Sean Howard
Econometrics
Final Project Paper

An Analysis of the Determinants and Factors of Physical Education Attendance in the Fourth Quarter

Introduction

This project attempted to gain a more complete understanding of the internal and external factors that drive or deter 4^{th} quarter attendance in Physical Education classes.

Data

The data was obtained from the sports center, in the form of attendance sheets and class rosters from the spring of 2004-2009 that contained information on the students themselves such as sex, status towards graduation and their major as well as gleaned class characteristics such as class size, time commitment, time of meeting, gender percentage, intensity of the activity in the class. Weather data was also obtained from the National Climatic Data Center, specifically the measuring station used was the Portland International Airport.

Variable	Obs	Mean	Std. Dev.	Min	Max
timer	837	4.373955	1.427663	1	6
attendrate	837	.5997544	.3508946	0	1
intensity	837	3.188769	.7467429	1	4
timecom	837	109.1398	24.6502	80	220
bf4	837	.3225806	.4677433	0	1
gender	837	.3357228	.4725249	0	1
status	837	2.139785	1.067487	1	4
classsize	837	14.67025	8.596134	1	41
cornc	837	.5746714	.4946883	0	1
genratio	837	.3369295	.2968043	0	1
isUnd	837	.1911589	.393449	0	1
prcp	837	23.20704	6.879986	12.65	41.55
avghigh	837	149.8975	13.94548	131.6098	187.65
Major	0				
ClassID	0				
ltc	837	4.672691	.1890199	4.382027	5.393628

Challenges

Though all the classes in the previously mentioned time frame contain information about whether people received credit, around 40% of the classes don't have information concerning the students themselves or the days they attended. I concluded that this would not be debilitating hindrance in attempting to answer the question because there is enough data to draw conclusion on a random eclectic mix of classes.

Analysis

OLS Explorations

The large t-statistic of the credit/no credit variable is certainly a red flag concerning it's admittedly special relationship with the dependent variable.

```
reg attendrate timecom intensity gender status prcp avghigh cornc

      Source |       SS       df       MS              Number of obs =     837
-------------+------------------------------           F(  7,   829) =  300.80
       Model |  73.8559378     7  10.5508483           Prob > F      =  0.0000
    Residual |  29.0782567   829  .035076305           R-squared     =  0.7175
-------------+------------------------------           Adj R-squared =  0.7151
       Total |  102.934195   836  .123127027           Root MSE      =  .18729

------------------------------------------------------------------------------
  attendrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     timecom |  -.0008133    .000291    -2.79   0.005    -.0013844   -.0002421
   intensity |  -.0146804   .0097521    -1.51   0.133    -.0338221    .0044613
      gender |   .0196603   .0140024     1.40   0.161     -.007824    .0471445
      status |   -.012279   .0061144    -2.01   0.045    -.0242806   -.0002774
        prcp |  -.0001472   .0009601    -0.15   0.878    -.0020317    .0017373
     avghigh |   .0013093   .0004772     2.74   0.006     .0003725     .002246
       cornc |   .5817143   .0134765    43.17   0.000     .5552623    .6081664
       _cons |   .2278634   .0873895     2.61   0.009     .0563327    .3993942
```

If one chooses to omit it one finds that similar variables stay statistically significant but the R^2 value becomes very small indicating poor explanatory power

```
reg attendrate intensity timecom bf4 gender status classsize genratio isUnd
prcp avghigh

      Source |       SS       df       MS              Number of obs =     837
-------------+------------------------------           F( 10,   826) =    7.90
       Model |  8.98703207    10  .898703207           Prob > F      =  0.0000
    Residual |  93.9471625   826  .113737485           R-squared     =  0.0873
-------------+------------------------------           Adj R-squared =  0.0763
       Total |  102.934195   836  .123127027           Root MSE      =  .33725

------------------------------------------------------------------------------
  attendrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   intensity |  -.0172774   .0193461    -0.89   0.372    -.0552507    .0206958
     timecom |   -.001294   .0005447    -2.38   0.018     -.002363   -.0002249
         bf4 |  -.0098879   .0276266    -0.36   0.720    -.0641146    .0443387
      gender |   .0434626   .0316177     1.37   0.170    -.0185978     .105523
      status |  -.0238113   .0118972    -2.00   0.046    -.0471636   -.0004591
   classsize |  -.0025306   .0017502    -1.45   0.149     -.005966    .0009049
    genratio |   .0487809   .0551416     0.88   0.377    -.0594533    .1570151
       isUnd |   .0019464   .0316236     0.06   0.951    -.0601256    .0640184
        prcp |   .0033157   .0017742     1.87   0.062    -.0001668    .0067982
```

```
     avghigh |   .0046773   .0008791     5.32   0.000     .0029518    .0064027
       _cons |   .0778785   .1738487     0.45   0.654    -.2633587    .4191157
------------------------------------------------------------------------------
```

However, if one also notices the significance of the constant term, it also appears dubious. Dropping the constant term yields a much more attractive R^2 value.

```
reg attendrate intensity timecom bf4 gender status classsize genratio isUnd pr
> cp avghigh, noconstant

      Source |       SS       df       MS              Number of obs =     837
-------------+------------------------------           F( 10,   827) =  272.85
       Model |   310.03762    10   31.003762           Prob > F      =  0.0000
    Residual |  93.9699867   827  .113627553           R-squared     =  0.7674
-------------+------------------------------           Adj R-squared =  0.7646
       Total |  404.007606   837  .482685312           Root MSE      =  .33709

------------------------------------------------------------------------------
  attendrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   intensity |  -.0162248   .0191935    -0.85   0.398    -.0538985     .021449
     timecom |  -.0012203    .000519    -2.35   0.019    -.0022389   -.0002016
         bf4 |  -.0107228   .0275504    -0.39   0.697    -.0647997    .0433541
      gender |   .0434083   .0316022     1.37   0.170    -.0186215    .1054382
      status |  -.0229114   .0117207    -1.95   0.051    -.0459172    .0000944
   classsize |  -.0022384   .0016234    -1.38   0.168    -.0054249     .000948
    genratio |     .05517   .0532394     1.04   0.300    -.0493303    .1596703
       isUnd |   .0037185     .03136     0.12   0.906    -.0578361    .0652731
        prcp |   .0036098   .0016475     2.19   0.029     .0003761    .0068436
     avghigh |   .0050167   .0004453    11.27   0.000     .0041428    .0058907
------------------------------------------------------------------------------
```

If one narrows the model specification search further we can see that we are at a bit of an impasse.

```
reg attendrate timecom status prcp avghigh, noconstant

      Source |       SS       df       MS              Number of obs =     837
-------------+------------------------------           F(  4,   833) =  667.88
       Model |  307.977667     4  76.9944167           Prob > F      =  0.0000
    Residual |  96.0299396   833   .11528204           R-squared     =  0.7623
-------------+------------------------------           Adj R-squared =  0.7612
       Total |  404.007606   837  .482685312           Root MSE      =  .33953

------------------------------------------------------------------------------
  attendrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     timecom |  -.0016631   .0004391    -3.79   0.000     -.002525   -.0008012
      status |  -.0262258   .0108604    -2.41   0.016    -.0475428   -.0049088
        prcp |   .0041404   .0016375     2.53   0.012     .0009263    .0073544
     avghigh |   .0049462   .0003789    13.05   0.000     .0042024    .0056899
------------------------------------------------------------------------------

.
.
. estimates store sub


. reg attendrate timecom status avghigh cornc
```

```
      Source |       SS       df       MS              Number of obs =     837
-------------+------------------------------           F(  4,   832) =  523.79
       Model |  73.6766459     4  18.4191615           Prob > F      =  0.0000
    Residual |  29.2575486   832  .035165323           R-squared     =  0.7158
-------------+------------------------------           Adj R-squared =  0.7144
       Total |  102.934195   836  .123127027           Root MSE      =  .18752

------------------------------------------------------------------------------
  attendrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     timecom |  -.0010106   .0002651    -3.81   0.000    -.0015309   -.0004903
      status |  -.0128664   .0060834    -2.11   0.035     -.024807   -.0009257
     avghigh |   .0013095    .000475     2.76   0.006     .0003771    .0022419
       cornc |   .5847748   .0133512    43.80   0.000     .5585689    .6109807
       _cons |   .2052378   .0802179     2.56   0.011     .0477846     .362691
------------------------------------------------------------------------------

. estimates store full

. estimates stats full sub

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
        full |    837   -310.5868    215.8662      5    -421.7325   -398.0833
         sub |    837           .   -281.5304      4     571.0608    589.9801
-----------------------------------------------------------------------------
```

Using only the statistically significant variables in each case leads to similar R^2 values if we look at the information criteria it appears that we'd have to make a choice between poor fit evidenced by negative Akiake & Bayesian values with the credit variable or lots of complexity evidenced by the large values of the two in the model with the omitted credit variable and dropped constant. Thankfully we are not forced to just rely on OLS.

A Problematic Variable

We are bound to encounter interesting phenomena with the credit/no credit variable. In practice instructors decide whether to award credit based on their degree of attendance which means that the dependent variable exogenously determines the credit or no credit variable. However, students can possibly know within three classes that they will not receive credit and still choose to come for reasons such as enjoyment or necessity or what have you. This may make the credit variable correlated with the error term.

Application of Two Stage Least Squares

After performing a Hausman test of endogeneity of cornc, it has been shown that there is collinearity between vhat, the error of the first regression with cornc being the dependent variable and the variable attendance rate. Dropping the cornc term from the second equation shoes that vhat is markedly statistically significant from zero.

It would probably not all that wise to use the two stage least squares at this time because any instrument we'd potentially use would need to not be correlated with the y of our scenario and given the overwhelming collinearity that cornc has on the attendance rate, such an outcome is unlikely

Probit & Logit

In applying the probit and logit model to the data, it became increasingly clear that it would be more appropriate to use a probit model to better capture the relationship between attendance and the other explanatory variables. If we are to use attendance rate as an explanatory variable for credit/no credit we get a very high pseudo r-squared values with the probit model, however if we use the logit model and use credit/no credit as a predict stata drops the variable because it successfully predicts a high attendance rate when they receive credit which is essentially a statistical tautology in this case and doesn’t tell us much else about it’s predictive ability with the rest of the data.

```
logit attendrate cornc timecom gender classsize prcp avghigh

note: cornc != 0 predicts success perfectly
      cornc dropped and 481 obs not used

Iteration 0:   log likelihood = -228.18209
Iteration 1:   log likelihood = -205.35721
Iteration 2:   log likelihood = -204.45754
Iteration 3:   log likelihood = -204.45204
Iteration 4:   log likelihood = -204.45204

Logistic regression                               Number of obs   =        356
                                                  LR chi2(5)      =      47.46
                                                  Prob > chi2     =     0.0000
Log likelihood = -204.45204                       Pseudo R2       =     0.1040

------------------------------------------------------------------------------
  attendrate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       cornc |          0  (omitted)
     timecom |  -.0060344   .0043628    -1.38   0.167    -.0145854    .0025165
      gender |  -.4549471   .2753227    -1.65   0.098    -.9945696    .0846754
   classsize |  -.0300788   .0142009    -2.12   0.034    -.0579121   -.0022455
        prcp |   .1143427    .028376     4.03   0.000     .0587268    .1699585
     avghigh |   .0217899    .008936     2.44   0.015     .0042756    .0393042
       _cons |  -3.720066   1.530148    -2.43   0.015    -6.719102   -.7210307
------------------------------------------------------------------------------
```

Further the probit model was able to illustrate a model of good fit even without an extremely collinear variable probably because such a variable was not very applicable.

```
probit attendrate status timecom  classsize prcp avghigh

Iteration 0:   log likelihood = -345.81709
Iteration 1:   log likelihood = -308.80956
Iteration 2:   log likelihood = -307.65697
Iteration 3:   log likelihood = -307.65449
Iteration 4:   log likelihood = -307.65449

Probit regression                                 Number of obs   =        837
```

```
                                                  LR chi2(5)      =      76.33
                                                  Prob > chi2     =     0.0000
Log likelihood = -307.65449                       Pseudo R2       =     0.1104

------------------------------------------------------------------------------
  attendrate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      status |  -.0827628   .0537173    -1.54   0.123    -.1880468    .0225211
     timecom |  -.0044244   .0020962    -2.11   0.035    -.0085328   -.000316
   classsize |  -.0127608   .0061527    -2.07   0.038    -.0248198   -.0007018
        prcp |   .0539664   .0121431     4.44   0.000     .0301663    .0777664
     avghigh |   .0193205   .0040677     4.75   0.000     .0113479     .027293
       _cons |  -2.113469   .7194963    -2.94   0.003    -3.523656   -.7032828
------------------------------------------------------------------------------
```

In this particular model it made sense not to drop status because the loss in an insignificant variable would lead to comparatively steep drop in R-squared value.

Discussion & Conclusion

If one examines the OLS observations and the Probit model observations we can conclude that status towards graduation, time commitment, precipitation and average high temperature throughout spring have similar influences on the attendance rates of Reedies in fourth quarter pe classes overall. If one chooses to look at results by major, the results are a bit more specific:

```
probit attendrate  timecom  classsize prcp avghigh if Major == "ENG"

Probit regression                                 Number of obs   =        102
                                                  LR chi2(4)      =       9.90
                                                  Prob > chi2     =     0.0421
Log likelihood = -31.995264                       Pseudo R2       =     0.1340

------------------------------------------------------------------------------
  attendrate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     timecom |  -.0110808   .0055558    -1.99   0.046    -.0219699   -.0001918
   classsize |   -.025218   .0233303    -1.08   0.280    -.0709445    .0205084
        prcp |   .0164544    .038404     0.43   0.668    -.0588161    .0917249
     avghigh |   .0218337   .0166862     1.31   0.191    -.0108706     .054538
       _cons |  -.6960805   2.640665    -0.26   0.792    -5.871689    4.479528
------------------------------------------------------------------------------

probit attendrate  timecom  classsize prcp avghigh if Major == "BIOL"

Probit regression                                 Number of obs   =         61
                                                  LR chi2(4)      =      11.18
                                                  Prob > chi2     =     0.0246
Log likelihood = -26.011143                       Pseudo R2       =     0.1769

------------------------------------------------------------------------------
  attendrate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     timecom |  -.0333493    .012573    -2.65   0.008     -.057992   -.0087067
   classsize |   .0114838   .0201506     0.57   0.569    -.0280107    .0509782
        prcp |   .0843066   .0387896     2.17   0.030     .0082803     .160333
     avghigh |   .0010203   .0151189     0.07   0.946    -.0286122    .0306528
       _cons |   2.085181   2.885376     0.72   0.470    -3.570051    7.740413
```

```
------------------------------------------------------------------------------

probit attendrate  timecom  classsize prcp avghigh if Major == "ANTH"

Probit regression                                 Number of obs   =         48
                                                  LR chi2(4)      =      14.36
                                                  Prob > chi2     =     0.0062
Log likelihood = -18.654384                       Pseudo R2       =     0.2780

------------------------------------------------------------------------------
  attendrate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     timecom |  -.0121093   .0073521    -1.65   0.100    -.0265192    .0023006
   classsize |  -.0649974   .0299915    -2.17   0.030    -.1237797   -.0062151
        prcp |   .0399769   .0523738     0.76   0.445    -.0626739    .1426276
     avghigh |   .0169272   .0216199     0.78   0.434     -.025447    .0593014
       _cons |   .0577167   3.399837     0.02   0.986    -6.605842    6.721275
------------------------------------------------------------------------------
```

While some of the effects carry over to other majors, it was difficult to achieve a comprehensive understanding of major effects in part because even in pe enrollment there is a noticeable pre-ponderance of some majors over others and we can expect to see the coefficient estimates for English more closely match the total population’s estimates because they form the largest contingent of students in 5 years of classes. It is rather hard understand and demonstrate the major effects of general literature.

Additionally females make up 556 of the 837 observations. Which may mean that the estimates of the effects may only be reflective of classes that are popular with females. But the fact that gender hasn’t been implicated in any statistically significant sense in any model is encouraging given that I have data for classes where guys predominate or are in equal portion.

Omitted Variable Bias

This was probably the biggest problem lurking in the background of this project, I was unable to get specific dates of Quals and Hum Papers and thesis due dates so I scaled down the specificity of the potential time data I’d look at. As a result, the hand-wavy solution was to hope that the status variable would take up the slack in determining how differing external factors affected upper and underclassmen which is practically an invitation to inefficient and biased estimators. This particular approach can say nothing of which class is worse in terms of hardship that would preclude going to attend p.e. Further, the weather variables were also averaged out in order to apply them in a blanket like fashion over each quarter instead of each day or week.

Heteroskedasticity

Also another problem especially in the probit model in that it causes the maximum likelihood estimator to inefficient. Greater error that occurs at different levels is a large problem in that the coefficients are constrained in order to fit a function that behaves close to asymptotically.

Autocorrelation

I did not see this as being much of a problem considering that the variables themselves seemed to have such little explanatory power right off the bat although they were found to be significant.

In conclusion, I'd posit that it would likely be better to examine students on a completely daily or weekly basis with weather being the changing explanatory variable and the class attributes being the unchanging dummy variables. Given that this approach would make more of the data actually usable, there would be more degrees of freedom and more possible classes with different attributes to use.