LECTURE: CLASSICAL LINEAR REGRESSION - ESTIMATION

Quote for the lecture

There are two things you are better off not watching in the making: sausages and econometric estimates --- Edward Leamer

Starting out simple: The simple linear regression (SLR) model

Basics

Note: The term "regression" has historical roots in the "regression-to-the-mean" phenomenon in genetics and was popularized by Sir Francis Galton during the late 19th century with the publication of Regression towards mediocrity in hereditary stature. Galton observed that extreme characteristics (e.g., height) in parents are not passed on completely to their offspring. Rather, the characteristics in the offspring regress towards a mediocre point (a point which has since been identified as the mean).

The simplest form of regression can be written as $y = \beta_0 + \beta_1 x + u$, where y is the dependent variable, and x the explanatory variable. This is called a **simple** regression model because it features only one explanatory variables. An example of this would be $wage = \beta_0 + \beta_1 educ + u$. Note that u contains "unobserved factors", e.g., ability of the individual in the wage-education example above. It called the **error term** or **disturbance**. It plays a very important role in econometrics, and can also capture measurement problems in one or more of the variables. Holding everything inside u constant (say, ability), we see that $\Delta y = \beta_1 \Delta x$, which means β_1 is the slope parameter representing the change in the dependent variable given a certain change in the independent variable. Note that we are assuming that β_1 is the effect of one more year of education on wage. A good qeustion to ask is: Is each year of education really worth the same dollar amount no matter how much education one starts with -- i.e., going from kindergarten to first grade versus going from ninth grade to tenth grade?

y and x are not treated symmetrically. We want to explain y in terms of x. From a causality standpoint, it makes no sense to *explain* past educational attainment in terms of future labor earnings. As another example, we want to explain student performance (y) in terms of class size (x), not the other way around. The y and x variables are referred to by the following terminology:

y	x
Dependent Variable	Independent Variable
Explained Variable	Explanatory Variable
Response Variable	Control Variable
Predicted Variable	Predictor Variable
Regressand	Regressor

The terms *explained* and *explanatory* are probably best, as they are the most descriptive and widely applicable. But *dependent* and *independent* are used often. (*Independent* here should not be confused with the notion of statistical independence.)

Assumptions

To get reliable estimators of β_0 and β_1, we need some assumptions about u. First, without loss of generality, we can assume $E(u) = 0$, as long as the regression contains an intercept term. To understand this, suppose on a certain scale, ability has a mean of 100. Then we can consider the distibution of ability shifted to the left by 100 units, and therefore centered around zero, while the 100 units can be absorbed by the β_0. Therefore this is quite an innocuous assumption.

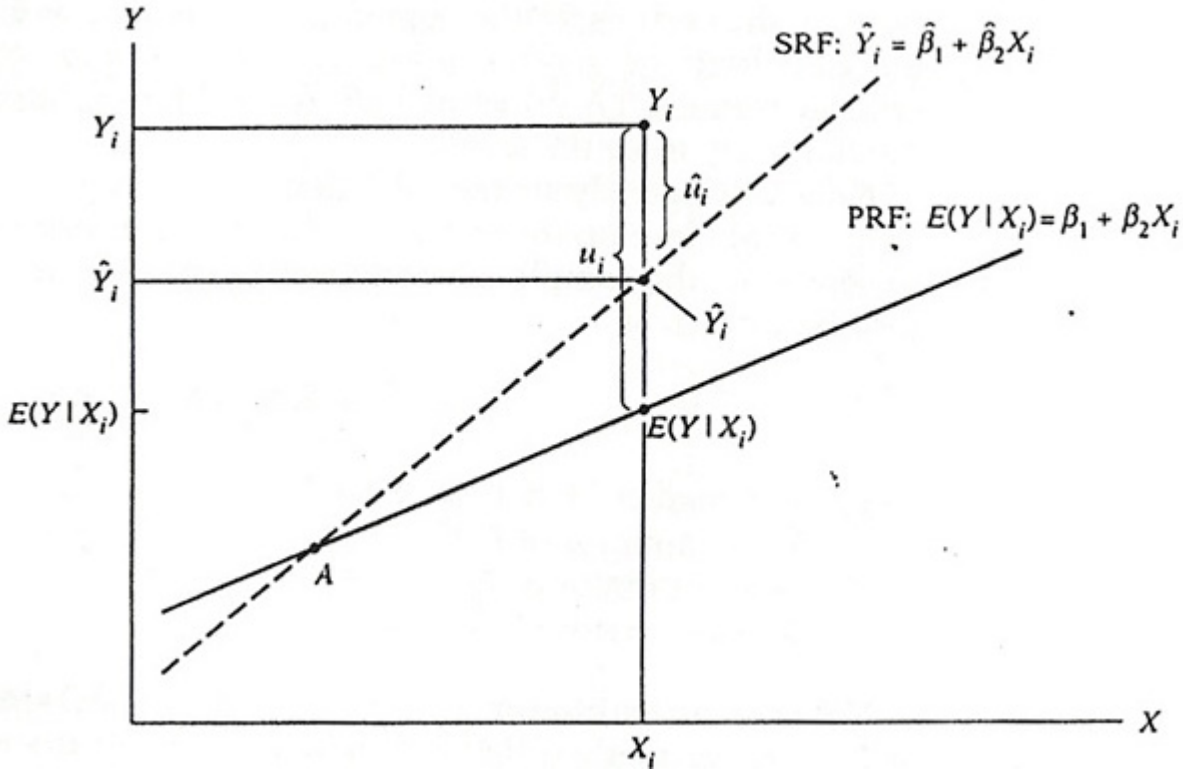
The crucial assumption is regarding the relationship between x and u. A natural assumption is $Cov(u, x) = 0$, meaning the unobserved factors are not correlated with the explanatory variable, e.g., (unobserved) ability is not correlated with education. It turns out this is not enough as $Cov(u, x) = 0$ does not imply that $Cov(u, g(x)) = 0$ for a function of x, $g(x)$. We therefore assume that $E(u \mid x) = E(u) = 0$. This is called the conditional mean independence of u. This assumption states that the average value of u does not depend on the value of x, e.g., average ability does not vary by education level (As people choose education levels partly based on ability, this assumption is almost certainly false.) The **population regression function (PRF)** is then: $E(y \mid x) = \beta_0 + \beta_1 x$.

Estimators

The estimators of the parameters β_0 and β_1 are given by the formulas:

- $\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

The above formulas can be derived using the FOCs from the previous lecture (method of moments, least squares, or Maximum Likelihood with normally distributed errors). Therefore, the **sample regression function (SRF)** is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, a.k.a. the regression line. The figure below shows the difference between a population regression function (PRF) and a sample regression function (SRF).



Remember that the PRF is fixed, yet unknown, but the SRF is dependent on the sample at hand. A new sample may produce a new SRF, but the underlying PRF remains unchanged.

Now for a bit of Ordinary Least Squares (OLS) miscellany. The algebraic properties of the estimator are:

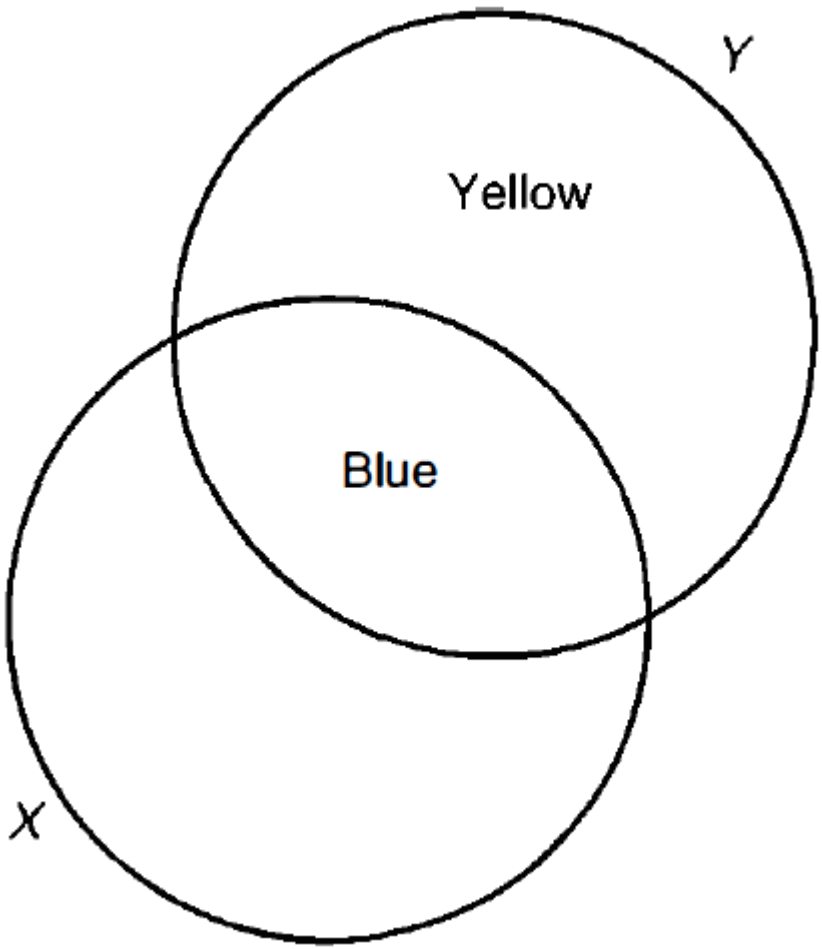
- $\sum_{i=1}^{N} \hat{u}_i = 0$, i.e., the sample average of the residuals is zero -- this is basically the first FOC for deriving the OLS estimators
- $\sum_{i=1}^{N} x_i \hat{u}_i = 0$, i.e., the sample covariance between the explanatory variable and the residuals is zero -- this is the second FOC for deriving the OLS estimators
- The point $(\bar{x}, \bar{y})$ always lies on the regression line -- this follows from the formula for $\hat{\beta}_0$

Goodness of fit

Decomposing the variance of the dependent variable y, we have:

- Total sum of squares: $SST = \sum_{i=1}^{N}(y_i - \bar{y})^2$
- Explained sum of squares: $SSE = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$
- Sum of squared residuals: $SST = \sum_{i=1}^{N}\hat{u}_i^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
- Can be proved that SST=SSE+SSR

A popular **goodness-of-fit** measure is R^2 defined as SSE/SST, i.e., the proportion of variation in dependent variable explained by that in the explanatory variable. A higher R^2 in general means a better explanation of the dependent variable by the explanatory variable, hence a better "fit" of the estimated econometric model. Although this is true in general, one must bear in mind the caveat that it is not uncommon to see R^2 values of less than 20% in cross-sectional regressions. It is a useful summary measure but tells us nothing about causality. Having a “high” R-squared is neither necessary nor sufficient to infer causality. In other words, R^2 is not everything, it is only a guidepost... A useful visual representation of R^2 as follows:



The circle labelled Y represents variation in Y, while that labelled X represents the variation in X. The are marked blue is the information employed in estimating β_X: the larger this area, the more information used to form the estimate, and hence the smaller the variance of the estimate.

Specification

Note that when we specify the model: $y = \beta_0 + \beta_1 x + u$, we deduce that $\Delta y = \beta_1 \Delta x$. What if the model is written as $\ln(y) = \beta_0 + \beta_1 x + u$?

From basic calculus, we know that $e^x \approx 1 + x$, for small x, which $\Rightarrow x \approx ln(1 + x) \Rightarrow ln(y) \approx y - 1$. Therefore, $ln(y_1/y_0) = \Delta ln(y) \approx \frac{y_1 - y_0}{y_0} = \Delta y/y$. From here it is easy to see that $100\Delta ln(y) \approx \%\Delta y. \Rightarrow \%\Delta y = 100\beta_1 \Delta x$. In the wage-education example above, since the percentage change in wage given one additional year of education is constant, the change in wage for an extra year of education increases as education increases: an increasing return to education, which is more realistic.

What if the specification is $\ln(y) = \beta_0 + \beta_1 \ln(x) + u$? Can be shown as above that $\%\Delta y = \beta_1 \%\Delta x$. Try this proof for homework.

Model	Dep. Var.	Indep. Var.	β_1
Level-Level	y	x	$\Delta y = \beta_1 \Delta x$
Level-Log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-Level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-Log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

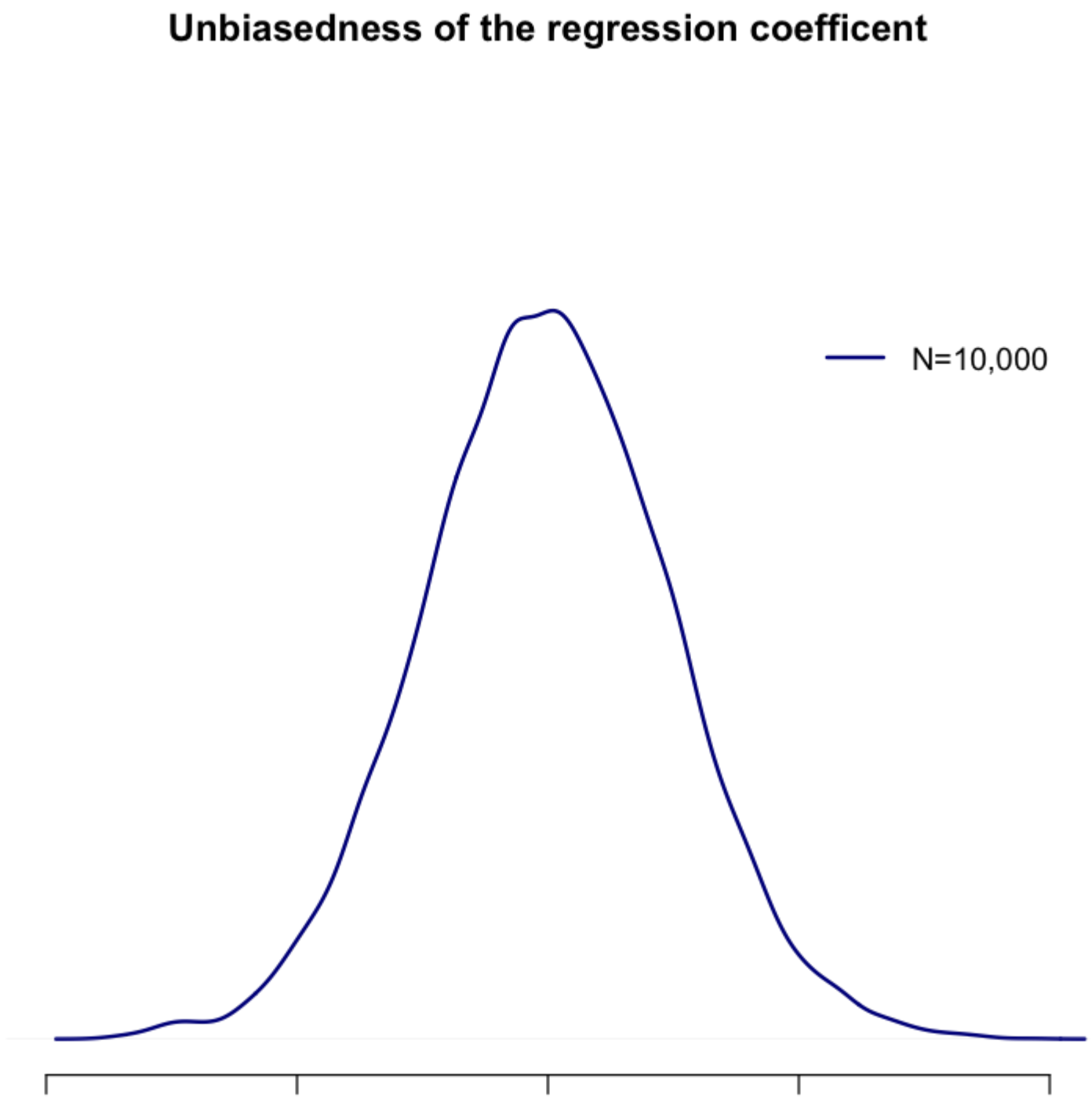
Finally, note that the possibility of using the natural log to get nonlinear relationships between y and x raises a question: What do we mean now by **linear** regression? The answer is that the model is linear in the *parameters* β_0 and β_1. We can use any transformations (e.g., x^2, $\sqrt{x}$, $ln(x)$) of the dependent and independent variables to get interesting interpretations for the parameters.

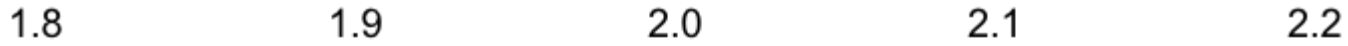
The formal SLR assumptions

- There is a linear relation between y and x: $y = \beta_0 + \beta_1 x + u$ ... SLR.1

In [1]:
```
#Visual proof of unbiasedness
#Define a function to regress y against x when beta_0=1 and beta_1=2 and return beta_1 coefficient
library(MASS)
basicreg <- function(v1,v2,m) {
#Posit joint distribution of x and u
Sigma <- matrix(c(v1,0,0,v2),m,m)
Xu<-mvrnorm(n=50,c(0,0),Sigma)
X <- Xu[,1]
u<-Xu[,2]
y <- 1+(2*X)+u
regdata <- lm(y ~ X)
coef<-summary(regdata)$coefficients[2,1]
return(coef)
}

#Repeat regression function N times
repreg<-function(N,v1,v2,m) {
nestims <- rep(1,N)
for(j in 1:N) {
    nestims[j]<-basicreg(v1,v2,m)
}
    return(nestims)
}

betahats<-repreg(10000,9,1,2)

#Plotting frequency distribution regression coefficients
plot(density(betahats), xlab="", yaxt='n', ann=FALSE, col="darkblue", bty="n", xlim=c(1.8, 2.2), ylim=c(0, 10
), lwd=2)
legend(2.1,8,"N=10,000",col="darkblue",lwd=2,bty="n")
title("Unbiasedness of the regression coefficent")
```

```
Warning message:
“package ‘MASS’ was built under R version 3.5.2”
```

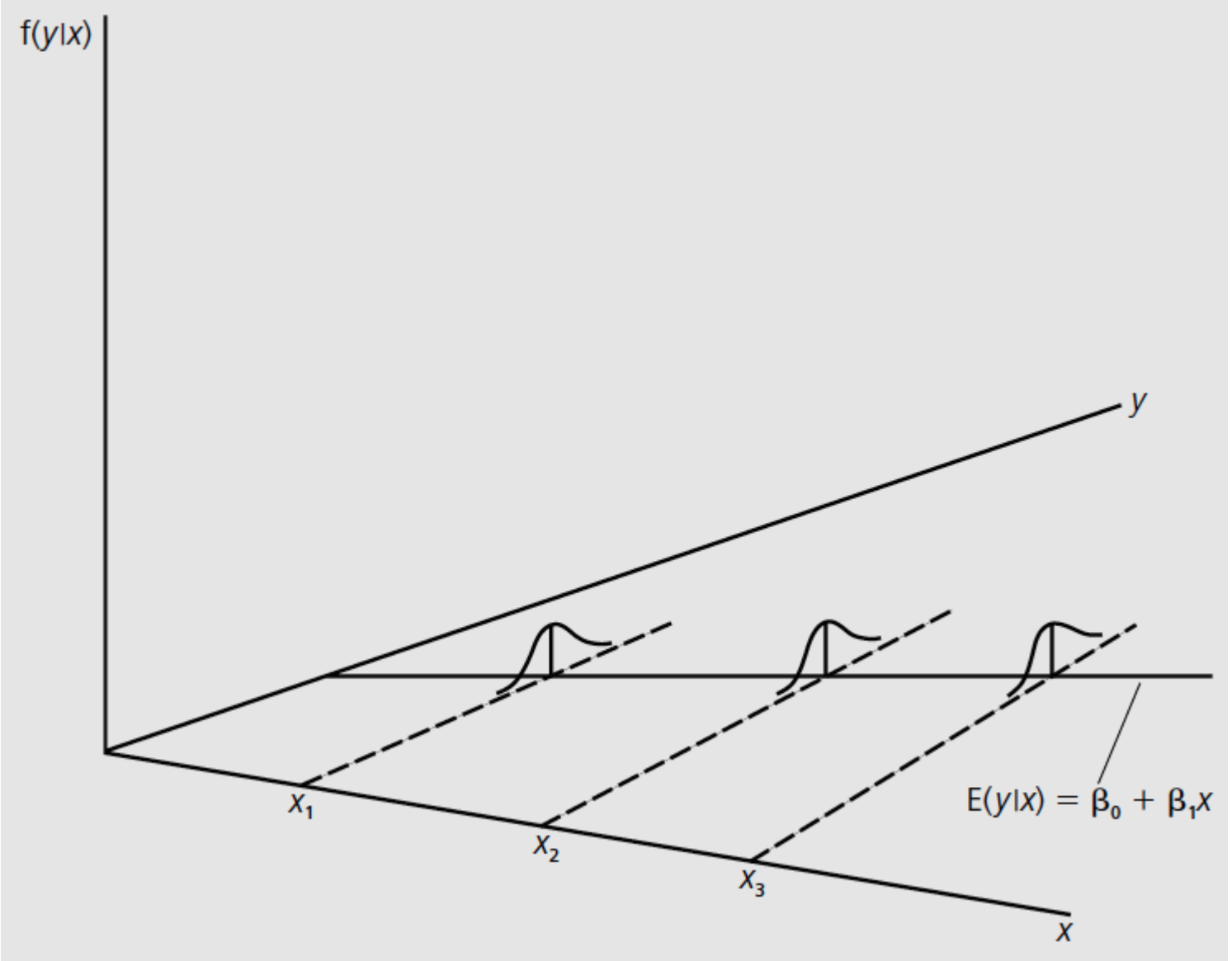

Unbiasedness of the regression coefficent
N=10,000



- Conclusion: The distribution of the beta is centered around the true value and hence confirms that OLS estimates are **unbiased**.

Another assumption

- The error terms have the same (conditional) variance: Homoskedasticity: $Var(u \mid x) = \sigma^2$ .. SLR.5



- This assumption is **NOT** required for unbiasedness
  - It will be critical for deriving the sample variance of the OLS estimator

The multiple regression model

Multiple regression is a straightforward extension to the simple regression model. We simply extend the one explanatory variable to k explanatory variables, and the model becomes $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + u$. As a counterpart to the formal assumptions of the simple regression model, we have the multiple regression model assumptions as follows:

- There is a linear relation between y and x: model is specified above ... MLR.1
- (Randomly sampled) Observations are of the form $y_i = \beta_0 + \beta_1 x_{i,i} + \cdots + \beta_k x_{i,k} + u_i$ ... MLR.2
- No perfect collinearity in the sample ... MLR.3. This means that no x_k is constant and there are no linear relationships between any x variables
  - Note that if in the population x does not change then we are not asking an interesting question. If the x_i are all the same value in the sample, we are unlucky and cannot proceed. This assumption ensures that the OLS estimators are unique and can be obtained from the first order conditions (minizing the sum of squared residuals).
- Conditional zero mean: $E(u \mid x_1, x_2, \cdots x_k) = E(u) = 0$ ... MLR.4. Note that this assumption states that the average value of the error does not change across different slices of the population defined by $x_1, \cdots, x_k$. Setting $E_u = 0$ essentially defines β_0.
  - When Assumption MLR.4 holds, we say $x_1, \cdots, x_k$ are **exogenous** explanatory variables. If x_j is correlated with u, we often say x_j is an **endogenous** explanatory variable. (This name makes more sense in more advanced contexts, but it is used generally.)
- MLR.1 through MLR.4 ensure that the OLS estimators of $\beta_0 \cdots \beta_k$ are **unbiased**
  - $E(\hat{\beta}_j) = \beta_j$, for $j = 0, 1, 2, \cdots k$

Multiple regression: OLS estimates

- Note that this methodology has the following notation:
- $x_{i,k}$ is an observation (or instance) of the k^{th} independent variable
- $\mathbf{x}_i$ is a $1 \times (k+1)$ vector of all independent variables
- $\mathbf{X}$ is all the x_is stacked on top of each other
- $\mathbf{y}$ is all the y_i observations stacked on top of each other
- It turns out that the OLS estimator is algebraically given by $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$
  - As a practical matter, we will rely on `R` rather than actually use this formula to calculate estimates

Ceteris Paribus

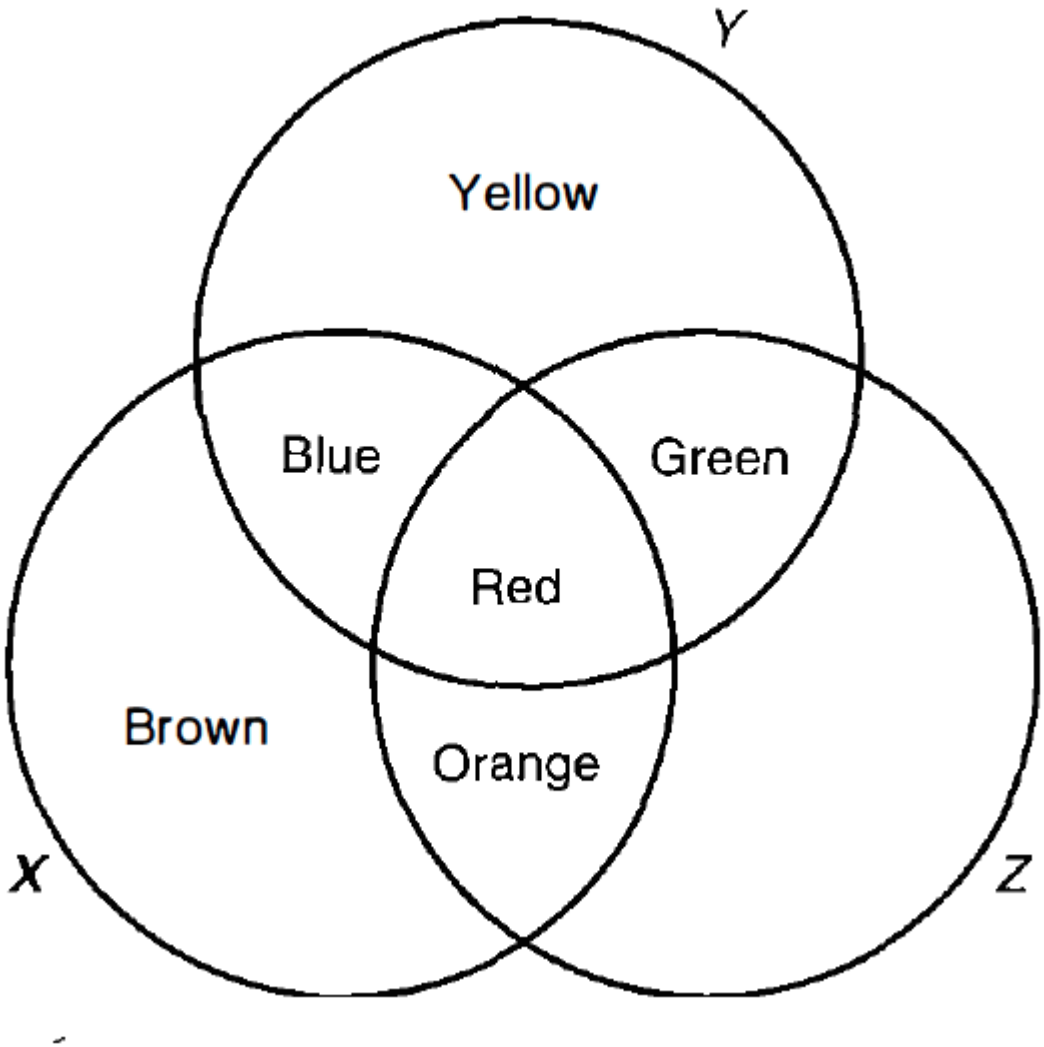
- Q: How many econometricians does it take to change a lightbulb?
- A: Eight. One to screw it in and seven to hold **everything else constant**

The ideal of econometrics is to run experiments where, for example, one might separate twins and give them separate treatments. Such experiments are impossible in social sciences, while also being morally repugnant. The beauty of multiple regression is that it gives us the *ceteris paribus* interpretation without having to find two people with the same value of IQ who differ in education by one year. The estimation method does that for us.

Multiple regression isolates the effects of a particular variable on the dependent variable, while pretending to hold everything else constant. This allows us to write things like: $\Delta y = \beta_k \Delta x_k$, holding $x_1 \cdots x_{k-1}$ constant.

Comparing simple and multiple regression: A "partialling out" interpretation

- Consider a two variable multiple regression: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$. It turns out one can write $\hat{\beta}_1 = \frac{\sum_{i=1}^{N} \hat{r}_{i,1} y_i}{\sum_{i=1}^{N} \hat{r}_{i,1}^2}$
  - $\hat{r}_{i,1}$ are the OLS residuals from a simple regression of x_1 on x_2. We regress x_1 on x_2 and obtain the residuals (y plays no role here)
- Above equation shows that we can then do a simple regression of y on r_1 to obtain $\hat{\beta}_1$
  - Essentially says that the estimate $\hat{\beta}_1$ measures the effect of x_1 on y after the effects of x_2 have been netted out (or *partialled out*

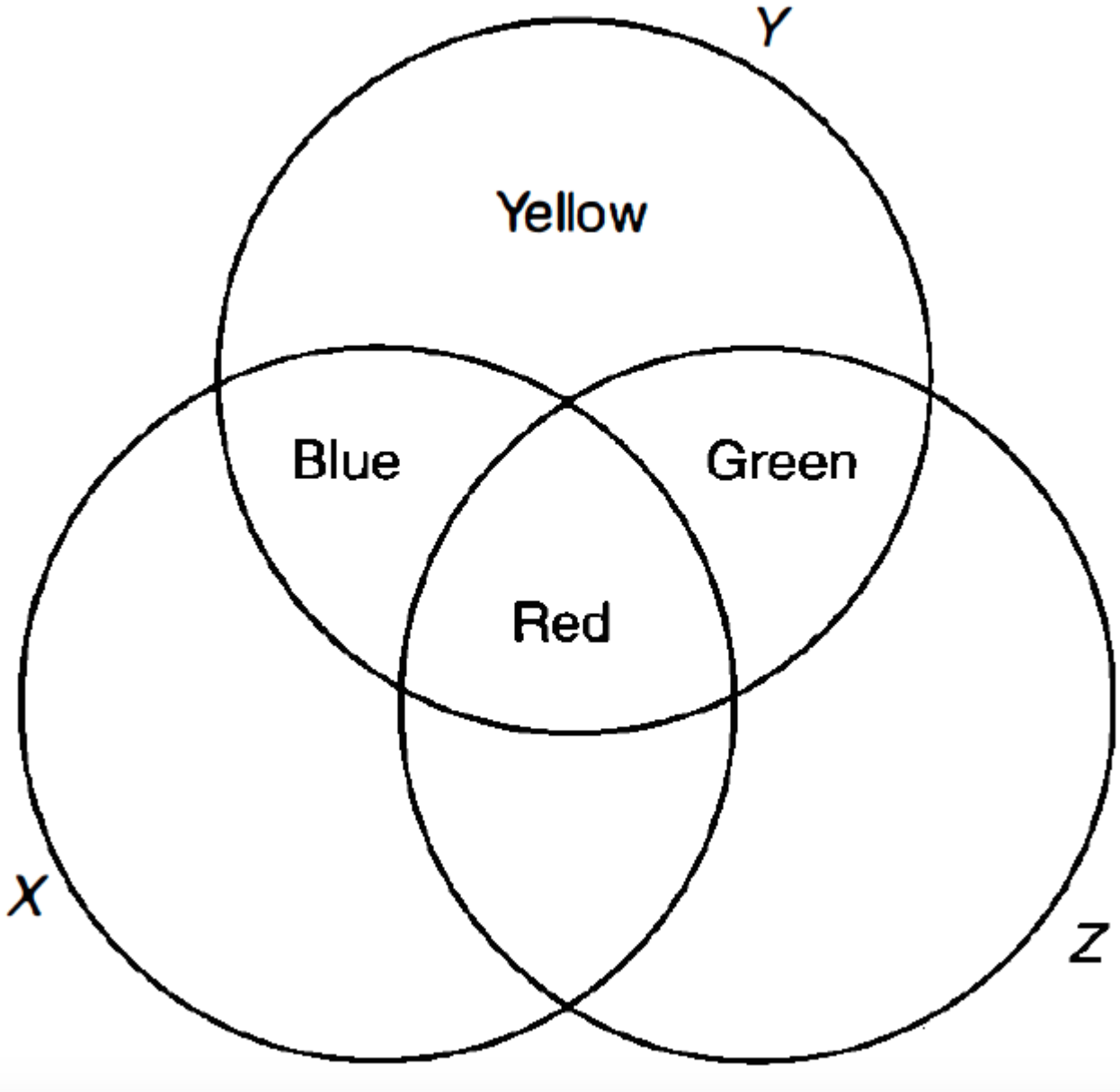


- Blue area= variation in Y uniquely determined by variation in X
- Information in the red area should not be used as it reflects variation in Y determined by variation in both X and Z
  - Partialling out removes the red area

Omitted variables

Say the true model is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$. If we could run this regression, the parameter estimates would be $\hat{\beta}_1$ and $\hat{\beta}_2$. For various reasons, say we omit x_2 and run the regression $y = \beta_0 + \beta_1 x_1 + v$. Call this estimate $\tilde{\beta}_1$

- Question: How does $\tilde{\beta}_1$ compare to $\hat{\beta}_1$ in general?
- Answer: $\tilde{\beta}_1$ is in general **biased**, i.e., $E(\tilde{\beta}_1) \neq \beta_1$



- Since Y is regressed on only X, the blue-plus-red area is used to estimate β_X. But red area reflects variation in Y due to both X and Z, so the resulting estimate of β_X will be biased.
- If Z had been included, only blue area would have been used in estimating β_X. Omitting Z thus increases the information used to estimate β_X by the red area
- Hence, the resulting estimate, although biased, will have a *smaller* variance

Omitted variable bias

In the two explanatory variables case, it is easy to derive the amount of omitted variable bias. It turns out that $E(\tilde{\beta}_1) = \beta_1 + \beta_2 \frac{\sum_{i=1}^{N}(x_{i,1}-\bar{x}_1)x_{i,2}}{\sum_{i=1}^{N}(x_{i,1}-\bar{x}_1)^2}$

Therefore $\tilde{\beta}_1$ is biased unless:

- $\beta_2 = 0$: x_2 is not part of the true model, or
- $Cov(x_1, x_2) = 0$ in the sample: Imagine the previous figure with no red area

We do know the sign of β_2 and might only have a vague idea about the size of $Cov(x_1, x_2)$. But we often can guess at the signs. In this simple case, we can see that the direction of the bias is as follows:

Bias	$Cov(x_1, x_2) > 0$	$Cov(x_1, x_2) < 0$
$\beta_2 > 0$	positive	negative
$\beta_2 < 0$	negative	positive

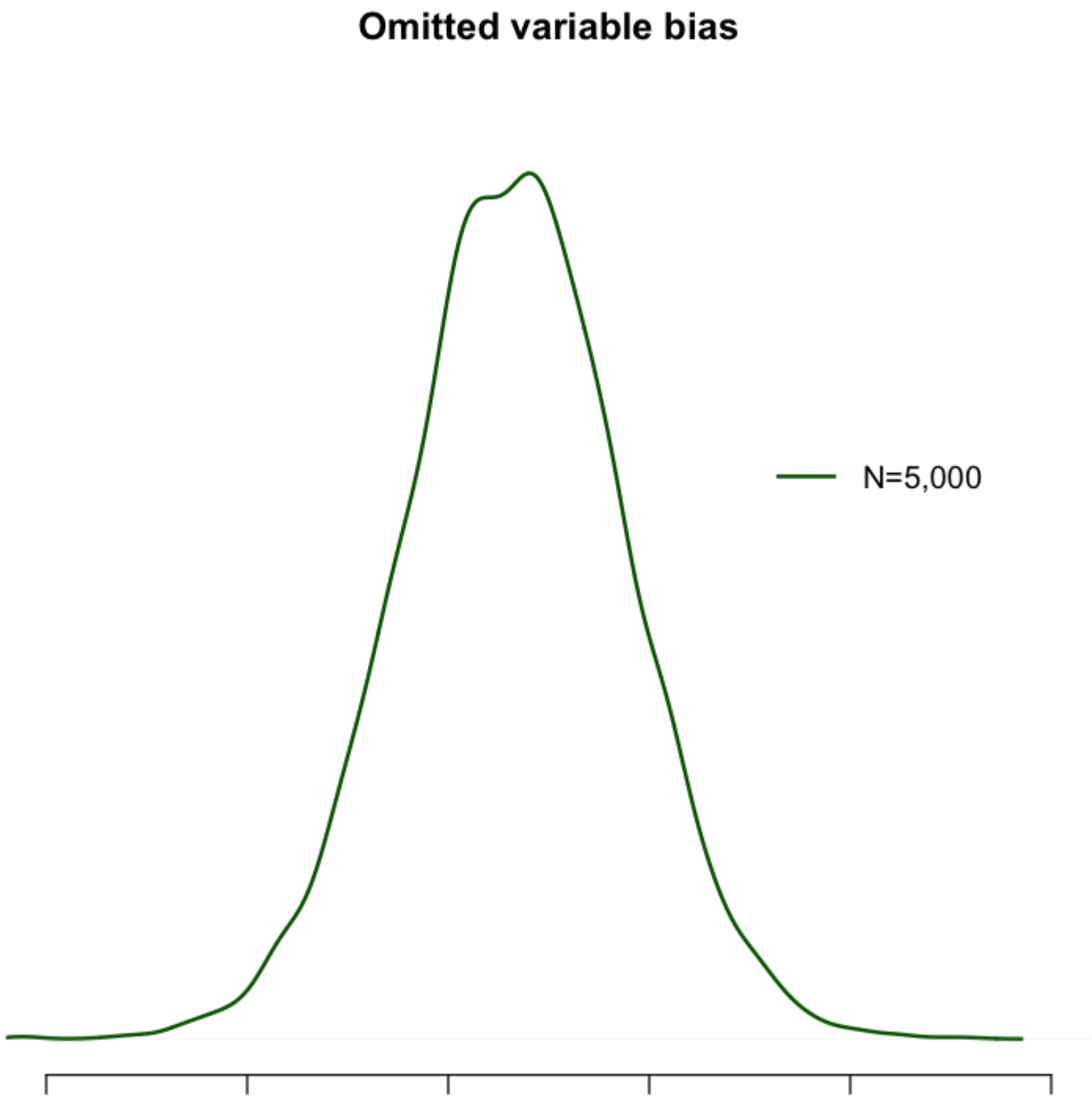
Consider the regression $log(wage) = \beta_0 + \beta_1 educ + \beta_2 ability + u$. Since we cannot measure ability, we have to necessarily run an omitted variable regression: $log(wage) = \beta_0 + \beta_1 educ + v$. We know that $\beta_2 > 0$, since the greater the ability, the higher should be the wage. We also know that all else equal, higher ability people get more education, on average, which means $Corr(educ, ability) > 0$. Therefore the bias is positive, i.e. $E(\tilde{\beta}_1) > \beta_1$. In summary, our failure to control for ability leads to (on average) overestimating the return to education. We attribute some of the effect of ability to education because ability and education are positively correlated. This is the essence of omitted variable bias.

In the general case with k explanatory variables...

- say x_3 is omitted and is correlated with x_2 but NOT with x_1...
- still, OLS estimator $\hat{\beta}_1$ will be biased

Let us run a simulation with $\beta_0 = 1$, $\beta_1 = 2$, and $\beta_2 = 3$. We draw sample sizes of 50, and repeat regression of y *only on an intercept and* x_1 5000 times. Look at the sampling distribution of $\hat{\beta}_1$

In [2]:
```
#Visual proof of omitted variable bias
library(MASS)
basicreg <- function(v1,v2,v3,cov) {
#Posit joint distribution of x and u
Sigma <- matrix(c(v1,cov,0,cov,v2,0,0,0,v3),3,3)
Xu<-mvrnorm(n=50,c(0,0,0),Sigma)
x1 <- Xu[,1]
x2 <- Xu[,2]
u <- Xu[,3]
y <- 1+(2*x1)+(3*x2)+u
regdata <- lm(y ~ x1)
coef<-summary(regdata)$coefficients[2,1]
return(coef)
}

#Repeat regression function N times
repreg<-function(N,v1,v2,v3,cov) {
nestims <- rep(1,N)
for(j in 1:N) {
    nestims[j]<-basicreg(v1,v2,v3,cov)
}
    return(nestims)
}

betahats<-repreg(5000,9,4,1,2)

#Plotting frequency distribution of regression coefficients
plot(density(betahats), xlab="", yaxt='n', ann=FALSE, col="darkgreen", bty="n", xlim=c(1.5,4), ylim=c(0, 1.5
), lwd=2)
legend(3.25,1,"N=5,000",col="darkgreen",lwd=2,bty="n")
title("Omitted variable bias")
```

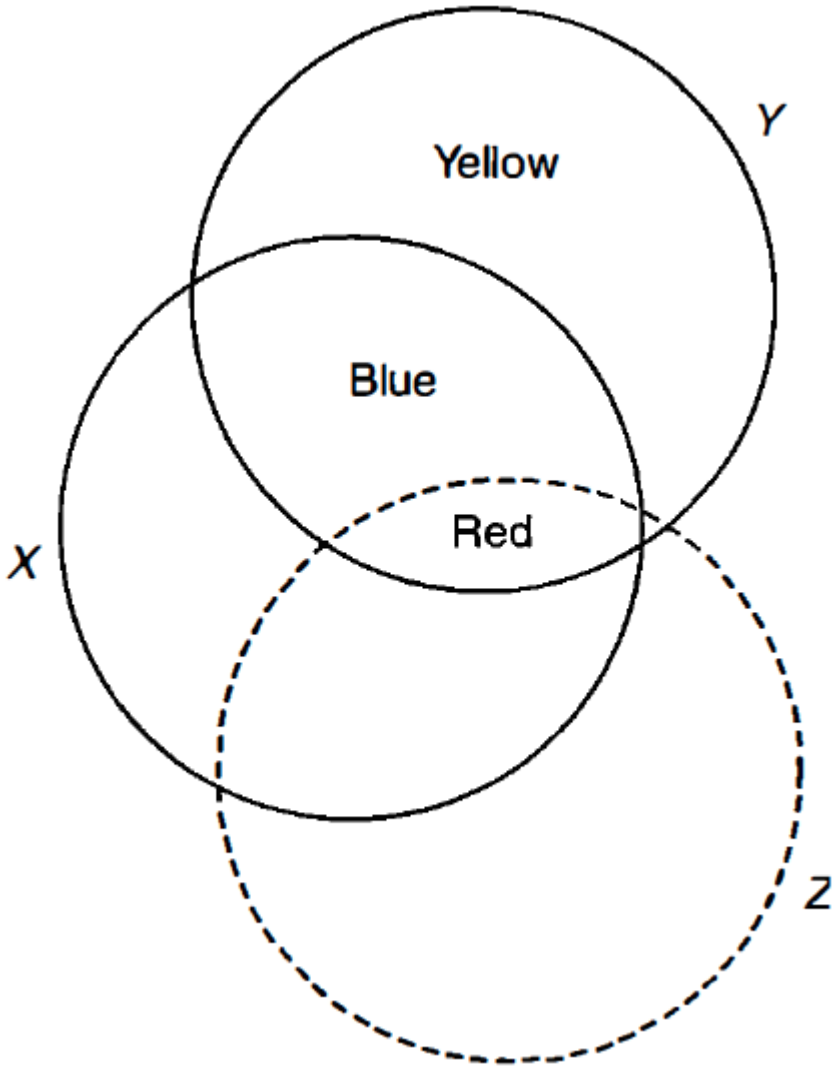

Omitted variable bias
N=5,000

- Conclusion: OLS estimator is **biased** when variables are omitted

Irrelevant variables

Say the true model is: $y = \beta_0 + \beta_1 x_1 + u$. If we ran this regression, the parameter estimate would be $\hat{\beta}_1$. For various reasons, we add an irrelevant variable x_2 and run the regression $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v$. Call these estimates $\tilde{\beta}_1$ and $\tilde{\beta}_2$

- Question: How does $\tilde{\beta}_1$ compare to $\hat{\beta}_1$ in general?
- Answer: $\tilde{\beta}_1$ is in general *unbiased*...
- ...which raises the question: why not include all available explanatory variables?



- Using the correct specification, if Y is regressed on X, the blue-plus-red area is employed to create an unbiased estimate of β_X
- Including Z in the regression implies that only the blue area is used to estimate β_Xá
- Since the blue area reflects variation in Y due entirely to X, the estimate of β_X (while including Z) is *unbiased*
- However, since the blue area is smaller than the blue-plus-red area, the variance of the estimate of β_X (while including Z) is larger
- This is why one must avoid "**kitchen sink**" regressions!