Austin Weisgrau
Jeff Parker
Metrics
Spring 2014
Final Project

Gender Differences in Performance Evolution: Physics 200 Grades Relative to Physics 100 Grades

Intro

This research project attempts to answer this question: "Do grades between Physics 100 and Physics 200 improve differently for men and women?" It examines whether the difference in grades between these courses for individual students is different, on average, between men and women, and also explores other explanatory variables. This question does not really measure performance overall at Reed, but rather captures a bit of the process of becoming a Reed student of whatever success level. Performance in Physics 100 can be thought of a baseline of initial performance at Reed. Performance in Physics 200 relative to Physics 100 is meant to capture how the Reed environment in a longer term affects student performance.

The physics department is one of a few departments notorious for having a disproportionate number of men students.[1] During the years 1989-2002, 31% of students who took Physics 100 were female, and only 20% of the declared physics majors were female. One reason for this discrepancy, and one danger of a long standing gender discrepancy, could be that female students studying physics are more poorly academically supported during their time at Reed. This project is meant to evaluate whether there is evidence of academic support differing based on gender at Reed, and especially in the Physics department.

Data

The data for this project come from Reed admissions & course data for 10,094 students at Reed during the academic years from 1989-1990 through 2002-2003. Admissions data, including information about high school performance and reader rating scores, are provided for each student. Reed performance, including major and gpa, are provided for each student. Every course taken by almost all of these students is provided - a total of 92,074 courses - as well as the term taken and grade received.

For the sake of this project, data were narrowed down to include only those students who completed two semesters of Physics 100 and two semesters of Physics 200. There were 187 students who fit this category over these thirteen years in the dataset - an average class size of 14, which seems reasonable given Reed's current physics class sizes of around 30 and the growth in Reed's student body and the physics program over time. There were six students who took only one semester of each class, or took more than two semesters to complete each course - these students were excluded from the dataset to simplify analysis and facilitate data

[1] This department was also chosen for study because of the standardized curriculum for freshman and sophomores (Physics 100 and 200), which makes evaluations of major-wide student performance simpler.

manipulation. This should not bias results significantly, because so few students are implicated, although it is likely that these students' performance and grades are related to how many semesters they took to complete the course. For this reason, if there were more of these students, it would become important to find a way to account for them in the analysis.

The following variables are used in the analysis:

grdiff - this variable is intended to measure the difference in grade between Physics 100 and Physics 200. Grades for each semester are first indexed on an eleven point scale, with "F" and withdrawal given a score of 1, "D" a score of 2, and then counting up with three numbers per letter grade to account for "-" and "+". "A+" is given a score of 11. Grades for semesters 1 and 2 of Physics 100 and 200 are averaged, and then the difference is found for each student to generate *grdiff*. This involved some nontrivial and not easily replicable spreadsheet work.

hsgpa - this variable measures GPA received in high school. this measure has some success in predicting student success in general at Reed, so it might be important in determining how students continue to succeed or respond to the Reed academic environment.

rdr - this variable is a five point score given by the admissions department to every student as a prediction of student success at Reed. Like hsgpa, it can be useful in predicting student success, and therefore may be important in determining the evolution of student success between Physics 100 and Physics 200

satm - this variable measures score received on the math portion of the SAT reported on admission application. Like hsgpa and rdr, this score is usually one of the more important determinants of student success at Reed

satv - this variable measures score received on the verbal portion of the SAT reported on admission application. This variable tends to be slightly less important, but often collinear with, satm in determining student success at Reed

yearterm - this variable reports the year a student begins at Reed. It roughly proxies for the years a student takes Physics 100 and Physics 200. Most students, but not all, take 100 their first year and 200 their second year. It would be better to account for these differences, but it was not possible in this project.

course200 - this dummy measures 1 for physics 200 and 0 for physics 100. This dummy is useful for differences in differences modeling as a time variable

fem - this is a dummy variable, with 1 indicating female and 0 indicating male

Variable	Obs	Mean	Std. Dev.	Min	Max
grdiff	187	-.5240642	2.105601	-7	6
hsgpa	151	3.796497	.3377119	2.49	4.83
rdr	186	2.591452	.5010611	1	4
yearterm	187	199380.7	300.9196	198901	199801
satv	181	679.779	80.47744	420	800
satm	181	698.8398	60.92146	550	800
fem	187	.1871658	.3910916	0	1

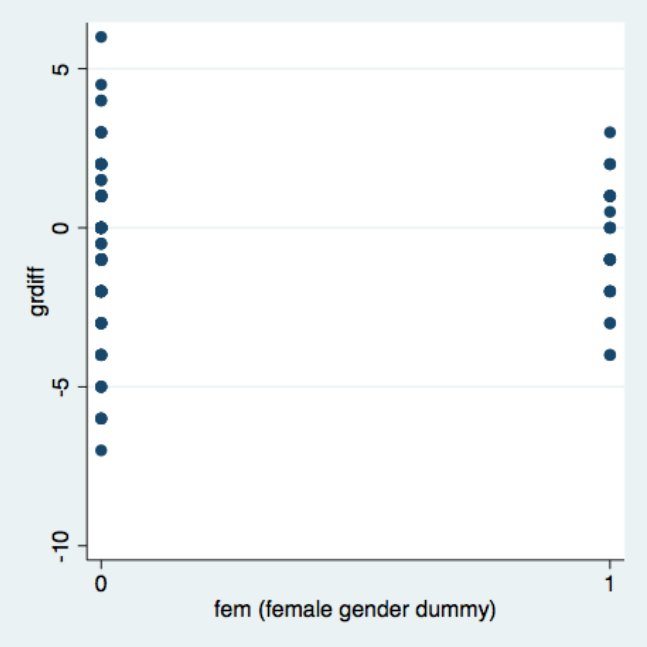
From this table we can tell that about 19% of the students in our sample are female. We have 187 observations of *grdiff*, and slightly less than that for most of the variables. Less than 150 observations may be available for regression, depending on how the gaps in the data line up. It is interesting to note that on average, students get slightly worse grades in Physics 200 than in Physics 100 - this is given by the negative mean3 for *grdiff*. This point estimate is not statistically significantly different from zero[2], however, so it is not statistically certain that this is negative on average.

Estimation

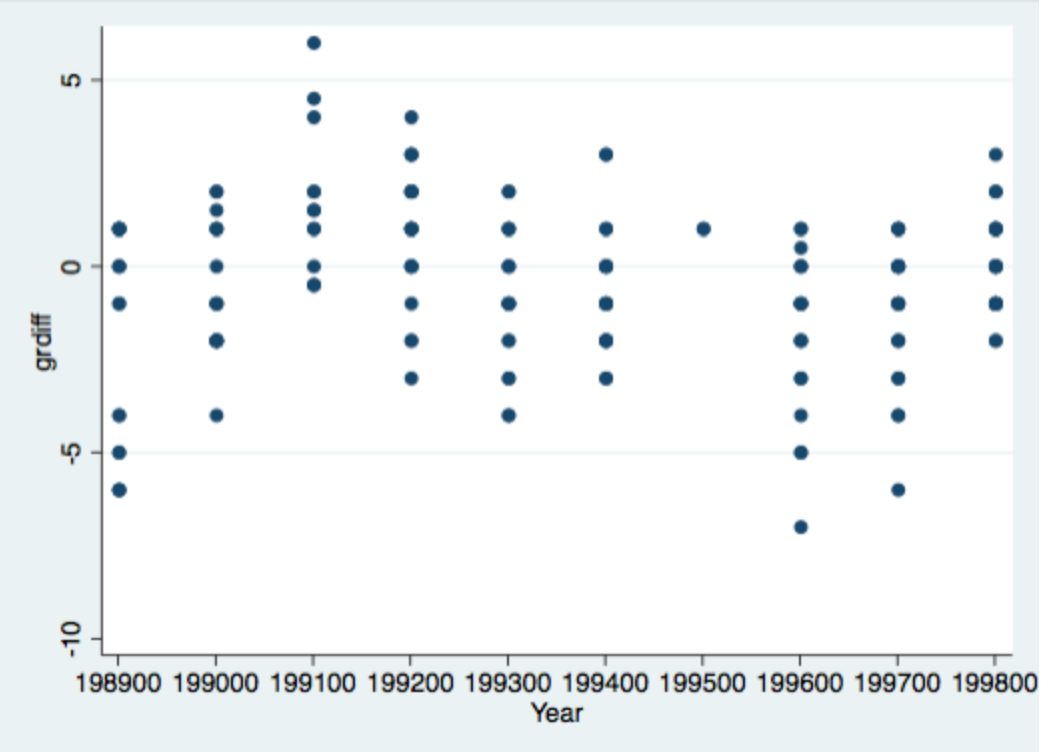
Diagnostics

I begin by examining a few scatterplots of the regressors against the dependent variable grdiff to check for heteroskedasticity and to roughly estimate an appropriate specification.

[2] This is obvious given that the standard deviation is an order of magnitude larger than the point estimate.



This is the scatterplot of gender against grdiff. It is unclear whether or not there is a significant difference in means, but it is obvious that the data is heteroskedastic. Part of this can be explained by the larger sample of male students, but using heteroskedasticity robust standard errors will be important for generating unbiased standard errors.



This is the scatterplot of year and term against grdiff. There is a huge, nonlinear variability from year to year. It is hard to imagine what can explain this, although if different professors are

teaching these courses from year to year, there may be nonstationary differences in grading systems from year to year. This is problematic, as it may invalidate comparing grades between Physics 100 and Physics 200 at all, especially over time. Using heteroskedasticity robust standard errors may account for some of the variability here to generate unbiased standard errors, and using time dummies for each term will be essential as well.

OLS

I use OLS with grdiff as the dependent variable with robust standard errors, the above mentioned explanatory variables, and time dummies for each year.

$$grdiff_i = \beta_1 + \beta\chi_{i,characteristics} + \beta(fem_i) + \beta\sum_{1990}^{1998}(yearterm_j) + e_i$$

```
Linear regression                               Number of obs =     146
                                                F( 13,   131) =       .
                                                Prob > F      =       .
                                                R-squared     =  0.2884
                                                Root MSE      =  1.8108

------------------------------------------------------------------------------
             |               Robust
      grdiff |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      hsgpa |  -.3567193   .4752897    -0.75   0.454               -.0592361
      rdr |    -.6388367   .3960509    -1.61   0.109               -.1507754
      satv |     .0035173   .0023777     1.48   0.141                .1304775
      satm |   -.0047706   .0029918    -1.59   0.113               -.1427597
      fem |      .4928822   .4128238     1.19   0.235                .092734
          |
     yearterm |
     199001  |   1.873441   .9593914   1.95   0.053                .2327169
     199101  |   3.874186   1.027711   3.77   0.000                .4812473
     199201  |   2.743106   .9846645   2.79   0.006                .3972025
     199301  |   1.324022   .954612    1.39   0.168                .2088493
     199401  |   2.077655   .9839787   2.11   0.037                .3359178
     199501  |   3.020785   .9314952   3.24   0.002                .1225243
     199601  |   .3676636   1.001769   0.37   0.714                .0608342
     199701  |   1.078411   .9704856   1.11   0.269                .1784355
     199801  |   2.366669   .9036448   2.62   0.010                .44528
             |
     _cons   |   1.516973   3.794747   0.40   0.690                     .
------------------------------------------------------------------------------
```

This model does not do a great job of explaining variability in the sample. None of the explanatory variables are statistically significantly different from zero at a five percent level of significance. The R-squared of the model is 28%, which is not trivial, but is likely mostly

explained by the few year dummies which are statistically significant. It seems that most of the variability in student grade changes between second and first year physics is determined by the years those classes were taken. Personal characteristics like gender and academic performance in high school have little effect on the evolution of student success between first and second year physics. Notably, the female dummy is not statistically significantly different from zero.

OLS with Interactions

Because grades are so variable from year to year, it is plausible that gendered differences are variable from year to year as well. I repeat this model including interaction terms between the gender dummy and each year term.

$$grdiff_i = \beta_1 + \beta\,\chi_{i,characteristics} + \beta * (fem_i) + \beta * \sum_{1990}^{1998}(yearterm_j) \;\ldots.$$
$$\ldots.. + \beta * < \sum_{1990}^{1998}(yearterm_j) \;\; \sum_{0}^{1}(fem_i) > + e_i$$

```
Linear regression                               Number of obs =     146
                                                F( 17,   125) =       .
                                                Prob > F      =       .
                                                R-squared     =  0.3149
                                                Root MSE      =  1.8189

------------------------------------------------------------------------------
             |               Robust
      grdiff |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      hsgpa  |  -.3538798   .4744847    -0.75   0.457               -.0587646
      rdr    |  -.6170824   .4144383    -1.49   0.139               -.145641
      satv   |   .004652    .0025124     1.85   0.066                .1725683
      satm   |  -.0051878   .0029278    -1.77   0.079               -.1552466
             |
      yearterm |
      199001 |   1.571101   1.002451     1.57   0.120                .1951605
      199101 |   3.778031   1.083082     3.49   0.001                .469303
      199201 |   2.435089   1.025673     2.37   0.019                .3526016
      199301 |   .7942173   .9991314     0.79   0.428                .1252786
      199401 |   1.829504   1.042811     1.75   0.082                .2957964
      199501 |   2.701008   .9679283     2.79   0.006                .109554
      199601 |  -.0094566   1.109617    -0.01   0.993               -.0015647
      199701 |   .7016052   1.078571     0.65   0.517                .1160887
      199801 |   2.060026   .9483714     2.17   0.032                .3875863
             |
      1.fem  |  -2.856399   .9413879    -3.03   0.003               -.537421
             |
yearterm#fem |
   199001 1  |      0  (empty)                                          .
```

```
   199101 1  |   1.370072   1.100355    1.25   0.215              .0555707
   199201 1  |      0  (empty)                                        .
   199301 1  |   4.198114   1.463983    2.87   0.005              .3370131
   199401 1  |   3.02748    1.462745    2.07   0.041              .1730597
   199501 1  |      0  (empty)                                        .
   199601 1  |   3.607128   1.258411    2.87   0.005              .3226075
   199701 1  |   3.613649   1.323409    2.73   0.007              .4044352
   199801 1  |   3.489248   1.161229    3.00   0.003              .3120648
             |
       _cons |   1.264869   3.851402    0.33   0.743                    .
------------------------------------------------------------------------------
```

First of all, it is worthwhile to use a Wald test to check if the relationship is nonstationary over time.

$$F_{9,125} = \frac{(SSE_R - SSE_U)/12}{SSE_U/(125)} = 4.146$$

This test evaluates the joint significance of all time variables and interaction terms. The F-statistic is 4.146, which is far beyond the critical value for this distribution at 1.601 for a 5% level of significance, indicating that the model is indeed nonstationary over time.

In this case, we are able to explain differences in student success using personal characteristics more effectively. Interestingly, SAT scores are nearly significant, and because they are highly collinear are nearly jointly significant[3]. This model is highly nonstationary, and it may be important to interact every explanatory variable against every year dummy. The interaction terms between gender and year are statistically significantly different from zero in every year but 1995 between 1993 and 1998. This means that the effect of gender is statistically significantly different in these years from its effect on grade difference for students who began at Reed in 1989. The uninteracted effect of gender on grade difference is -2.86 and statistically significant, meaning that women's grades improved by half a standard deviation less than men's between physics 100 and physics 200 for those who entered Reed in 1989. In later years, this gap became much smaller, with women perhaps improving more than men[4]. If the effect of gender moves from negative to positive over the time period of the sample, then any regression which averages over the time period may find insignificant coefficient estimates for a gender dummy.

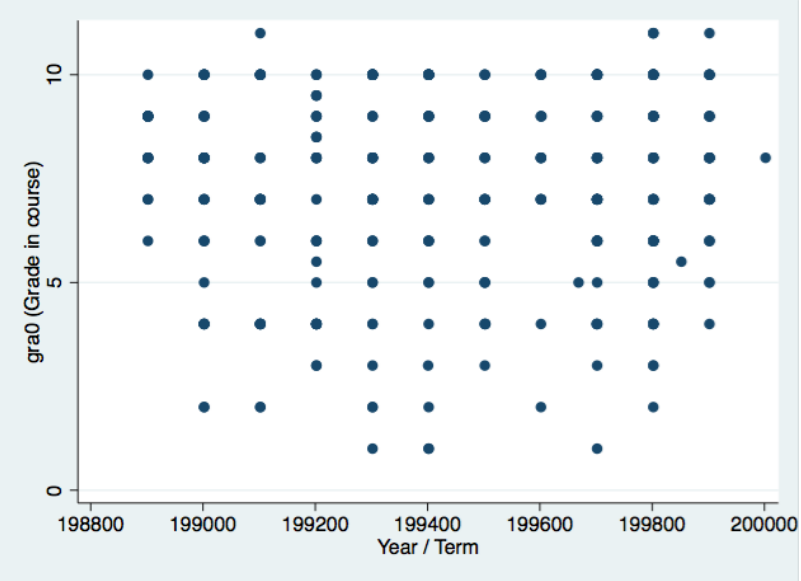
Differences in Differences

We can also use a difference-in-difference estimator to estimate the effect of being a female versus being a male in the second year of physics versus the first year of physics. This model estimates effects on grades in physics 100 or 200, most importantly as a result of being female and in relation to grade in physics 100. OLS and fixed effects regression can be used to estimate differences in differences, and I use an OLS model because of its simplicity.

[3] These variables are jointly significant at a 7% level of confidence.

[4] This effect is not statistically significantly different from zero. The gender gap may have closed, but not reversed, by the late 1990s. Results are inconclusive.

First I scatter grades against the year and term, to see whether it would be important to include a dummy in this model as in the previous models.



Interestingly, there does not appear to be very much difference at all from year to year, as there was in the grdiff time scatterplot. This may be because differences between physics 100 and physics 200 grading tend to cancel eachother out from year to year. The relative stability of grade levels makes the variability of difference in grades between the two years of physics much more mystifying.

In any case, it does not seem important to include dummies for every year and term.

$$gra0_i = \beta_1 + \beta_2(fem_i) + \beta_3(course200_i) + \beta_4(fem_i * course200_i) + \beta\,\chi_{i,characteristics} + e_i$$

```
Linear regression                               Number of obs =     301
                                                F(  6,   294) =   14.89
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1918
                                                Root MSE      =  2.0554

------------------------------------------------------------------------------
             |               Robust
        gra0 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
       hsgpa |    1.116081   .394846     2.83   0.005                .1668666
         rdr |   -1.083387   .3032014   -3.57   0.000                -.2306487
        satm |    .0066156   .0022117    2.99   0.003                .1765951
       1.fem |   -.5882418   .3989489   -1.47   0.141                -.100608
  1.course200 |  -.6300813   .2625018   -2.40   0.017                -.1394308
              |
fem#course200 |
       1 1    |    .168051   .6114314    0.27   0.784                .0212533
```

```
             |
       _cons |   1.755898     2.668176   0.66    0.511                     .
```

First of all, in this model, hsgpa, rdr, and satm all are statistically significant, because here they are predicting grades rather than differences in grades between two years of physics. Their coefficients are in the expected direction as well. We can also read that females do worse in Physics 100 and Physics 200 than males, although this difference is statistically insignificant.

The estimated coefficient on the reaction term, β_4, is what we are interested in here.

$$\beta_4 = (\bar{y}_{fem,200} - \bar{y}_{fem,100}) - (\bar{y}_{male,200} - \bar{y}_{male,100})$$

This coefficient estimates the difference between females and males in grades between physics 200 and physics 100. This coefficient estimate is not only estimated to be relatively small, but it is also not statistically significantly different from zero. The course200 dummy in this model is effectively a time variable, with 0 meaning the grade is for Physics 100 and 1 meaning the grade is for Physics 200. The interaction term on this model between *fem* and *course200* measures the difference in gender within differences in grades between physics courses. It is not statistically significantly different from zero, which is consistent with the previous models' results.

Discussion

Results

The sample data of differences for individuals between Physics 100 and Physics 200 does not seem to be well explained by gender over the period 1989-2002. Highschool and admissions predictors of student success *are* useful for predicting grades in Physics 100 and Physics 200, but not for predicting the change in performance between these two courses. The evolution of performance at Reed does not seem to be determined by the same factors that determine overall performance.

There is a little evidence to suggest that the effect of gender on changing performance between years of physics changed after 1989, but evaluating the functional nature of this change was not in the purview of this project. Our model *can* conclude that women performed more poorly than men in Physics 200 relative to performance in Physics 100 in the early 1990s, and that by the end of the decade this gender gap had closed and perhaps even reversed. This suggests that during the 1990s Reed's physics department became more successful at offering academic support to previously underserved female students.

External Validity

These results are unlikely to have much external validity, either over time, or departments at Reed, or academic departments at other schools. It seems that the change in performance from Physics 100 to Physics 200, as well as the effects of gender on that change in performance, changed a lot from year to year. Although there is a little evidence of a trend in the direction of the change for the effect of gender, this trend was not statistically tested. Evaluating

whether this trend exists would be a good direction for future research, and could be tested ex post against Reed's data for enrollment and grades after 2002.

Every department is structured differently, and the role of gender between departments likely differs a lot as well. We might expect trends toward closing gender gaps in every department over time, but this likely looks very different in different departments. It is also not obvious what a corollary between changing performance between Physics 100 and Physics 200 would be in other departments. Further, the changing academic environment for female students in physics over the 1990s could have been constituted by any of a variety of highly particular phenomena, including student groups and support and the hiring of supportive faculty or staff. These may or may not be duplicated in other departments, at Reed or elsewhere, over the same or future time periods.

Internal Validity

Omitted Variables - It is nearly certain that we are missing nearly *all* of the variables which determine the evolution of performance over time at Reed. The evolution of performance over time is not determined by most metrics of high school academic performance, the exception being SAT score. Evolution of performance over freshman and sophomore years at Reed is likely determined by access to and use of support services, peer environment, study and recreation habits, etc. Results about gender, however, are not invalidated by omitted variables. The use of gender in analysis is intended to proxy for a variety of other factors, and points to how the collection of those factors are systematically gendered.

Sample selection bias - It is possible that narrowing the data sample down to only students that took both Physics 100 and Physics 200 introduced bias. Students that performed poorly in Physics 100 were probably less likely to take Physics 200. Students that began Physics 100 and encountered an unsupportive academic environment probably were unlikely to complete the class. Both of these types of sample selection problems would bias my results to show *less* of an effect, so any effects I find are likely an underestimate.

Heteroskedasiticity - There was clear evidence of heteroskedasticity in data scatterplots, so robust standard errors were used. While inefficient, my results should be consistent. Inefficient estimators will also deflate the significance of my results, so the t scores of the results I find may be slightly underestimated.

Autocorrelation - Students are not totally independent, and their performance is likely autocorrelated to a certain extent. Our models attempted to account for autocorrelation of students in time by using time dummies, but accounting for autocorrelation of students in friend groups or dorms, etc. is not possible. Autocorrelation does not seem like a very important concern for this model.

Cardinality - Grades were assumed to be cardinal, such that the difference between an F or a W and a D is the same as the difference between a B+ and an A-. This does not seem logically reasonable, but is reasonably valid for statistical purposes and is a common modelling strategy.

Future Research

As mentioned above, it would be useful to explore the trend in the effect of gender over time, perhaps by constraining that effect to fall in a line over time and testing for statistical validity of the constraint with a Wald test.

Choosing to evaluate gendered academic environments by looking at performance in Physics 200 relative to performance in Physics 100 was a somewhat arbitrary choice and not necessarily the best method. For one thing, it loses a lot of absolute difference in performance between genders by looking at the change in relative performance over time. This method assumes that the gendered academic environment affects performance in Physics 200 more than it affects performance in Physics 100, which is arguably invalid. A more effective method could compare high school performance to performance in Physics 100, which would also benefit from having more generalizability to other departments and colleges. This method would effectively evaluate the gendered difference in academic support between high school and college.

```
------------------------------------------------------------------------------------------------------------------------
      name:  <unnamed>
       log:  /Users/austinweisgrau/Documents/academics/jun/metrics/final/newplan/final1.log
  log type:  text
 opened on:   7 May 2014, 18:06:17

.
. insheet using "grdiff.csv"
(2 vars, 748 obs)

.
. collapse grdiff, by(case)

.
. replace grdiff = grdiff*4
(153 real changes made)

.
. merge 1:1 case using "infob.dta"
case was int now long

    Result                           # of obs.
    -----------------------------------------
    not matched                         9,907
        from master                         0  (_merge==1)
        from using                      9,907  (_merge==2)

    matched                               187  (_merge==3)
    -----------------------------------------

. drop if _merge==2
(9907 observations deleted)

.
. g fem = 0

. replace fem = 1 if gender == "F"
(35 real changes made)

.
. drop case

. drop majs

. drop majd

. drop majp

. drop _merge

. drop hstype

. drop hsrank

. drop hssize

. drop gender

. drop citz

. drop ethnic

. drop degree

. drop uggpa

. drop finaid

.
. summ

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      grdiff |       187   -.5240642    2.105601         -7          6
       hsgpa |       151    3.796497    .3377119       2.49       4.83
         rdr |       186    2.591452    .5010611          1          4
```

```
    yearterm |       187    199380.7    300.9196     198901     199801
        satv |       181     679.779    80.47744        420        800
-------------+--------------------------------------------------------
        satm |       181    698.8398    60.92146        550        800
         fem |       187    .1871658    .3910916          0          1

.
. graph twoway scatter grdiff fem

. graph save "grXfem", replace
(file grXfem.gph saved)

.
. graph twoway scatter grdiff yearterm

. graph save "grXyear", replace
(file grXyear.gph saved)

.
. reg grdiff hsgpa rdr satv satm fem i.(yearterm), robust beta

Linear regression                                       Number of obs =     146
                                                        F( 13,   131) =       .
                                                        Prob > F      =       .
                                                        R-squared     =  0.2884
                                                        Root MSE      =  1.8108

------------------------------------------------------------------------------
             |               Robust
      grdiff |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
       hsgpa |  -.3567193   .4752897    -0.75   0.454                -.0592361
         rdr |  -.6388367   .3960509    -1.61   0.109                -.1507754
        satv |   .0035173   .0023777     1.48   0.141                 .1304775
        satm |  -.0047706   .0029918    -1.59   0.113                -.1427597
         fem |   .4928822   .4128238     1.19   0.235                  .092734
             |
    yearterm |
      199001 |   1.873441   .9593914     1.95   0.053                 .2327169
      199101 |   3.874186   1.027711     3.77   0.000                 .4812473
      199201 |   2.743106   .9846645     2.79   0.006                 .3972025
      199301 |   1.324022    .954612     1.39   0.168                 .2088493
      199401 |   2.077655   .9839787     2.11   0.037                 .3359178
      199501 |   3.020785   .9314952     3.24   0.002                 .1225243
      199601 |   .3676636   1.001769     0.37   0.714                 .0608342
      199701 |   1.078411   .9704856     1.11   0.269                 .1784355
      199801 |   2.366669   .9036448     2.62   0.010                   .44528
             |
       _cons |   1.516973   3.794747     0.40   0.690                        .
------------------------------------------------------------------------------

.
. reg grdiff hsgpa rdr satv satm fem##i.(yearterm), robust beta
note: 1.fem#199001.yearterm identifies no observations in the sample
note: 1.fem#199201.yearterm identifies no observations in the sample
note: 1.fem#199501.yearterm identifies no observations in the sample

Linear regression                                       Number of obs =     146
                                                        F( 17,   125) =       .
                                                        Prob > F      =       .
                                                        R-squared     =  0.3149
                                                        Root MSE      =  1.8189

------------------------------------------------------------------------------
             |               Robust
      grdiff |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
       hsgpa |  -.3538798   .4744847    -0.75   0.457                -.0587646
         rdr |  -.6170824   .4144383    -1.49   0.139                 -.145641
        satv |    .004652   .0025124     1.85   0.066                 .1725683
        satm |  -.0051878   .0029278    -1.77   0.079                -.1552466
      1.fem |  -2.856399   .9413879    -3.03   0.003                 -.537421
             |
    yearterm |
      199001 |   1.571101   1.002451     1.57   0.120                 .1951605
      199101 |   3.778031   1.083082     3.49   0.001                  .469303
```

```
     199201  |   2.435089   1.025673     2.37   0.019                .3526016
     199301  |   .7942173   .9991314     0.79   0.428                .1252786
     199401  |   1.829504   1.042811     1.75   0.082                .2957964
     199501  |   2.701008   .9679283     2.79   0.006                 .109554
     199601  |  -.0094566   1.109617    -0.01   0.993               -.0015647
     199701  |   .7016052   1.078571     0.65   0.517                .1160887
     199801  |   2.060026   .9483714     2.17   0.032                .3875863
             |
fem#yearterm |
   1 199001  |          0  (empty)                                        .
   1 199101  |   1.370072   1.100355     1.25   0.215                .0555707
   1 199201  |          0  (empty)                                        .
   1 199301  |   4.198114   1.463983     2.87   0.005                .3370131
   1 199401  |    3.02748   1.462745     2.07   0.041                .1730597
   1 199501  |          0  (empty)                                        .
   1 199601  |   3.607128   1.258411     2.87   0.005                .3226075
   1 199701  |   3.613649   1.323409     2.73   0.007                .4044352
   1 199801  |   3.489248   1.161229     3.00   0.003                .3120648
             |
       _cons |   1.264869   3.851402     0.33   0.743                      .
------------------------------------------------------------------------------

. display(e(rss))
413.55931

.
. reg grdiff hsgpa rdr satv satm fem, robust beta

Linear regression                               Number of obs =     146
                                                F(  5,   140) =    1.15
                                                Prob > F      =  0.3384
                                                R-squared     =  0.0423
                                                Root MSE      =  2.0322

------------------------------------------------------------------------------
             |               Robust
      grdiff |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
       hsgpa |   -.208406   .5108815    -0.41   0.684               -.0346075
         rdr |  -.2610787   .4838455    -0.54   0.590               -.0616186
        satv |     .00394    .002747     1.43   0.154                .1461584
        satm |  -.0059813   .0031381    -1.91   0.059               -.1789894
         fem |   .1448047   .4200863     0.34   0.731                .0272445
       _cons |   2.294138   4.374887     0.52   0.601                      .
------------------------------------------------------------------------------

. display(e(rss))
578.15496

.
end of do-file

. ls

total 10656
-rw-r--r--@ 1 austinweisgrau  staff     2081 May  6 23:25 100_200_f.do
-rwxrwxrwx  1 austinweisgrau  staff  4514420 Apr 28 18:49 courseb.dta*
drwxr-xr-x  6 austinweisgrau  staff      204 May  6 20:14 did/
-rw-r--r--  1 austinweisgrau  staff      735 May  6 23:25 final.do
-rw-r--r--  1 austinweisgrau  staff     8482 May  7 18:06 final1.log
-rw-r--r--  1 austinweisgrau  staff     1263 May  7 18:06 flesh.smcl
-rw-r--r--  1 austinweisgrau  staff     6445 May  7 18:06 grXfem.gph
-rw-r--r--  1 austinweisgrau  staff     6485 May  7 18:06 grXyear.gph
-rw-r--r--  1 austinweisgrau  staff     9593 May  7 17:47 graXyear.gph
-rw-r--r--  1 austinweisgrau  staff     5192 May  6 16:56 grdiff.csv
-rwxrwxrwx  1 austinweisgrau  staff   875545 Apr 28 18:50 infob.dta*

. do 100_200_f.do

. clear all

. log close
      name:  <unnamed>
       log:  /Users/austinweisgrau/Documents/academics/jun/metrics/final/newplan/final1.log
  log type:  text
 closed on:   7 May 2014, 18:06:46
```

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

-------------------------------------------------------------------------------------------------------------------
      name:  <unnamed>
       log:  /Users/austinweisgrau/Documents/academics/jun/metrics/final/newplan/final2.log
  log type:  text
 opened on:   7 May 2014, 18:06:46

. 
. use "infob.dta"

. sort case

. clear all

. 
. use "courseb.dta"

. sort case

. 
. merge m:1 case using "infob.dta"
case was int now long

    Result                           # of obs.
    -----------------------------------------
    not matched                         6,556
        from master                         0  (_merge==1)
        from using                      6,556  (_merge==2)

    matched                            92,074  (_merge==3)
    -----------------------------------------

. 
. drop if _merge == 2
(6556 observations deleted)

. 
. drop if subject!="PHYS"
(88143 observations deleted)

. 
. keep if coursenumber == "100" | coursenumber == "200"
(1730 observations deleted)

. 
. //I only want students who took both physics 100 & physics 200 in this sample
. g course200 = 0

. replace course200 = 1 if coursenumber == "200"
(571 real changes made)

. 
. g course300 = 0

. replace course300 = -1 if coursenumber == "100"
(1630 real changes made)

. replace course300 = 1 if coursenumber == "200"
(571 real changes made)

. 
. collapse course300, by(case)

. keep if course300 == 0
(718 observations deleted)

. 
. drop course300

. 
. merge 1:m case using "courseb.dta"
```

```
    Result                           # of obs.
    -----------------------------------------
    not matched                         85,894
        from master                          0  (_merge==1)
        from using                      85,894  (_merge==2)

    matched                              6,180  (_merge==3)
    -----------------------------------------

.
. // now we're back to the beginning, but only with the 193 students
. // who took physics 100 & physics 200
.
. drop if _merge == 2
(85894 observations deleted)

. drop _merge

. merge m:1 case using "infob.dta"

    Result                           # of obs.
    -----------------------------------------
    not matched                          9,901
        from master                          0  (_merge==1)
        from using                       9,901  (_merge==2)

    matched                              6,180  (_merge==3)
    -----------------------------------------

.
. drop if _merge == 2
(9901 observations deleted)

. drop _merge

.
. drop if subject!="PHYS"
(4225 observations deleted)

.
. keep if coursenumber == "100" | coursenumber == "200"
(1185 observations deleted)

.
. g course200 = 0

. replace course200 = 1 if coursenumber == "200"
(385 real changes made)

.
.
. //I don't think it's valid to exclude withdrawals from the sample...
. //It might be more appropriate to treat withdrawals as failures
.
. //A-*, AA, B+*, B-*, C+*, C-*, CR, MG, NC and UW all have less than five observations
. //so I'm going to ignore them for now
.
. g gra0 = 0

. replace gra0 = 1 if grade == "F"
(8 real changes made)

. replace gra0 = 1 if grade == "W"
(3 real changes made)

. replace gra0 = 2 if grade == "D"
(17 real changes made)

. replace gra0 = 3 if grade == "C-"
(15 real changes made)

. replace gra0 = 4 if grade == "C"
(80 real changes made)

. replace gra0 = 5 if grade == "C+"
```

```
(31 real changes made)

. replace gra0 = 6 if grade == "B-"
(85 real changes made)

. replace gra0 = 7 if grade == "B"
(128 real changes made)

. replace gra0 = 8 if grade == "B+"
(116 real changes made)

. replace gra0 = 9 if grade == "A-"
(113 real changes made)

. replace gra0 = 10 if grade == "A"
(159 real changes made)

. replace gra0 = 11 if grade == "A+"
(11 real changes made)

. drop if gra0 == 0
(4 observations deleted)

.
. g fem = 0

. replace fem = 1 if gender == "F"
(146 real changes made)

.
. collapse yearterm gra0 hsgpa rdr satm satv fem, by(case course200)

. //this averages gra0 for the first and second semester of physics 200
. //one observation must be deleted because more than two semesters are spent on each
. replace yearterm = 199401 in 123
(1 real change made)

.
. graph twoway scatter gra0 yearterm

. graph save "graXyear", replace
(file graXyear.gph saved)

.
. reg gra0 hsgpa rdr satm fem##course200, robust beta

Linear regression                               Number of obs =     301
                                                F(  6,   294) =   14.89
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1918
                                                Root MSE      =  2.0554

------------------------------------------------------------------------------
             |               Robust
        gra0 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
       hsgpa |   1.116081    .394846     2.83   0.005                .1668666
         rdr |  -1.083387   .3032014    -3.57   0.000               -.2306487
        satm |   .0066156   .0022117     2.99   0.003                .1765951
       1.fem |  -.5882418   .3989489    -1.47   0.141                -.100608
 1.course200 |  -.6300813   .2625018    -2.40   0.017               -.1394308
             |
fem#course200 |
         1 1 |    .168051   .6114314     0.27   0.784                .0212533
             |
       _cons |   1.755898   2.668176     0.66   0.511                      .
------------------------------------------------------------------------------

.
.
.
end of do-file

. log close
      name:  <unnamed>
       log:  /Users/austinweisgrau/Documents/academics/jun/metrics/final/newplan/final2.log
```

```
  log type:  text
 closed on:   7 May 2014, 18:07:04
-------------------------------------------------------------------------------------------------------------
```