.20	285 1.80	15,862 100.00
Total	24,656 98.00	504 2.00	25,160 100.00

Binary Variables - College (Logistic regression)

Logistic regression

Number of obs	=	25,160	
LR chi2(1)	=	9.13	
Prob > chi2	=	0.0025	
Log likelihood = -2465.2122	Pseudo R2	=	0.0018

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
college						
Engineering	.7584989	.0688913	-3.04	0.002	.6348102	.9062875
_cons	.0241216	.0016495	-54.47	0.000	.0210959	.0275813

Note: _cons estimates baseline odds.

Categorical Variables with More than Two Groups

Our dataset also contains variables that are categorical in nature, but unlike binary variables, these variables contain more than one group.

Categorical Variables - Semester (Contingency table)

Key
frequency *row percentage*

Semester the course was taken	Withdraw from course? No withdr	Withdraw	Total
Fall	10,447 98.20	191 1.80	10,638 100.00
Spring	10,638 97.76	244 2.24	10,882 100.00
Summer	3,571 98.10	69 1.90	3,640 100.00
Total	24,656 98.00	504 2.00	25,160 100.00

Categorical Variables - Semester (Logistic regression)

Logistic regression — Number of obs = 25,160
LR chi2(2) = 5.69
Prob > chi2 = 0.0581
Log likelihood = -2466.931 — Pseudo R2 = 0.0012

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
semester						
Spring	1.25455	.1224308	2.32	0.020	1.036143	1.518995
Summer	1.05686	.1498511	0.39	0.697	.8004358	1.395431
_cons	.0182828	.0013349	-54.81	0.000	.0158449	.0210957

Note: _cons estimates baseline odds.

Categorical Variables - Semester (Collapsing the variable)

If you look at the output in which the fall semester is the base, you will notice that the p-value for the category spring is less than 0.05 while the p-value for the category summer is greater than 0.05. This means that the difference between the spring semester and the fall semester is significant, while the difference between the summer semester and the fall semester is not. Given this result, we might want to consider collapsing the variable semester. Since the odds ratio for summer when compared to fall is not significant, it might be better if we just treated these two as a single group. In other words, we can create a binary variable that takes a value of zero when the semester is fall or summer, and takes a value of one when the semester is spring:

Categorical Variables - Semester (Collapsing the variable)

Semester the course was taken	Spring semester Fall or S	Spring	Total
Fall	10,638	0	10,638
Spring	0	10,882	10,882
Summer	3,640	0	3,640
Total	14,278	10,882	25,160

Categorical Variables - Semester (Logistic regression)

Logistic regression — Number of obs = 25,160
LR chi2(1) = 5.54
Prob > chi2 = 0.0186
Log likelihood = -2467.0064 — Pseudo R2 = 0.0011

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
spring						
Spring	1.236638	.1113651	2.36	0.018	1.036544	1.475358
_cons	.0185476	.0011609	-63.71	0.000	.0164063	.0209683

Note: _cons estimates baseline odds.

Categorical Variables - Level

We now perform the same analysis on the level variable.

Categorical Variables - Level (Contingency table)

Key
frequency *row percentage*

The level of the course	Withdraw from course? No withdr	Withdraw	Total
remedial	450 98.90	5 1.10	455 100.00
100 level course	1,204 98.69	16 1.31	1,220 100.00
200 level course	10,085 96.94	318 3.06	10,403 100.00
300 level course	8,516 98.26	151 1.74	8,667 100.00
400 level course	3,245 99.60	13 0.40	3,258 100.00
500 level course	1,156 99.91	1 0.09	1,157 100.00
Total	24,656 98.00	504 2.00	25,160 100.00

Categorical Variables - Level (Logistic regression)

Logistic regression — Number of obs = 25,160
LR chi2(5) = 161.47
Prob > chi2 = 0.0000
Log likelihood = -2389.0396 — Pseudo R2 = 0.0327

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
level						
100 level course	1.196013	.6163273	0.35	0.728	.4356086	3.283792
200 level course	2.837878	1.286364	2.30	0.021	1.167234	6.899688
300 level course	1.59582	.7294871	1.02	0.307	.6514473	3.909203
400 level course	.3605547	.1906013	-1.93	0.054	.1279374	1.01612
500 level course	.0778559	.085396	-2.33	0.020	.009071	.6682342
_cons	.0111111	.0049966	-10.01	0.000	.0046023	.0268247

Note: _cons estimates baseline odds.

Categorical Variables - Semester (Collapsing the variable)

We see that the result for 200-level courses and 500-level courses is significant, with 200-level courses having odds that are 2.84 times higher than the odds of intensive courses (the base category), and 500-level courses having odds that are 0.08 times the odds of intensive courses. The other categories have p-values that are less than 0.05. Looking at this output, we might deduce that once students have reached the very end of their studies, the probability that they will withdraw from a course decreases significantly since such a decision will probably postpone their graduation. We can also deduce that students who have just enrolled in a major face the largest uncertainty in terms of not being sure whether this is the correct major for them, thus leading to a higher probability of withdrawal. Given that the other categories are not significant, we might choose to collapse this variable as well by creating a new three-group variable that contains the groups 200-level courses, 500-level courses, and the remaining courses.

Categorical Variables - Semester (Collapsing the variable)

The level of the course	RECODE of level (The level of the course)			Total
	Other cou	200 level	500 level	
remedial	455	0	0	455
100 level course	1,220	0	0	1,220
200 level course	0	10,403	0	10,403
300 level course	8,667	0	0	8,667
400 level course	3,258	0	0	3,258
500 level course	0	0	1,157	1,157
Total	13,600	10,403	1,157	25,160

Categorical Variables - Semester (Logistic regression)

Logistic regression

Number of obs	=	25,160	
LR chi2(2)	=	121.49	
Prob > chi2	=	0.0000	
Log likelihood = -2409.0301	Pseudo R2	=	0.0246

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
level3						
200 level courses	2.286495	.213561	8.85	0.000	1.904	2.745828
500 level courses	.062729	.0629272	-2.76	0.006	.0087817	.4480835
_cons	.0137905	.0010209	-57.87	0.000	.011928	.0159438

Note: _cons estimates baseline odds.

Multiple Regression

Logistic regression — Number of obs = 25,160; LR chi2(7) = 539.85; Prob > chi2 = 0.0000; Log likelihood = -2199.8495; Pseudo R2 = 0.1093

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	1.217514	.1043718	2.30	0.022	1.029211	1.440269
gpa2	.9976925	.0006185	-3.73	0.000	.9964811	.9989054
gender						
male	1.528905	.2032302	3.19	0.001	1.178241	1.983931
college						
Engineering	1.011602	.0971431	0.12	0.904	.83805	1.221096
spring						
Spring	1.223419	.111875	2.21	0.027	1.022675	1.463569
level3						
200 level courses	1.979354	.1880548	7.19	0.000	1.643056	2.384485
500 level courses	.0726959	.0730008	-2.61	0.009	.0101564	.520333
_cons	.0016024	.0047537	-2.17	0.030	4.78e-06	.5369355

Note: _cons estimates baseline odds.

Multiple Regression

- Notice that we include the quadratic term of the variable GPA, since we had uncovered that the logit function is not linear with respect to GPA.
- We also include the collapsed versions of the variables semester and level.
- We see that all of the variables are significant except for the variable college. Therefore, it seems like a good idea to remove this variable from the model.

Multiple Regression

Logistic regression — Number of obs = 25,160; LR chi2(6) = 539.84; Prob > chi2 = 0.0000; Log likelihood = -2199.8567; Pseudo R2 = 0.1093

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	1.217242	.1043142	2.29	0.022	1.029038	1.439868
gpa2	.9976954	.0006179	-3.73	0.000	.9964851	.9989072
gender						
male	1.534498	.1985629	3.31	0.001	1.190752	1.977476
spring						
Spring	1.2234	.111873	2.20	0.027	1.022659	1.463545
level3						
200 level courses	1.9789	.1879736	7.19	0.000	1.642741	2.383848
500 level courses	.0729881	.073254	-2.61	0.009	.0102082	.5218603
_cons	.0016093	.0047734	-2.17	0.030	4.81e-06	.538841

Note: _cons estimates baseline odds.

Likelihood Ratio Test

Logistic regression — Number of obs = 25,160
LR chi2(6) = 539.84
Prob > chi2 = 0.0000
Log likelihood = -2199.8567 — Pseudo R2 = 0.1093

withdraw	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	1.217242	.1043142	2.29	0.022	1.029038	1.439868
gpa2	.9976954	.0006179	-3.73	0.000	.9964851	.9989072
gender						
male	1.534498	.1985629	3.31	0.001	1.190752	1.977476
spring						
Spring	1.2234	.111873	2.20	0.027	1.022659	1.463545
level3						
200 level courses	1.9789	.1879736	7.19	0.000	1.642741	2.383848
500 level courses	.0729881	.073254	-2.61	0.009	.0102082	.5218603
_cons	.0016093	.0047734	-2.17	0.030	4.81e-06	.538841

Note: _cons estimates baseline odds.

Hosmer-Lemeshow Test

```
Logistic model for withdraw, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
      number of observations =     25160
            number of groups =        10
     Hosmer-Lemeshow chi2(8) =      6.76
                Prob > chi2 =      0.5623
```

Classification Table

- The classification table allows us to compare the observed outcome with the outcome as predicted by our model.
- Before producing the classification table, we first need to determine the cutoff probability.
- The ideal cutoff value is the point at which the sensitivity and the specificity are equal.
- We can produce a graph of the sensitivity and the specificity in order for us to see where the graphs intersect.

Classification Table - Determing the cutoff value

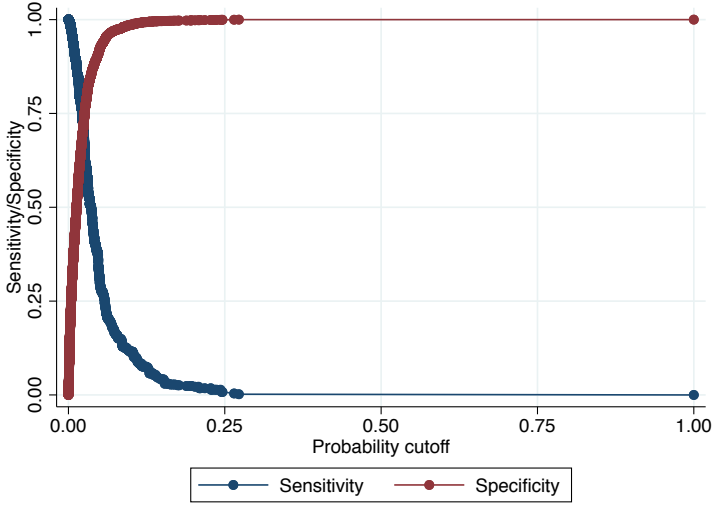

Figure: The sensitivity-specificity curves.

Classification Table - Determing the cutoff value

We see that the sensitivity and the specificity curves intersect at a very small probability. The exact value of the point of intersection is 0.024088. We can now produce the classification table.

Classification Table

Logistic model for withdraw

Classified	True D	True ~D	Total
+	361	6973	7334
-	143	17683	17826
Total	504	24656	25160

Classified + if predicted Pr(D) >= .024088
True D defined as withdraw != 0

Sensitivity	Pr(+| D)	71.63%
Specificity	Pr(-|~D)	71.72%
Positive predictive value	Pr(D| +)	4.92%
Negative predictive value	Pr(~D| -)	99.20%
False + rate for true ~D	Pr(+|~D)	28.28%
False - rate for true D	Pr(-| D)	28.37%
False + rate for classified +	Pr(~D| +)	95.08%
False - rate for classified -	Pr(D| -)	0.80%
Correctly classified		71.72%

ROC Curve

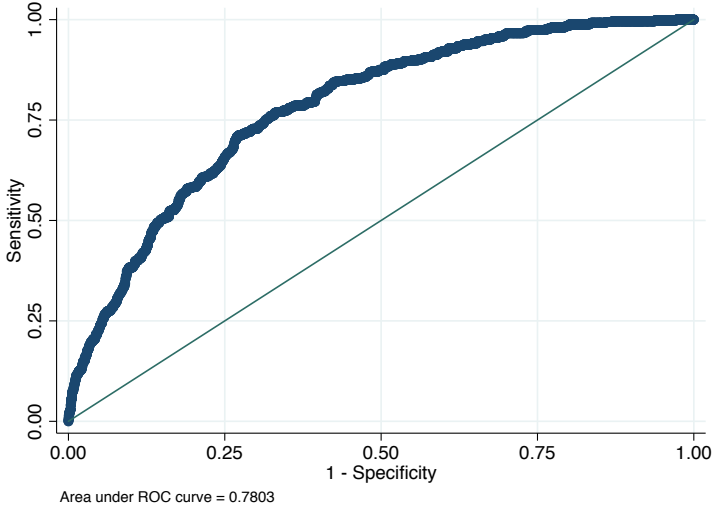

Figure: The ROC curve.

Residual Analysis

When discussing the theory of logistic regression, it was mentioned that the three most commonly used ones are the standardized residuals, the deviance residuals, and the DeltaX residuals. At this point in the analysis, we need to calculate these residuals and produce the appropriate plots.

Standardized Residuals

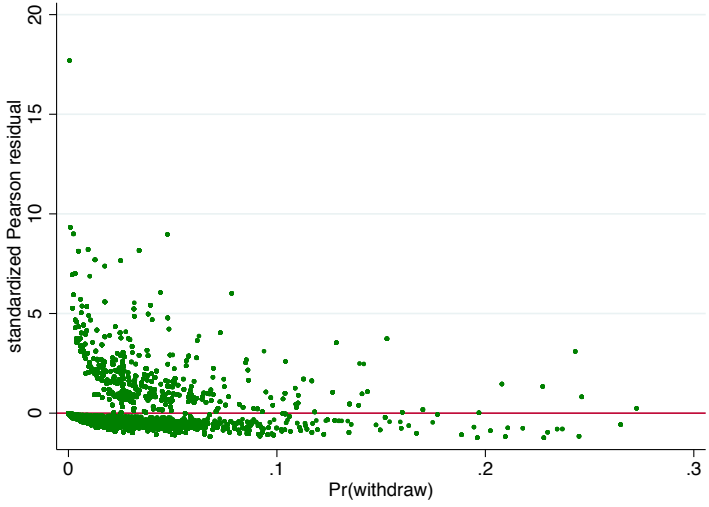

Figure: Plotting the standardized residuals against the predicted probabilities.

Deviance Residuals

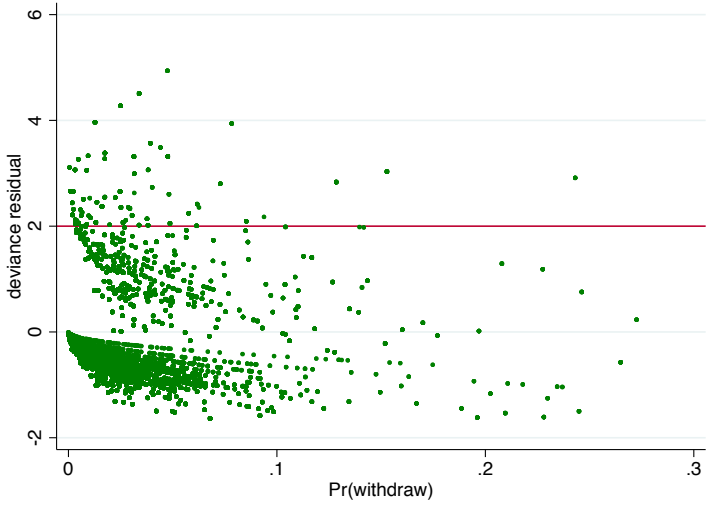

Figure: Plotting the deviance residuals against the predicted probabilities (values above two are considered to be outliers).

DeltaX

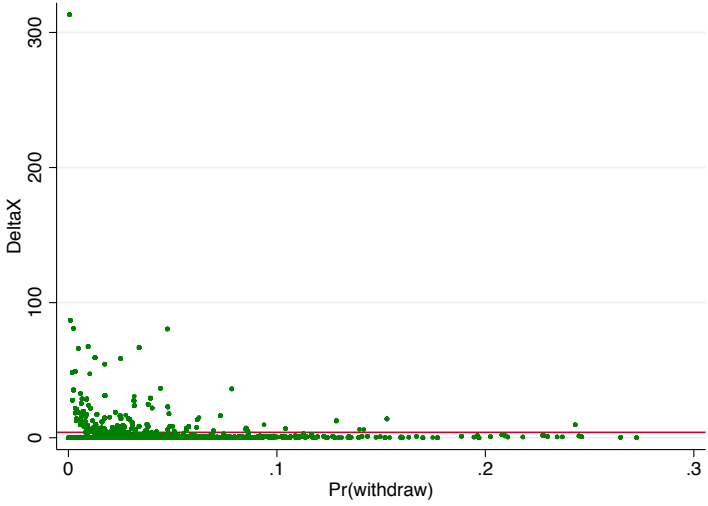

Figure: Plotting the DeltaX residuals against the deviance residuals (values above four are considered to be outliers).

The Hat Diagonal Statistic

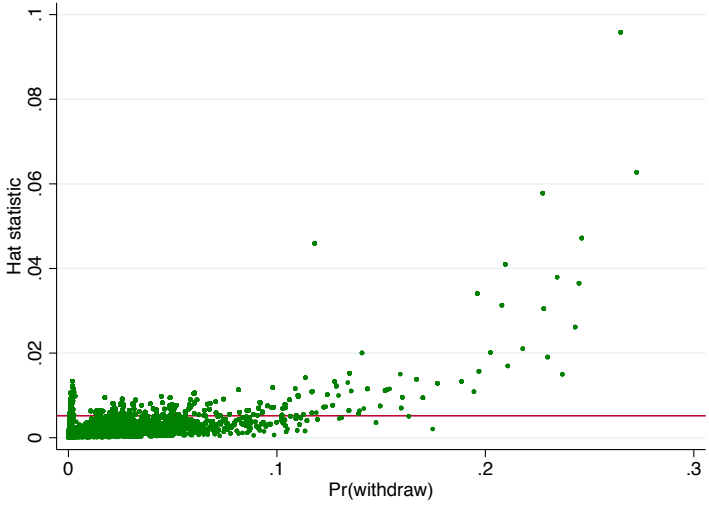

Figure: Plotting the hat statistic residuals against the predicted probabilities (values that are more than two times greater than the average are considered to be influential).

The Hat Diagonal Statistic

We draw a horizontal line at the y equal 0.00519402 point since values that are more than two times greater than the average are considered to be influential (the mean of the variable hat is 0.002597).

Delta-Beta Statistic

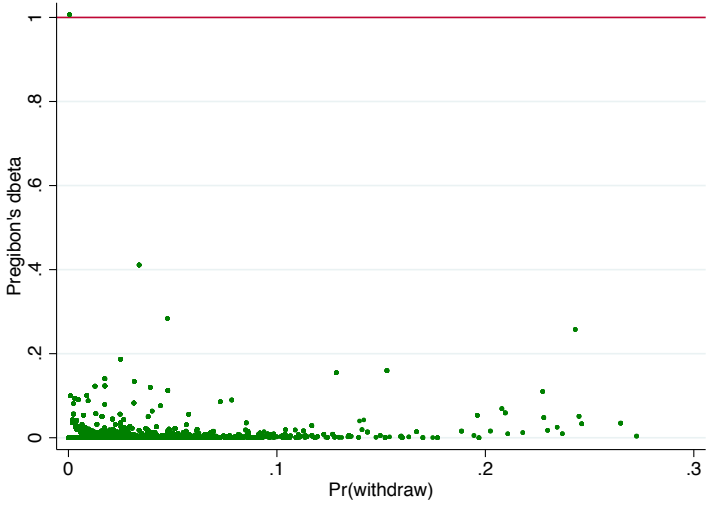

Figure: Plotting the Delta-Beta statistic against the predicted probabilities.

Delta-Beta Statistic

We draw a horizontal line at the y $= 1$ point because values greater than one indicate that the observation is influential.

Combining the Statistics

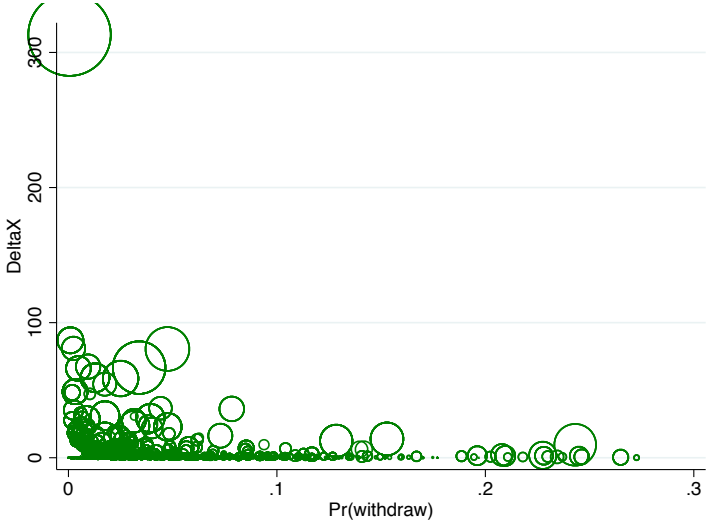

Figure: Plot of DeltaX versus estimated probability weighted by the variable delta-beta.

Combining the Statistics

What we are doing here is that we are producing a scatter plot of the DeltaX residuals against the predicted probabilities. We have actually already done this in the residual section. This time however, the size of the dots are weighted by the value of the delta-beta statistic. Since DeltaX is a residual measure, the larger the value, the worse the fit of the observation, since residuals are a measure of the difference between the observed value and the predicted value. Since delta-measure is a measure of influence, and since we are weighing the dots by this variable, the larger the dot, the more influential it is. What this means is that when we produce this plot, the most problematic points are the large points in the upper left corner. This means that they are influential (hence their large size) and they are not a good fit with the model (high value of the residual which leads to them being near the top).

Taking a closer look at the problematc points

withdraw	gender	gpa	college	spring	level3
No withdraw	male	79.3	Engineering	Fall or Summer	500 level courses
Withdraw	male	79.3	Engineering	Fall or Summer	500 level courses
No withdraw	male	79.3	Engineering	Fall or Summer	500 level courses
No withdraw	male	79.3	Engineering	Fall or Summer	500 level courses
No withdraw	male	79.3	Engineering	Fall or Summer	500 level courses
No withdraw	male	79.3	Engineering	Fall or Summer	500 level courses

Taking a closer look at the problematc points

We see that there are six observations (the noobs option tells Stata not to list the observation numbers). We also notice that they all have a similar pattern: male engineering student with a GPA of 79.3, taking a 500-level course in a semester other than the spring semester. It seems that our model is not doing a good job of predicting the probability for observations with these covariate patterns. What would happen if we fit the model without including these observations?

Comparing the models

Table: Comparing estimates of both models

	(1)	(2)
Withdraw from course?		
Student overall GPA	0.208*	0.200*
gpa2	-0.00244***	-0.00233***
Spring semester	0.200*	0.206*
Other courses	0	0
200 level courses	0.677***	0.682***
500 level courses	-2.577*	0
Gender		0.425**
Constant	-6.241*	-6.525*
Observations	25160	24003

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Visualizing the Result

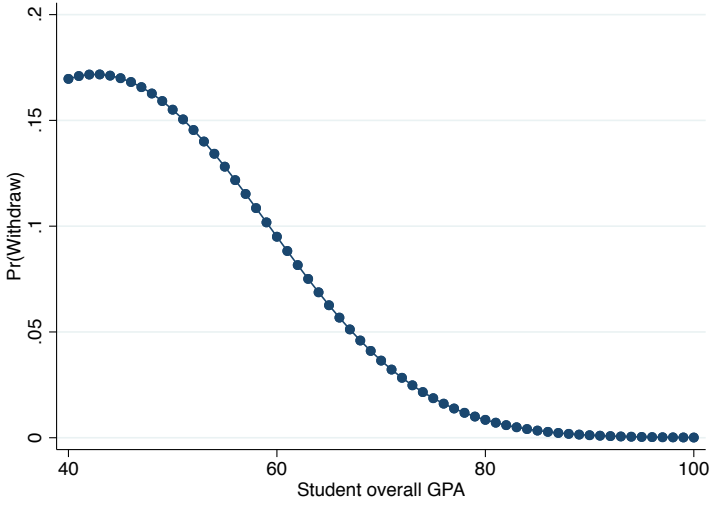

Figure: Graphing the probability of withdrawing from a course for each value of GPA.

Visualizing the Result

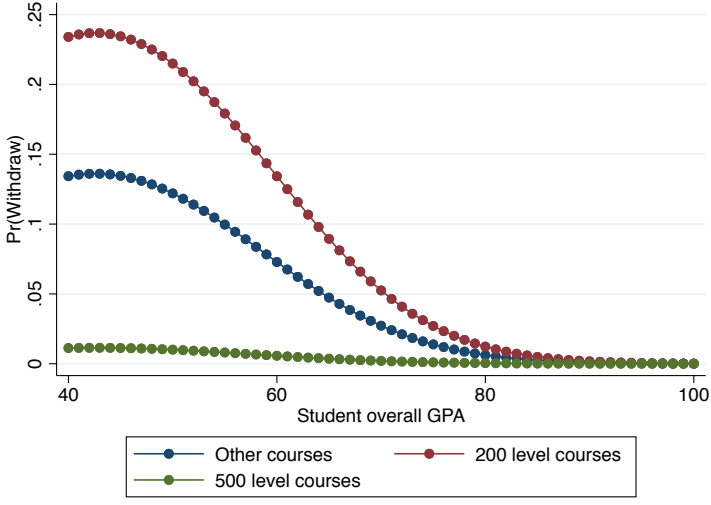

Figure: Graphing the probability of withdrawing from a course for different values of GPA for different level courses.

Visualizing the Result

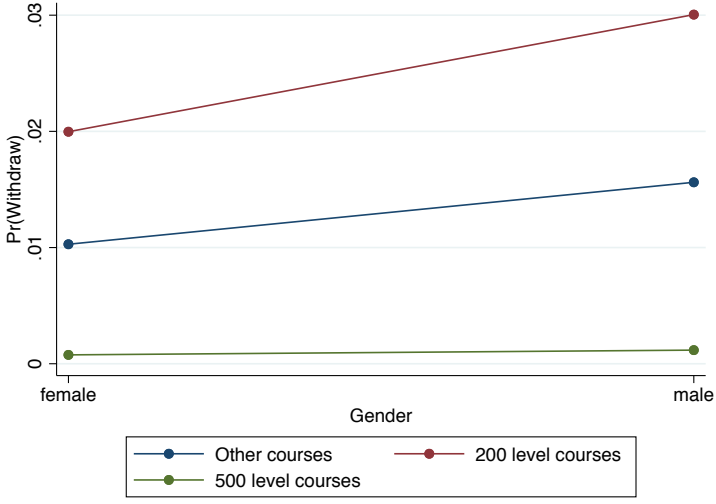

Figure: Graphing the probability of withdrawing from a course for different values of gender for different level courses.