2. Regression estimation and inference (review)

LPO 8852: Regression II

Sean P. Corcoran

Last time

Regression as a tool for estimating or approximating a *conditional expectation function* ($E[y|x]$)

- The CEF may or may not be *linear*
- The CEF may or may not be *causal*
- We often estimate (linear) population regression functions, either because we have reason to believe they represent the CEF (e.g., saturated regression) or as an approximation. The PRF is the best linear predictor of the CEF.

Last time

Regression as a tool for estimating or approximating a *conditional expectation function* ($E[y|x]$)

- The CEF may or may not be *linear*
- The CEF may or may not be *causal*
- We often estimate (linear) population regression functions, either because we have reason to believe they represent the CEF (e.g., saturated regression) or as an approximation. The PRF is the best linear predictor of the CEF.
- A *causal* CEF describes differences in *average potential outcomes*
- Example of the return to attending a selective private college, and omitted variables bias.

What econometric analysis is all about

Key activities in econometric analysis:

- Defining *estimands*: what is the quantity you are trying to estimate? Is it a *causal* parameter?
- *Estimators*: how (mechanically) will you estimate this quantity?
- *Inference*: what is the sampling distribution of your estimator in finite samples? In large samples? Needed to construct confidence intervals, conduct hypothesis tests

See handout on Github: *Wooldridge results* for a compact summary of key results from Wooldrige/Regression I

Sampling distributions: basics

Statistical properties of the OLS estimator of regression coefficients

- Since estimated regression coefficients are calculated from a random sample, *they are also random variables* and thus have their own distribution: a *sampling distribution*.
- This sampling distribution has its own mean and variance. The standard deviation of the sampling distribution is called the *standard error.*
- Assumptions about the distribution of u allows us to determine what this sampling distribution looks like.

Taking us back to $\bar{x}$

Suppose x is a random variable with mean μ and standard deviation σ_x. Now suppose we draw n observations at random from this distribution and calculate $\bar{x}$. The Central Limit Theorem tells us:

- If x has a normal distribution, then the sampling distribution of $\bar{x}$ is normal with a mean of μ and a standard error of $\sigma/\sqrt{n}$. Since its mean is the population μ, the estimator $\bar{x}$ is *unbiased*.

Taking us back to $\bar{x}$

Suppose x is a random variable with mean μ and standard deviation σ_x. Now suppose we draw n observations at random from this distribution and calculate $\bar{x}$. The Central Limit Theorem tells us:

- If x has a normal distribution, then the sampling distribution of $\bar{x}$ is normal with a mean of μ and a standard error of $\sigma/\sqrt{n}$. Since its mean is the population μ, the estimator $\bar{x}$ is *unbiased*.
- Even if x does not have a normal distribution, with a large enough n, the sampling distribution of $\bar{x}$ will be *approximately* normal with a mean of μ and a standard error of $\sigma/\sqrt{n}$. The estimator has an asymptotic normal distribution.

These are statements about the distribution of $\bar{x}$ in *repeated samples*. This is important to keep in mind. In practice, we only have one sample, and thus one draw of $\bar{x}$.

Taking us back to $\bar{x}$

Knowledge of the sampling distribution is powerful, as it allows us to make statements about the likelihood $\bar{x}$ takes on certain values.

- Hypothesis testing: we can calculate the likelihood that an observed value of $\bar{x}$ takes on a particular value under the null hypothesis that $\mu = \mu_0$.

Taking us back to $\bar{x}$

Knowledge of the sampling distribution is powerful, as it allows us to make statements about the likelihood $\bar{x}$ takes on certain values.

- Hypothesis testing: we can calculate the likelihood that an observed value of $\bar{x}$ takes on a particular value under the null hypothesis that $\mu = \mu_0$.
- Confidence intervals: we can calculate a range of likely values for μ. For example, a 95% CI will contain the true population μ in 95% of random samples.

Smaller standard errors give one *more precise* confidence intervals and *greater power* to reject the null hypothesis.

Confidence interval for μ

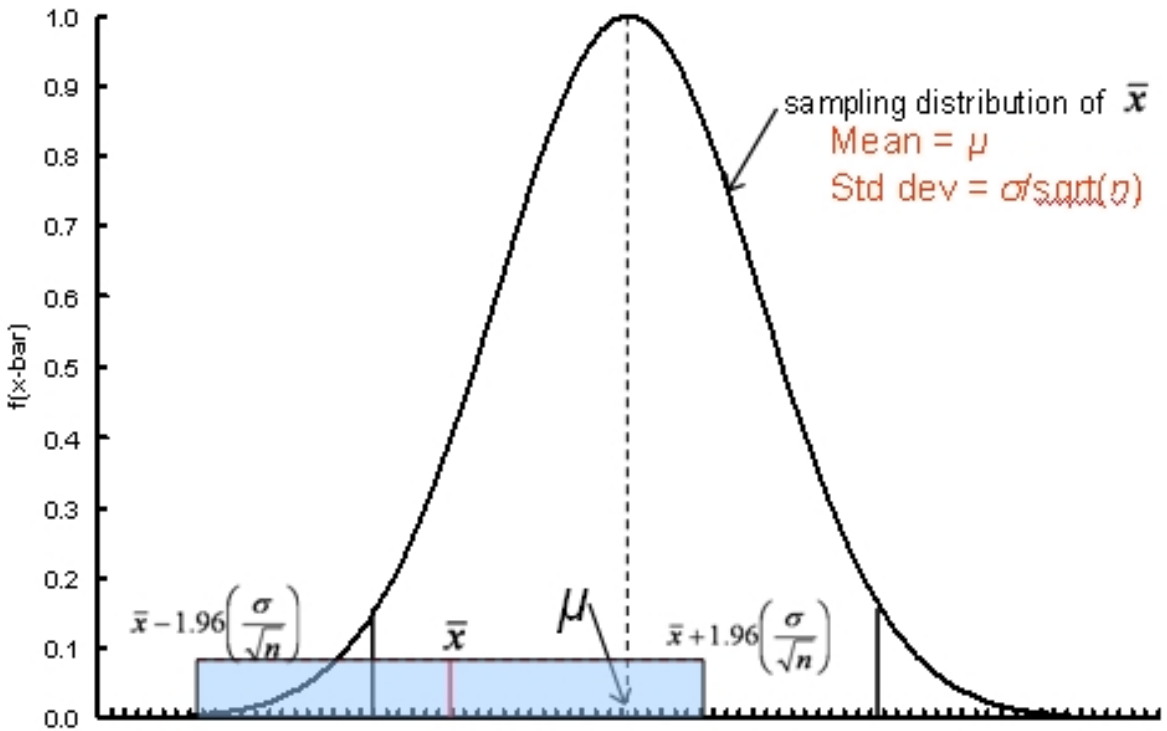


Figure: Confidence interval around $\bar{x}$

Note: assumed normal sampling distribution (e.g., large n)

Confidence interval for μ

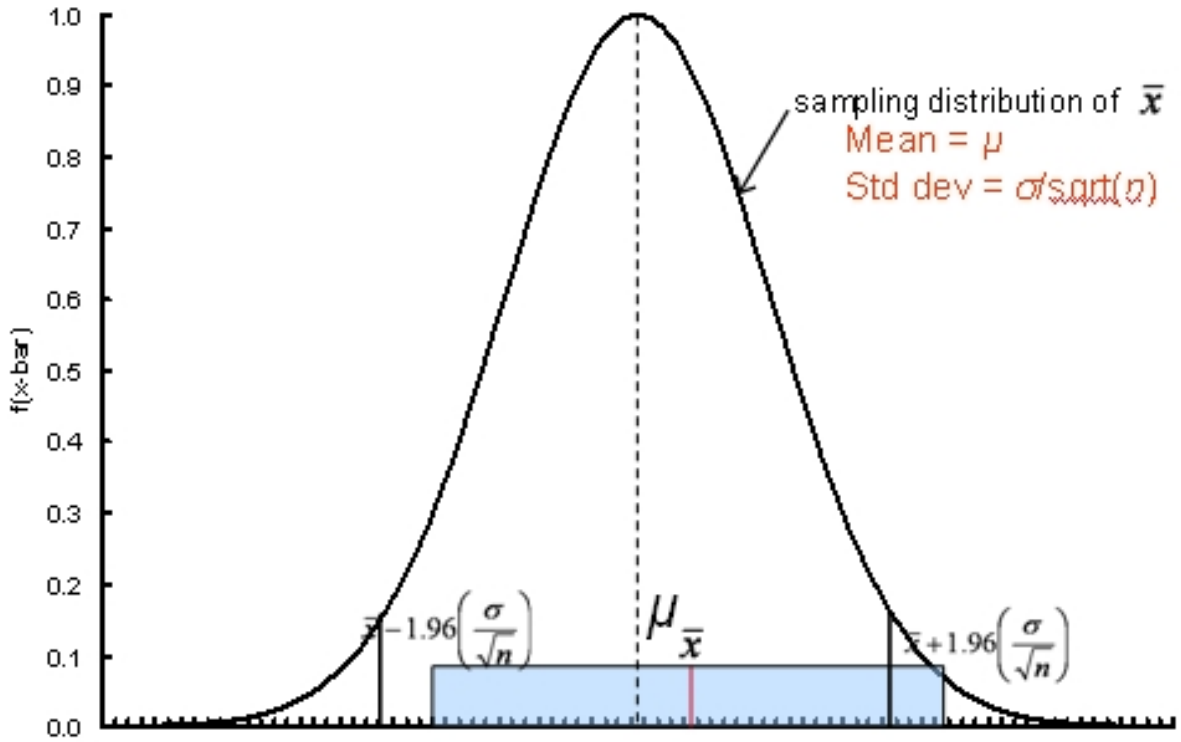


Figure: Confidence interval around $\bar{x}$

Confidence interval for μ

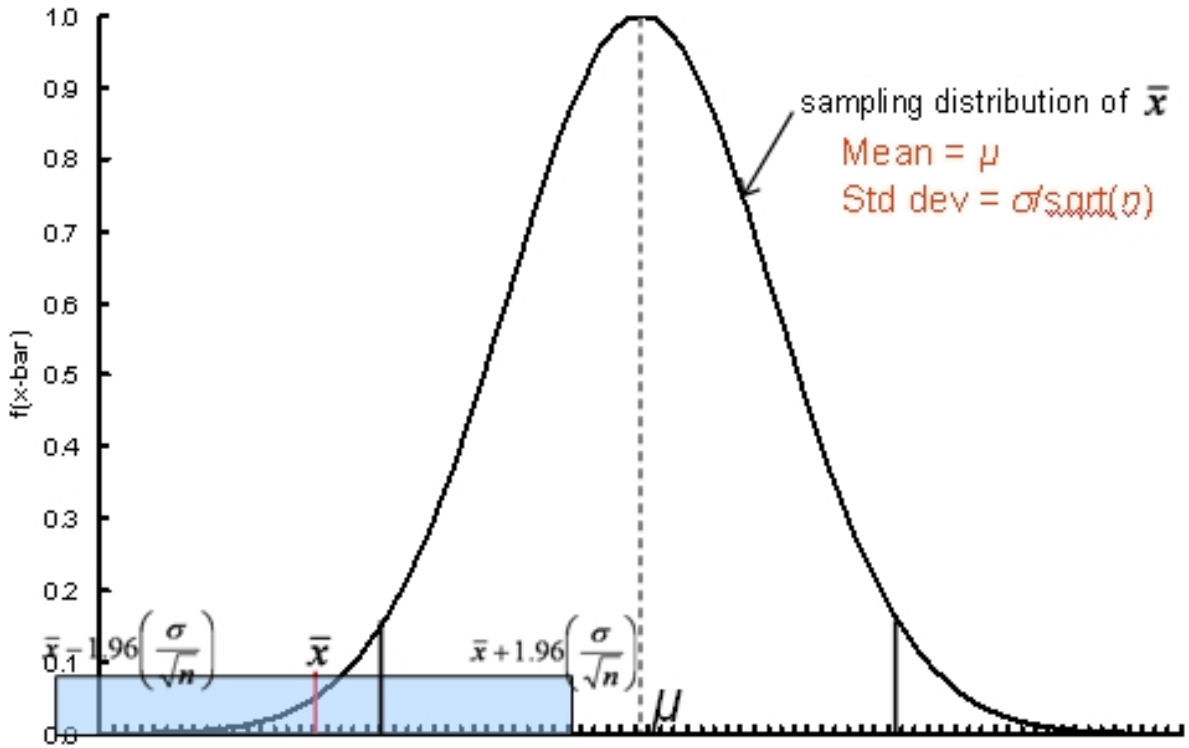


Figure: Confidence interval around $\bar{x}$

Simple linear regression: Gauss-Markov assumptions

- SLR1: $y = \beta_0 + \beta_1 x + u$ is the population model.
- SLR2: Data represent a random sample of n draws of (x_i, y_i) from the population.
- SLR3: $\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0$ (there is variation in x_i).
- SLR4: $E(u|x) = 0$ (zero conditional mean; exogenous explanatory variable)
- SLR5: $Var(u|x) = \sigma_u^2$ (homoskedasticity)
- SLR6: The error u has a *normal* distribution.

Simple linear regression: properties of $\hat{\beta}_1$

- Under SLR1-SLR4, $\hat{\beta}_1$ is *unbiased* and *consistent*.
- Under SLR1-SLR5: $Var(\hat{\beta}_1) = \frac{\sigma_u^2}{SST_x}$ and $se(\hat{\beta}_1) = \frac{\sigma_u}{\sqrt{SST_x}}$
  - In practice we estimate σ_u^2 using $\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{n-2}$ so then $\widehat{se(\hat{\beta}_1)} = \frac{\hat{\sigma}_u}{\sqrt{SST_x}}$
- Under SLR1-SLR5, the OLS estimator is BLUE (minimum variance).

Simple linear regression: properties of $\hat{\beta}_1$

- Under SLR1-SLR4, $\hat{\beta}_1$ is *unbiased* and *consistent*.
- Under SLR1-SLR5: $Var(\hat{\beta}_1) = \frac{\sigma_u^2}{SST_x}$ and $se(\hat{\beta}_1) = \frac{\sigma_u}{\sqrt{SST_x}}$
  - In practice we estimate σ_u^2 using $\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{n-2}$ so then $\widehat{se(\hat{\beta}_1)} = \frac{\hat{\sigma}_u}{\sqrt{SST_x}}$
- Under SLR1-SLR5, the OLS estimator is BLUE (minimum variance).
- Knowledge of the sampling distribution of $\hat{\beta}_1$ allows us to make statements about the likelihood $\hat{\beta}_1$ takes on certain values. E.g.:
  - Can use t-statistic to conduct hypothesis tests. In small samples requires SLR6.
  - Can construct $(1-\alpha)\%$ confidence intervals for β_1. In small samples requires SLR6.

Simple linear regression: properties of $\hat{\beta}_1$

- Under SLR1-SLR4, $\hat{\beta}_1$ is *unbiased* and *consistent*.
- Under SLR1-SLR5: $Var(\hat{\beta}_1) = \frac{\sigma_u^2}{SST_x}$ and $se(\hat{\beta}_1) = \frac{\sigma_u}{\sqrt{SST_x}}$
  - In practice we estimate σ_u^2 using $\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{n-2}$ so then $\widehat{se(\hat{\beta}_1)} = \frac{\hat{\sigma}_u}{\sqrt{SST_x}}$
- Under SLR1-SLR5, the OLS estimator is BLUE (minimum variance).
- Knowledge of the sampling distribution of $\hat{\beta}_1$ allows us to make statements about the likelihood $\hat{\beta}_1$ takes on certain values. E.g.:
  - Can use t-statistic to conduct hypothesis tests. In small samples requires SLR6.
  - Can construct $(1-\alpha)\%$ confidence intervals for β_1. In small samples requires SLR6.

Note: SST_x is the total variation in x: $\sum_{i=1}^{n}(x_i - \bar{x})^2$. We can also write the population standard error as $se(\hat{\beta}_1) = \frac{\sigma_u}{nVar(x)}$

Multiple linear regression: Gauss-Markov assumptions

- MLR1: $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$ is the population model.
- MLR2: Data represent a random sample of n draws of $(x_{1i}, x_{2i}, ..., x_{ki}, y_i)$ from the population.
- MLR3: There is no perfect collinearity in the x. (Fails if perfect collinearly between any two x_j or if $n < k + 1$)
- MLR4: $E(u|x_1, x_2, ..., x_k) = 0$ (exogenous explanatory variables)
- MLR5: $Var(u|x_1, x_2, ..., x_k) = \sigma_u^2$ (homoskedasticity)
- MLR6: The error u has a *normal* distribution.

Multiple linear regression: properties of $\hat{\beta}_j$

- Under MLR1-MLR4, $\hat{\beta}_j$ is *unbiased* and *consistent*.
- Under MLR1-MLR5: $Var(\hat{\beta}_j) = \frac{\sigma_u^2}{SST_j(1-R_j^2)}$ and $se(\hat{\beta}_1) = \frac{\sigma_u}{\sqrt{SST_j(1-R_j^2)}}$
  - In practice we estimate σ_u^2 using $\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{n-2}$
- Knowledge of the sampling distribution of $\hat{\beta}_j$ allows us to make statements about the likelihood $\hat{\beta}_j$ takes on certain values. E.g.:
  - Can use t-statistic to conduct hypothesis tests. In small samples requires MLR6.
  - Can construct $(1-\alpha)\%$ confidence intervals for β_j. In small samples requires MLR6.

Multiple linear regression: properties of $\hat{\beta}_j$

- Under MLR1-MLR4, $\hat{\beta}_j$ is *unbiased* and *consistent*.
- Under MLR1-MLR5: $Var(\hat{\beta}_j) = \frac{\sigma_u^2}{SST_j(1-R_j^2)}$ and

  $se(\hat{\beta}_1) = \frac{\sigma_u}{\sqrt{SST_j(1-R_j^2)}}$

  - In practice we estimate σ_u^2 using $\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{n-2}$
- Knowledge of the sampling distribution of $\hat{\beta}_j$ allows us to make statements about the likelihood $\hat{\beta}_j$ takes on certain values. E.g.:
  - Can use t-statistic to conduct hypothesis tests. In small samples requires MLR6.
  - Can construct $(1-\alpha)\%$ confidence intervals for β_j. In small samples requires MLR6.

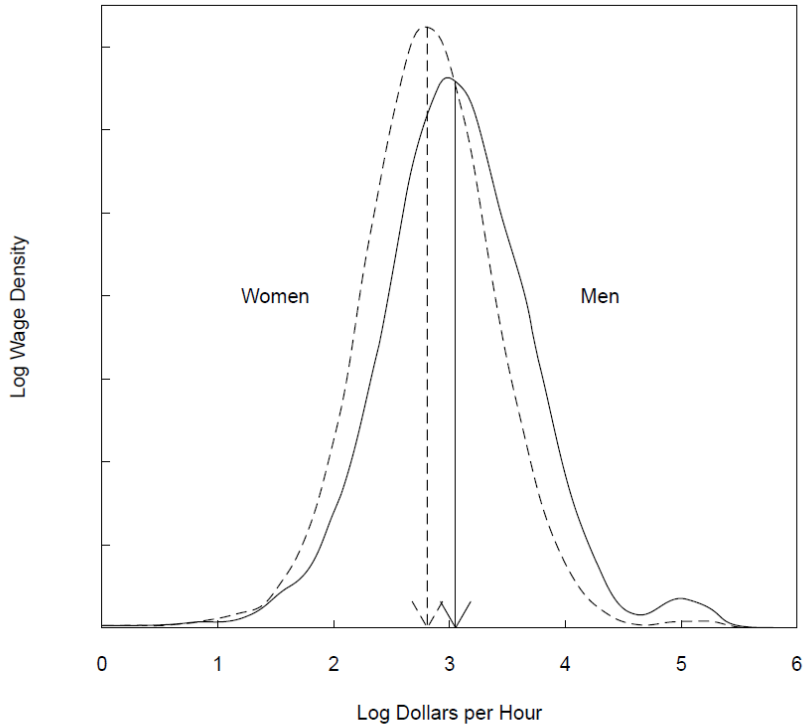
Note: SST_j is the total variation in x_j: $\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$. R_j^2 is the R-squared from a regression of x_j on all other x's

BLUE

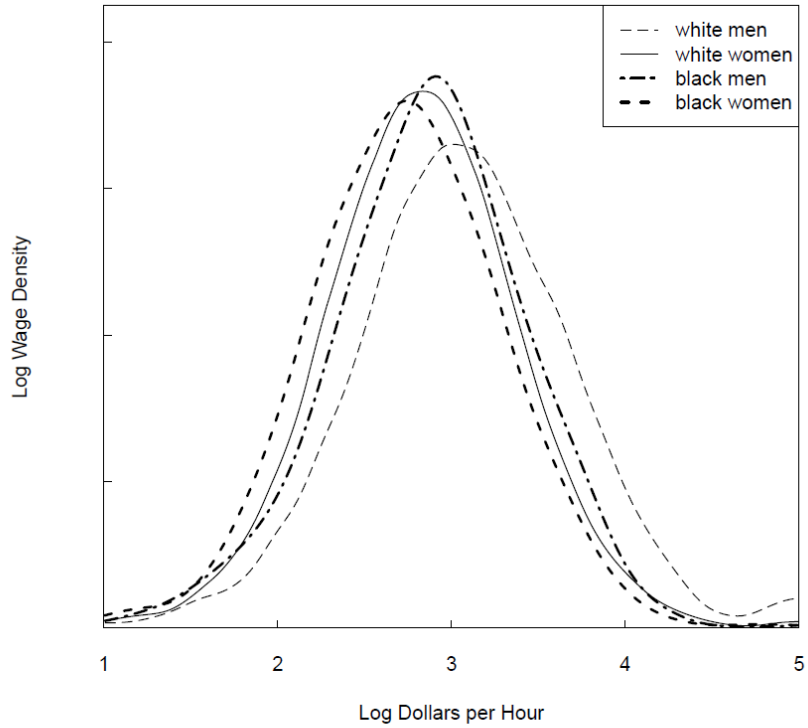
Under MLR1-MLR5, the OLS estimators are the *best linear unbiased estimators* of the population parameters, in that they have the smallest variance. (They are *efficient*).

Heteroskedasticity

The constant variance assumption (SLR5, MLR5) was made for *convenience*. It is the exception, not the rule. In practice, u is unlikely to have a fixed variance. *Heteroskedasticity* is when the variance of u changes across values of x.



(a) Women and Men



(b) By Sex and Race

Heteroskedasticity

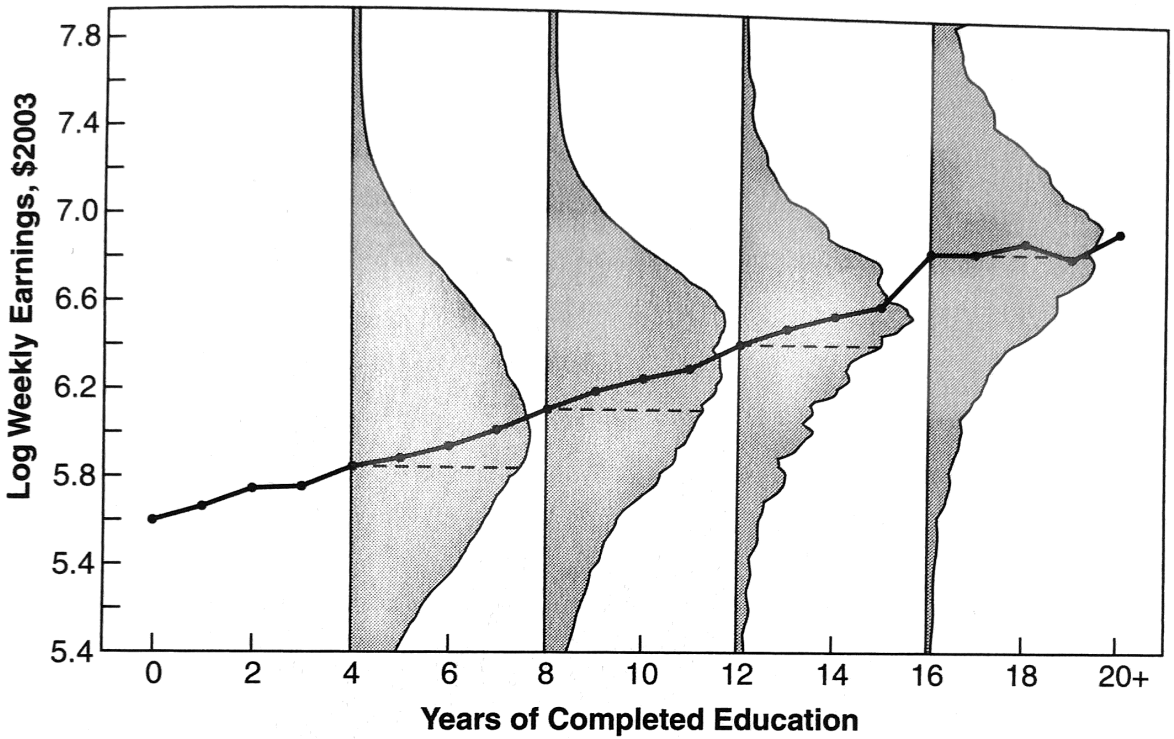


Figure 3.1.1 Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40–49 in the 1980 IPUMS 5 percent file.

Heteroskedasticity

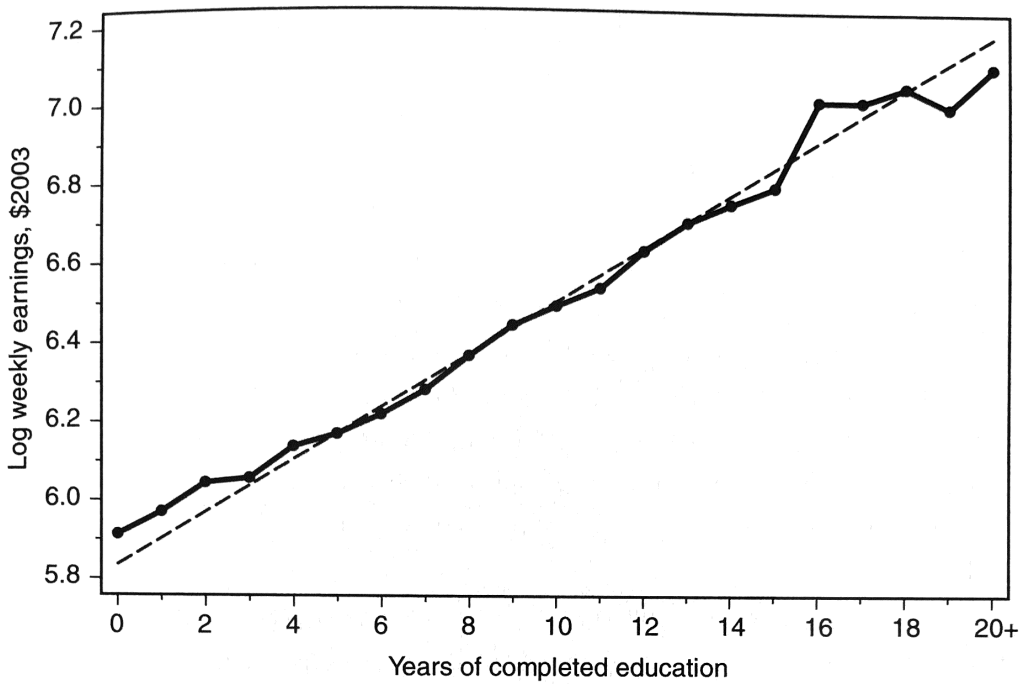


Figure 3.1.2 Regression threads the CEF of average weekly wages given schooling (dots = CEF; dashes = regression line).

Another reason for heteroskedasticity: CEF is not linear, so approximation is not perfect.

Consequences of heteroskedasticity

When the constant variance assumption is wrong, this creates problems!

- Traditional standard error formula is incorrect
- Confidence intervals are incorrect
- t-statistic and F-statistics are incorrect
- OLS is no longer BLUE

Consequences of heteroskedasticity

When the constant variance assumption is wrong, this creates problems!

- Traditional standard error formula is incorrect
- Confidence intervals are incorrect
- t-statistic and F-statistics are incorrect
- OLS is no longer BLUE
- However, OLS remains *unbiased* and *consistent*

Consequences of heteroskedasticity

When the constant variance assumption is wrong, this creates problems!

- Traditional standard error formula is incorrect
- Confidence intervals are incorrect
- t-statistic and F-statistics are incorrect
- OLS is no longer BLUE
- However, OLS remains *unbiased* and *consistent*

We can relax the *normality* assumption (MLR6) if our sample size is large, but a large sample will not fix the standard errors if MLR5 does not hold.

Addressing heteroskedasticity

Two approaches:

- Weighted least squares (WLS) when form of heteroskedasticity is known (e.g., $Var(u_i|x_i) = \sigma^2 x_i$). An example of generalized least squares (GLS).
- Calculating heteroskedasticity-robust statistics (standard errors, t-statistics, etc.). In Stata: `robust` option.

Weighted least squares

Suppose we know exactly how the error variance relates to x, e.g. $\sigma_i^2 = Var(u_i|x_i) = \sigma^2 h_i$ (where h_i is a function of x).

- GLS: transform the PRF with heteroskedastic errors into one with homoskedastic errors.
- The variance of $\frac{u_i}{\sqrt{h_i}}$ is σ^2, so divide through by $\sqrt{h_i}$.
- OLS estimator of the transformed model will be BLUE. In effect we are minimizing the *weighted* sum of squared residuals, where each squared residual is weighted by $1/h_i$.
- Less weight is given to observations with higher error variance

Weighted least squares

We rarely know exactly how the error variance relates to x, although there are exceptions:

- Grouped data, where observations represent group averages. In this case group averages have greater variance when the same size is smaller. (We know the variance of a sample mean is σ^2/n_g). $h_g = 1/n_g$ so weight is $1/h_g = n_g$.
- Linear probability model: where outcome is binary.

Recommended: Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge. (2015). "What Are We Weighting For?". *Journal of Human Resources* 50(2): 301—16.
`http://jhr.uwpress.org/content/50/2/301.abstract`

Heteroskedasticity-robust standard errors

Suppose there is heteroskedasticity of unknown form in a simple regression with SLR1-4: $Var(u_i|x_i) = \sigma_i^2$. Then:

$$Var(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_i^2}{SST_x^2}$$

Since the population σ_i^2 are unknown, we can estimate then using sample residuals $\hat{u}_i^2$. The square root of this is White's heteroskedasticity-robust standard error for $\hat{\beta}_1$.

Heteroskedasticity-robust standard errors

Suppose there is heteroskedasticity of unknown form in a multiple regression with MLR1-4: $Var(u_i|x_i) = \sigma_i^2$. Then one can estimate:

$$Var(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}$$

Where $\hat{r}_{ij}^2$ is the residual from a regrssion of x_j on all other x, and SSR_j is the SSR from this regression. The square root of the above is White's heteroskedasticity-robust standard error for $\hat{\beta}_j$.

Tests for heteroskedasticity

There are various methods for testing for the presence of heteroskedasticity:

- Breusch-Pagan test (regressing squared residuals on x's) and conducting an F-test or Lagrange multiplier (LM) test
- White test (regressing squared residuals on $\hat{y}$ and $\hat{y}^2$, since together these will be a nonlinear function of the x's)
- Visual inspection of residual plots

Heteroskedasticity-robust statistics are a common workaround, but keep in mind that a better-fitting model will increase precision. It pays to look for specification errors, correct functional form, etc.

Bootstrapping

- Bootstrapping is an *approach* to estimating sampling variance (i.e., standard errors), confidence intervals, and other properties of statistics
- It relies on repeated re-sampling *with replacement* from the observed data.
- The observed data "assume the role of an underlying population."
- It is most useful for estimators without a straightforward expression for the standard error, or settings in which the assumptions made by this expression are questionable (e.g., smaller samples)

Bootstrapping

- Repeated re-sampling of *observations* for a multi-variate analysis is sometimes called a "paired bootstrap" or "nonparametric bootstrap" since the data for each observation remain intact and no information or assumption is used about the conditional distribution of (say) y given x.
- There are *residual* bootstrap methods which involve re-sampling the residuals from a model and calculating a new y value. This is a type of "parametric" bootstrap.

Bootstrapping logic: the sample mean

Suppose we take a random sample of size n from a population with mean μ and variance σ^2. We estimate μ with $\bar{x} = \sum_i x_i/n$. The Central Limit Theorem tells us that, over repeated samples:

$$E(\bar{x}) = \mu$$
$$Var(\bar{x}) = E(\bar{x} - \mu)^2 = \sigma^2/n$$

The standard deviation of $\bar{x}$ is $\sigma/\sqrt{n}$.

Bootstrapping logic: the sample mean

Suppose we take a random sample of size n from a population with mean μ and variance σ^2. We estimate μ with $\bar{x} = \sum_i x_i/n$. The Central Limit Theorem tells us that, over repeated samples:

$$E(\bar{x}) = \mu$$
$$Var(\bar{x}) = E(\bar{x} - \mu)^2 = \sigma^2/n$$

The standard deviation of $\bar{x}$ is $\sigma/\sqrt{n}$.

When σ^2 is unknown, we estimate it with s^2. The *standard error* of $\bar{x}$ is an *estimate* of how much $\bar{x}$ varies from sample to sample.

$$se(\bar{x}) = \frac{s}{\sqrt{n}}$$

Bootstrapping logic: the sample mean

How well the estimated sampling variance s^2/n estimates the true sampling variance σ^2/n depends on how close the distribution of x_i is to normal. For non-normal populations, s^2/n may not perform well. Bootstrapping is an alternative method for estimating σ^2/n.

Bootstrapping logic: the sample mean

How well the estimated sampling variance s^2/n estimates the true sampling variance σ^2/n depends on how close the distribution of x_i is to normal. For non-normal populations, s^2/n may not perform well. Bootstrapping is an alternative method for estimating σ^2/n.

σ^2/n is the sampling variance we would observe under repeated sampling from the population. We don't observe the CDF of x, $F(x) = Pr(x_i < x)$. But we can view our observed data as providing an approximation of $F(x)$. Bootstrapping draws repeated samples from this approximated $F(x)$.

Bootstrapping logic: the sample mean

How well the estimated sampling variance s^2/n estimates the true sampling variance σ^2/n depends on how close the distribution of x_i is to normal. For non-normal populations, s^2/n may not perform well. Bootstrapping is an alternative method for estimating σ^2/n.

σ^2/n is the sampling variance we would observe under repeated sampling from the population. We don't observe the CDF of x, $F(x) = Pr(x_i < x)$. But we can view our observed data as providing an approximation of $F(x)$. Bootstrapping draws repeated samples from this approximated $F(x)$.

For $\bar{x}$: draw B repeated samples with replacement from the observed data. For each sample compute $\bar{x}^*$. The standard deviation of these $\bar{x}^*$ across the B samples is $se^*(\bar{x})$, the bootstrap standard error of $\bar{x}$. This standard error can then be used for confidence intervals and test statistics.

Bootstrapping logic: the sample mean

Clearly, the success of bootstrapping hinges on our observed data providing a good approximation of $F(x)$. "The key analogy is that the *resampling* properties of $(\bar{x}^* - \bar{x})$ must be similar to the *sampling* properties of $(\bar{x} - \mu)$" (Stine, 1989).

Bootstrapping logic: the sample mean

Clearly, the success of bootstrapping hinges on our observed data providing a good approximation of $F(x)$. "The key analogy is that the *resampling* properties of $(\bar{x}^* - \bar{x})$ must be similar to the *sampling* properties of $(\bar{x} - \mu)$" (Stine, 1989).

$\bar{x}$ is not the kind of statistic for which bootstrapping is useful, since theory already tells us much of what we need to know about the sampling distribution of $\bar{x}$. It is also worth noting that simulation is not necessary to get $se^*(\bar{x})$ in this case. Mathematically, the variance of the bootstrap means will be s^2/n.

Still, it is useful for illustration purposes.

`bootstrap` command in Stata

`bootstrap` *`exp_list`* `[,` *`reps(#) otheroptions`*`]` `:` *`command`*

- *exp_list* are the quantities/estimators you want to bootstrap
- *command* is the estimation or other command (or user-written program) you want bootstrapped
- The default number of replications (*reps*) is 50
- Can bootstrap an expression, e.g. `range=(r(max)-r(min))`
- Can save bootstrap results (one observation per replication) with the `saving()` option

Be sure to `set seed #` for replicability. Stata recommends you drop unnecessary variables and drop observations with missing values prior to bootstrap, given how it draws bootstrap samples. (It may not be attentive to which observations have 100% non-missing values for the *command*)

Bootstrapping the sample mean

See handout Examples 1-4: obtaining $se^*(\bar{x})$ and bootstrap confidence intervals for $\bar{x}$. Bootstrap *confidence intervals* can be constructed in a number of ways:

- Normal-based confidence interval: uses observed $\bar{x}$, the bootstrapped standard error, and assumes the distribution of $\bar{x}$ is normal. $\bar{x} \pm z * se^*(\bar{x})$
- Bootstrap percentile confidence interval: uses the empirical distribution of the $\bar{x}^*$

For the latter (and other CI calculations) can use `estat bootstrap`. Note that bootstrap standard errors can often be requested as an option in an estimation command, e.g. `vce(bootstrap)`.

Bootstrapping the sample mean

Some takeaways from handout Examples 1-4:

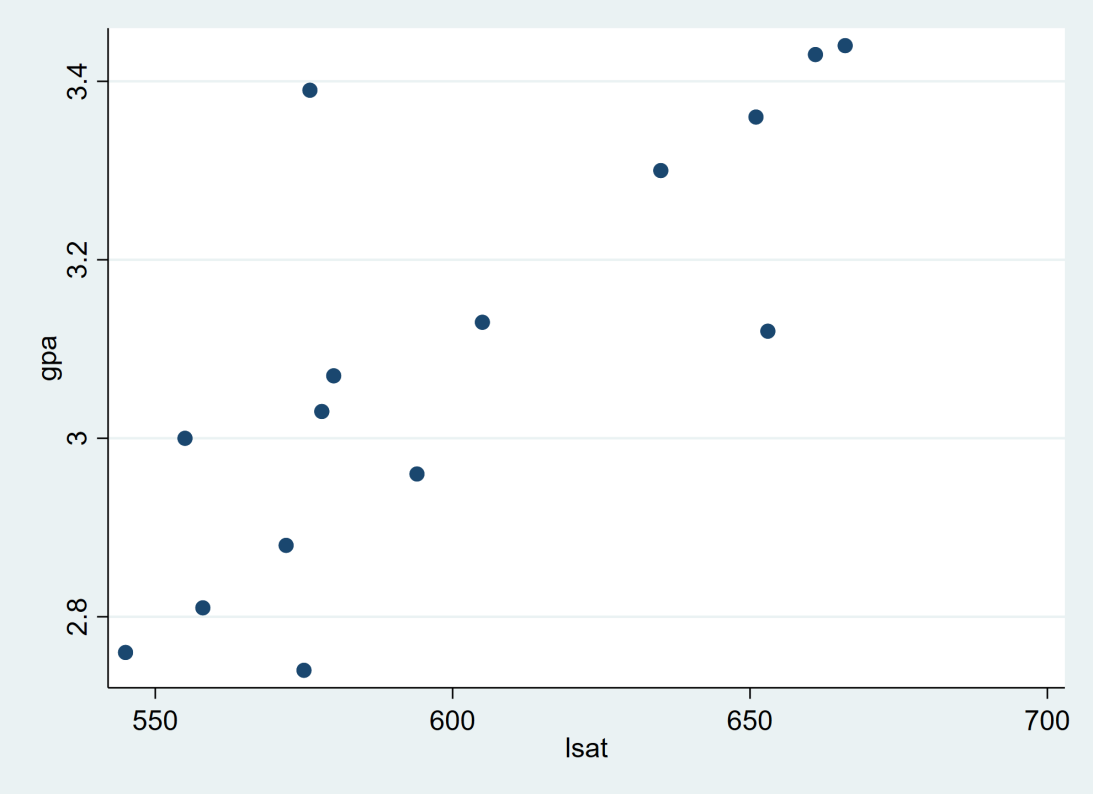
- The bootstrap standard error performs remarkably well, even in this example of $n = 20$! Part of its good performance in this case is the normal population from which the sample was drawn.
- More bootstrap replications has little to no effect on the bootstrap standard error, although increasing the number of replications reduces variability in this estimate.
- For the standard error, $B = 99$ performs about as well as $B = 499$.
- For the confidence interval—especially the percentile CI—there is more to be gained from increasing the number of replications. In general, confidence interval estimation requires more replications.

Bootstrapping the sample mean

Stine (1989): "even with only 19 bootstrap samples, the bootstrap intervals nearly obtain the performance of the best interval in this case, that given by *knowing* the data are from a normal distribution" (p. 250).

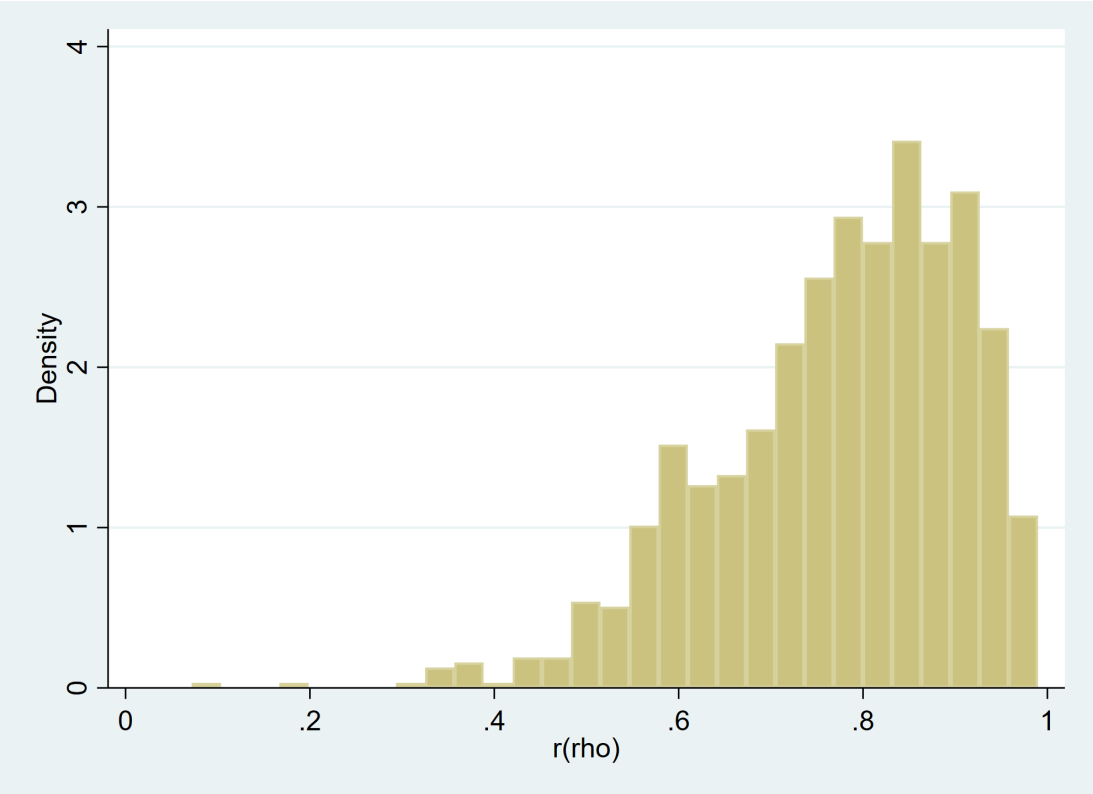
Bootstrapping the correlation coefficient

See Example 5: obtaining $se^*(r_{xy})$ and bootstrap confidence intervals for r_{xy}. Uses 1982 data from 15 law schools (Efron).



Bootstrapping the correlation coefficient

Takeaways from Example 5: the distribution of bootstrap r^*_{xy} is very skewed. Suggests the percentile CI is more appropriate than one assuming a symmetric (e.g., normal) distribution.



The jackknife

The jackknife is the most well-known repeated sample predecessor to the bootstrap. It relies on dividing the sample into S disjoint subsets with the same n. The statistic of interest is then calculated S times, each time *omitting one of the subsets*. The variance across these is the jackknife variance estimate.

- A common jackknife estimator is the "leave-one-out" method where there is one observation in each set S
- Unlike the bootstrap, the jackknife replicates are not independent of one another. This requires an adjustment to the variance calculation (vs. the bootstrap).

See Stata command `jackknife` and the variance estimation option `vce(jackknife)`.