Propensity Score Matching
Regression Discontinuity
Limited Dependent Variables

Christopher F Baum

ECON 8823: Applied Econometrics

Boston College, Spring 2015

Propensity score matching

Policy evaluation seeks to determine the effectiveness of a particular intervention. In economic policy analysis, we rarely can work with experimental data generated by purely random assignment of subjects to the treatment and control groups. Random assignment, analogous to the 'randomized clinical trial' in medicine, seeks to ensure that participation in the intervention, or treatment, is the only differentiating factor between treatment and control units.

In non-experimental economic data, we observe whether subjects were treated or not, but in the absence of random assignment, must be concerned with differences between the treated and non-treated. For instance, do those individuals with higher aptitude self-select into a job training program? If so, they are not similar to corresponding individuals along that dimension, even though they may be similar in other aspects.

Propensity score matching

Policy evaluation seeks to determine the effectiveness of a particular intervention. In economic policy analysis, we rarely can work with experimental data generated by purely random assignment of subjects to the treatment and control groups. Random assignment, analogous to the 'randomized clinical trial' in medicine, seeks to ensure that participation in the intervention, or treatment, is the only differentiating factor between treatment and control units.

In non-experimental economic data, we observe whether subjects were treated or not, but in the absence of random assignment, must be concerned with differences between the treated and non-treated. For instance, do those individuals with higher aptitude self-select into a job training program? If so, they are not similar to corresponding individuals along that dimension, even though they may be similar in other aspects.

The key concern is that of similarity. How can we find individuals who are similar on all observable characteristics in order to match treated and non-treated individuals (or plants, or firms...) With a single measure, we can readily compute a measure of distance between a treated unit and each candidate match. With multiple measures defining similarity, how are we to balance similarity along each of those dimensions?

The method of *propensity score matching* (PSM) allows this matching problem to be reduced to a single dimension: that of the propensity score. That score is defined as the probability that a unit in the full sample receives the treatment, given a set of observed variables. If all information relevant to participation and outcomes is observable to the researcher, the propensity score will produce valid matches for estimating the impact of an intervention. Thus, rather than matching on all values of the variables, individual units can be compared on the basis of their propensity scores alone.

The key concern is that of similarity. How can we find individuals who are similar on all observable characteristics in order to match treated and non-treated individuals (or plants, or firms...) With a single measure, we can readily compute a measure of distance between a treated unit and each candidate match. With multiple measures defining similarity, how are we to balance similarity along each of those dimensions?

The method of *propensity score matching* (PSM) allows this matching problem to be reduced to a single dimension: that of the propensity score. That score is defined as the probability that a unit in the full sample receives the treatment, given a set of observed variables. If all information relevant to participation and outcomes is observable to the researcher, the propensity score will produce valid matches for estimating the impact of an intervention. Thus, rather than matching on all values of the variables, individual units can be compared on the basis of their propensity scores alone.

An important attribute of PSM methods is that they do not require the functional form to be correctly specified. If we used OLS methods such as

$$y = X\beta + D\gamma + \epsilon$$

where y is the outcome, X are covariates and D is the treatment indicator, we would be assuming that the effects of treatment are constant across individuals. We need not make this assumption to employ PSM. As we will see, a crucial assumption is made on the contents of X, which should include all variables that can influence the probability of treatment.

Why use matching methods?

The greatest challenge in evaluating a policy intervention is obtaining a credible estimate of the *counterfactual:* what would have happened to participants (treated units) had they not participated? Without a credible answer, we cannot rule out that whatever successes have occurred among participants could have happened anyway. This relates to the *fundamental problem of causal inference*: it is impossible to observe the outcomes of the same unit in both treatment conditions at the same time.

The impact of a treatment on individual i, δ_i, is the difference between potential outcomes with and without treatment:

$$\delta_i = Y_{1i} - Y_{0i}$$

where states 0 and 1 corrrespond to non-treatment and treatment, respectively.

Why use matching methods?

The greatest challenge in evaluating a policy intervention is obtaining a credible estimate of the *counterfactual:* what would have happened to participants (treated units) had they not participated? Without a credible answer, we cannot rule out that whatever successes have occurred among participants could have happened anyway. This relates to the *fundamental problem of causal inference*: it is impossible to observe the outcomes of the same unit in both treatment conditions at the same time.

The impact of a treatment on individual i, δ_i, is the difference between potential outcomes with and without treatment:

$$\delta_i = Y_{1i} - Y_{0i}$$

where states 0 and 1 corrrespond to non-treatment and treatment, respectively.

To evaluate the impact of a program over the population, we may compute the average treatment effect (ATE):

$$ATE = E[\delta_i] = E(Y_1 - Y_0)$$

Most often, we want to compute the average treatment effect on the treated (ATT):

$$ATT = E(Y_1 - Y_0|D = 1)$$

where $D = 1$ refers to the treatment.

To evaluate the impact of a program over the population, we may compute the average treatment effect (ATE):

$$ATE = E[\delta_i] = E(Y_1 - Y_0)$$

Most often, we want to compute the average treatment effect on the treated (ATT):

$$ATT = E(Y_1 - Y_0|D = 1)$$

where $D = 1$ refers to the treatment.

The problem is that not all of these parameters are observable, as they rely on counterfactual outcomes. For instance, we can rewrite ATT as

$$ATT = E(Y_1|D = 1) - E(Y_0|D = 1)$$

The second term is the average outcome of treated individuals had they not received the treatment. We cannot observe that, but we do observe a corresponding quantity for the untreated, and can compute

$$\Delta = E(Y_1|D = 1) - E(Y_0|D = 0)$$

The difference between ATT and Δ can be defined as

$$\Delta = ATT + SB$$

where SB is the selection bias term: the difference between the counterfactual for treated units and observed outcomes for untreated units.

For the computable quantity Δ to be useful, the SB term must be zero. But selection bias in a non-experimental context is often sizable. For instance, those who voluntarily sign up for a teacher-training program may be the more motivated teachers, who might be more likely to do well (in terms of student test scores) even in the absence of treatment.

In other cases, the bias may not arise due to individuals self-selecting into treatment, but being selected for treatment on the basis of an interview or evaluation of their willingness to cooperate with the program. This gives rise to administrative selection bias or program placement bias.

Even in the case of a randomized experiment, participants selected for treatment may choose not to be treated, or may not comply with all aspects of the treatment regime. In this sense, even a randomized trial may involve bias in evaluating the effects of treatment, and nonexperimental methods may be required to adjust for that bias.

For the computable quantity Δ to be useful, the SB term must be zero. But selection bias in a non-experimental context is often sizable. For instance, those who voluntarily sign up for a teacher-training program may be the more motivated teachers, who might be more likely to do well (in terms of student test scores) even in the absence of treatment.

In other cases, the bias may not arise due to individuals self-selecting into treatment, but being selected for treatment on the basis of an interview or evaluation of their willingness to cooperate with the program. This gives rise to administrative selection bias or program placement bias.

Even in the case of a randomized experiment, participants selected for treatment may choose not to be treated, or may not comply with all aspects of the treatment regime. In this sense, even a randomized trial may involve bias in evaluating the effects of treatment, and nonexperimental methods may be required to adjust for that bias.

For the computable quantity Δ to be useful, the SB term must be zero. But selection bias in a non-experimental context is often sizable. For instance, those who voluntarily sign up for a teacher-training program may be the more motivated teachers, who might be more likely to do well (in terms of student test scores) even in the absence of treatment.

In other cases, the bias may not arise due to individuals self-selecting into treatment, but being selected for treatment on the basis of an interview or evaluation of their willingness to cooperate with the program. This gives rise to administrative selection bias or program placement bias.

Even in the case of a randomized experiment, participants selected for treatment may choose not to be treated, or may not comply with all aspects of the treatment regime. In this sense, even a randomized trial may involve bias in evaluating the effects of treatment, and nonexperimental methods may be required to adjust for that bias.

Requirements for PSM validity

Two key assumptions underly the use of matching methods, and PSM in particular:

1. Conditional independence: there exists a set X of observable covariates such that after controlling for these covariates, the potential outcomes are independent of treatment status:

$$(Y_1, Y_0) \perp D|X$$

2. Common support: for each value fo X, there is a positive probability of being both treated and untreated:

$$0 < P(D = 1|X) < 1$$

Requirements for PSM validity

Two key assumptions underly the use of matching methods, and PSM in particular:

1. Conditional independence: there exists a set X of observable covariates such that after controlling for these covariates, the potential outcomes are independent of treatment status:

$$(Y_1, Y_0) \perp D|X$$

2. Common support: for each value fo X, there is a positive probability of being both treated and untreated:

$$0 < P(D = 1|X) < 1$$

The conditional independence assumption

$$(Y_1, Y_0) \perp D|X$$

implies that after controlling for X, the assignment of units to treatment is 'as good as random.' This assumption is also known as *selection on observables*, and it requires that all variables relevant to the probability of receiving treatment may be observed and included in X. This allows the untreated units to be used to construct an unbiased counterfactual for the treatment group.

The common support assumption

$$0 < P(D = 1|X) < 1$$

implies that the probability of receiving treatment for each possible value of the vector X is strictly within the unit interval: as is the probability of not receiving treatment. This assumption of common support ensures that there is sufficient overlap in the characteristics of treated and untreated units to find adequate matches.

When these assumptions are satisfied, the treatment assignment is said to be *strongly ignorable* in the terminology of Rosenbaum and Rubin (*Biometrika*, 1983).

The common support assumption

$$0 < P(D = 1|X) < 1$$

implies that the probability of receiving treatment for each possible value of the vector X is strictly within the unit interval: as is the probability of not receiving treatment. This assumption of common support ensures that there is sufficient overlap in the characteristics of treated and untreated units to find adequate matches.

When these assumptions are satisfied, the treatment assignment is said to be *strongly ignorable* in the terminology of Rosenbaum and Rubin (*Biometrika*, 1983).

Basic mechanics of matching

The procedure for estimating the impact of a program can be divided into three steps:

1. Estimate the propensity score
2. Choose a matching algorithm that will use the estimated propensity scores to match untreated units to treated units
3. Estimate the impact of the intervention with the matched sample and calculate standard errors

Basic mechanics of matching

The procedure for estimating the impact of a program can be divided into three steps:

1. Estimate the propensity score
2. Choose a matching algorithm that will use the estimated propensity scores to match untreated units to treated units
3. Estimate the impact of the intervention with the matched sample and calculate standard errors

Basic mechanics of matching

The procedure for estimating the impact of a program can be divided into three steps:

1. Estimate the propensity score
2. Choose a matching algorithm that will use the estimated propensity scores to match untreated units to treated units
3. Estimate the impact of the intervention with the matched sample and calculate standard errors

To estimate the propensity score, a logit or probit model is usually employed. It is essential that a flexible functional form be used to allow for possible nonlinearities in the participation model. This may involve the introduction of higher-order terms in the covariates as well as interaction terms.

There will usually be no comprehensive list of the clearly relevant variables that would assure that the matched comparison group will provide an unbiased estimate of program impact. Obviously explicit criteria that govern project or program eligibility should be included, as well as factors thought to influence self-selection and administrative selection.

To estimate the propensity score, a logit or probit model is usually employed. It is essential that a flexible functional form be used to allow for possible nonlinearities in the participation model. This may involve the introduction of higher-order terms in the covariates as well as interaction terms.

There will usually be no comprehensive list of the clearly relevant variables that would assure that the matched comparison group will provide an unbiased estimate of program impact. Obviously explicit criteria that govern project or program eligibility should be included, as well as factors thought to influence self-selection and administrative selection.

In choosing a matching algorithm, you must consider whether matching is to be performed with or without replacement. Without replacement, a given untreated unit can only be matched with one treated unit. A criterion for assessing the quality of the match must also be defined. The number of untreated units to be matched with each treated unit must also be chosen.

Early matching estimators paired each treated unit with one unit from the control group, judged most similar. Researchers have found that estimators are more stable if a number of comparison cases are considered for each treated case, usually implying that the matching will be done with replacement.

In choosing a matching algorithm, you must consider whether matching is to be performed with or without replacement. Without replacement, a given untreated unit can only be matched with one treated unit. A criterion for assessing the quality of the match must also be defined. The number of untreated units to be matched with each treated unit must also be chosen.

Early matching estimators paired each treated unit with one unit from the control group, judged most similar. Researchers have found that estimators are more stable if a number of comparison cases are considered for each treated case, usually implying that the matching will be done with replacement.

The matching criterion could be as simple as the absolute difference in the propensity score for treated vs. non-treated units. However, when the sampling design oversamples treated units, it has been found that matching on the log odds of the propensity score ($p/(1-p)$) is a superior criterion.

The *nearest neighbor* matching algorithm merely evaluates absolute differences between propensity scores (or their log odds), where you may choose to use 1, 2, ... K nearest neighbors in the match. A variation, *radius matching*, specifies a 'caliper' or maximum propensity score difference. Larger differences will not result in matches, and all units whose differences lie within the caliper's radius will be chosen.

The matching criterion could be as simple as the absolute difference in the propensity score for treated vs. non-treated units. However, when the sampling design oversamples treated units, it has been found that matching on the log odds of the propensity score ($p/(1-p)$) is a superior criterion.

The *nearest neighbor* matching algorithm merely evaluates absolute differences between propensity scores (or their log odds), where you may choose to use 1, 2, ... K nearest neighbors in the match. A variation, *radius matching*, specifies a 'caliper' or maximum propensity score difference. Larger differences will not result in matches, and all units whose differences lie within the caliper's radius will be chosen.

In many-to-one radius matching with replacement, the estimator of program impact may be written as

$$E(\Delta Y) = \frac{1}{N} \sum_{i=1}^{N} \left[Y_{1i} - \bar{Y}_{0j(i)} \right]$$

where $\bar{Y}_{0j(i)}$ is the average outcome for all comparison individuals matched with case i, Y_{1i} is the outcome for treated case i, and N is the number of treated cases.

As an alternative to radius matching, which rules out matches beyond the threshold of the caliper, the *kernel* and *local-linear* methods are nonparametric methods that compare each treated unit to a weighted average of the outcomes of all untreated units, with higher weights being placed on the untreated units with scores closer to that of the treated individual. These methods exhibit lower variance, but may suffer from the inclusion of information from poor matches. To use these methods, a kernel function must be chosen, and its bandwidth parameter must be specified.

The usual tradeoff between bias and efficiency arises in selecting a matching algorithm. By choosing only one nearest neighbor, we minimize bias by using the most similar observation. However, this ignores a great deal of information, and thus may yield less efficient estimates.

As an alternative to radius matching, which rules out matches beyond the threshold of the caliper, the *kernel* and *local-linear* methods are nonparametric methods that compare each treated unit to a weighted average of the outcomes of all untreated units, with higher weights being placed on the untreated units with scores closer to that of the treated individual. These methods exhibit lower variance, but may suffer from the inclusion of information from poor matches. To use these methods, a kernel function must be chosen, and its bandwidth parameter must be specified.

The usual tradeoff between bias and efficiency arises in selecting a matching algorithm. By choosing only one nearest neighbor, we minimize bias by using the most similar observation. However, this ignores a great deal of information, and thus may yield less efficient estimates.

Evaluating the validity of matching assumptions

The conditional independence assumption cannot be directly tested, but several guidelines for model specification should be considered. The more transparent and well-controlled is the selection process, the more confidence you may have in arguing that all relevant variables have been included. Measures included in the PSM model should be stable over time, or deterministic (e.g., age), or measured before participation, so that they are not confounded with outcomes or the anticipation of treatment. The specification should allow for nonlinear covariate effects and potential interactions in order to avoid inappropriate constraints on the functional form.

Balancing tests consider whether the estimated propensity score adequately balances characteristics between the treatment and control group units. The assumption

$$D \perp X|p(X)$$

is testable. If it is supported by the data, then after conditioning on the estimated propensity score $p(X)$, there should be no other variable that could be added to the conditioning set X that would improve the estimation, and after the application of matching, there should be no statistically significant differences between covariate means of the treated and comparison units. These mean comparisons can be contrasted with the unconditional means of the treatment and control groups, which are likely to be statistically significant in most applications.

Finally, the common support or overlap condition

$$0 < P(D = 1|X) < 1$$

should be tested. This can be done by visual inspection of the densities of propensity scores of treated and non-treated groups, or more formally via a comparison test such as the Kolmogorov–Smirnov nonparametric test. If there are sizable differences between the maxima and minima of the density distributions, it may be advisable to remove cases that lie outside the support of the other distribution. However, as with any trimming algorithm, this implies that results of the analysis are strictly valid only for the region of common support.

An empirical example

As an example of propensity score matching techniques, we follow Sianesi's 2010 presentation at the German Stata Users Group meetings (`http://ideas.repec.org/p/boc/dsug10/02.html`) and employ the `nsw_psid` dataset that has been used in several articles on PSM techniques. This dataset combines 297 treated individuals from a randomised evaluation of the NSW Demonstration job-training program with 2,490 non-experimental untreated individuals drawn from the Panel Study of Income Dynamics (PSID), all of whom are male. The outcome of interest is `re78`, 1978 earnings. Available covariates include age, ethnic status (black, Hispanic or white), marital status, years of education, an indicator for no high school degree and 1975 earnings (in 1978 dollars).

We use Leuven and Sianesi's `psmatch2` routine, available from SSC.

An empirical example

As an example of propensity score matching techniques, we follow Sianesi's 2010 presentation at the German Stata Users Group meetings (`http://ideas.repec.org/p/boc/dsug10/02.html`) and employ the `nsw_psid` dataset that has been used in several articles on PSM techniques. This dataset combines 297 treated individuals from a randomised evaluation of the NSW Demonstration job-training program with 2,490 non-experimental untreated individuals drawn from the Panel Study of Income Dynamics (PSID), all of whom are male. The outcome of interest is `re78`, 1978 earnings. Available covariates include age, ethnic status (black, Hispanic or white), marital status, years of education, an indicator for no high school degree and 1975 earnings (in 1978 dollars).

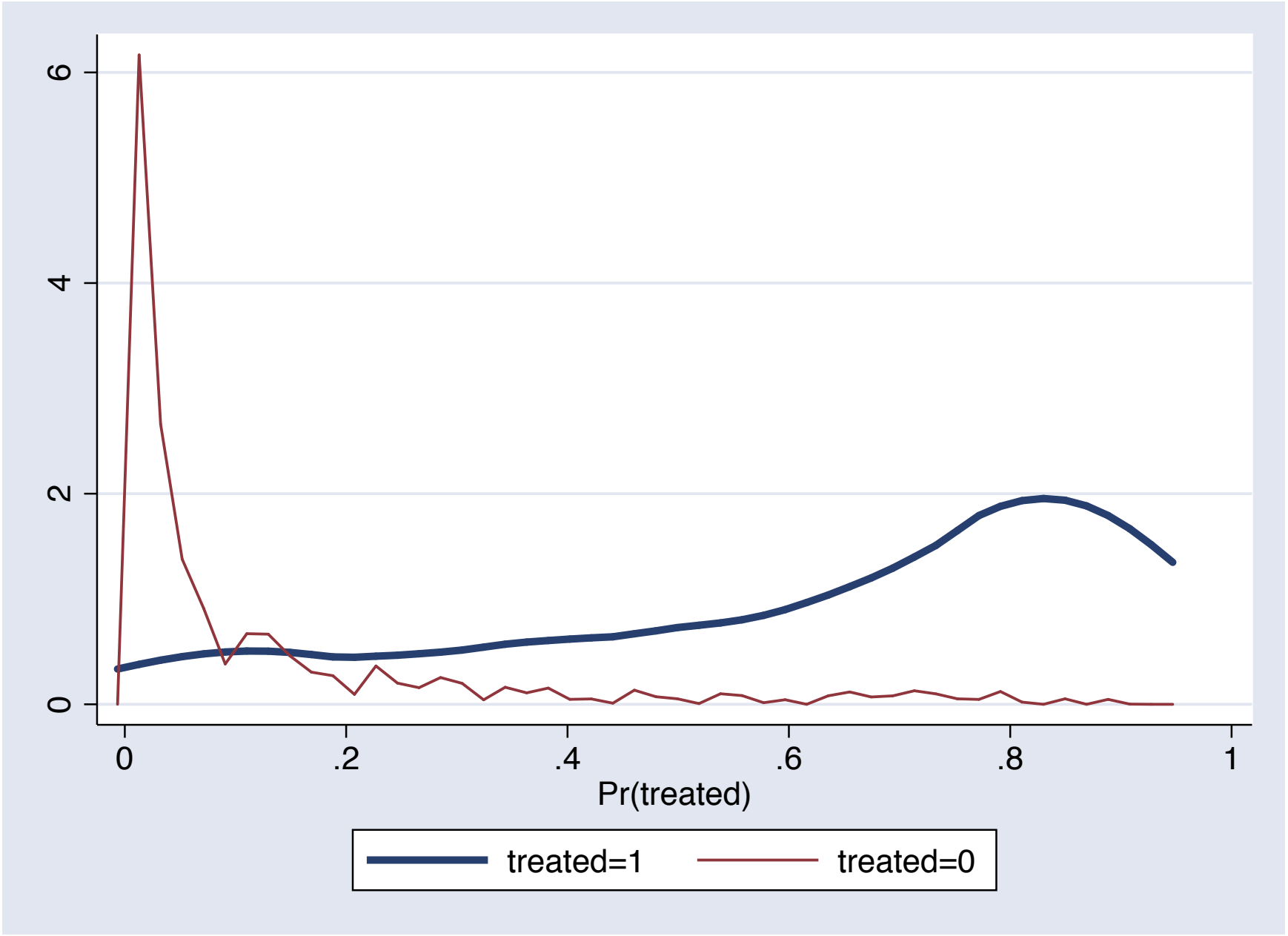
We use Leuven and Sianesi's `psmatch2` routine, available from SSC.

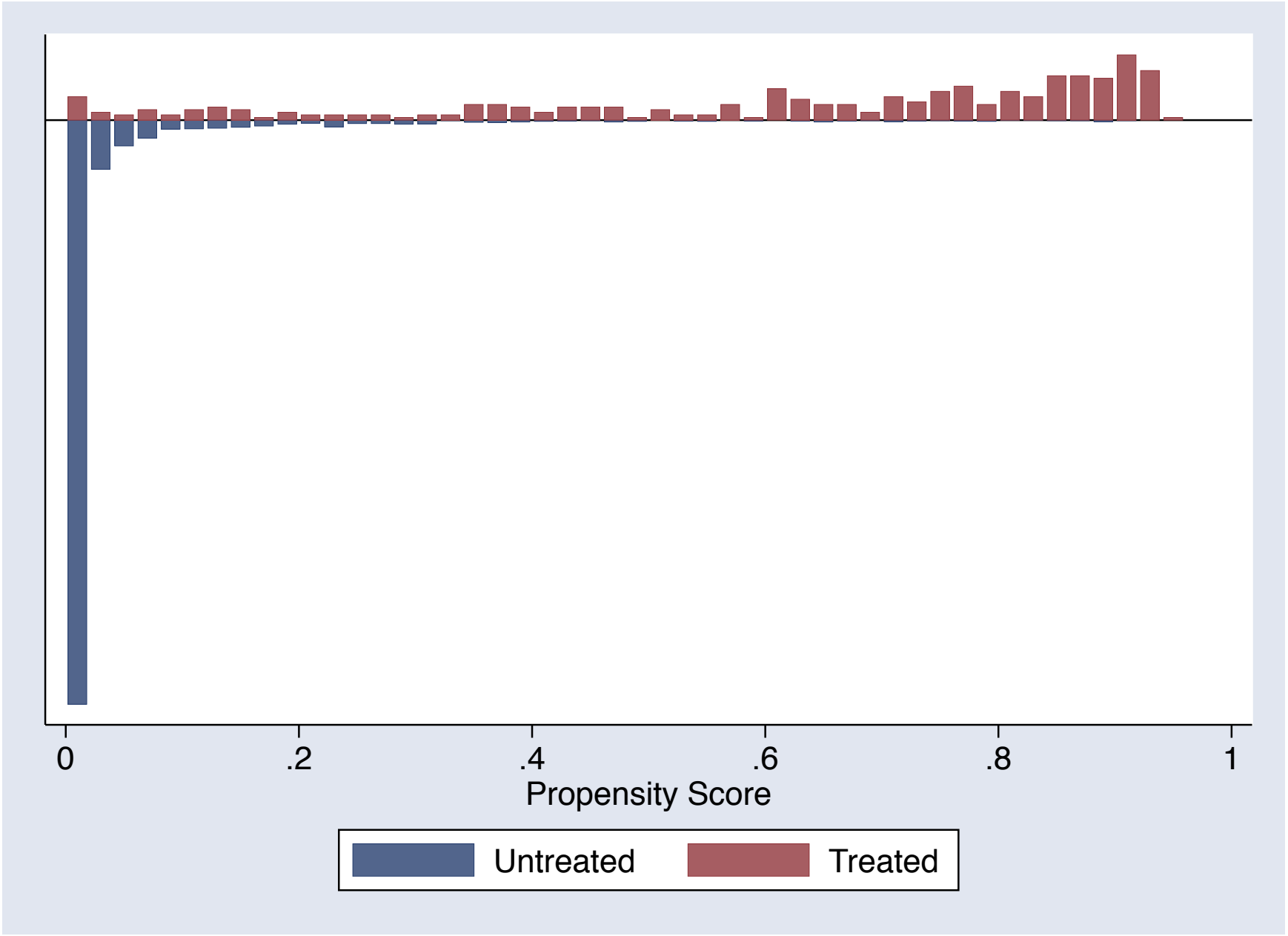
```
 . use nsw_psid, clear
(NSW treated and PSID non-treated)
. qui probit treated age black hispanic married educ nodegree re75
. margins, dydx(_all)
Average marginal effects                        Number of obs   =       2787
Model VCE    : OIM
Expression   : Pr(treated), predict()
dy/dx w.r.t. : age black hispanic married educ nodegree re75
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0035844	.000462	-7.76	0.000	-.0044899	-.002679
black	.0766501	.0088228	8.69	0.000	.0593577	.0939426
hispanic	.0831734	.0157648	5.28	0.000	.0522751	.1140718
married	-.0850743	.0070274	-12.11	0.000	-.0988478	-.0713009
educ	.0003458	.0023048	0.15	0.881	-.0041716	.0048633
nodegree	.0418875	.0108642	3.86	0.000	.0205942	.0631809
re75	-6.89e-06	5.89e-07	-11.71	0.000	-8.04e-06	-5.74e-06

```
. // compute the propensity score
. predict double score
(option pr assumed; Pr(treated))
```

```
 . // compare the densities of the estimated propensity score over groups
. density2 score, group(treated) saving(psm2a, replace)
(file psm2a.gph saved)
. graph export psm2a.pdf, replace
(file /Users/cfbaum/Documents/Stata/StataWorkshops/psm2a.pdf written in PDF for
> mat)
. psgraph, treated(treated) pscore(score) bin(50) saving(psm2b, replace)
(file psm2b.gph saved)
. graph export psm2b.pdf, replace
(file /Users/cfbaum/Documents/Stata/StataWorkshops/psm2b.pdf written in PDF for
> mat)
```


0
2
4
6
0
.2
.4
.6
.8
1
Pr(treated)
treated=1
treated=0


0
.2
.4
.6
.8
1
Propensity Score
Untreated
Treated

```
1 . // compute nearest-neighbor matching with caliper and replacement
2 . psmatch2 treated, pscore(score) outcome(re78) caliper(0.01)
```

There are observations with identical propensity score values.
The sort order of the data could affect your results.
Make sure that the sort order is random before calling psmatch2.

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	**5976.35202**	**21553.9209**	**-15577.5689**	**913.328457**	**-17.06**
	ATT	**6067.8117**	**5768.70099**	**299.110712**	**1078.28065**	**0.28**

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support		Total
	Off suppo	On suppor	
Untreated	**0**	**2,490**	**2,490**
Treated	**26**	**271**	**297**
Total	**26**	**2,761**	**2,787**

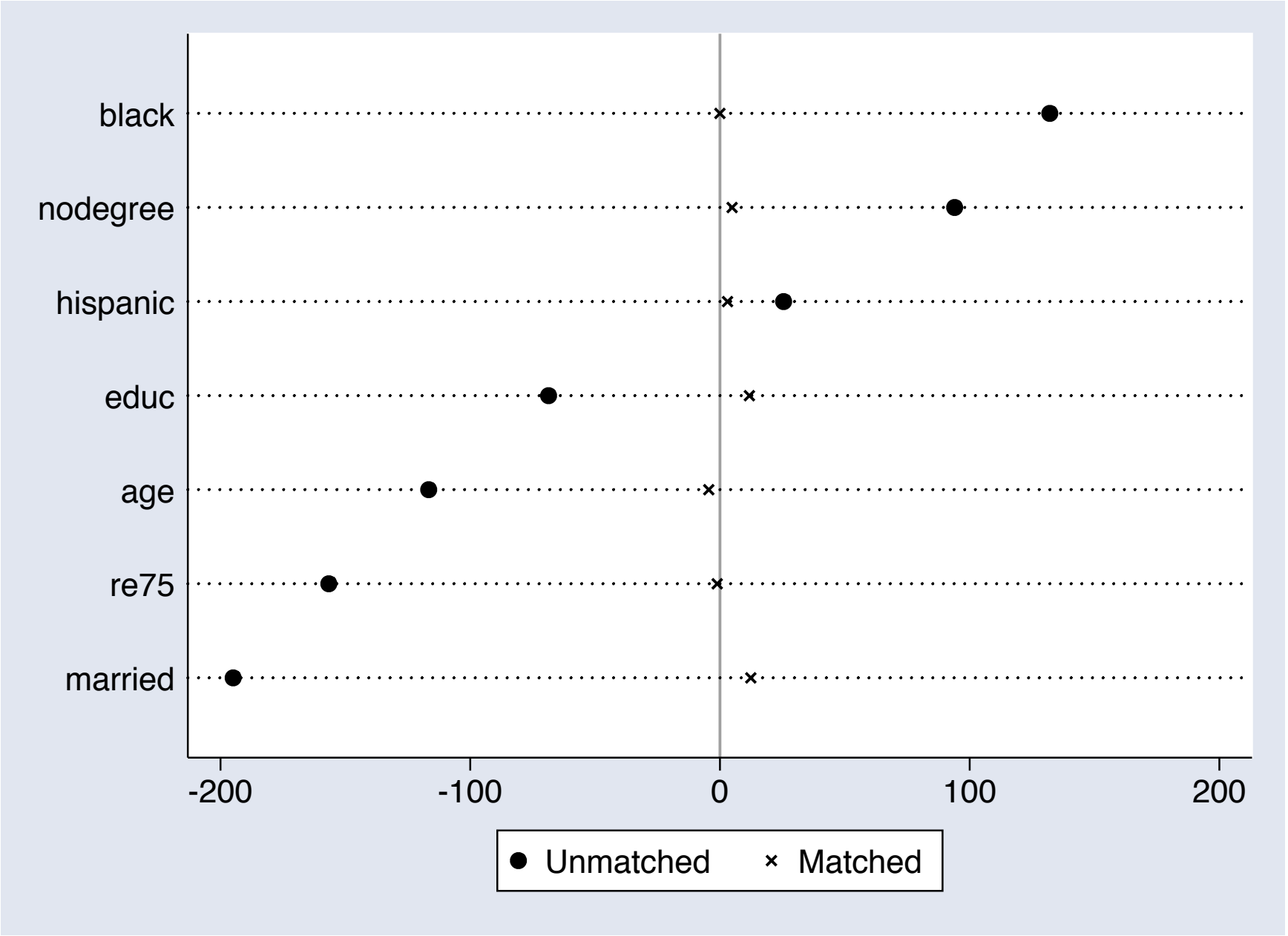
```
3 . // evaluate common support
4 . summarize _support if treated
```

Variable	Obs	Mean	Std. Dev.	Min	Max
_support	**297**	**.9124579**	**.2831048**	**0**	**1**

```
5 . qui log close
```

```
1 . // evaluate quality of matching
2 . pstest2 age black hispanic married educ nodegree re75, sum graph
```

Variable	Sample	Mean Treated	Mean Control	%bias	%reduct |bias|	t-test t	t-test p>|t|
age	Unmatched	24.626	34.851	-116.6		-16.48	0.000
	Matched	25.052	25.443	-4.5	96.2	-0.61	0.545
black	Unmatched	.80135	.2506	132.1		20.86	0.000
	Matched	.78967	.78967	0.0	100.0	-0.00	1.000
hispanic	Unmatched	.09428	.03253	25.5		5.21	0.000
	Matched	.09594	.08856	3.0	88.0	0.30	0.767
married	Unmatched	.16835	.86627	-194.9		-33.02	0.000
	Matched	.1845	.14022	12.4	93.7	1.40	0.163
educ	Unmatched	10.38	12.117	-68.6		-9.51	0.000
	Matched	10.465	10.166	11.8	82.8	1.54	0.125
nodegree	Unmatched	.73064	.30522	94.0		15.10	0.000
	Matched	.71587	.69373	4.9	94.8	0.56	0.573
re75	Unmatched	3066.1	19063	-156.6		-20.12	0.000
	Matched	3197.4	3307.8	-1.1	99.3	-0.28	0.778


black
nodegree
hispanic
educ
age
re75
married
-200
-100
0
100
200
Unmatched
Matched

Alternatively, we can perform PSM with a kernel-based method. Notice that the estimate of ATT switches sign relative to that produced by the nearest-neighbor matching algorithm.

```
1 . // compute kernel-based matching with normal kernel
2 . psmatch2 treated, pscore(score) outcome(re78) kernel k(normal) bw(0.01)
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	**5976.35202**	**21553.9209**	**-15577.5689**	**913.328457**	**-17.06**
	ATT	**5976.35202**	**6882.18396**	**-905.831935**	**2151.26377**	**-0.42**

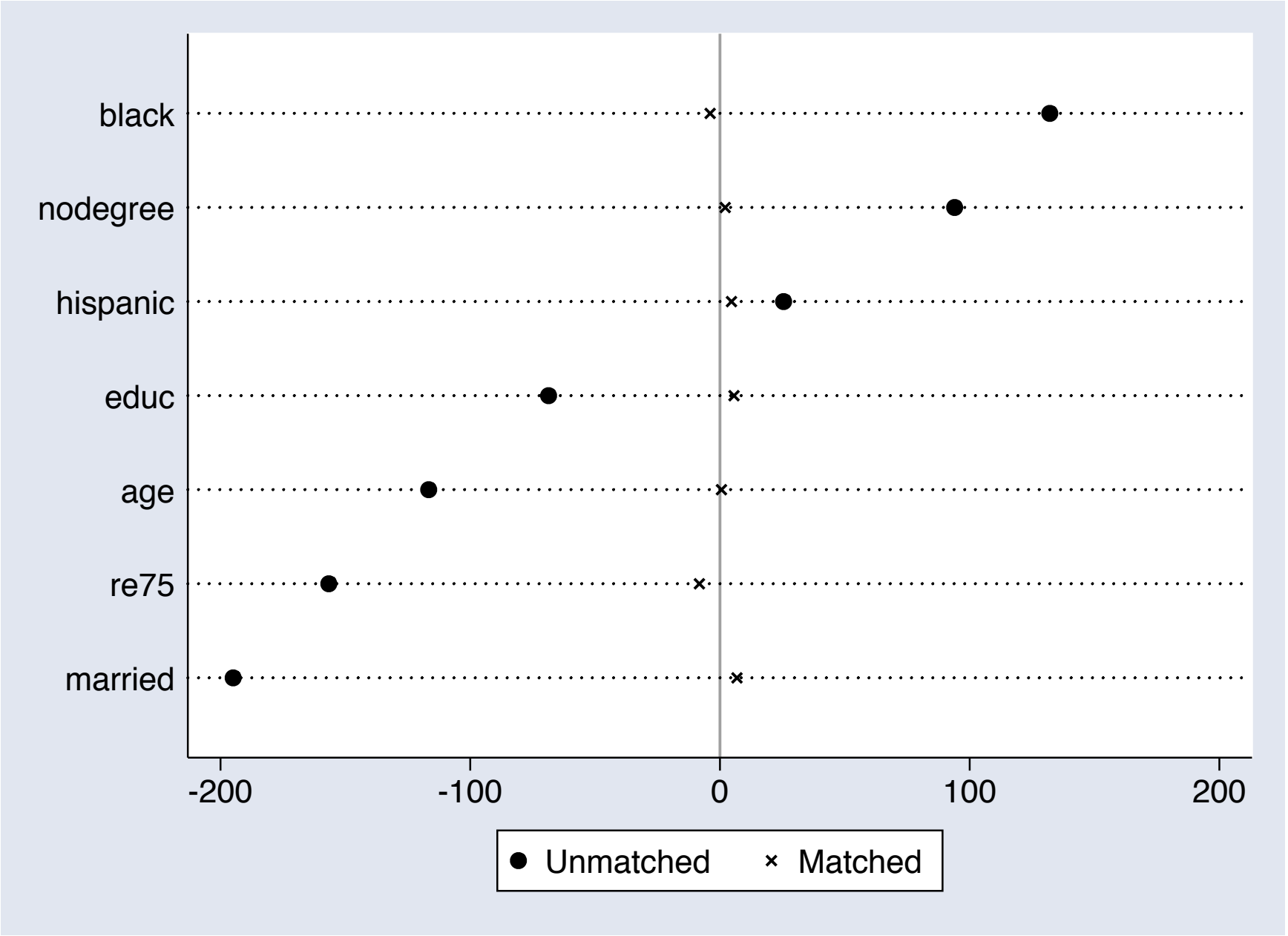
Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support On suppor	Total
Untreated	**2,490**	**2,490**
Treated	**297**	**297**
Total	**2,787**	**2,787**

```
3 . qui log close
```

```
1 . // evaluate quality of matching
2 . pstest2 age black hispanic married educ nodegree re75, sum graph
```

Variable	Sample	Mean Treated	Mean Control	%bias	%reduct |bias|	t-test t	t-test p>|t|
age	Unmatched	24.626	34.851	-116.6		-16.48	0.000
	Matched	24.626	24.572	0.6	99.5	0.09	0.926
black	Unmatched	.80135	.2506	132.1		20.86	0.000
	Matched	.80135	.81763	-3.9	97.0	-0.50	0.614
hispanic	Unmatched	.09428	.03253	25.5		5.21	0.000
	Matched	.09428	.08306	4.6	81.8	0.48	0.631
married	Unmatched	.16835	.86627	-194.9		-33.02	0.000
	Matched	.16835	.1439	6.8	96.5	0.82	0.413
educ	Unmatched	10.38	12.117	-68.6		-9.51	0.000
	Matched	10.38	10.238	5.6	91.8	0.81	0.415
nodegree	Unmatched	.73064	.30522	94.0		15.10	0.000
	Matched	.73064	.72101	2.1	97.7	0.26	0.793
re75	Unmatched	3066.1	19063	-156.6		-20.12	0.000
	Matched	3066.1	3905.8	-8.2	94.8	-1.99	0.047


black
nodegree
hispanic
educ
age
re75
married
-200
-100
0
100
200
Unmatched
Matched

We could also employ Mahalanobis matching, which matches on the whole vector of X values (and possibly the propensity score as well), using a different distance metric.

An additional important issue: how might we address unobserved heterogeneity, as we do in a panel data context with fixed effects models? A *differences-in-differences matching estimator* (DID) has been proposed, in which rather than evaluating the effect on the outcome variable, you evaluate the effect on the change in the outcome variable, before and after the intervention. Akin to DID estimators in standard policy evaluation, this allows us to control for the notion that there may be substantial unobserved differences between treated and untreated units, relaxing the 'selection on observables' assumption.

We could also employ Mahalanobis matching, which matches on the whole vector of X values (and possibly the propensity score as well), using a different distance metric.

An additional important issue: how might we address unobserved heterogeneity, as we do in a panel data context with fixed effects models? A *differences-in-differences matching estimator* (DID) has been proposed, in which rather than evaluating the effect on the outcome variable, you evaluate the effect on the change in the outcome variable, before and after the intervention. Akin to DID estimators in standard policy evaluation, this allows us to control for the notion that there may be substantial unobserved differences between treated and untreated units, relaxing the 'selection on observables' assumption.

Regression discontinuity models

The idea of Regression Discontinuity (RD) design, due to Thistlewaite and Campbell (*J. Educ. Psych.*, 1960) and Hahn et al. (*Econometrica*, 2001) is to use a discontinuity in the level of treatment related to some observable to get a consistent estimate of the LATE: the local average treatment effect. This compares those just eligible for the treatment (above the threshold) to those just ineligible (below the threshold).

Among non-experimental or quasi-experimental methods, RD techniques are considered to have the highest internal validity (the ability to identify causal relationships in this research setting). Their external validity (ability to generalize findings to similar contexts) may be less impressive, as the estimated treatment effect is local to the discontinuity.

Regression discontinuity models

The idea of Regression Discontinuity (RD) design, due to Thistlewaite and Campbell (*J. Educ. Psych.*, 1960) and Hahn et al. (*Econometrica*, 2001) is to use a discontinuity in the level of treatment related to some observable to get a consistent estimate of the LATE: the local average treatment effect. This compares those just eligible for the treatment (above the threshold) to those just ineligible (below the threshold).

Among non-experimental or quasi-experimental methods, RD techniques are considered to have the highest internal validity (the ability to identify causal relationships in this research setting). Their external validity (ability to generalize findings to similar contexts) may be less impressive, as the estimated treatment effect is local to the discontinuity.

What could give rise to a RD design? In 1996, a number of US states adopted a policy that while immigrants were generally ineligible for food stamps, a form of welfare assistance, those who had been in the country legally for at least five years would qualify. At a later date, one could compare self-reported measures of dietary adequacy, or measures of obesity, between those immigrants who did and did not qualify for this assistance. The sharp discontinuity in this example relates to those on either side of the five-year boundary line.

Currently, US states are eligible for additional Federal funding if their unemployment rate is above 8% (as most currently are). This funding permits UI recipients to receive a number of additional weeks of benefit. Two months ago, the Massachusetts unemployment rate dropped below the threshold: good news for those who are employed, but bad news for those still seeking a job, as the additional weeks of benefit are now not available to current recipients.

What could give rise to a RD design? In 1996, a number of US states adopted a policy that while immigrants were generally ineligible for food stamps, a form of welfare assistance, those who had been in the country legally for at least five years would qualify. At a later date, one could compare self-reported measures of dietary adequacy, or measures of obesity, between those immigrants who did and did not qualify for this assistance. The sharp discontinuity in this example relates to those on either side of the five-year boundary line.

Currently, US states are eligible for additional Federal funding if their unemployment rate is above 8% (as most currently are). This funding permits UI recipients to receive a number of additional weeks of benefit. Two months ago, the Massachusetts unemployment rate dropped below the threshold: good news for those who are employed, but bad news for those still seeking a job, as the additional weeks of benefit are now not available to current recipients.

Other examples of RD designs arise in terms of taxation. In my home state of Massachusetts, items of clothing are not taxable if they cost less than US$250.00. An item selling for that price or higher is taxable at 6.25%. During a recent weekend 'sales tax holiday', items purchased on a single invoice were not taxable if the total was below US$2,500.00. Thus, a one-dollar increase in the invoice would incur an additional US$156.25 in taxes, as the entire sale is then taxable.

As Austin Nichols pointed out, the sequence of US estate tax rates on large inheritances: 45% in 2009, zero in 2010, and 55% in 2011 may have caused some perverse incentives among potential heirs! That may be a difficult hypothesis to test, though, without the assistance of the homicide squad.

Other examples of RD designs arise in terms of taxation. In my home state of Massachusetts, items of clothing are not taxable if they cost less than US$250.00. An item selling for that price or higher is taxable at 6.25%. During a recent weekend 'sales tax holiday', items purchased on a single invoice were not taxable if the total was below US$2,500.00. Thus, a one-dollar increase in the invoice would incur an additional US$156.25 in taxes, as the entire sale is then taxable.

As Austin Nichols pointed out, the sequence of US estate tax rates on large inheritances: 45% in 2009, zero in 2010, and 55% in 2011 may have caused some perverse incentives among potential heirs! That may be a difficult hypothesis to test, though, without the assistance of the homicide squad.

RD design elements

There are four crucial elements to a RD design:

1. Treatment is not randomly assigned, but dependent at least in part on an observable assignment variable Z.
2. There is a discontinuity at some cutoff value of the assignment variable in the level of treatment.
3. Individuals cannot manipulate their status to affect whether they fall on one side of the cutoff or the other. Those near the cutoff are assumed to be *exchangeable* or otherwise identical.
4. Other variables are smooth functions of the assignment variable, conditional on treatment. That is, a jump in the outcome variable should be due to the discontinuity in the level of treatment.

RD design elements

There are four crucial elements to a RD design:

1. Treatment is not randomly assigned, but dependent at least in part on an observable assignment variable Z.
2. There is a discontinuity at some cutoff value of the assignment variable in the level of treatment.
3. Individuals cannot manipulate their status to affect whether they fall on one side of the cutoff or the other. Those near the cutoff are assumed to be *exchangeable* or otherwise identical.
4. Other variables are smooth functions of the assignment variable, conditional on treatment. That is, a jump in the outcome variable should be due to the discontinuity in the level of treatment.

RD design elements

There are four crucial elements to a RD design:

1. Treatment is not randomly assigned, but dependent at least in part on an observable assignment variable Z.
2. There is a discontinuity at some cutoff value of the assignment variable in the level of treatment.
3. Individuals cannot manipulate their status to affect whether they fall on one side of the cutoff or the other. Those near the cutoff are assumed to be *exchangeable* or otherwise identical.
4. Other variables are smooth functions of the assignment variable, conditional on treatment. That is, a jump in the outcome variable should be due to the discontinuity in the level of treatment.

RD design elements

There are four crucial elements to a RD design:

1. Treatment is not randomly assigned, but dependent at least in part on an observable assignment variable Z.
2. There is a discontinuity at some cutoff value of the assignment variable in the level of treatment.
3. Individuals cannot manipulate their status to affect whether they fall on one side of the cutoff or the other. Those near the cutoff are assumed to be *exchangeable* or otherwise identical.
4. Other variables are smooth functions of the assignment variable, conditional on treatment. That is, a jump in the outcome variable should be due to the discontinuity in the level of treatment.

RD methodology

There is considerable art involved in choosing some continuous function of the assignment variable Z for treatment and outcomes. A high-order polynomial in Z is often used to estimate separately on both sides of the discontinuity. Better yet, a local polynomial, local linear regression model or local mean smoother may be used, where as in other nonparametric settings one must choose a kernel and bandwidth parameter. It is probably best to choose several different bandwidth parameters to analyze the sensitivity of the results to the choice of bandwidth.

The first test performed should be a test that the hypothesized cutoff in the assignment variable produces a jump in the level of treatment. In the case of an election with two candidates on the ballot, for instance, the probability of winning the election jumps from zero to one at the 50% cutoff. A local linear regression of X (the treatment variable) on Z in the vicinity of the cutoff should identify the magnitude of the jump.

It should also be verified that there are no extraneous discontinuities in the level of treatment or the outcome variable at other points, where no hypothesized cutoff exists. Likewise, there should be no discontinuities in other variables in the vicinity of the cutoff.

The first test performed should be a test that the hypothesized cutoff in the assignment variable produces a jump in the level of treatment. In the case of an election with two candidates on the ballot, for instance, the probability of winning the election jumps from zero to one at the 50% cutoff. A local linear regression of X (the treatment variable) on Z in the vicinity of the cutoff should identify the magnitude of the jump.

It should also be verified that there are no extraneous discontinuities in the level of treatment or the outcome variable at other points, where no hypothesized cutoff exists. Likewise, there should be no discontinuities in other variables in the vicinity of the cutoff.

RD empirical example

We make use of Austin Nichols' `rd` package, available from SSC. `rd` estimates local linear or kernel regression models on both sides of the cutoff, using a triangle kernel. Estimates are sensitive to the choice of bandwidth, so by default several estimates are constructed using different bandwidths.

In the simplest case, assignment to treatment depends on a variable Z being above a cutoff Z_0. Frequently, Z is defined so that $Z_0 = 0$. In this case, treatment is 1 for $Z >= 0$ and 0 for $Z < 0$, and we estimate local linear regressions on both sides of the cutoff to obtain estimates of the outcome at $Z = 0$. The difference between the two estimates (for the samples where $Z >= 0$ and where $Z < 0$) is the estimated effect of treatment.

RD empirical example

We make use of Austin Nichols' `rd` package, available from SSC. `rd` estimates local linear or kernel regression models on both sides of the cutoff, using a triangle kernel. Estimates are sensitive to the choice of bandwidth, so by default several estimates are constructed using different bandwidths.

In the simplest case, assignment to treatment depends on a variable Z being above a cutoff Z_0. Frequently, Z is defined so that $Z_0 = 0$. In this case, treatment is 1 for $Z >= 0$ and 0 for $Z < 0$, and we estimate local linear regressions on both sides of the cutoff to obtain estimates of the outcome at $Z = 0$. The difference between the two estimates (for the samples where $Z >= 0$ and where $Z < 0$) is the estimated effect of treatment.

For example, having a Democratic representative in the US Congress may be considered a treatment applied to a Congressional district, and the assignment variable Z is the vote share garnered by the Democratic candidate. At Z=50%, the probability of treatment=1 jumps from zero to one. Suppose we are interested in the effect a Democratic representative has on the federal spending within a Congressional district.

The `votex` dataset contains information for 349 of the 435 Congressional districts in the 102^{nd} US Congress. `lne` is the logarithm of Federal expenditures in the district (evidence of the member of Congress 'bringing home the bacon'.) Variable `d` is the Democratic vote share minus 0.5, so that it is positive for Democratic districts and negative for Republican districts.

For example, having a Democratic representative in the US Congress may be considered a treatment applied to a Congressional district, and the assignment variable Z is the vote share garnered by the Democratic candidate. At Z=50%, the probability of treatment=1 jumps from zero to one. Suppose we are interested in the effect a Democratic representative has on the federal spending within a Congressional district.

The `votex` dataset contains information for 349 of the 435 Congressional districts in the 102^{nd} US Congress. `lne` is the logarithm of Federal expenditures in the district (evidence of the member of Congress 'bringing home the bacon'.) Variable `d` is the Democratic vote share minus 0.5, so that it is positive for Democratic districts and negative for Republican districts.

We may now estimate the RD model:

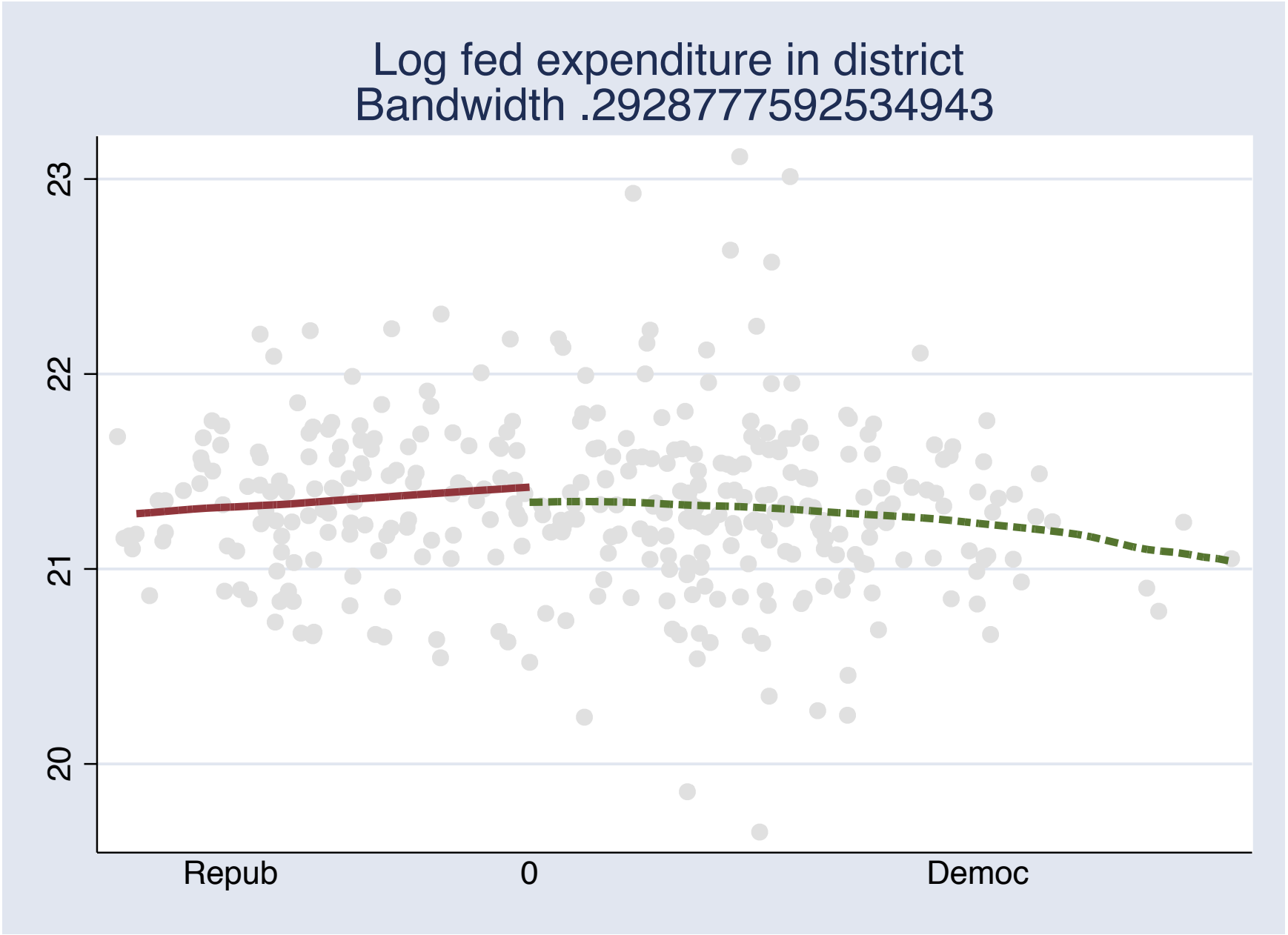
```
1 . rd lne d, gr mbw(100) line(`"xla(-.2 "Repub" 0 .3 "Democ", noticks)"')
  Two variables specified; treatment is
  assumed to jump from zero to one at Z=0.

   Assignment variable Z is d
   Treatment variable X_T unspecified
   Outcome variable y is lne

  Command used for graph: lpoly; Kernel used: triangle (default)
  Bandwidth: .29287776; loc Wald Estimate: -.07739553
  Estimating for bandwidth .2928777592534943
```

lne	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwald	-.0773955	.1056062	-0.73	0.464	-.28438	.1295889

The estimate of the LATE is not significantly different from zero in this case. Interestingly, as we move from ‘safe’ seats in either party toward the contested boundary, expenditures modestly increase.


Log fed expenditure in district
Bandwidth .2928777592534943
23
22
21
20
Repub
0
Democ

To evaluate the sensitivity of these results to the estimation technique, we may vary the bandwidth:

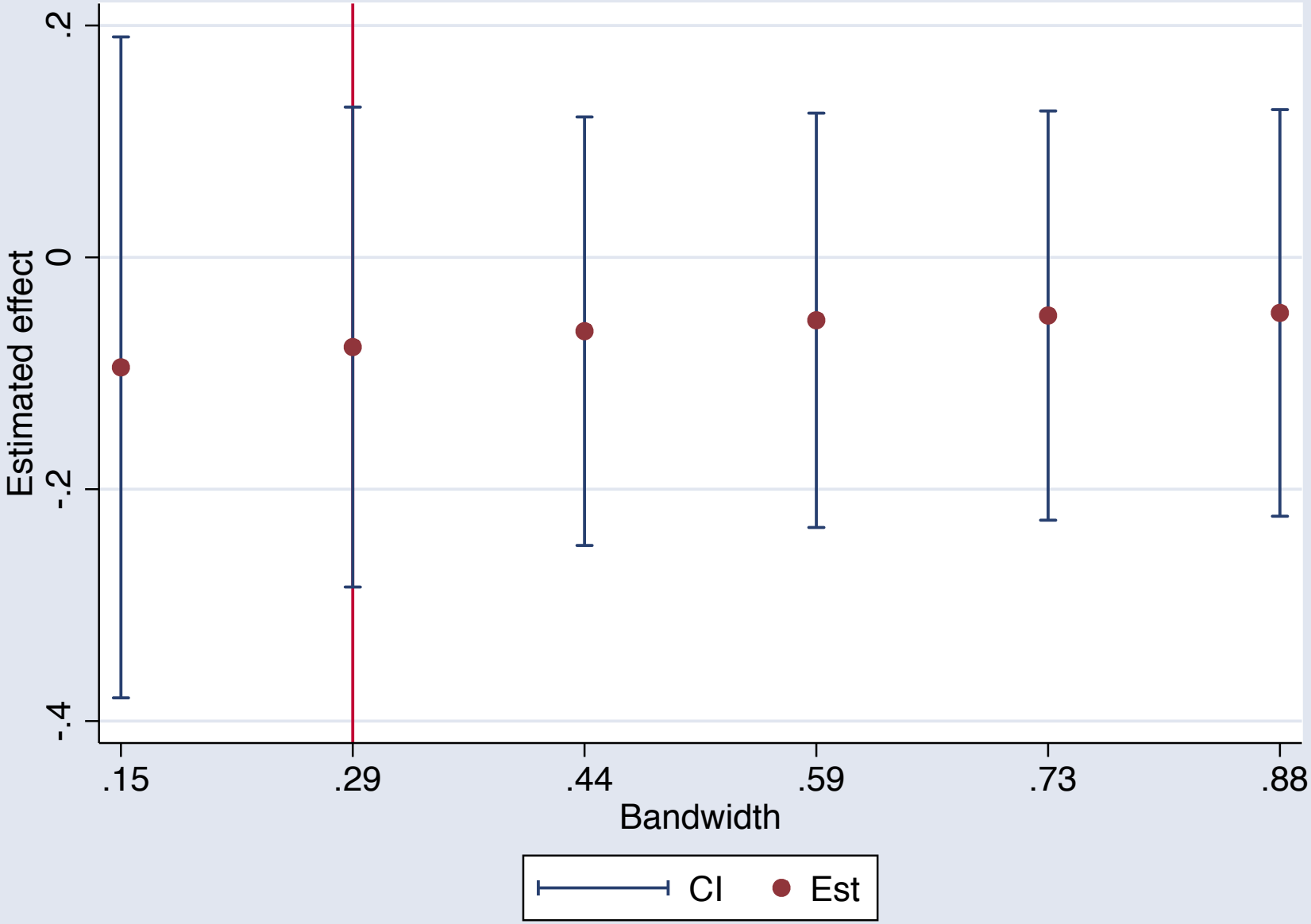
```
1 . rd lne d, mbw(50(50)300) bdep ox
  Two variables specified; treatment is
  assumed to jump from zero to one at Z=0.

   Assignment variable Z is d
   Treatment variable X_T unspecified
   Outcome variable y is lne

  Estimating for bandwidth .2928777592534943
  Estimating for bandwidth .1464388796267471
  Estimating for bandwidth .4393166388802414
  Estimating for bandwidth .5857555185069886
  Estimating for bandwidth .7321943981337358
  Estimating for bandwidth .8786332777604828
```

lne	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwald	-.0773955	.1056062	-0.73	0.464	-.28438	.1295889
lwald50	-.0949149	.1454442	-0.65	0.514	-.3799804	.1901505
lwald150	-.0637113	.0942934	-0.68	0.499	-.248523	.1211004
lwald200	-.0543086	.0911788	-0.60	0.551	-.2330157	.1243985
lwald250	-.0502168	.0900457	-0.56	0.577	-.2267032	.1262696
lwald300	-.0479296	.0894768	-0.54	0.592	-.2233009	.1274417

```
2 . qui log close
```


Estimated effect
.2
0
-.2
-.4
.15
.29
.44
.59
.73
.88
Bandwidth
CI
Est

As we can see, the conclusion of no meaningful difference in the outcome variable is not sensitive to the choice of bandwidth in the local linear regression estimator.

For a more detailed discussion of RD and other quasi-experimental methods, see Austin Nichols, 2007, 'Causal Inference with Observational Data,' *Stata Journal* 7(4): 507–541. Freely downloadable from `http://www.stata-journal.com`.

As we can see, the conclusion of no meaningful difference in the outcome variable is not sensitive to the choice of bandwidth in the local linear regression estimator.

For a more detailed discussion of RD and other quasi-experimental methods, see Austin Nichols, 2007, 'Causal Inference with Observational Data,' *Stata Journal* 7(4): 507–541. Freely downloadable from `http://www.stata-journal.com`.

Limited dependent variables

We consider models of *limited dependent variables* in which the economic agent's response is limited in some way. The dependent variable, rather than being continuous on the real line (or half–line), is restricted. In some cases, we are dealing with *discrete choice*: the response variable may be restricted to a Boolean or binary choice, indicating that a particular course of action was or was not selected.

In others, it may take on only integer values, such as the number of children per family, or the ordered values on a Likert scale. Alternatively, it may appear to be a continuous variable with a number of responses at a threshold value. For instance, the response to the question "how many hours did you work last week?" will be recorded as zero for the non-working respondents. None of these measures are amenable to being modeled by the linear regression methods we have discussed.

Limited dependent variables

We consider models of *limited dependent variables* in which the economic agent's response is limited in some way. The dependent variable, rather than being continuous on the real line (or half–line), is restricted. In some cases, we are dealing with *discrete choice*: the response variable may be restricted to a Boolean or binary choice, indicating that a particular course of action was or was not selected.

In others, it may take on only integer values, such as the number of children per family, or the ordered values on a Likert scale. Alternatively, it may appear to be a continuous variable with a number of responses at a threshold value. For instance, the response to the question "how many hours did you work last week?" will be recorded as zero for the non-working respondents. None of these measures are amenable to being modeled by the linear regression methods we have discussed.

We first consider models of Boolean response variables, or *binary choice*. In such a model, the response variable is coded as 1 or 0, corresponding to responses of True or False to a particular question. A behavioral model of this decision could be developed, including a number of "explanatory factors" (we should not call them regressors) that we expect will influence the respondent's answer to such a question. But we should readily spot the flaw in the *linear probability model*:

$$R_i = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + u_i \tag{1}$$

where we place the Boolean response variable in R and regress it upon a set of X variables. All of the observations we have on R are either 0 or 1. They may be viewed as the *ex post* probabilities of responding "yes" to the question posed. But the predictions of a linear regression model are unbounded, and the model of Equation (1), estimated with `regress`, can produce negative predictions and predictions exceeding unity, neither of which can be considered probabilities.

Because the response variable is bounded, restricted to take on values of {0,1}, the model should be generating a predicted *probability* that individual *i* will choose to answer Yes rather than No. In such a framework, if $\beta_j > 0$, those individuals with high values of X_j will be more likely to respond Yes, but their probability of doing so must respect the upper bound.

For instance, if higher disposable income makes new car purchase more probable, we must be able to include a very wealthy person in the sample and still find that the individual's predicted probability of new car purchase is no greater than 1.0. Likewise, a poor person's predicted probability must be bounded by 0.

Because the response variable is bounded, restricted to take on values of {0,1}, the model should be generating a predicted *probability* that individual *i* will choose to answer Yes rather than No. In such a framework, if $\beta_j > 0$, those individuals with high values of X_j will be more likely to respond Yes, but their probability of doing so must respect the upper bound.

For instance, if higher disposable income makes new car purchase more probable, we must be able to include a very wealthy person in the sample and still find that the individual's predicted probability of new car purchase is no greater than 1.0. Likewise, a poor person's predicted probability must be bounded by 0.

A useful approach to motivate such a model is that of a *latent variable*. Express the model of Equation (1) as:

$$y_i^* = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + u_i \tag{2}$$

where y^* is an unobservable magnitude which can be considered the net benefit to individual i of taking a particular course of action (e.g., purchasing a new car). We cannot observe that net benefit, but can observe the outcome of the individual having followed the decision rule

$$\begin{aligned} y_i &= 0 \text{ if } y_i^* < 0 \\ y_i &= 1 \text{ if } y_i^* \geq 0 \end{aligned} \tag{3}$$

That is, we observe that the individual did or did not purchase a new car in 2005. If she did, we observed $y_i = 1$, and we take this as evidence that a rational consumer made a decision that improved her welfare. We speak of y^* as a *latent variable*, linearly related to a set of factors X and a disturbance process u.

In the latent variable model, we must make the assumption that the disturbance process has a known variance σ_u^2. Unlike the regression problem, we do not have sufficient information in the data to estimate its magnitude. Since we may divide Equation (2) by any positive σ without altering the estimation problem, the most useful strategy is to set $\sigma_u = \sigma_u^2 = 1$.

That is, we observe that the individual did or did not purchase a new car in 2005. If she did, we observed $y_i = 1$, and we take this as evidence that a rational consumer made a decision that improved her welfare. We speak of y^* as a *latent variable*, linearly related to a set of factors X and a disturbance process u.

In the latent variable model, we must make the assumption that the disturbance process has a known variance σ_u^2. Unlike the regression problem, we do not have sufficient information in the data to estimate its magnitude. Since we may divide Equation (2) by any positive σ without altering the estimation problem, the most useful strategy is to set $\sigma_u = \sigma_u^2 = 1$.

In the latent model framework, we model the probability of an individual making each choice. Using equations (2) and (3) we have

$$\begin{aligned} Pr[y^* > 0|X] &= \\ Pr[u > -X\beta|X] &= \\ Pr[u < X\beta|X] &= \\ Pr[y = 1|X] &= \Psi(y_i^*) \end{aligned} \tag{4}$$

The function $\Psi(\cdot)$ is a cumulative distribution function (*CDF*) which maps points on the real line $\{-\infty, \infty\}$ into the probability measure $\{0, 1\}$. The explanatory variables in X are modeled in a linear relationship to the latent variable y^*. If $y = 1$, $y^* > 0$ implies $u < X\beta$.

In the latent model framework, we model the probability of an individual making each choice. Using equations (2) and (3) we have

$$\begin{aligned} Pr[y^* > 0|X] &= \\ Pr[u > -X\beta|X] &= \\ Pr[u < X\beta|X] &= \\ Pr[y = 1|X] &= \Psi(y_i^*) \end{aligned} \tag{4}$$

The function $\Psi(\cdot)$ is a cumulative distribution function (*CDF*) which maps points on the real line $\{-\infty, \infty\}$ into the probability measure $\{0, 1\}$. The explanatory variables in X are modeled in a linear relationship to the latent variable y^*. If $y = 1$, $y^* > 0$ implies $u < X\beta$.

Consider a case where $u_i = 0$. Then a positive y^* would correspond to $X\beta > 0$, and *vice versa*. If u_i were now negative, observing $y_i = 1$ would imply that $X\beta$ must have outweighed the negative u_i (and *vice versa*). Therefore, we can interpret the outcome $y_i = 1$ as indicating that the explanatory factors and disturbance faced by individual *i* have combined to produce a positive net benefit.

For example, an individual might have a low income (which would otherwise suggest that new car purchase was not likely) but may have a sibling who works for Toyota and can arrange for an advantageous price on a new vehicle. We do not observe that circumstance, so it becomes a large positive u_i, explaining how $(X\beta + u_i) > 0$ for that individual.

Consider a case where $u_i = 0$. Then a positive y^* would correspond to $X\beta > 0$, and *vice versa*. If u_i were now negative, observing $y_i = 1$ would imply that $X\beta$ must have outweighed the negative u_i (and *vice versa*). Therefore, we can interpret the outcome $y_i = 1$ as indicating that the explanatory factors and disturbance faced by individual *i* have combined to produce a positive net benefit.

For example, an individual might have a low income (which would otherwise suggest that new car purchase was not likely) but may have a sibling who works for Toyota and can arrange for an advantageous price on a new vehicle. We do not observe that circumstance, so it becomes a large positive u_i, explaining how $(X\beta + u_i) > 0$ for that individual.

The two common estimators of the binary choice model are the *binomial probit* and *binomial logit* models. For the probit model, $\Psi(\cdot)$ is the *CDF* of the Normal distribution function (Stata's `norm` function):

$$Pr[y = 1|X] = \int_{-\infty}^{X\beta} \psi(t)dt = \Psi(X\beta) \tag{5}$$

where $\psi(\cdot)$ is the probability density function (*PDF*) of the Normal distribution: Stata's `normden` function.

For the logit model, $\Psi(\cdot)$ is the *CDF* of the Logistic distribution:

$$Pr[y = 1|X] = \frac{\exp(X\beta)}{1 + \exp(X\beta)} \tag{6}$$

The two common estimators of the binary choice model are the *binomial probit* and *binomial logit* models. For the probit model, $\Psi(\cdot)$ is the *CDF* of the Normal distribution function (Stata's `norm` function):

$$Pr[y = 1|X] = \int_{-\infty}^{X\beta} \psi(t)dt = \Psi(X\beta) \tag{5}$$

where $\psi(\cdot)$ is the probability density function (*PDF*) of the Normal distribution: Stata's `normden` function.

For the logit model, $\Psi(\cdot)$ is the *CDF* of the Logistic distribution:

$$Pr[y = 1|X] = \frac{\exp(X\beta)}{1 + \exp(X\beta)} \tag{6}$$

The two models will produce quite similar results if the distribution of sample values of y_i is not too extreme. However, a sample in which the proportion $y_i = 1$ (or the proportion $y_i = 0$) is very small will be sensitive to the choice of *CDF*. Neither of these cases are really amenable to the binary choice model.

If a very unusual event is being modeled by y_i, the "naïve model" that it will not happen in any event is hard to beat. The same is true for an event that is almost ubiquitous: the naïve model that predicts that everyone has eaten a candy bar at some time in their lives is quite accurate.

The two models will produce quite similar results if the distribution of sample values of y_i is not too extreme. However, a sample in which the proportion $y_i = 1$ (or the proportion $y_i = 0$) is very small will be sensitive to the choice of *CDF*. Neither of these cases are really amenable to the binary choice model.

If a very unusual event is being modeled by y_i, the "naïve model" that it will not happen in any event is hard to beat. The same is true for an event that is almost ubiquitous: the naïve model that predicts that everyone has eaten a candy bar at some time in their lives is quite accurate.

We may estimate these binary choice models in Stata with the commands `probit` and `logit`, respectively. Both commands assume that the response variable is coded with zeros indicating a negative outcome and a positive, non-missing value corresponding to a positive outcome (i.e., I purchased a new car in 2005). These commands do not require that the variable be coded {0,1}, although that is often the case.

Because any positive value (including all missing values) will be taken as a positive outcome, it is important to ensure that missing values of the response variable are excluded from the estimation sample either by dropping those observations or using an `if !mi(`*depvar*`)` qualifier.

We may estimate these binary choice models in Stata with the commands `probit` and `logit`, respectively. Both commands assume that the response variable is coded with zeros indicating a negative outcome and a positive, non-missing value corresponding to a positive outcome (i.e., I purchased a new car in 2005). These commands do not require that the variable be coded {0,1}, although that is often the case.

Because any positive value (including all missing values) will be taken as a positive outcome, it is important to ensure that missing values of the response variable are excluded from the estimation sample either by dropping those observations or using an `if !mi(` *depvar* `)` qualifier.

One of the major challenges in working with limited dependent variable models is the complexity of explanatory factors' marginal effects on the result of interest. That complexity arises from the nonlinearity of the relationship. In Equation (4), the latent measure is translated by $\Psi(y_i^*)$ to a probability that $y_i = 1$. While Equation (2) is a linear relationship in the β parameters, Equation (4) is not. Therefore, although X_j has a linear effect on y_i^*, it will not have a linear effect on the resulting probability that $y = 1$:

$$\frac{\partial Pr[y = 1|X]}{\partial X_j} = \frac{\partial Pr[y = 1|X]}{\partial X\beta} \cdot \frac{\partial X\beta}{\partial X_j} =$$
$$\Psi'(X\beta) \cdot \beta_j = \psi(X\beta) \cdot \beta_j.$$

The probability that $y_i = 1$ is not constant over the data. Via the chain rule, we see that the effect of an increase in X_j on the probability is the product of two factors: the effect of X_j on the latent variable and the derivative of the *CDF* evaluated at y_i^*. The latter term, $\psi(\cdot)$, is the probability density function (*PDF*) of the distribution.

One of the major challenges in working with limited dependent variable models is the complexity of explanatory factors' marginal effects on the result of interest. That complexity arises from the nonlinearity of the relationship. In Equation (4), the latent measure is translated by $\Psi(y_i^*)$ to a probability that $y_i = 1$. While Equation (2) is a linear relationship in the β parameters, Equation (4) is not. Therefore, although X_j has a linear effect on y_i^*, it will not have a linear effect on the resulting probability that $y = 1$:

$$\frac{\partial Pr[y = 1|X]}{\partial X_j} = \frac{\partial Pr[y = 1|X]}{\partial X\beta} \cdot \frac{\partial X\beta}{\partial X_j} =$$
$$\Psi'(X\beta) \cdot \beta_j = \psi(X\beta) \cdot \beta_j.$$

The probability that $y_i = 1$ is not constant over the data. Via the chain rule, we see that the effect of an increase in X_j on the probability is the product of two factors: the effect of X_j on the latent variable and the derivative of the *CDF* evaluated at y_i^*. The latter term, $\psi(\cdot)$, is the probability density function (*PDF*) of the distribution.

In a binary choice model, the marginal effect of an increase in factor X_j *cannot* have a constant effect on the conditional probability that $(y = 1|X)$ since $\Psi(\cdot)$ varies through the range of X values. In a linear regression model, the coefficient β_j and its estimate b_j measures the marginal effect $\partial y/\partial X_j$, and that effect is constant for all values of X. In a binary choice model, where the probability that $y_i = 1$ is bounded by the {0,1} interval, the marginal effect *must* vary.

For instance, the marginal effect of a one dollar increase in disposable income on the conditional probability that $(y = 1|X)$ must approach zero as X_j increases. Therefore, the marginal effect in such a model varies continuously throughout the range of X_j, and must approach zero for both very low and very high levels of X_j.

In a binary choice model, the marginal effect of an increase in factor X_j *cannot* have a constant effect on the conditional probability that $(y = 1|X)$ since $\Psi(\cdot)$ varies through the range of X values. In a linear regression model, the coefficient β_j and its estimate b_j measures the marginal effect $\partial y/\partial X_j$, and that effect is constant for all values of X. In a binary choice model, where the probability that $y_i = 1$ is bounded by the {0,1} interval, the marginal effect *must* vary.

For instance, the marginal effect of a one dollar increase in disposable income on the conditional probability that $(y = 1|X)$ must approach zero as X_j increases. Therefore, the marginal effect in such a model varies continuously throughout the range of X_j, and must approach zero for both very low and very high levels of X_j.

When using Stata's `probit` (or `logit`) command, the reported coefficients (computed via maximum likelihood) are *b*, corresponding to β. You can use `margins` to compute the marginal effects. If a `probit` estimation is followed by the command `margins, dydx(_all)`, the `dF/dx` values will be calculated.

The `margins` command's `at()` option can be used to compute the effects at a particular point in the sample space. The `margins` command may also be used to calculate elasticities and semi-elasticities.

When using Stata's `probit` (or `logit`) command, the reported coefficients (computed via maximum likelihood) are *b*, corresponding to β. You can use `margins` to compute the marginal effects. If a `probit` estimation is followed by the command `margins, dydx(_all)`, the `dF/dx` values will be calculated.

The `margins` command's `at()` option can be used to compute the effects at a particular point in the sample space. The `margins` command may also be used to calculate elasticities and semi-elasticities.

After estimating a probit model, the `predict` command may be used, with a default option `p`, the predicted probability of a positive outcome. The `xb` option may be used to calculate the *index function* for each observation: that is, the predicted value of y_i^* from Equation (4), which is in z-units (those of a standard Normal variable). For instance, an index function value of 1.69 will be associated with a predicted probability of 0.95 in a large sample.

We use a modified version of the `womenwk` Reference Manual dataset, which contains information on 2,000 women, 657 of which are not recorded as wage earners. The indicator variable *work* is set to zero for the non-working and to one for those reporting positive wages.

```
. summarize work age married children education
```

Variable	Obs	Mean	Std. Dev.	Min	Max
work	2000	.6715	.4697852	0	1
age	2000	36.208	8.28656	20	59
married	2000	.6705	.4701492	0	1
children	2000	1.6445	1.398963	0	5
education	2000	13.084	3.045912	10	20

We estimate a probit model of the decision to work depending on the woman's age, marital status, number of children and level of education.

```
. probit work age  married children education, nolog
Probit regression                                 Number of obs   =       2000
                                                  LR chi2(4)      =     478.32
                                                  Prob > chi2     =     0.0000
Log likelihood = -1027.0616                       Pseudo R2       =     0.1889
```

work	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0347211	.0042293	8.21	0.000	.0264318	.0430105
married	.4308575	.074208	5.81	0.000	.2854125	.5763025
children	.4473249	.0287417	15.56	0.000	.3909922	.5036576
education	.0583645	.0109742	5.32	0.000	.0368555	.0798735
_cons	-2.467365	.1925635	-12.81	0.000	-2.844782	-2.089948

Surprisingly, the effect of additional children in the household increases the likelihood that the woman will work.

Average marginal effects (AMEs) are computed via `margins`.

```
. margins, dydx(_all)
Average marginal effects                        Number of obs   =       2000
Model VCE    : OIM
Expression   : Pr(work), predict()
dy/dx w.r.t. : age married children education
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0100768	.0011647	8.65	0.000	.0077941	.0123595
married	.1250441	.0210541	5.94	0.000	.0837788	.1663094
children	.1298233	.0068418	18.98	0.000	.1164137	.1432329
education	.0169386	.0031183	5.43	0.000	.0108269	.0230504

The marginal effects imply that married women have a 12.5% higher probability of labor force participation, while the addition of a child is associated with an 13% increase in participation.

Average marginal effects (AMEs) are computed via `margins`.

```
. margins, dydx(_all)
Average marginal effects                        Number of obs   =       2000
Model VCE    : OIM
Expression   : Pr(work), predict()
dy/dx w.r.t. : age married children education
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0100768	.0011647	8.65	0.000	.0077941	.0123595
married	.1250441	.0210541	5.94	0.000	.0837788	.1663094
children	.1298233	.0068418	18.98	0.000	.1164137	.1432329
education	.0169386	.0031183	5.43	0.000	.0108269	.0230504

The marginal effects imply that married women have a 12.5% higher probability of labor force participation, while the addition of a child is associated with an 13% increase in participation.

When the Logistic *CDF* is employed, the probability (π_i) of $y = 1$, conditioned on X, is $\exp(X\beta)/(1 + \exp(X\beta)$. Unlike the *CDF* of the Normal distribution, which lacks an inverse in closed form, this function may be inverted to yield

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = X_i\beta. \tag{7}$$

This expression is termed the *logit* of π_i, with that term being a contraction of the *log of the odds ratio*. The *odds ratio* reexpresses the probability in terms of the odds of $y = 1$.

As the logit of $\pi_i = X_i\beta$, it follows that the *odds ratio* for a one-unit change in the j^{th} X, holding other X constant, is merely $\exp(\beta_j)$. When we estimate a `logit` model, the `or` option specifies that odds ratios are to be displayed rather than coefficients.

If the odds ratio exceeds unity, an increase in that X increases the likelihood that $y = 1$, and vice versa. Estimated standard errors for the odds ratios are calculated via the delta method.

As the logit of $\pi_i = X_i\beta$, it follows that the *odds ratio* for a one-unit change in the j^{th} X, holding other X constant, is merely $\exp(\beta_j)$. When we estimate a `logit` model, the `or` option specifies that odds ratios are to be displayed rather than coefficients.

If the odds ratio exceeds unity, an increase in that X increases the likelihood that $y = 1$, and vice versa. Estimated standard errors for the odds ratios are calculated via the delta method.

We can define the logit, or log of the odds ratio, in terms of grouped data (averages of microdata). For instance, in the 2004 U.S. presidential election, the *ex post* probability of a Massachusetts resident voting for John Kerry was 0.62, with a logit of $\log(0.62/(1-0.62)) = 0.4895$. The probability of that person voting for George Bush was 0.37, with a logit of -0.5322. Say that we had such data for all 50 states. It would be inappropriate to use linear regression on the probabilities *voteKerry* and *voteBush*, just as it would be inappropriate to run a regression on individual voter's *voteKerry* and *voteBush* indicator variables.

In this case, Stata's `glogit` (grouped logit) command may be used to produce weighted least squares estimates for the model on state-level data. Alternatively, the `blogit` command may be used to produce maximum-likelihood estimates of that model on grouped (or "blocked") data.

The equivalent commands `gprobit` and `bprobit` may be used to fit a probit model to grouped data.

Estimation with ordinal data

We earlier discussed the issues related to the use of *ordinal variables*: those which indicate a ranking of responses, rather than a cardinal measure, such as the codes of a Likert scale of agreement with a statement. Since the values of such an ordered response are arbitrary, an ordinal variable should not be treated as if it was measurable in a cardinal sense and entered into a regression, either as a response variable or as a regressor.

However, what if we want to model an ordinal variable as the response variable, given a set of explanatory factors? Just as we can use binary choice models to evaluate the factors underlying a decision without being able to quantify the net benefit of making that choice, we may employ a generalization of the binary choice framework to model an ordinal variable using *ordered probit* or *ordered logit* estimation techniques.

Estimation with ordinal data

We earlier discussed the issues related to the use of *ordinal variables*: those which indicate a ranking of responses, rather than a cardinal measure, such as the codes of a Likert scale of agreement with a statement. Since the values of such an ordered response are arbitrary, an ordinal variable should not be treated as if it was measurable in a cardinal sense and entered into a regression, either as a response variable or as a regressor.

However, what if we want to model an ordinal variable as the response variable, given a set of explanatory factors? Just as we can use binary choice models to evaluate the factors underlying a decision without being able to quantify the net benefit of making that choice, we may employ a generalization of the binary choice framework to model an ordinal variable using *ordered probit* or *ordered logit* estimation techniques.

In the latent variable approach to the binary choice model, we observe $y_i = 1$ if the individual's net benefit is positive: i.e., $y_i^* > 0$. The ordered choice model generalizes this concept to the notion of multiple thresholds. For instance, a variable recorded on a five-point Likert scale will have four thresholds. If $y^* \leq \kappa_1$, we observe $y = 1$. If $\kappa_1 < y^* \leq \kappa_2$, we observe $y = 2$. If $\kappa_2 < y^* \leq \kappa_3$, we observe $y = 3$, and so on, where the κ values are the thresholds. In a sense, this can be considered imprecise measurement: we cannot observe y^* directly, but only the range in which it falls.

The parameters to be estimated are a set of coefficients β corresponding to the explanatory factors in X as well as a set of $(I-1)$ threshold coefficients κ corresponding to the I alternatives. In Stata's implementation of these estimators via commands `oprobit` and `ologit`, the actual values of the response variable are not relevant. Larger values are taken to correspond to higher outcomes. If there are I possible outcomes (e.g., 5 for the Likert scale), a set of threshold coefficients or *cut points* $\{\kappa_1, \kappa_2, \ldots, \kappa_{I-1}\}$ is defined, where $\kappa_0 = -\infty$ and $\kappa_I = \infty$.

Then the model for the j^{th} observation defines:

$$Pr[y_j = i] = Pr[\kappa_{i-1} < \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_k X_{kj} + u_j < \kappa_i]$$

where the probability that individual j will choose outcome i depends on the product $X\beta$ falling between cut points $(i-1)$ and i. This is a direct generalization of the two-outcome binary choice model, which has a single threshold at zero. As in the binomial probit model, we assume that the error is normally distributed with variance unity (or distributed Logistic with variance $\pi^2/3$ in the case of ordered logit).

We may estimate these binary choice models in Stata with the commands `oprobit` and `ologit`, respectively. We illustrate the ordered probit and logit techniques with a model of automobile reliability. The `fullauto` data set contains information on 66 automobiles' repair records, on a five-point scale (1=poor, 5=excellent).

```
. tab rep77
```

Repair Record 1977	Freq.	Percent	Cum.
Poor	3	4.55	4.55
Fair	11	16.67	21.21
Average	27	40.91	62.12
Good	20	30.30	92.42
Excellent	5	7.58	100.00
Total	66	100.00	

We estimate the model with `oprobit`; the model's predictions are quantitatively similar if `ologit` is employed.

```
. ologit rep77 foreign length mpg, nolog
Ordered logistic regression                     Number of obs   =         66
                                                LR chi2(3)      =      23.29
                                                Prob > chi2     =     0.0000
Log likelihood = -78.250719                     Pseudo R2       =     0.1295
```

rep77	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
foreign	2.896807	.7906411	3.66	0.000	1.347179	4.446435
length	.0828275	.02272	3.65	0.000	.0382972	.1273579
mpg	.2307677	.0704548	3.28	0.001	.0926788	.3688566
/cut1	17.92748	5.551191			7.047344	28.80761
/cut2	19.86506	5.59648			8.896161	30.83396
/cut3	22.10331	5.708936			10.914	33.29262
/cut4	24.69213	5.890754			13.14647	36.2378

We find that all three explanatory factors have significant, positive effects on the repair record.

Following the `ologit` estimation, we employ `predict` to compute the predicted probabilities of achieving each repair record. We then examine the automobiles who were classified as most likely to have a poor rating and an excellent rating, respectively.

```
. predict poor fair avg good excellent if e(sample)
(option pr assumed; predicted probabilities)
. summarize poor, meanonly
. list poor fair avg good excellent rep77 make if poor==r(max), noobs
```

poor	fair	avg	good	excell~t	rep77	make
.4195219	.4142841	.14538	.01922	.001594	Poor	AMC

```
. summarize excellent, meanonly
. list poor fair avg good excellent rep77 make if excellent==r(max), noobs
```

poor	fair	avg	good	excell~t	rep77	make
.0006963	.0041173	.0385734	.3331164	.6234967	Good	VW

The AMC Pacer and VW Diesel were those vehicles, respectively.

Truncation

We turn now to a context where the response variable is not binary nor necessarily integer, but subject to *truncation*. This is a bit trickier, since a truncated or censored response variable may not be obviously so. We must fully understand the context in which the data were generated. Nevertheless, it is quite important that we identify situations of *truncated* or *censored* response variables. Utilizing these variables as the dependent variable in a regression equation without consideration of these qualities will be misleading.

In the case of *truncation* the sample is drawn from a subset of the population so that only certain values are included in the sample. We lack observations on both the response variable and explanatory variables. For instance, we might have a sample of individuals who have a high school diploma, some college experience, or one or more college degrees. The sample has been generated by interviewing those who completed high school.

This is a *truncated* sample, relative to the population, in that it excludes all individuals who have not completed high school. The characteristics of those excluded individuals are not likely to be the same as those in our sample. For instance, we might expect that average or median income of dropouts is lower than that of graduates.

In the case of *truncation* the sample is drawn from a subset of the population so that only certain values are included in the sample. We lack observations on both the response variable and explanatory variables. For instance, we might have a sample of individuals who have a high school diploma, some college experience, or one or more college degrees. The sample has been generated by interviewing those who completed high school.

This is a *truncated* sample, relative to the population, in that it excludes all individuals who have not completed high school. The characteristics of those excluded individuals are not likely to be the same as those in our sample. For instance, we might expect that average or median income of dropouts is lower than that of graduates.

The effect of truncating the distribution of a random variable is clear. The expected value or mean of the truncated random variable moves away from the truncation point and the variance is reduced. Descriptive statistics on the level of education in our sample should make that clear: with the minimum years of education set to 12, the mean education level is higher than it would be if high school dropouts were included, and the variance will be smaller.

In the subpopulation defined by a truncated sample, we have no information about the characteristics of those who were excluded. For instance, we do not know whether the proportion of minority high school dropouts exceeds the proportion of minorities in the population.

The effect of truncating the distribution of a random variable is clear. The expected value or mean of the truncated random variable moves away from the truncation point and the variance is reduced. Descriptive statistics on the level of education in our sample should make that clear: with the minimum years of education set to 12, the mean education level is higher than it would be if high school dropouts were included, and the variance will be smaller.

In the subpopulation defined by a truncated sample, we have no information about the characteristics of those who were excluded. For instance, we do not know whether the proportion of minority high school dropouts exceeds the proportion of minorities in the population.

A sample from this truncated population cannot be used to make inferences about the entire population without correction for the fact that those excluded individuals are not randomly selected from the population at large. While it might appear that we could use these truncated data to make inferences about the subpopulation, we cannot even do that.

A regression estimated from the subpopulation will yield coefficients that are biased toward zero—or *attenuated*—as well as an estimate of σ_u^2 that is biased downward.

A sample from this truncated population cannot be used to make inferences about the entire population without correction for the fact that those excluded individuals are not randomly selected from the population at large. While it might appear that we could use these truncated data to make inferences about the subpopulation, we cannot even do that.

A regression estimated from the subpopulation will yield coefficients that are biased toward zero—or *attenuated*—as well as an estimate of σ_u^2 that is biased downward.

If we are dealing with a truncated Normal distribution, where $y = X\beta + u$ is only observed if it exceeds τ, we may define:

$$\begin{aligned} \alpha_i &= (\tau - X_i\beta)/\sigma_u \\ \lambda(\alpha_i) &= \frac{\phi(\alpha_i)}{(1 - \Phi(\alpha_i))} \end{aligned} \tag{8}$$

where σ_u is the standard error of the untruncated disturbance u, $\phi(\cdot)$ is the Normal density function (*PDF*) and $\Phi(\cdot)$ is the Normal *CDF*. The expression $\lambda(\alpha_i)$ is termed the *inverse Mills ratio*, or *IMR*.

If a regression is estimated from the truncated sample, we find that

$$[y_i|y_i > \tau, X_i] = X_i\beta + \sigma_u\lambda(\alpha_i) + u_i \tag{9}$$

These regression estimates suffer from the exclusion of the term $\lambda(\alpha_i)$. This regression is misspecified, and the effect of that misspecification will differ across observations, with a heteroskedastic error term whose variance depends on X_i. To deal with these problems, we include the *IMR* as an additional regressor. This allows us to use a truncated sample to make consistent inferences about the subpopulation.

If we can justify making the assumption that the regression errors in the *population* are Normally distributed, then we can estimate an equation for a truncated sample with the Stata command `truncreg`. Under the assumption of normality, inferences for the population may be made from the truncated regression model. The estimator used in this command assumes that the regression errors are Normal.

The `truncreg` option `ll(#)` is used to indicate that values of the response variable less than or equal to `#` are truncated. We might have a sample of college students with *yearsEduc* truncated from below at 12 years. Upper truncation can be handled by the `ul(#)` option: for instance, we may have a sample of individuals whose income is recorded up to $200,000. Both lower and upper truncation can be specified by combining the options.

If we can justify making the assumption that the regression errors in the *population* are Normally distributed, then we can estimate an equation for a truncated sample with the Stata command `truncreg`. Under the assumption of normality, inferences for the population may be made from the truncated regression model. The estimator used in this command assumes that the regression errors are Normal.

The `truncreg` option `ll(#)` is used to indicate that values of the response variable less than or equal to `#` are truncated. We might have a sample of college students with *yearsEduc* truncated from below at 12 years. Upper truncation can be handled by the `ul(#)` option: for instance, we may have a sample of individuals whose income is recorded up to $200,000. Both lower and upper truncation can be specified by combining the options.

The coefficient estimates and marginal effects from `truncreg` may be used to make inferences about the entire population, whereas the results from the misspecified regression model should not be used for any purpose.

We consider a sample of married women from the `laborsub` dataset whose hours of work are truncated from below at zero.

```
. use laborsub,clear
. summarize whrs kl6 k618 wa we
```

Variable	Obs	Mean	Std. Dev.	Min	Max
whrs	250	799.84	915.6035	0	4950
kl6	250	.236	.5112234	0	3
k618	250	1.364	1.370774	0	8
wa	250	42.92	8.426483	30	60
we	250	12.352	2.164912	5	17

The coefficient estimates and marginal effects from `truncreg` may be used to make inferences about the entire population, whereas the results from the misspecified regression model should not be used for any purpose.

We consider a sample of married women from the `laborsub` dataset whose hours of work are truncated from below at zero.

```
. use laborsub,clear
. summarize whrs kl6 k618 wa we
```

Variable	Obs	Mean	Std. Dev.	Min	Max
whrs	250	799.84	915.6035	0	4950
kl6	250	.236	.5112234	0	3
k618	250	1.364	1.370774	0	8
wa	250	42.92	8.426483	30	60
we	250	12.352	2.164912	5	17

To illustrate the consequences of ignoring truncation we estimate a model of hours worked with OLS, including only working women. The regressors include measures of the number of preschool children (`kl6`), number of school-age children (`k618`), age (`wa`) and years of education (`we`).

```
. regress whrs kl6 k618 wa we if whrs>0
```

Source	SS	df	MS
Model	7326995.15	4	1831748.79
Residual	94793104.2	145	653745.546
Total	102120099	149	685369.794

Number of obs	150
F(4, 145)	2.80
Prob > F	0.0281
R-squared	0.0717
Adj R-squared	0.0461
Root MSE	808.55

whrs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
kl6	-421.4822	167.9734	-2.51	0.013	-753.4748	-89.48953
k618	-104.4571	54.18616	-1.93	0.056	-211.5538	2.639668
wa	-4.784917	9.690502	-0.49	0.622	-23.9378	14.36797
we	9.353195	31.23793	0.30	0.765	-52.38731	71.0937
_cons	1629.817	615.1301	2.65	0.009	414.0371	2845.597

We now reestimate the model with `truncreg`, taking into account that 100 of the 250 observations have zero recorded `whrs`:

```
. truncreg whrs kl6 k618 wa we, ll(0) nolog
(note: 100 obs. truncated)
Truncated regression
Limit:   lower =          0                      Number of obs =     150
         upper =       +inf                      Wald chi2(4)  =   10.05
Log likelihood = -1200.9157                      Prob > chi2   =  0.0395
```

whrs	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
eq1						
kl6	-803.0042	321.3614	-2.50	0.012	-1432.861	-173.1474
k618	-172.875	88.72898	-1.95	0.051	-346.7806	1.030579
wa	-8.821123	14.36848	-0.61	0.539	-36.98283	19.34059
we	16.52873	46.50375	0.36	0.722	-74.61695	107.6744
_cons	1586.26	912.355	1.74	0.082	-201.9233	3374.442
sigma						
_cons	983.7262	94.44303	10.42	0.000	798.6213	1168.831

The effect of truncation in the subsample is quite apparent. Some of the attenuated coefficient estimates from `regress` are no more than half as large as their counterparts from `truncreg`. The parameter `sigma _cons`, comparable to `Root MSE` in the OLS regression, is considerably larger in the truncated regression reflecting its downward bias in a truncated sample.

Censoring

Let us now turn to another commonly encountered issue with the data: *censoring*. Unlike truncation, in which the distribution from which the sample was drawn is a non-randomly selected subpopulation, censoring occurs when a response variable is set to an arbitrary value above or below a certain value: the *censoring point*. In contrast to the truncated case, we have observations on the explanatory variables in this sample. The problem of censoring is that we do not have observations on the response variable for certain individuals. For instance, we may have full demographic information on a set of individuals, but only observe the number of hours worked per week for those who are employed.

As another example of a censored variable, consider that the numeric response to the question “How much did you spend on a new car last year?” may be zero for many individuals, but that should be considered as the expression of their choice not to buy a car.

Such a censored response variable should be considered as being generated by a mixture of distributions: the binary choice to purchase a car or not, and the continuous response of how much to spend conditional on choosing to purchase. Although it would appear that the variable *caroutlay* could be used as the dependent variable in a regression, it should not be employed in that manner, since it is generated by a censored distribution.

As another example of a censored variable, consider that the numeric response to the question “How much did you spend on a new car last year?” may be zero for many individuals, but that should be considered as the expression of their choice not to buy a car.

Such a censored response variable should be considered as being generated by a mixture of distributions: the binary choice to purchase a car or not, and the continuous response of how much to spend conditional on choosing to purchase. Although it would appear that the variable *caroutlay* could be used as the dependent variable in a regression, it should not be employed in that manner, since it is generated by a censored distribution.

A solution to this problem was first proposed by Tobin (1958) as the *censored regression* model; it became known as "Tobin's probit" or the *tobit* model.The model can be expressed in terms of a latent variable:

$$\begin{aligned} y_i^* &= X\beta + u \\ y_i &= 0 \ \texttt{if} \ y_i^* \leq 0 \\ y_i &= y_i^* \ \texttt{if} \ y_i^* > 0 \end{aligned} \tag{10}$$

As in the prior example, our variable y_i contains either zeros for non-purchasers or a dollar amount for those who chose to buy a car last year. The model combines aspects of the binomial probit for the distinction of $y_i = 0$ versus $y_i > 0$ and the regression model for $[y_i | y_i > 0]$. Of course, we could collapse all positive observations on y_i and treat this as a binomial probit (or logit) estimation problem, but that would discard the information on the dollar amounts spent by purchasers. Likewise, we could throw away the $y_i = 0$ observations, but we would then be left with a truncated distribution, with the various problems that creates.

To take account of all of the information in y_i properly, we must estimate the model with the `tobit` estimation method, which employs maximum likelihood to combine the probit and regression components of the log-likelihood function.

As in the prior example, our variable y_i contains either zeros for non-purchasers or a dollar amount for those who chose to buy a car last year. The model combines aspects of the binomial probit for the distinction of $y_i = 0$ versus $y_i > 0$ and the regression model for $[y_i | y_i > 0]$. Of course, we could collapse all positive observations on y_i and treat this as a binomial probit (or logit) estimation problem, but that would discard the information on the dollar amounts spent by purchasers. Likewise, we could throw away the $y_i = 0$ observations, but we would then be left with a truncated distribution, with the various problems that creates.

To take account of all of the information in y_i properly, we must estimate the model with the `tobit` estimation method, which employs maximum likelihood to combine the probit and regression components of the log-likelihood function.

Tobit models may be defined with a threshold other than zero. Censoring from below may be specified at any point on the *y* scale with the `ll(#)` option for *left censoring*. Similarly, the standard tobit formulation may employ an upper threshold (censoring from above, or *right censoring*) using the `ul(#)` option to specify the upper limit. Stata's `tobit` also supports the *two-limit tobit* model where observations on *y* are censored from both left and right by specifying both the `ll(#)` and `ul(#)` options.

Even in the case of a single censoring point, predictions from the tobit model are quite complex, since one may want to calculate the regression-like `xb` with `predict`, but could also compute the predicted probability that $[y|X]$ falls within a particular interval (which may be open-ended on left or right).This may be specified with the `pr(a,b)` option, where arguments `a, b` specify the limits of the interval; the missing value code (`.`) is taken to mean infinity (of either sign).

Another `predict` option, `e(a,b)`, calculates the expectation $Ey = E[X\beta + u]$ conditional on $[y|X]$ being in the a, b interval. Last, the `ystar(a,b)` option computes the prediction from Equation (10): a censored prediction, where the threshold is taken into account.

Even in the case of a single censoring point, predictions from the tobit model are quite complex, since one may want to calculate the regression-like `xb` with `predict`, but could also compute the predicted probability that $[y|X]$ falls within a particular interval (which may be open-ended on left or right).This may be specified with the `pr(a,b)` option, where arguments `a, b` specify the limits of the interval; the missing value code (`.`) is taken to mean infinity (of either sign).

Another `predict` option, `e(a,b)`, calculates the expectation $Ey = E[X\beta + u]$ conditional on $[y|X]$ being in the a, b interval. Last, the `ystar(a,b)` option computes the prediction from Equation (10): a censored prediction, where the threshold is taken into account.

The marginal effects of the tobit model are also quite complex. The estimated coefficients are the marginal effects of a change in X_j on y^* the unobservable latent variable:

$$\frac{\partial E(y^*|X_j)}{\partial X_j} = \beta_j \tag{11}$$

but that is not very useful. If instead we evaluate the effect on the observable y, we find that:

$$\frac{\partial E(y|X_j)}{\partial X_j} = \beta_j \times Pr[a < y_i^* < b] \tag{12}$$

where a, b are defined as above for `predict`. For instance, for left-censoring at zero, $a = 0, b = +\infty$. Since that probability is at most unity (and will be reduced by a larger proportion of censored observations), the marginal effect of X_j is attenuated from the reported coefficient toward zero.

An increase in an explanatory variable with a positive coefficient will imply that a left-censored individual is less likely to be censored. Their predicted probability of a nonzero value will increase. For a non-censored individual, an increase in X_j will imply that $E[y|y > 0]$ will increase. So, for instance, a decrease in the mortgage interest rate will allow more people to be homebuyers (since many borrowers' income will qualify them for a mortgage at lower interest rates), and allow prequalified homebuyers to purchase a more expensive home.

The marginal effect captures the combination of those effects. Since the newly-qualified homebuyers will be purchasing the cheapest homes, the effect of the lower interest rate on the average price at which homes are sold will incorporate both effects. We expect that it will increase the average transactions price, but due to attenuation, by a smaller amount than the regression function component of the model would indicate.

An increase in an explanatory variable with a positive coefficient will imply that a left-censored individual is less likely to be censored. Their predicted probability of a nonzero value will increase. For a non-censored individual, an increase in X_j will imply that $E[y|y > 0]$ will increase. So, for instance, a decrease in the mortgage interest rate will allow more people to be homebuyers (since many borrowers' income will qualify them for a mortgage at lower interest rates), and allow prequalified homebuyers to purchase a more expensive home.

The marginal effect captures the combination of those effects. Since the newly-qualified homebuyers will be purchasing the cheapest homes, the effect of the lower interest rate on the average price at which homes are sold will incorporate both effects. We expect that it will increase the average transactions price, but due to attenuation, by a smaller amount than the regression function component of the model would indicate.

We return to the `womenwk` data set used to illustrate binomial probit. We generate the log of the wage (`lw`) for working women and set `lwf` equal to `lw` for working women and zero for non-working women. This could be problematic if recorded wages below $1.00 were present in the data, but in these data the minimum wage recorded is $5.88. We first estimate the model with OLS ignoring the censored nature of the response variable.

```
. use womenwk,clear
. regress lwf age married children education
```

Source	SS	df	MS
Model	937.873188	4	234.468297
Residual	3485.34135	1995	1.74703827
Total	4423.21454	1999	2.21271363

Number of obs	= 2000
F(4, 1995)	= 134.21
Prob > F	= 0.0000
R-squared	= 0.2120
Adj R-squared	= 0.2105
Root MSE	= 1.3218

lwf	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0363624	.003862	9.42	0.000	.0287885	.0439362
married	.3188214	.0690834	4.62	0.000	.1833381	.4543046
children	.3305009	.0213143	15.51	0.000	.2887004	.3723015
education	.0843345	.0102295	8.24	0.000	.0642729	.1043961
_cons	-1.077738	.1703218	-6.33	0.000	-1.411765	-.7437105

Reestimating the model as a tobit and indicating that `lwf` is left-censored at zero with the `ll` option yields:

```
. tobit lwf age married children education, ll(0)
Tobit regression                                  Number of obs   =       2000
                                                  LR chi2(4)      =     461.85
                                                  Prob > chi2     =     0.0000
Log likelihood = -3349.9685                       Pseudo R2       =     0.0645

------------------------------------------------------------------------------
         lwf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .052157   .0057457     9.08   0.000     .0408888    .0634252
     married |   .4841801   .1035188     4.68   0.000     .2811639    .6871964
    children |   .4860021   .0317054    15.33   0.000     .4238229    .5481812
   education |   .1149492   .0150913     7.62   0.000     .0853529    .1445454
       _cons |  -2.807696   .2632565   -10.67   0.000    -3.323982   -2.291409
-------------+----------------------------------------------------------------
      /sigma |   1.872811    .040014                      1.794337    1.951285
------------------------------------------------------------------------------

  Obs. summary:        657  left-censored observations at lwf<=0
                      1343     uncensored observations
                         0 right-censored observations
```

The tobit estimates of `lwf` show positive, significant effects for age, marital status, the number of children and the number of years of education. Each of these factors is expected to both increase the probability that a woman will work as well as increase her wage conditional on employed status.

Following tobit estimation, we first generate the marginal effects of each explanatory variable on the probability that an individual will have a positive log(wage): the `pr(a,b)` option of `predict`.

```
. margins, dydx(*) predict(pr(0,.))
Average marginal effects                        Number of obs   =       2000
Model VCE    : OIM
Expression   : Pr(lwf>0), predict(pr(0,.))
dy/dx w.r.t. : age married children education
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0071483	.0007873	9.08	0.000	.0056052	.0086914
married	.0663585	.0142009	4.67	0.000	.0385254	.0941917
children	.0666082	.0044677	14.91	0.000	.0578516	.0753649
education	.0157542	.0020695	7.61	0.000	.0116981	.0198103

We then calculate the marginal effect of each explanatory variable on the expected log wage, given that the individual has not been censored (i.e., was working). These effects, unlike the estimated coefficients from `regress`, properly take into account the censored nature of the response variable.

```
. margins, dydx(*) predict(e(0,.))
Average marginal effects                        Number of obs   =       2000
Model VCE    : OIM
Expression   : E(lwf|lwf>0), predict(e(0,.))
dy/dx w.r.t. : age married children education
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0315183	.00347	9.08	0.000	.0247172	.0383194
married	.2925884	.0625056	4.68	0.000	.1700797	.4150971
children	.2936894	.0189659	15.49	0.000	.2565169	.3308619
education	.0694634	.0091252	7.61	0.000	.0515784	.0873484

Note, for instance, the much smaller marginal effects associated with number of children and level of education in tobit vs. regress.