

Political Science 150B/350B
Final Exam
Winter 2005
Answer Guide

Question 1: Needless to say, running a marathon is hard work, and doing so in hot weather is likely to reduce runners' performance. Data from the New York City marathon confirms this fact. The male winners' time from the 1978-1998 NYC marathons (in minutes) is regressed on temperature and temperature squared (temperature is measured in Fahrenheit, F), yielding the following results:

$$E(\text{Time}_t|F_t) = 148.51 - .71F_t + .0065F_t^2$$

where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)' = (148.51, -.71, .0065)'$ and are all more than twice the size of their standard errors, and $t = 1978, \dots, 1998$ indexes the marathons.¹

(a): (5 points) What is the optimal temperature for running the NYC marathon? [Hint: this is a math problem, not a statistics problem]

Answer: Simple calculus. Differentiate time wrt F , set to zero, solve for F^* :

$$\begin{aligned}\frac{\partial T}{\partial F} &= -.71 + 2 \times .0065F \\ F^* &= \frac{.71}{2 \times .0065} \\ F^* &= 54.6\end{aligned}$$

(b): (5 points) How would your answer change if $\hat{\beta}_2$ was only half the size of its standard error?

Answer: Then we're left with a linear term in temperature, with a statistically significant negative sign, suggesting that as it gets hotter (and hotter), times get faster (and faster).

(c): (5 points) How would you augment the estimated regression model to test for the possibility that (net of year-to-year fluctuations

¹See David E. Martin and John F. Buoncristiani, "The Effect of Temperature on Marathon Runners' Performance", *Chance*, Vol. 12, No. 4 (Fall 1999), pp. 20-24; my source is Orley Ashenfelter, Phillip B. Levine and David J. Zimmerman (2003), *Statistics and Econometrics: Methods and Applications*, Wiley, New York, p193.

in wining times due to temperature) winning times have been improving over the years?

Answer: Include a time trend in the model; i.e., let t , the time index, be a variable in the model. If this variable picks up a negative sign, then winning times are going down over the years, net of the temperature effect.

(d): (5 points) How would you test the possibility that over time, winning times have become less sensitive to race-day temperature?

Answer: Interact F_t and F_t^2 with the time counter. So the full model would be

$$\text{Time}_t = \beta_0 + \beta_1 F_t + \beta_2 F_t^2 + \beta_3 t + \beta_4 (F_t \times t) + \beta_5 (F_t^2 \times t) + u_t$$

Question 2: Economists have conjectured that cigarette smokers may be penalized with lower wages in the job market: for instance, smokers may receive lower levels of pay because they are less productive (if the act of smoking takes a worker away from his or her job); because they may be absent from work more often (due to their greater risk of respiratory infections); because it is more expensive to provide them with health insurance; or simply because firms discriminate against them. A regression analysis of log wages on a dummy variable for smoking and other predictors of wages is summarized in the following table (standard errors in parentheses):²

	Model 1	Model 2	Model 3
Smoking	-.176 (.021)	-.080 (.021)	-.069 (.019)
Education		.070 (.004)	.045 (.005)
Other factors included	no	no	yes

(a): (10 points) In two or three sentences, explain why the estimated impact of smoking on wages decreases in magnitude across the three models?

²See Philip B. Levine, Tara A. Gustafson and Ann D. Velenchik, "More Bad News for Smokers? The Effects of Cigarette Smoking on Wages," *Industrial and Labor Relations Review*, Vol. 50, No. 3 (April 1997), pp. 493-509; my source is Orley Ashenfelter, Phillip B. Levine and David J. Zimmerman (2003), *Statistics and Econometrics: Methods and Applications*, Wiley, New York, p190.

Answer: Omitted variable bias. When appearing on its own, the smoking variable is proxying for other predictors of wages that are correlated with smoking. In particular, education and “other factors” are probably negatively correlated with smoking, but have positive impacts on wages, and so the effect of smoking diminishes (is pushed up towards zero) as education and “other factors” enter the model. From my lecture notes:

$X_1 X_2$	β_2	Bias: $E(\hat{\beta}_1) - \beta_1$	$\hat{\beta}_1$ is
positive	positive	positive	over-estimated
positive	negative	negative	under-estimated
negative	positive	negative	under-estimated
negative	negative	positive	over-estimated

The situation here corresponds to the 3rd row.

(b): (5 points) Who incurs the higher wage penalty for smoking in absolute terms, relatively higher-paid or lower-paid workers?

Answer: Simple calculus. Note that the coefficient on smoking is always negative, so

$$\frac{\partial \log w}{\partial s} = \lambda < 0$$

But

$$\begin{aligned} \frac{\partial \log w}{\partial s} &= \frac{\partial \log w}{\partial w} \frac{\partial w}{\partial s} \\ &= \frac{1}{w} \frac{\partial w}{\partial s} = \lambda < 0. \end{aligned}$$

So the change in wages as a function of smoking is

$$\frac{\partial w}{\partial s} = \lambda w$$

which is obviously getting larger in magnitude (larger wage penalties) as w increases. Indeed,

$$\frac{\partial \frac{\partial w}{\partial s}}{\partial w} = \lambda < 0$$

shows that the wage decrement is not constant with respect to w , but increases in magnitude over w at rate λ .

As an aside, note that one of the nice features of the log model is that relative change is constant. That is, if

$$\frac{\partial w}{\partial s} = \lambda w.$$

then as a proportion of w

$$\frac{\lambda w}{w} = \lambda$$

Question 3: (4 points) Let $\mathbf{y} = \Xi\boldsymbol{\beta} + \mathbf{u}$ be a statistical model of substantive interest. A researcher analyzes the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{w}$, using least squares to estimate $\boldsymbol{\beta}^*$, where $\mathbf{X} = \Xi + \mathbf{v}$, and

$$\begin{aligned} E(\mathbf{v}'\mathbf{w}) &= \mathbf{0} \\ E(\Xi'\mathbf{v}) &= \mathbf{0} \\ E(\Xi'\mathbf{w}) &= \mathbf{0} \\ E(\mathbf{v}) &= \mathbf{0} \\ \text{var}(\mathbf{w}) &= \sigma_w^2 \mathbf{I} \\ \text{var}(\mathbf{v}) &= \sigma_v^2 \mathbf{I} \end{aligned}$$

Is $\hat{\boldsymbol{\beta}}^*$ an unbiased estimator of $\boldsymbol{\beta}^*$?

Answer: No. In general, measurement error in an independent variable generates bias in the estimated regression coefficients. In the case of a single independent variable, the bias is downwards.

Question 4: A colleague is interested in the effects of campaigns on U.S. presidential elections. She has data on Clinton's vote margin in 1992 and in 1996 (Clinton vote share minus the Republican candidate vote share) for each state, denoted CVM_{it} , where i indexes states, and $t \in \{92, 96\}$. She also has data on states targeted in a media campaign in 1996 by Clinton's 1996 political consultant Dick Morris, coded as a binary indicator $D_i = 1$ if state i was a target state, and 0 otherwise.

(a): (10 points) What *substantive* assumptions are being made with the model

$$CVM_{i96} = CVM_{i92} + \varepsilon_i$$

Answer: On average, Clinton's vote margin is constant (the same in 1996 as it was in 1992).

- (b): (7 points) Offer a *substantive* interpretation of the intercept in the following model

$$CVM_{i96} = \alpha + \beta CVM_{i92} + \varepsilon_i$$

Answer: Consider $\alpha = 0$, which means that $E(CVM_{i96} | CVM_{i92} = 0) = 0$, or, on average, states which evenly split between Clinton and Bush also evenly split between Clinton and Dole. If $\alpha > 0$ it means that there has been an overall “swing” towards Clinton from 1992 to 1996, such that states that evenly split in 1992 now have an average Clinton vote margin of $\alpha > 0$; indeed, *irrespective* of the 1992 vote margin, Clinton does better in 1996, on average.

- (c): (7 points) Assume CVM_{i92} is a reasonable measure of a baseline level of support for Clinton in state i . How would you augment the model in the previous question so as to test for the effects of Dick Morris’ media campaign?

Answer: Add D_i as a regressor; the coefficient picks up an additional intercept shift for targeted states.

Question 5: (5 points) Multicollinearity means

- (a): Regression analysis can not proceed because the matrix $\mathbf{X}'\mathbf{X}$ can not be inverted.
- (b): Conditional on the predictors \mathbf{X} , the disturbances \mathbf{u} are not iid.
- (c): Because the predictors \mathbf{X} are correlated with one another, estimates of their partial impact on \mathbf{y} (i.e., “controlling for” one another) are less precise than if the predictors were uncorrelated with one another.
- (d): A critical predictor of \mathbf{y} has been omitted from the regression model.

Answer: C.

Question 6: (5 points) When the disturbances are not iid, it is well known that OLS estimates of the regression coefficients $\hat{\beta}_{OLS}$ that are unbiased but not BLUE. GLS can be used for this situation, yielding estimated coefficients $\hat{\beta}_{GLS}$ that are also unbiased. Thus, if OLS and GLS are applied to a data set with disturbances that are not iid:

- (a): $\hat{\beta}_{OLS} = \hat{\beta}_{GLS}$
- (b): $\hat{\beta}_{OLS} \neq \hat{\beta}_{GLS}$

- (c): each element of the vector $\hat{\beta}_{OLS}$ is smaller than the corresponding element of $\hat{\beta}_{GLS}$
- (d): each element of the vector $\hat{\beta}_{OLS}$ is larger than the corresponding element of $\hat{\beta}_{GLS}$

Answer: B.

Question 7: (5 points) A regression's residuals are likely to be highly autocorrelated if

- (a): the estimated t -statistics are smaller than 2 in absolute value
- (b): the X variables have distinct time trends
- (c): the dependent variable y rises and falls over time but none of the X variables have a similar pattern
- (d): we include a linear time trend as one of the predictors in the model

Answer: C.

Question 8: Figure 1 presents the results of a Monte Carlo experiment, which assessed the repeated sampling properties of two linear estimators of θ , $\hat{\theta}_A$ and $\hat{\theta}_B$, at a fixed sample size n . θ is known to be 1.0.

- (a): (10 points) Based on Figure 1, what can you say about the statistical properties of estimators A and B ?

Answer: Both estimators A and B are unbiased. B has smaller sampling variation than A and so can be said to be "more efficient" than A .

- (b): (5 points) Do you have sufficient information to conclude that either estimator is BLUE?

Answer: No. While B is better than A (or more efficient, in the sense of having smaller sampling variability) it could well be that there is some other estimator C that is better, and is the best linear unbiased estimator of θ . Quite simply, we haven't been told enough to rule out this possibility.

Question 9: Consider the following regression analysis, evaluating a remedial reading project in a large school district. The data are a random sample of 125 students from across the school district who completed the fifth grade and who had been identified as "slow readers" in the previous year (more than a year behind the average 4th grade students). The project

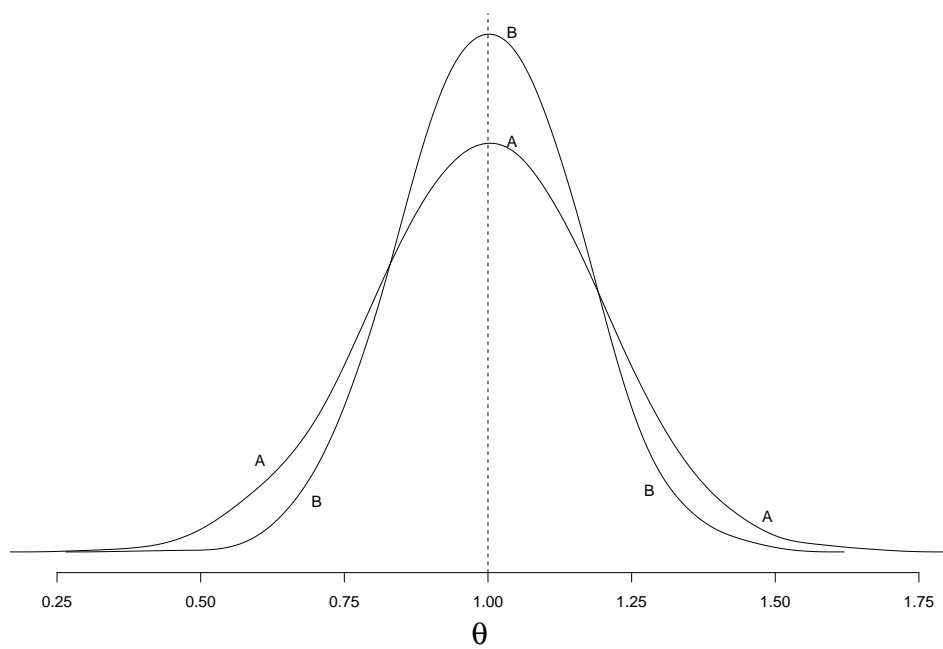


Figure 1: Results of Monte Carlo Experiment, Linear Estimators of θ .

evaluation measure (the dependent variable of the regression analysis, *RCHANGE*) is the number of reading points gained in the sample year. Average students gain 100 reading points a year.

Two treatments were administered to the students: remedial group reading sessions, and individual tutoring sessions. These two treatments are measured in hours per week, for each student, with the variables *GROUP* and *TUTOR*, respectively. The variable *COHORT* measures a “baseline growth” or “maturation effect”: i.e., $COHORT_i$ is defined as the average gain in reading points in the classroom of slow reader i .

<i>Parameter</i>	<i>Estimate</i>	<i>Std Err</i>
<i>COHORT</i>	.81	.19
<i>GROUP</i>	5.66	.68
<i>TUTOR</i>	12.52	.90
<i>Intercept</i>	-33.32	19.11
<i>Adjusted r²</i>	.63	
$\hat{\sigma}$	12.05	
<i>Mean of dep. var</i>	87.91	

(a): (18 points) Fully interpret each of the three slope parameters. Think carefully about the effect of the *COHORT* variable (8 points for the interpretation of this coefficient; 5 each for the other slope coefficients).

Answer:

- *TUTOR*. Highly statistically significant, and large in substantive terms. Net of other effects, each hour per week of tutoring increases a student’s reading performance by about 12.5 points (on average -- recall this is just a point estimate, and the standard error gives a sense of our uncertainty of the actual magnitude of the effect).
- *GROUP*. Also highly statistically significant. Net of other effects, each hour per week spent in remedial group sessions increases reading performance by about 5.7 points, on average.
- *COHORT*. Slightly harder to interpret. This variables measures how much of the gains by *all* students translates into gains for students in the remedial reading projects. We are told that students usually gain 100 points a year, probably through maturation, but the coefficient on *COHORT* suggests that only .81 of this gain translates into gains for remedial students. That

is, net of the interventions through tutoring and group sessions (and ignoring the large, negative, though statistically insignificant intercept), slow readers exhibit an 81 point increase in reading performance over the year (on average), while average students will improve 100 points. This estimated difference is not statistically distinguishable from parity with average students; note that the standard error on the point estimate of .81 is .19. This means we can not reject the null hypothesis that the coefficient on *COHORT* is 1.0 at conventional levels of statistical significance.

(b): (5 points) Briefly describe how to test the null hypothesis that both treatments are ineffective.

Answer: An *F* test for the joint null hypothesis, $\beta_{COHORT} = 0$ and $\beta_{TUTOR} = 0$. Operationalize by running a restricted regression with *COHORT* and *TUTOR* variables omitted, compare the sum of the squared residuals or r^2 from this regression with that from the unrestricted regression (reported above, in the question).

(c): (7 points) Test the null hypothesis that the effects of each tutoring hour per week is greater than 11.5.

Answer: The null hypothesis can be expressed as

$$H_0 : \beta_{TUTOR} = 11.5$$

or, equivalently,

$$H_0 : \beta_{TUTOR} - 11.5 = 0$$

The relevant *t*-statistic is

$$\begin{aligned} t &= \frac{\hat{\beta}_{TUTOR} - 11.5}{se(\hat{\beta}_{TUTOR})} \\ &= \frac{12.52 - 11.5}{.90} \\ &= \frac{1.02}{.90} \\ &\approx 1.13 \end{aligned}$$

Since the alternative hypothesis is one sided, we are interested in the proportion of the *t* distribution that lies to the right of 1.13. With $n - k = 125 - 4 = 121$ degrees of freedom, the critical $\alpha = .05$ *t*-statistic for a one-sided test is about 1.66. We therefore fail to reject the null

that the estimated effects of hours of tutoring per week is 11.5 in favor of the one-sided alternative. The actual p -value for the t -test here is approximately .13 (i.e., 13% of the t -distribution lies to the right of 1.13, whereas only 5% lies to the right of about 1.66).

Another way to approach this problem is to construct a 90% confidence bound around the point estimate of β_{TUTOR} . We use a 90% confidence interval because we want lower 5% of the distribution to lie to the left of the confidence interval, since this is a one-sided test (versus putting 2.5% of the probability mass on either side of a confidence interval for a two-sided test). The point estimate is 12.52, and the standard error is .90. The critical value of t for $df = 121$ and $\alpha = .05$ (one-tailed) is 1.66, and so a lower 90% bound is $12.52 - 1.66 \times .90 \approx 12.52 - 1.49 = 11.03$. The hypothesized value of 11.5 lies within the confidence interval and so we again fail to reject the null at this level of statistical significance.

- (d): (10 points) Each hour of tutoring costs about \$9 per student, and each hour of group sessions costs about \$3 per student. Test the null hypothesis that the two programs are equally cost effective. The following information will be helpful:

Lower-triangle of $var(\hat{\beta})$.

	<i>COHORT</i>	<i>GROUP</i>	<i>TUTOR</i>	<i>Intercept</i>
<i>COHORT</i>	.034783			
<i>GROUP</i>	.0074	.456218		
<i>TUTOR</i>	.003512	.419009	.806414	
<i>Intercept</i>	-3.50749	-2.66247	-3.05492	365.024

Answer: The null hypothesis about cost-effectiveness means we want to see if the per dollar effect of a tutoring hour equals the per dollar effect of a group session hour. Since tutoring costs \$9/hr and group sessions cost \$3/hr, we are interested in the null hypothesis

$$H_0 : \frac{\beta_{TUTOR}}{9} = \frac{\beta_{GROUP}}{3}$$

or

$$H_0 : \frac{\beta_{TUTOR}}{9} - \frac{\beta_{GROUP}}{3} = 0$$

or still more simply,

$$H_0 : \beta_{TUTOR} - 3\beta_{GROUP} = 0$$

This can be tested with a t -test. Let q be our sample estimate of the quantity to be tested; i.e.,

$$q = \frac{\hat{\beta}_{TUTOR}}{9} - \frac{\hat{\beta}_{GROUP}}{3}$$

Then the test statistic here is simply $t = q/se(q)$. The real work here is to derive the standard error of q . To do this we first require $var(q)$, noting that

$$q = k_1 a - k_2 b$$

where k_1 and k_2 are constants (here I will use $1/9$ and $1/3$, respectively) and a and b are random variables ($\hat{\beta}_{TUTOR}$ and $\hat{\beta}_{GROUP}$, respectively). Substituting yields

$$\begin{aligned} q &= \frac{1}{9}12.52 - \frac{1}{3}5.66 \\ &\approx 1.40 - 1.87 \\ &\approx -.49556 \end{aligned}$$

An elementary result in mathematical statistics is that

$$var(q) = k_1^2 var(a) + k_2^2 var(b) - 2 k_1 k_2 cov(a, b)$$

Substituting for k_1 , k_2 , and from the variance-covariance matrix supplied in the question yields

$$\begin{aligned} var(q) &= (1/9)^2 .80614 + (1/3)^2 .456218 - 2(1/9)(1/3) .419009 \\ &\approx .02961 \end{aligned}$$

$$\text{and so } se(q) = \sqrt{var(q)} \approx \sqrt{.02961} \approx .17201$$

Accordingly, the appropriate t -statistic here is $-.49556/.17201 \approx -2.88$. We can thus reject the null hypothesis that the two programs are cost effective. In fact, the large negative t -statistic allows us to be quite confident that group sessions are more cost effective than tutoring ($p \approx .002$, one-tailed).

Question 10: (10 points) A researcher estimate a regression model with panel data. She finds that without fixed effects for the observational units, the sign of the regression coefficient for an important predictor is negative, and the r^2 is about .15. However, when the researcher includes fixed effects for each unit, the regression coefficient on the important predictor

is positive, and the r^2 is about .85. The researcher comes to your office hours and seeking methodological guidance. What is going on in her data?

Answer: (3 points for this first part of between-variation). The jump in r^2 from .15 to .85 suggests that much of the variation in the data is between-unit variation, rather than within-unit variation. That is, the values of the dependent variable display more variation across the units than within the units; a variable that taps differences across units will fit well (e.g., the unit-specific fixed effects).

(7 points; partial credit for getting pieces of this; full credit if the student could sign the correlation between the unit-specific effects and the predictor, note that the full proof isn't necessary). The regression coefficient on the important predictor changes sign because the predictor must be correlated with the unit-specific effects. In particular, with the unit-specific effects omitted from the regression model we obtain an underestimate of the effect relative to what we obtain with the unit-specific effects included in the model. That is, we have two models:

$$\begin{aligned} y_{ij} &= x_{ij}\beta + u_{ij} \\ y_{ij} &= x_{ij}\tilde{\beta} + \alpha_i + \varepsilon_{ij} \end{aligned}$$

that is, $u_{ij} = \alpha_i + \varepsilon_{ij}$. The bias in a least squares regression estimate of β is

$$E(\hat{\beta}) - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\alpha + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$$

but under the usual assumptions, $E(\mathbf{X}'\varepsilon) = 0$ and so the bias term is simply $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\alpha$. Since $\hat{\beta}$ is an underestimate of $\tilde{\beta}$ (the regression coefficient that results from *including* the unit-specific effects), the bias term is negative, which in turn means that $E(\mathbf{X}'\alpha) < 0$, or, in other words, the unit-specific effects are negatively correlated with the predictor. This is an instance of Simpson's paradox; Figure 2 provides a visual display of what is going on here.

Question 11: Suppose a researcher is interested in whether attending class is useful or not and estimates the following model:

$$\text{GRADGPA} = \beta_0 + \beta_1\text{COLGPA} + \beta_2\text{GRE} + \beta_3\text{SKIPPED} + u$$

where GRADGPA is GPA in graduate school, COLGPA is college GPA, GRE is the student's GRE score, and SKIPPED is the average number of classes skipped per week. We believe that u includes, among other things, how lazy the student is.

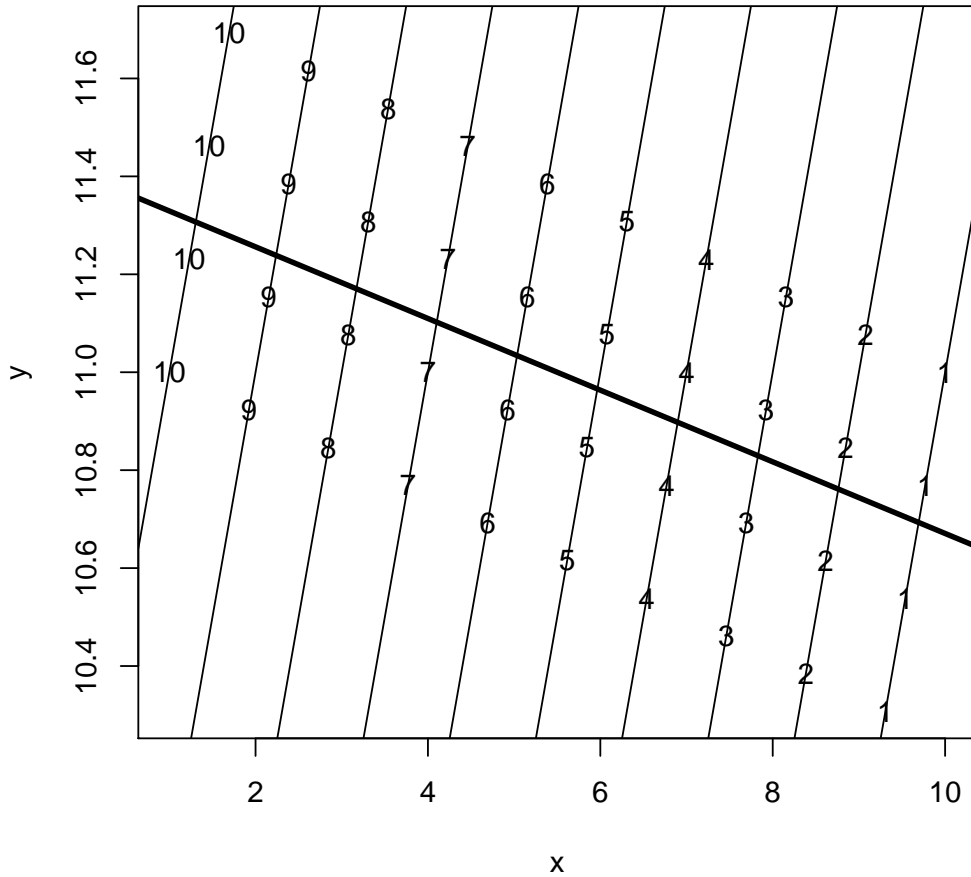


Figure 2: **Simpson's Paradox. 10 units, each with 4 observations.** There is a positive relationship between y and x within each unit. But the naive regression of y on x yields a negative slope, because the units differ systematically with respect to x , in a way that masks the positive within-unit relationship between x and y .

- (a):** (7 points) Discuss whether OLS estimates of this equation will be biased or not. If biased, can you say anything about the direction of bias?
- (b):** (10 points) You also have data on the distance (in miles) that each student lives from campus, denoted DIST. Describe how you might use this information to obtain consistent estimates of the effects of class attendance.