Intro to Multivariate Regression

1 Matchmaker, Matchmaker

- We use *multivariate regression* to control for confounding factors in an effort to create *ceteris paribus* comparisons
- Multivariate regression is an automatic matchmaker

We're often interested in the relationship between a dependent variable, Y_i, and another variable, X_{1i}, in a scenario where the connection between Y_i and X_{1i} can be explained (in a statistical sense) by the fact that X_{1i} is associated with another variable, X_{2i}, that also predicts Y_i (the association between health insurance and health in the NHIS might be explained by the higher schooling of the insured). In treatment effects problems, this is called selection bias. In a regression context, we call it *omitted variables bias*.

To keep things simple, suppose that X_{1i} is Bernoulli. "Holding things constant" in this case means we replace the unconditional comparison,

$$E[Y_i \mid X_{1i} = 1] - E[Y_i \mid X_{1i} = 0],$$

with conditional comparisons,

$$E[Y_i \mid X_{1i} = 1, X_{2i} = x] - E[Y \mid X_{1i} = 0, X_{2i} = x]. \tag{1}$$

In other words, we look at the CEF of Y given X_{1i}, *conditional on* $X_{2i} = x$.

- Such comparisons are said to be (not necessarily causal) "effects" of X_{1i}, computed while *matching* on values of X_{2i}.
  - Matching doesn't produce 100% *ceteris paribus* comparisons, but it takes us some way on the path to this. Matching on X_{2i} ensures that our comparison of averages across values of X_{1i} have the same value of X_{2i}
- Note that $E[Y_i \mid X_{1i} = 1, X_{2i} = x] - E[Y_i \mid X_{1i} = 0, X_{2i} = x]$ takes on as many values as there are values of X_{2i}
- As we'll soon see, multiple regression neatly combines sets of matched comparisons into a single controlled average effect, while also giving us the necessary standard errors for this single average effect

1.1 Multivariate Regression Makes Me a Match

- Our controls, X_{2i}, often take on many values (either because there is more than one thing to be controlled or because the individual controls take on many values, like SAT scores in *MM* Chapter 2). This threatens to overwhelm us with a multitude of conditional comparisons.
- Regression methods solve this problem by fitting a linear model with a single conditional effect.

As an expedient, assume the CEF given X_{1i} and X_{2i} is linear:

$$E[Y_i \mid X_{1i}, X_{2i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \tag{2}$$

Equivalently, write

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \qquad E[\varepsilon_i \mid X_{1i}, X_{2i}] = 0 \tag{3}$$

Equation (3) reminds us that CEF residuals are mean zero and mean-independent of conditioning variables. Consequently, β_0, β_1, β_2 solve

$$E[Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}] = E[\varepsilon_i] = 0 \tag{4}$$

$$E[(Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})X_{1i}] = E[\varepsilon_i X_{1i}] = 0$$

$$E[(Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})X_{2i}] = E[\varepsilon_i X_{2i}] = 0$$

Coefficients derived by solving this system define the *multivariate regression* of Y_i on X_{1i} and X_{2i}.

- What if the CEF is nonlinear? Then, as detailed in MHE Chpt 3 and the third set of regressions notes, multivariate regression provides a best-in-class linear approximation to *any* CEF
  - An important consequence of approximation awesomeness, which we'll "prove" by computer, is that regression is an *automatic matchmaker*

1.1.1 Asians and Whites Under Control

- In a sample of prime age male high school grads in the 2016 American Community Survey, Asians (75% foreign-born) earn more than whites

```
59 . summarize

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        agep |     57,696      44.632    2.834972         40         49
        wagp |     57,696    85197.99    88589.83          0     714000
        wkhp |     57,696    45.16632     10.0141          1         99
      racasn |     57,696    .0987243    .2982941          0          1
       racpi |     57,696    .0009706    .0311397          0          1
-------------+---------------------------------------------------------
      racwht |     57,696    .9083299    .2885622          0          1
         uhe |     56,924    34.81603    29.37749          0   201.5789
      loguhe |     53,750    3.361481    .7235695  -6.437752   5.306181
       immig |     57,696    .1748475    .3798399          0          1
     yearsEd |     57,696    14.53616     2.42775         12         21
-------------+---------------------------------------------------------
      hsgrad |     57,696           1           0          1          1
     somecol |     57,696    .5348551     .498788          0          1
     colgrad |     57,696    .4422664    .4966599          0          1
    asianpac |     57,696    .0987243    .2982941          0          1
       white |     57,289    .9076786    .2894816          0          1

60 . bys asianpac: summarize loguhe yearsEd colgrad immig

----------------------------------------------------------------------------------------------
-> asianpac = 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      loguhe |     48,411    3.345458    .7155705  -6.437752   5.306181
     yearsEd |     52,000    14.40617    2.377595         12         21
     colgrad |     52,000    .4188462    .4933748          0          1
       immig |     52,000      .11025    .3132041          0          1

----------------------------------------------------------------------------------------------
-> asianpac = 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      loguhe |      5,339    3.506776    .7775642  -3.912023    5.30231
     yearsEd |      5,696    15.72279    2.555968         12         21
     colgrad |      5,696    .6560744    .4750583          0          1
       immig |      5,696    .7645716    .4243035          0          1
```

- Asians in this sample are mostly immigrants yet immigrants earn less and Asians earn more - what's up w/that?
- Is the Asian effect causal? (Ponder potential outcomes). Either way, ethnicity gaps in college graduation rates might explain it
- Compare the college-controlled reg estimate of 0.0167 to the average conditional-on-college Asian effect:

$$-.0756\frac{29903}{53750}(= .556) + .0689\frac{23847}{53750}(= .444) \simeq -.01$$

```
-------------+----------------------------------   F(1, 53748)     =    240.08
       Model |  125.139748         1  125.139748   Prob > F        =    0.0000
    Residual |  28015.2987    53,748  .521234253   R-squared       =    0.0044
-------------+----------------------------------   Adj R-squared   =    0.0044
       Total |  28140.4384    53,749  .523552781   Root MSE        =    .72197

------------------------------------------------------------------------------
      loguhe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    asianpac |   .1613188   .0104113    15.49   0.000     .1409126    .1817249
       _cons |   3.345458   .0032813  1019.56   0.000     3.339026    3.351889
------------------------------------------------------------------------------

63 . reg loguhe asianpac colgrad

      Source |       SS           df       MS      Number of obs   =    53,750
-------------+----------------------------------   F(2, 53747)     =   5623.46
       Model |  4869.57938         2  2434.78969   Prob > F        =    0.0000
    Residual |   23270.859    53,747   .43297038   R-squared       =    0.1730
-------------+----------------------------------   Adj R-squared   =    0.1730
       Total |  28140.4384    53,749  .523552781   Root MSE        =      .658

------------------------------------------------------------------------------
      loguhe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    asianpac |    .016507   .0095892     1.72   0.085     -.002288    .0353019
     colgrad |   .6043304   .0057731   104.68   0.000      .593015    .6156457
       _cons |   3.091722   .0038496   803.14   0.000     3.084176    3.099267
------------------------------------------------------------------------------

64 . reg loguhe asianpac yearsEd

      Source |       SS           df       MS      Number of obs   =    53,750
-------------+----------------------------------   F(2, 53747)     =   6283.13
       Model |  5332.56924         2  2666.28462   Prob > F        =    0.0000
    Residual |  22807.8692    53,747  .424356134   R-squared       =    0.1895
-------------+----------------------------------   Adj R-squared   =    0.1895
       Total |  28140.4384    53,749  .523552781   Root MSE        =    .65143

------------------------------------------------------------------------------
      loguhe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    asianpac |  -.0109312   .0095219    -1.15   0.251    -.0295942    .0077317
     yearsEd |   .1301684   .0011751   110.78   0.000     .1278653    .1324715
       _cons |   1.469512   .0171914    85.48   0.000     1.435816    1.503207
------------------------------------------------------------------------------

65 .
66 . bys colgrad: reg loguhe asianpac

-------------------------------------------------------------------------------------------------------------------
-> colgrad = 0

      Source |       SS           df       MS      Number of obs   =    29,903
-------------+----------------------------------   F(1, 29901)     =     25.12
       Model |  9.74715785         1  9.74715785   Prob > F        =    0.0000
    Residual |  11600.8634    29,901  .387975766   R-squared       =    0.0008
-------------+----------------------------------   Adj R-squared   =    0.0008
       Total |  11610.6105    29,902  .388288761   Root MSE        =    .62288

------------------------------------------------------------------------------
      loguhe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    asianpac |  -.0755548   .0150739    -5.01   0.000    -.1051003   -.0460093
       _cons |   3.097319   .0037168   833.34   0.000     3.090034    3.104604
------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------------------
-> colgrad = 1

      Source |       SS           df       MS      Number of obs   =    23,847
-------------+----------------------------------   F(1, 23845)     =     29.15
       Model |  14.2407638         1  14.2407638   Prob > F        =    0.0000
    Residual |  11647.2907    23,845  .488458407   R-squared       =    0.0012
-------------+----------------------------------   Adj R-squared   =    0.0012
       Total |  11661.5315    23,846   .48903512   Root MSE        =     .6989

------------------------------------------------------------------------------
      loguhe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    asianpac |    .068885   .0127577     5.40   0.000     .0438791    .0938908
       _cons |   3.688318   .0049022   752.39   0.000      3.67871    3.697927
------------------------------------------------------------------------------
```

1.2 Regression Anatomy

- Equations (4) don't immediately reveal just how multivariate regression works its matching magic.
- Here's a better way. Start with

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \tag{5}$$

- Consider the following two *auxiliary regressions*:

$$\begin{aligned} X_{1i} &= \delta_{10} + \delta_{12} X_{2i} + \tilde{x}_{1i} \\ X_{2i} &= \delta_{20} + \delta_{21} X_{1i} + \tilde{x}_{2i} \end{aligned}$$

where the δ's are bivariate regression coefficients [e.g., $\delta_{12} = COV(X_{1i},\, X_{2i})/V(X_{2i})$]

Regression-anatomy theorem.

$$\begin{aligned} \beta_1 &= COV(Y_i, \tilde{x}_{1i})/V(\tilde{x}_{1i}) \\ \beta_2 &= COV(Y_i, \tilde{x}_{2i})/V(\tilde{x}_{2i}) \end{aligned}$$

Proof. Substitute for Y using $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$, where ε_i is mean-zero and uncorrelated with the regressors by definition.

- The multivariate β_1 captures the effect of $\tilde{x}_{1i}$, the part of X_1 that is not explained (in a regression sense) by X_2
- The multivariate β_2 captures the effect of $\tilde{x}_{2i}$, the part of X_2 that is not explained (in a regression sense) by X_1

```
70 . reg loguhe asianpac yearsEd age

      Source |       SS           df       MS      Number of obs   =    53,750
-------------+----------------------------------   F(3, 53746)     =   4223.15
       Model |  5368.08594         3  1789.36198   Prob > F        =    0.0000
    Residual |  22772.3525    53,746  .423703205   R-squared       =    0.1908
-------------+----------------------------------   Adj R-squared   =    0.1907
       Total |  28140.4384    53,749  .523552781   Root MSE        =    .65092

------------------------------------------------------------------------------
      loguhe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    asianpac |   -.008028   .0095198    -0.84   0.399    -.0266869    .0106309
     yearsEd |   .1302794   .0011742   110.95   0.000     .1279779    .1325808
        agep |   .0090692   .0009906     9.16   0.000     .0071276    .0110107
       _cons |   1.062983   .0476094    22.33   0.000     .9696681    1.156298
------------------------------------------------------------------------------

71 .
72 . **step 1
73 .
74 . reg asianpac age yearsEd if e(sample)==1

      Source |       SS           df       MS      Number of obs   =    53,750
-------------+----------------------------------   F(2, 53747)     =    766.95
       Model |  133.428423         2  66.7142115   Prob > F        =    0.0000
    Residual |  4675.24747    53,747  .086986203   R-squared       =    0.0277
-------------+----------------------------------   Adj R-squared   =    0.0277
       Total |  4808.67589    53,749  .089465402   Root MSE        =    .29493

------------------------------------------------------------------------------
    asianpac |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        agep |   -.003466   .0004486    -7.73   0.000    -.0043452   -.0025868
     yearsEd |   .0200877   .0005249    38.27   0.000     .0190588    .0211166
       _cons |  -.0381703   .0215712    -1.77   0.077      -.08045    .0041095
------------------------------------------------------------------------------

75 . predict ap_resid, residuals

76 .
77 . **step 2
78 .
79 . reg loguhe ap_resid

      Source |       SS           df       MS      Number of obs   =    53,750
-------------+----------------------------------   F(1, 53748)     =      0.58
       Model |   .30131172         1   .30131172   Prob > F        =    0.4481
    Residual |  28140.1371    53,748  .523556915   R-squared       =    0.0000
-------------+----------------------------------   Adj R-squared   =   -0.0000
       Total |  28140.4384    53,749  .523552781   Root MSE        =    .72357

------------------------------------------------------------------------------
      loguhe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    ap_resid |   -.008028   .0105823    -0.76   0.448    -.0287693    .0127134
       _cons |   3.361481    .003121  1077.06   0.000     3.355364    3.367599
------------------------------------------------------------------------------

80 .
81 . log close
```

- It works! Phew!

2 Estimation and Inference

- Multivariate regression is our bread and butter! It is our version of the clinician's stratified RCT and the laboratory scientist's "controlled experiment" (but cheaper, no gloves needed, and much easier clean-up when we're done)
- We construct estimators by replacing sample moments with population moments (In practice, Stata does this for us)
- The tools of regression inference include:
  1. t-tests and coefficient standard errors
  2. F-statistics for joint tests
- Details done in MM and MHE

3 Regression, Causality, and Control

The Dale and Krueger (DK; 2002) study looks at difference in earnings between graduates of more and less selective colleges, as measured by the average SAT scores at their schools. To make this into a Bernoulli treatment, we look here (and in *MM*, Chpt 2) at a dummy for graduation from a private institution (which are also more selective than public, on average). Two of my former Ph.D. students were admitted to Harvard yet attended their local state (public) schools. Today, these students are professors in top econ departments - not bad! But perhaps they would have done better if they attended (private) Harvard instead. Who knows, they might even have found jobs on Wall Street!

These are just two data points, of course. But in larger and more representative samples, comparisons between private and state school graduates consistently show higher earnings for those who went private. No surprise! *Something* must justify the many thousands of dollars these schools collect from their students.

On the other hand, part of the difference in earnings between private and public college grads is surely attributable to differences in the characteristics ($Y_{0i}'s$) of people who did and didn't attend private schools. Variables that are likely to differ with school type include students' own SAT scores (which are correlated with their earnings), the kinds of school they applied to (which says something about students' own judgements of their ability) and family income (which is also correlated with earnings).

- We'd like to hold these things constant, that is, to "control" for them when comparing groups of students who went to different types of schools
- Such control brings us one giant step closer to an ideal experiment

3.1 The Payoff to Private College

The DK research design, as implemented in Chapter 2 of MM, looks at students who applied to and were admitted to schools of similar selectivity.

- Consider a hypothetical set of applicants, all of whom applied to one or more schools among three, Ivy, Leafy, and Smart. The matching matrix these students face appears below:

		Private			Public			
Applicant Group	Student	Ivy	Leafy	Smart	All State	Ball State	Altered State	1996 Earnings
	1		Reject	**Admit**		Admit		110,000
A	2		Reject	**Admit**		Admit		100,000
	3		Reject	Admit		**Admit**		110,000
B	4	**Admit**			Admit		Admit	60,000
	5	Admit			Admit		**Admit**	30,000
C	6		**Admit**					115,000
	7		**Admit**					75,000
D	8	Reject			**Admit**	Admit		90,000
	9	Reject			Admit	**Admit**		60,000

Notes: Students enroll at the college indicated in **bold**; enrollment decisions are also highlighted in grey.

- Five of nine students (numbers 1,2,4,6,7) attended private schools. Average earnings in this group are \$92,000. The other four, with average earnings of \$72,500, went to a public school. The almost \$20,000 gap between these two groups suggests a large private school advantage.

The hypothesis motivating a DK-style analysis is that, conditional on the identity (or selectivity) of schools that I've applied to, and the identity (or selectivity) of schools that have admitted me, comparisons of students who went to different schools (say, one to public and one to private) are more likely to be "apples to apples." In other words, we uncover the effects of private school attendance by ...

- Comparing students 1 and 2 with student 3 in group A and by comparing student 4 and student 5 in Group B
- Discarding students in groups C and D (why?)
- The average of the -5 thousand dollars gap for group A and the 30,000 gap dollars for group B is \$12,500. This is a good estimate of the effect of private school attendance on average earnings because it controls (at least partially) for applicants' ambition and ability
- Notice that overall earnings in Group A are much higher than overall average earnings in group B. Our within-group matching estimate of 12,500 eliminates this source of bias in our causal inquiry

Instead of averaging these group-specific contrasts by hand, regress!

- With only one control variable, A_i, the regression of interest can be written:

$$Y_i = \alpha + \beta P_i + \gamma A_i + \varepsilon_i \tag{6}$$

- The distinction between the causal variable, P_i, and the control variable, A_i, in equation (6) is conceptual, not formal: there is nothing in equation (6) to indicate which is which.
- Using data for the five students in Groups A and B generates $\beta = 10,000$ and $\gamma = 60,000$. The private school coefficient in this case is $10,000$, close to the we got by averaging the public-private contrasts within groups A and B and well below the raw public-private difference of almost 20,000.

Public-Private Face-Off

The *College and Beyond* (C&B) data set includes over 14,000 college graduates who attended 30 schools. We can increase the number of useful comparisons by deeming schools to be "matched" if they are equally selective instead of insisting on identical matches.

- To fatten up the selectivity categories, we'll call schools comparable if they fall into the same Barron's selectivity categories

In the College and Beyond data, 9,202 students can be matched in this way. Because we're interested in public-private comparisons, however, our Barron's matched sample is also limited to matched applicant groups that contain both public and private school graduates. This leaves 5,583 matched applicants for analysis. These matched applicants fall into 151 different selectivity groups containing both public and private graduates.

Our operational regression model for the Barron's selectivity-matched sample includes many control variables, while the stylized example controls only for the dummy variable A_i, indicating students in group A. The key controls in the operational model consist of a set of many dummy variables indicating all Barron's matches represented in the sample (with one group left out as a reference category). These controls capture the relative selectivity of the schools to which applicants have applied and been admitted in the real world, where many combinations of schools are possible. The resulting regression model looks like this:

$$\ln Y_i = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j GROUP_{ji} + \delta_1 SAT_i + \delta_2 \ln PI_i + \varepsilon_i \tag{7}$$

- The parameter β in this model is still the coefficient of interest, an estimate of the causal effect of attendance at a private school
- This model controls for 151 groups instead of the two groups in our stylized example. The parameters γ_j, for $j = 1$ to 150, are the coefficients on 150 selectivity-group dummies, denoted $GROUP_{ji}$
- The variable $GROUP_{ji}$ equals 1 whenever student i is in group j and is 0 otherwise; the summation symbol, $\sum_{j=1}^{150}$, indicates a sum from $j = 1$ to 150
- We add two further control variables: individual SAT scores and the log of parental income, plus a few more we haven't bother to write out

	No Selection Controls			Selection Controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private School	0.135	0.095	0.086	0.007	0.003	0.013
	(0.055)	(0.052)	(0.034)	(0.038)	(0.039)	(0.025)
Own SAT score/100		0.048	0.016		0.033	0.001
		(0.009)	(0.007)		(0.007)	(0.007)
Predicted log(Parental Income)			0.219			0.190
			(0.022)			(0.023)
Female			-0.403			-0.395
			(0.018)			(0.021)
Black			0.005			-0.040
			(0.041)			(0.042)
Hispanic			0.062			0.032
			(0.072)			(0.070)
Asian			0.170			0.145
			(0.074)			(0.068)
Other/Missing Race			-0.074			-0.079
			(0.157)			(0.156)
High School Top 10 Percent			0.095			0.082
			(0.027)			(0.028)
High School Rank Missing			0.019			0.015
			(0.033)			(0.037)
Athlete			0.123			0.115
			(0.025)			(0.027)
Selection Controls	N	N	N	Y	Y	Y

Notes: Columns (1)-(3) include no selection controls. Columns (4)-(6) include a dummy for each group formed by matching students according to schools at which they were accepted or rejected. Each model is estimated using only observations with Barron's matches for which different students attended both private and public schools. The sample size is 5,583. Standard errors are shown in parentheses.

- Perhaps it's enough to control linearly for the average SAT scores of the schools to which I'm admitted, as well as the number to which I apply. Here's how that comes out:

	No Selection Controls			Selection Controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private School	0.212	0.152	0.139	0.034	0.031	0.037
	(0.060)	(0.057)	(0.043)	(0.062)	(0.062)	(0.039)
Own SAT Score/100		0.051	0.024		0.036	0.009
		(0.008)	(0.006)		(0.006)	(0.006)
Predicted log(Parental Income)			0.181			0.159
			(0.026)			(0.025)
Female			-0.398			-0.396
			(0.012)			(0.014)
Black			-0.003			-0.037
			(0.031)			(0.035)
Hispanic			0.027			0.001
			(0.052)			(0.054)
Asian			0.189			0.155
			(0.035)			(0.037)
Other/Missing Race			-0.166			-0.189
			(0.118)			(0.117)
High School Top 10 Percent			0.067			0.064
			(0.020)			(0.020)
High School Rank Missing			0.003			-0.008
			(0.025)			(0.023)
Athlete			0.107			0.092
			(0.027)			(0.024)
Average SAT Score of				0.110	0.082	0.077
Schools Applied to/100				(0.024)	(0.022)	(0.012)
Sent Two Application				0.071	0.062	0.058
				(0.013)	(0.011)	(0.010)
Sent Three Applications				0.093	0.079	0.066
				(0.021)	(0.019)	(0.017)
Sent Four or more Applications				0.139	0.127	0.098
				(0.024)	(0.023)	(0.020)

Note: Standard errors are shown in parentheses. The sample size is 14,238.

- This buys us a larger sample and doesn't much change the results

- What about school selectivity instead of the public/private distinction? Here's a model much like DK's original:

	No Selection Controls			Selection Controls		
	(1)	(2)	(3)	(4)	(5)	(6)
School Avg. SAT Score/100	0.109	0.071	0.076	-0.021	-0.031	0.000
	(0.026)	(0.025)	(0.016)	(0.026)	(0.026)	(0.018)
Own SAT score/100		0.049	0.018		0.037	0.009
		(0.007)	(0.006)		(0.006)	(0.006)
Predicted log(Parental Income)			0.187			0.161
			(0.024)			(0.025)
Female			-0.403			-0.396
			(0.015)			(0.014)
Black			-0.023			-0.034
			(0.035)			(0.035)
Hispanic			0.015			0.006
			(0.052)			(0.053)
Asian			0.173			0.155
			(0.036)			(0.037)
Other/Missing Race			-0.188			-0.193
			(0.119)			(0.116)
High School Top 10 Percent			0.061			0.063
			(0.018)			(0.019)
High School Rank Missing			0.001			-0.009
			(0.024)			(0.022)
Athlete			0.102			0.094
			(0.025)			(0.024)
Average SAT Score of				0.138	0.116	0.089
Schools Applied To/100				(0.017)	(0.015)	(0.013)
Sent Two Application				0.082	0.075	0.063
				(0.015)	(0.014)	(0.011)
Sent Three Applications				0.107	0.096	0.074
				(0.026)	(0.024)	(0.022)
Sent Four or more Applications				0.153	0.143	0.106
				(0.031)	(0.030)	(0.025)

Note: Standard errors are shown in parentheses. The sample size is 14,238.

- Pity my poor parents, whom I made a little poorer by attending Oberlin, a pricey private college. It seems I could just as well have gone to Penn State!