# Logistic Regression, Part I:
# Problems with the Linear Probability Model (LPM)

Richard Williams, University of Notre Dame, https://academicweb.nd.edu/~rwilliam/
Last revised August 24, 2024

This handout steals heavily from Linear probability, logit, and probit models, by John Aldrich and Forrest Nelson, paper # 45 in the Sage series on Quantitative Applications in the Social Sciences.

**INTRODUCTION**. We are often interested in qualitative dependent variables:

- Voting (does or does not vote)
- Marital status (married or not)
- Fertility (have children or not)
- Immigration attitudes (opposes immigration or supports it)

In the next few handouts, we will examine different techniques for analyzing qualitative dependent variables; in particular, dichotomous dependent variables. We will first examine the problems with using OLS, and then present logistic regression as a more desirable alternative.

**OLS AND DICHOTOMOUS DEPENDENT VARIABLES**. While estimates derived from regression analysis may be robust against violations of some assumptions, other assumptions are crucial, and violations of them can lead to unreasonable estimates. Such is often the case when the *dependent* variable is a qualitative measure rather than a continuous, interval measure. If OLS Regression is done with a qualitative dependent variable

- it may seriously misestimate the magnitude of the effects of IVs
- all of the standard statistical inferences (e.g. hypothesis tests, construction of confidence intervals) are unjustified
- regression estimates will be highly sensitive to the range of particular values observed (thus making extrapolations or forecasts beyond the range of the data especially unjustified)

**OLS REGRESSION AND THE LINEAR PROBABILITY MODEL (LPM).** The regression model places no restrictions on the values that the *independent* variables take on. They may be continuous, interval level (net worth of a company), they may be only positive or zero (percent of vote a party received) or they may be dichotomous (dummy) variable (1 = married, 0 = not married).

The dependent variable, however, is assumed to be continuous. Because there are no restrictions on the IVs, the DVs must be free to range in value from negative infinity to positive infinity.

In practice, only a small range of Y values will be observed. Since it is also the case that only a small range of X values will be observed, the assumption of continuous, interval measurement is usually not problematic. That is, even though regression assumes that Y can range from negative

infinity to positive infinity, it usually won't be too much of a disaster if, say, it really only ranges from 1 to 17.

However, it does become a problem when Y can only take on 2 values, say, 0 and 1. If Y can only equal 0 or 1, then

$$E(Yi) = P(Yi = 1).$$

However, in a linear regression, recall that it is also the case that

$$E(Yi) = \alpha + \Sigma\beta_k X_k.$$

Combining these last 2 equations, we get

$$E(Yi) = P(Yi = 1) = \alpha + \Sigma\beta_k X_k.$$

Note: E(Yi) stands for the *Expected Value* of Yi. It is the value we expect Yi to have given its values on the X variables.

From this we conclude that the right hand side of the regression equation must be interpreted as a *probability*, i.e. restricted to between 0 and 1. For example, if the predicted value for a case is .70, this means the case has a 70% chance of having a score of 1. In other words, we would expect that 70% of the people who have this particular combination of values on X would fall into category 1 of the dependent variable, while the other 30% would fall into category 0.

For this reason, a linear regression model with a dependent variable that is either 0 or 1 is called the *Linear Probability Model*, or *LPM*. The LPM predicts the probability of an event occurring, and, like other linear models, says that the effects of X's on the probabilities are linear.

**AN EXAMPLE.** Spector and Mazzeo examined the effect of a teaching method known as PSI on the performance of students in a course, intermediate macro economics. The question was whether students exposed to the method scored higher on exams in the class. They collected data from students in two classes, one in which PSI was used and another in which a traditional teaching method was employed. For each of 32 students, they gathered data on

- GPA — Grade Point Average before taking the class. Observed values range from a low of 2.06 to a high of 4.0 with mean 3.12.
- TUCE (Test of Understanding of College Economics) — the score on an exam given at the beginning of the term to test entering knowledge of the material. In the sample, TUCE ranges from a low of 12 to a high of 29 with a mean of 21.94.
- PSI — a dummy variable indicating the teaching method used (1 = used Psi, 0 = other method). 14 of the 32 sample members (43.75%) are in PSI.
- GRADE — coded 1 if the final grade was an A, 0 if the final grade was a B or C. 11 sample members (34.38%) got As and are coded 1.

GRADE was the dependent variable, and of particular interest was whether PSI had a significant effect on GRADE. TUCE and GPA are included as control variables.

Here are the descriptive statistics and a Stata OLS regression analyses of these data:

```
. use https://academicweb.nd.edu/~rwilliam/statafiles/logist.dta, clear
. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       grade |        32      .34375    .4825587          0          1
         gpa |        32    3.117188    .4667128       2.06          4
        tuce |        32     21.9375    3.901509         12         29
         psi |        32      .4375    .5040161          0          1

. reg  grade gpa tuce i.psi

      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  3,    28) =    6.65
       Model |  3.00227631     3  1.00075877           Prob > F      =  0.0016
    Residual |  4.21647369    28  .150588346           R-squared     =  0.4159
-------------+------------------------------           Adj R-squared =  0.3533
       Total |     7.21875    31  .232862903           Root MSE      =  .38806

------------------------------------------------------------------------------
       grade |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   .4638517   .1619563     2.86   0.008     .1320992    .7956043
        tuce |   .0104951   .0194829     0.54   0.594    -.0294137    .0504039
       1.psi |   .3785548   .1391727     2.72   0.011     .0934724    .6636372
       _cons |  -1.498017   .5238886    -2.86   0.008    -2.571154   -.4248801
------------------------------------------------------------------------------
```

**INTERPRETING PARAMETERS IN THE LPM**. The coefficients can be interpreted as in regression with a continuous dependent variable *except* that they refer to the probability of a grade of A rather than to the level of the grade itself. Specifically, the model states that

$$P(Y=1) = -1.498 + .464 * GPA + .379 * PSI + .010 * TUCE$$

For example, according to these results, a student with a grade point of 3.0, taught by traditional methods, and scoring 20 on the TUCE exam would earn an A with probability of

$$P(Y=1) = -1.498 + .464 * 3 + .379 * 0 + .010 * 20 = .094$$

i.e. this person, who is a little below average on both GPA and TUCE,would have about a 9.4% chance of getting an A.

Or, if you had two otherwise identical individuals, the one taught with the PSI method would have a 37.9% greater chance of getting an A. i.e. if we take the last example and change the value of PSI to 1 but leave other X values the same,

$$P(Y=1) = -1.498 + .464 * 3 + .379 * 1 + .010 * 20 = .473$$

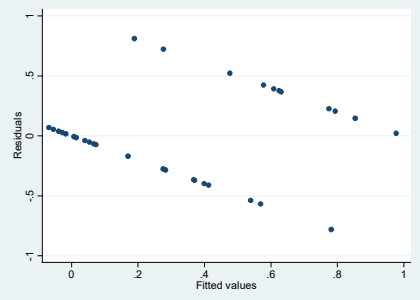
Here are the actual observed values for the data and the predicted probability that Y = 1.

```
. quietly predict yhat
. sort yhat
. list
```

| | grade | gpa | tuce | psi | yhat |
|---|---|---|---|---|---|
| 1. | 0 | 2.63 | 20 | 0 | -.0681847 |
| 2. | 0 | 2.66 | 20 | 0 | -.0542692 |
| 3. | 0 | 2.76 | 17 | 0 | -.0393694 |
| 4. | 0 | 2.74 | 19 | 0 | -.0276562 |
| 5. | 0 | 2.92 | 12 | 0 | -.0176287 |
| 6. | 0 | 2.86 | 17 | 0 | .0070157 |
| 7. | 0 | 2.83 | 19 | 0 | .0140904 |
| 8. | 0 | 2.75 | 25 | 0 | .039953 |
| 9. | 0 | 2.87 | 21 | 0 | .0536347 |
| 10. | 0 | 2.06 | 22 | 1 | .0669647 |
| 11. | 0 | 2.89 | 22 | 0 | .073407 |
| 12. | 0 | 3.03 | 25 | 0 | .1698315 |
| 13. | 1 | 2.39 | 19 | 1 | .1885505 |
| 14. | 0 | 3.28 | 24 | 0 | .2752993 |
| 15. | 1 | 3.26 | 25 | 0 | .2765174 |
| 16. | 0 | 3.32 | 23 | 0 | .2833582 |
| 17. | 0 | 2.89 | 14 | 1 | .3680008 |
| 18. | 0 | 2.67 | 24 | 1 | .3709046 |
| 19. | 0 | 3.57 | 23 | 0 | .3993212 |
| 20. | 0 | 3.53 | 26 | 0 | .4122525 |
| 21. | 1 | 2.83 | 27 | 1 | .4766062 |
| 22. | 0 | 3.1 | 21 | 1 | .5388754 |
| 23. | 0 | 3.12 | 23 | 1 | .5691426 |
| 24. | 1 | 4 | 21 | 0 | .5777872 |
| 25. | 1 | 3.16 | 25 | 1 | .608687 |
| 26. | 1 | 3.92 | 29 | 0 | .6246401 |
| 27. | 1 | 3.39 | 17 | 1 | .631412 |
| 28. | 1 | 3.54 | 24 | 1 | .7744555 |
| 29. | 0 | 3.51 | 26 | 1 | .7815303 |
| 30. | 1 | 3.65 | 21 | 1 | .7939939 |
| 31. | 1 | 3.62 | 28 | 1 | .8535441 |
| 32. | 1 | 4 | 23 | 1 | .9773322 |

Here is what the scatterplot of the predicted values by the residual values looks like:

```
. rvfplot
```



*Optional*. Why does the plot of residuals versus fitted values (i.e. yhat versus e) look the way it does? Recall that e = y – yhat, i.e. the observed value minus the predicted value. Ergo, when y is a 0-1 dichotomy, it must be the case that either

$$e = 0 - \text{yhat (which occurs when } y = 0)$$

or

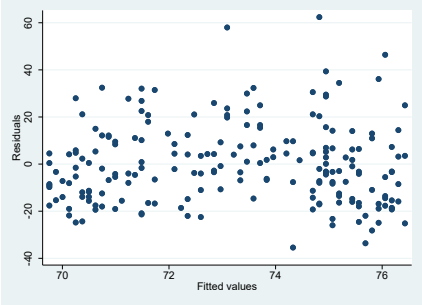
$$e = 1 - \text{yhat (which occurs when } y = 1).$$

These are equations for 2 parallel lines, which is what you see reflected in the residuals versus fitted plot.

This is in sharp contrast to the case when y is continuous and can take on an infinite number of values (or at least a lot more than two).

The above results suggest several potential problems with OLS regression using a binary dependent variable:

**VIOLATION I: HETEROSKEDASTICITY**. A residuals versus fitted plot in OLS ideally looks like a random scatter of points, e.g. something like this:




Clearly, the plot from our LPM regression does not look like this. This suggests that heteroskedasticity may be a problem, and this can be formally proven.

*Optional Proof 1:* Recall that one of the assumptions of OLS is that $V(\varepsilon_i) = \sigma^2_\varepsilon$, i.e. all disturbances have the same variance; there is just as much "error" when Y is large as when Y is small or somewhere in-between. This assumption is violated in the case of a dichotomous dependent variable. The variance of a dichotomy is pq, where p = the probability of the event occurring and q is the probability of it not occurring. Unless p is the same for all individuals, the variances will not be the same across cases. Hence, the assumption of homoscedasticity is violated. As a result, standard errors will be wrong, and hypothesis tests will be incorrect.

*Optional Proof 2:* If $X_i$ is coded 0/1, then $X_i^2 = X_i$. Thus, $V(X_i) = E(X_i^2) - E(X_i)^2 = p - p^2 = p(1 - p) = pq$.

*Optional Proof 3*: See what happens when we test for heteroskedasticity after using OLS with a binary dependent variable.

```
. webuse nhanes2f, clear
. quietly reg diabetes black female age
. estat hettest

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of diabetes

H0: Constant variance

    chi2(1) = 3099.12
Prob > chi2 =  0.0000
```

This test clearly demonstrates that the OLS assumption of homoskedasticity is severely violated in this case. Other possible tests show the same thing. (If you want to know more about heteroskedasticity and what to do about it, see https://academicweb.nd.edu/~rwilliam/stats2/l25.pdf.)

NOTE: Weighted Least Squares (WLS) is sometimes used to deal with problems of heteroskedasticity. The optional Appendix to this handout explains why that isn't a good idea in this case.

**VIOLATION II: ERRORS ARE NOT NORMALLY DISTRIBUTED**. Also, OLS assumes that, for each set of values for the k independent variables, the residuals are normally distributed. This is equivalent to saying that, for any given value of yhat, the residuals should be normally distributed. As the above rvfplot shows, this assumption is also clearly violated, i.e. you can't have a normal distribution when the residuals are only free to take on two possible values.

**VIOLATION III: LINEARITY**. These first two problems suggest that the estimated standard errors will be wrong when using OLS with a dichotomous dependent variable. However, the predicted values also suggest that there may be problems with the plausibility of the model and/or its coefficient estimates. As noted before, yhat can be interpreted as the estimated probability of success. Probabilities can only range between 0 and 1. However, in OLS, there is no constraint that the yhat estimates fall in the 0-1 range; indeed, yhat is free to vary between negative infinity and positive infinity. In this particular example, the yhat values include negative numbers (implying probabilities of success that are less than zero). In other examples there could be predicted values greater than 1 (implying that success is more than certain).

This problem, in and of itself, would not be too serious, at least if there were not too many out of range predicted values. However, this points to a much bigger problem: the OLS assumptions of linearity and additivity are almost certainly unreasonable when dealing with a dichotomous dependent variable.

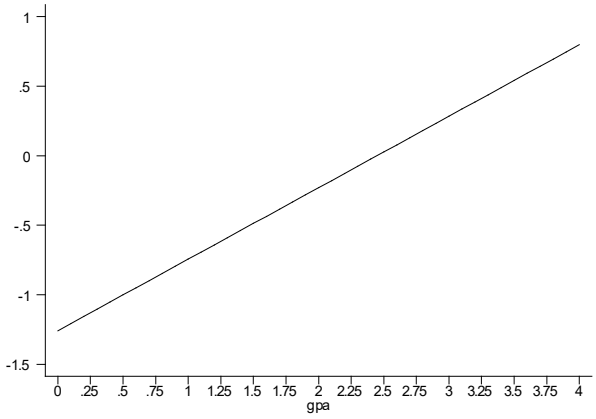
This is most easily demonstrated in the case of a bivariate regression. If you simply regress GRADE ON GPA, you get the following:

```
. reg grade gpa

      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  1,    30) =    9.85
       Model |  1.78415459     1  1.78415459           Prob > F      =  0.0038
    Residual |  5.43459541    30   .18115318           R-squared     =  0.2472
-------------+------------------------------           Adj R-squared =  0.2221
       Total |     7.21875    31  .232862903           Root MSE      =  .42562

------------------------------------------------------------------------------
       grade |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   .5140267   .1637919     3.14   0.004      .179519    .8485343
       _cons |  -1.258568   .5160841    -2.44   0.021    -2.312552   -.2045832
------------------------------------------------------------------------------
```

Here is what the plot looks like. The Y axis is the predicted probability of an A:



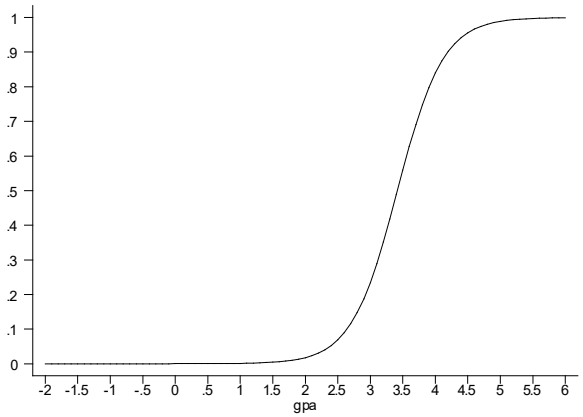

As you see, a linear regression predicts that those with GPAs of about 2.25 or below have a negative probability of getting an A. In reality, their chances may not be good, but they can't be that bad! Further, if the effect of GPA had been just a little bit stronger, the best students would have been predicted to have better than a 100% chance of getting an A. They may be good, but they can't be that good.

Even if predicted probabilities did not take on impossible values, the assumption that there will be a straight linear relationship between the IV and the DV is also very questionable when the DV is a dichotomy. If one student already has a great chance of getting an A, how much higher can the chances go for a student with an even better GPA? And, if a C student has very little chance for an A, how much worse can the chances for a D student be?

Another example will help to illustrate this. Suppose that the DV is home ownership, and one of the IVs is income. According to the LPM, an increase in wealth of $50,000 will have the same effect on ownership regardless of whether the family starts with 0 wealth or wealth of $1 million. Certainly, a family with $50,000 is more likely to own a home than one with $0. But, a millionaire is very likely to own a home, and the addition of $50,000 is not going to increase the likelihood of home ownership very much.

This explains why we get out of range values. If somebody has a 50% chance of home ownership, then an additional $10,000 could increase their chances to 60%. But, if somebody already has a 98% chance of home ownership, their probability of success can't increase to 108%. Yet, there is nothing that keeps OLS predicted values within the 0-1 range.

A more intuitive specification would express P(Yi = 1) as a *nonlinear* function of Xi, one which approaches 0 at slower and slower rates as Xi gets small and approaches 1 at slower and slower rates as Xi gets larger and larger. Such a specification has an S-Shaped curve:




As I'll explain more later, I created this graph by doing a bivariate *logistic* regression of GRADE on GPA. The Y axis is the predicted probability of getting an A. Note that

- The probabilities never go lower than 0 or above 1 (even though I've included values of GPA that are impossibly small and large)

- A person with a 0.0 GPA has only a slightly worse chance of getting an A than a person with a 2.0 GPA. The C students may be a lot better than the F students, but they're still not very likely to get an A.

- However, people with 4.0 GPAs are far more likely to get As than people with a 2.0 GPA. Indeed, inbetween 2.0 and 4.0, you see that increases in GPA produce steady increases in the probability of getting an A.

- After a while though, increases in GPA produce very little change in the probability of getting an A. After a certain point, a higher GPA can't do much to increase already very good chances.

- In short, very weak students are not that much less likely to get an A than are weak students. Terrific students are not that much more likely to get an A than are very good students. It is primarily in the middle of the GPA range you see that the better the past grades, the more likely the student is to get an A.

**SUMMARY: THE EFFECT OF AN INCORRECT LINEARITY ASSUMPTION**. Suppose that the true relationship between Y and X, or more correctly, between the expected value of Y and X, is nonlinear, but in our ignorance of the "true" relationship we adopt the LPM as an approximation. What happens?

- The OLS and WLS estimates will tend to give the correct sign of the effect of X on Y
- But, none of the distributional properties holds, so statistical inferences will have no statistical justification
- Estimates will be highly sensitive to the range of data observed in the sample. Extrapolations outside the range will generally not be valid.
    - For example, if your sample is fairly wealthy, you'll likely conclude that income has very little effect on the probability of buying a house (because, once you reach a certain income level, it is extremely likely that you'll own a house and more income won't have much effect).
    - Conversely, if your sample is more middle income, income will probably have a fairly large effect on home ownership.
    - Finally, if the sample is very poor, income may have very little effect, because you need some minimum income to have a house.
- The usual steps for improving quality of OLS estimates may in fact have adverse effects when there is a qualitative DV. For example, *if* the LPM was correct, a WLS correction would be desirable. But, the assumptions of linearity generally do *not* hold. Hence, correcting for heteroscedasticity, which is desirable when the model specification is correct, actually makes things worse when the model specification is incorrect. See the appendix for details.

In short, the incorrect assumption of linearity will lead to least squares estimates which

- have no known distributional properties
- are sensitive to the range of the data
- may grossly understate (or overstate) the magnitude of the true effects
- systematically yield probability predictions outside the range of 0 to 1
- get worse as standard statistical practices for improving the estimates are employed.

Hence, a specification that does not assume linearity is usually called for. Further, in this case, there is no way to simply "fix" OLS. Alternative estimation techniques are called for.

## *APPENDIX (Optional): Goldberger's WLS procedure*

This appendix is adapted from Section 1.2 of Linear probability, logit, and probit models, by John Aldrich and Forrest Nelson, paper # 45 in the Sage series on Quantitative Applications in the Social Sciences.

With continuous dependent variables, heteroscedasticity can sometimes be dealt with through the use of weighted least squares. That is, through appropriate weighting schemes, errors can be made homoskedastic. See my handout on heteroskedasticity at https://academicweb.nd.edu/~rwilliam/stats2/l25.pdf.

Goldberger (1964) laid out the 2 step procedure when the DV is dichotomous. [WARNING IN ADVANCE: YOU'LL PROBABLY NEVER WANT TO ACTUALLY USE THIS PROCEDURE, BECAUSE UNLESS CERTAIN ASSUMPTIONS ARE MET IT WILL ACTUALLY MAKE THINGS WORSE. BUT IT DOES HELP TO ILLUSTRATE SOME KEY IDEAS]

1. Run the usual OLS regression of Yi on the X's. From these estimates, construct the following weights:

$$w_i = \sqrt{\frac{1}{\hat{Y}_i * (1 - \hat{Y}_i)}}$$

2. Use these weights and again regress Yi on the X's.

*Assuming that other OLS assumptions are met,* the betas produced by this procedure are unbiased and have the smallest possible sampling variance. The standard errors of the beta estimates are the correct ones for hypothesis tests. Unfortunately, as explained above, it is highly unlikely other OLS assumptions are met.

Here is an SPSS program that illustrates Goldberger's approach. First, the data are read in.

```
data list free / gpa tuce psi grade.
begin data.
  2.66  20.00   .00   .00
  2.89  22.00   .00   .00
 [Rest deleted; but data are shown below]
end data.
```

Then, the OLS regression (Stage 1 in Goldberger's procedure) is run. The predicted values for each case are saved as a new variable called YHAT. (This saves us the trouble of having to write compute statements based on the regression output).

```
* Regular OLS. Save predicted values to get weights.
REGRESSION
 /DEPENDENT grade
 /METHOD=ENTER gpa psi tuce
 /SAVE PRED (yhat) /Scatterplot=(*resid *pred).
```

Also, recall that probabilities should only range between 0 and 1. However, there is no requirement that the predicted values from OLS will range between 0 and 1. Indeed, it works out

that 5 of the predicted values are negative. Hence, in Goldberger's procedure, "out of range" values (values less than 0 or greater than 1) are recoded to legitimate values. The weighting variable is then computed.

```
* recode out of range values.
recode yhat (lo thru .001 = .001) (.999 thru hi = .999).
* Compute weights.
COMPUTE WGT = 1/(yhat * (1 - yhat)).
```

Finally, the weighted least squares regression is run. This is stage 2 of Goldberger's procedure.

```
REGRESSION
 /REGWGT=wgt
 /DEPENDENT grade
 /METHOD=ENTER gpa psi tuce .
```

Following are the results.

**Regular OLS**

**Model Summary[b]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .645[a] | .416 | .353 | .38806 |

a. Predictors: (Constant), TUCE, PSI, GPA

b. Dependent Variable: GRADE

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | -1.498 | .524 | | -2.859 | .008 |
| | GPA | .464 | .162 | .449 | 2.864 | .008 |
| | PSI | .379 | .139 | .395 | 2.720 | .011 |
| | TUCE | .010 | .019 | .085 | .539 | .594 |

a. Dependent Variable: GRADE

**Weighted Least Squares**

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .873[a] | .762 | .737 | .90117 |

a. Predictors: (Constant), TUCE, PSI, GPA

**Coefficients[a,b]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | -1.309 | .288 | | -4.536 | .000 |
| | GPA | .398 | .088 | .540 | 4.533 | .000 |
| | PSI | .388 | .105 | .434 | 3.687 | .001 |
| | TUCE | .012 | .005 | .287 | 2.676 | .012 |

a. Dependent Variable: GRADE

b. Weighted Least Squares Regression - Weighted by WGT

Or, in Stata we can do

```
* APPENDIX (Optional): Goldberger's WLS procedure
use https://academicweb.nd.edu/~rwilliam/statafiles/logist, clear
reg grade gpa psi tuce, cformat(%9.4f) sformat(%8.3f)
est store lpm
predict yhat
recode yhat (min/.001 = .001) (.999 / max = .999)
gen wgt = 1/(yhat * (1 - yhat))
* In Stata, use aweights
reg grade gpa psi tuce [aweight=wgt], cformat(%9.4f) sformat(%8.3f)
est store goldberger
* esttab from SSC must be installed
esttab lpm goldberger, mtitles(LPM Goldberger) nonumbers b(4) t(3) r2 ar2
```

```
. esttab lpm goldberger, mtitles(LPM Goldberger) nonumbers b(4) t(3) r2 ar2

--------------------------------------------
                      LPM      Goldberger
--------------------------------------------
gpa                0.4639**        0.3982***
                  (2.864)         (4.533)

psi                0.3786*         0.3878***
                  (2.720)         (3.687)

tuce               0.0105          0.0122*
                  (0.539)         (2.676)

_cons             -1.4980**       -1.3087***
                 (-2.859)        (-4.536)
--------------------------------------------
N                     32              32
R-sq               0.416           0.762
adj. R-sq          0.353           0.737
--------------------------------------------
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

Comparing the LPM and the Goldberger approaches, we see that the beta estimates change, but not by much. However, the standard errors and T values change considerably. TUCE does not have a significant effect in the OLS estimation, but does when using WLS. The effect of PSI, which we are most interested in, is more significant under WLS than it is with OLS.

Unfortunately, the usual steps for improving quality of OLS estimates may in fact have adverse effects when there is a qualitative DV. For example, *if* the LPM was correct, the WLS correction used above would be desirable. However, note that the weights are largest when P is near 0 or 1, and smallest when p is near .5. This means that observations at the extremes will receive more weight than those near the center. If the true relationship is nonlinear, those extreme values will be the worst fit, exacerbating problems rather than alleviating them. Hence, correcting for heteroscedasticity, which is desirable when the model specification is correct, actually makes things worse when the model specification is incorrect.