Section 1: Regression review

Yotam Shem-Tov

Fall 2015

Contact information

- Yotam Shem-Tov, PhD student in economics.
- E-mail: shemtov@berkeley.edu.
- Office: Evans hall room 650
- Office hours: To be announced - any preferences or constraints? Feel free to e-mail me.

R resources - Prior knowledge in **R** is assumed

- In the link here is an excellent introduction to statistics in **R** .
- There is a free online course in coursera on programming in R. Here is a link to the course.
- An excellent book for implementing econometric models and method in **R** is Applied Econometrics with R. This is my favourite for starting to learn **R** for someone who has some background in statistic methods in the social science.
- The book Regression Models for Data Science in R has a nice online version. It covers the implementation of regression models using **R** .
- A great link to **R** resources and other cool programming and econometric tools is here.

Outline

There are two general approaches to regression

1. Regression as a model: a data generating process (DGP)
2. Regression as an algorithm (i.e., an estimator without assuming an underlying structural model).

Overview of this slides:

1. The conditional expectation function - a motivation for using OLS.
2. OLS as the *minimum mean squared error* linear approximation (projections).
3. Regression as a structural model (i.e., assuming the conditional expectation is linear).
4. Applications, examples and additional topics.

This slides relay heavily on Mostly Harmless Econometrics, Hansen's book Econometrics and Pat Kline's lecture notes.

Notation and theoretical framework

- The researcher observes an $i.i.d$ sample of N observations of (Y_i, X_i), denoted by $S_N \equiv (Y_1, X_1), \ldots (Y_N, X_N)$.
- $X_i' = (X_{1i}, X_{2i}, \ldots, X_{pi})$ is a $1 \times p$ random vector, where p is the number of covariates. Y_i is a scalar random variable.
- (Y_i, X_i) are sampled $i.i.d$ from the distribution function $F_{Y,X}(y, x)$.

The Conditional expectations function (CEF)

- The conditional expectation function is,

$$\mathbb{E}[Y_i|X_i = x] = \int t dF_{Y|X}(t|X_i = x) = \int t f_{Y|X}(t|X_i = x)dt$$

- For any two variables X_i and Y_i we can always compute the conditional expectation of one given the other.
- Additional assumptions are required to give anything other than a purely descriptive interpretation to the conditional expectation function.
- Descriptive tools are important and can be valuable.
- The *law of iterated expectations (LIE)* is,

$$\mathbb{E}[Y_i] = \mathbb{E}\left[\mathbb{E}\left[Y_i|X_i = x\right]\right]$$

- More in detail LIE is: $\mathbb{E}_Y[Y_i] = \mathbb{E}_X\left[\mathbb{E}_{Y|X=x}\left[Y_i|X_i = x\right]\right]$
- Another way of writing LIE is,

$$\mathbb{E}[Y_i|X_{1i} = x_1] = \mathbb{E}\left[\mathbb{E}\left[Y_i|X_{1i} = x_1, X_{2i} = x_2\right]|X_{1i} = x_1\right]$$

Properties of the CEF

- The LIE has several useful implications,

$$\begin{aligned}\mathbb{E}\left[Y_i - \mathbb{E}\left[Y_i|X_i\right]\right] &= 0\\ \mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i|X_i\right]\right)\cdot X_i\right] &= 0\\ \mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i|X_i\right]\right)\cdot g(X_i)\right] &= 0 \quad \forall\, g(\cdot)\end{aligned}$$

- Therefore we can decompose any scalar random variable Y_i to two parts,

$$Y_i = \mathbb{E}\left[Y_i|X_i\right] + u_i, \quad \mathbb{E}\left[u_i|X_i\right] = 0.$$

- Question: Does $\mathbb{E}\left[u_i|X_i\right]=0$ implies $\mathbb{E}\left[u_i\right]=0$? Yes, the proof follows from the LIE. The statement $\mathbb{E}\left[u_i|X_i\right]=0$ is more general than the statement $\mathbb{E}\left[u_i\right]=0$.
- Question: Does $\mathbb{E}\left[u_i\right]=0$ imply $\mathbb{E}\left[u_i|X_i\right]=0$? No.

Properties of the CEF

The *ANOVA theorem* (variance decomposition)

$$\mathbb{V}[Y_i] = \mathbb{V}[\mathbb{E}[Y_i|X_i]] + \mathbb{E}[\mathbb{V}[Y_i|X_i]]$$

The CEF prediction property

Let $m(X_i)$ be any function of X_i. The CEF solves,

$$\mathbb{E}[Y_i|X_i] = \underset{m(X_i)}{\operatorname{argmin}} \mathbb{E}\left[(Y_i - m(X_i))^2\right]$$

- The CEF is the best predictor of Y_i using the loss function $L(Y_i, \hat{Y}_i) = \mathbb{E}\left[(Y_i - \hat{Y}_i)^2\right]$. The CEF is also refereed to as the *minimum mean squares error* (MMSE) estimator of Y_i.

Homoskedasticity and Heteroskedasticity

Homoskedasticity

The error is homoskedastic if $\mathbb{E}\left[u^2|X\right] = \sigma^2$, does not depend on X.

Hetroskedasticity

The error is hetroskedastic if $\mathbb{E}\left[u^2|X = x\right] = \sigma^2(x)$, depends on x.

- The concepts of homoskedasticity and heteroskedasticity concern the conditional variance, not the unconditional variance.
- By definition the unconditional variance σ^2 is constant and independent of the regressors.
- The unconditional variance is a number, not a random variable or a function of X. The conditional variance depends on X and if X is a random variable it is also a random variable.

Projections

- Denote the population linear projection of Y_i on X_i by,

$$\mathbb{E}^*\left[Y_i|X_i\right] = X_i'\beta = \beta_0 + \sum_{j=1}^{p} X_{ji}\beta_j$$

where β is defined as the *minimum mean squared error* linear projection,

$$\beta = \underset{b}{\text{argmin}}\ \mathbb{E}\left[\left(Y_i - X_i'b\right)^2\right]$$

- The F.O.C of the minimization problem are,

$$\mathbb{E}\left[X_i \cdot \left(Y_i - X_i'\beta\right)\right] = 0 \Rightarrow \beta = \mathbb{E}\left[X_i X_i'\right]^{-1}\mathbb{E}\left[X_i Y_i\right]$$

- Denote by e_i the projection residual, $e_i = Y_i - \mathbb{E}^*\left[Y_i|X_i\right]$.
- It follows from the F.O.C that, $\mathbb{E}\left[X_i \cdot e_i\right] = 0$.
- The residual of projecting Y_i on X_i is not correlated with X_i.

Projections

Theorem (Projection is the MMSE linear approximation to the CEF)

$$\beta = \underset{b}{argmin}\, \mathbb{E}\left[\left(\mathbb{E}\left[Y_i|X_i\right] - X_i'b\right)^2\right]$$

If the conditional expectation is linear, $\mathbb{E}\left[Y_i|X_i\right] = X_i'\beta$ then, $\mathbb{E}^*\left[Y_i|X_i\right] = \mathbb{E}\left[Y_i|X_i\right]$. Proof

- Projections and conditional expectations are related also without imposing any structural assumptions,

$$\mathbb{E}^*\left[Y_i|X_i\right] = \mathbb{E}^*\left[\mathbb{E}\left[Y_i|Z_i, X_i\right]|X_i\right]$$

Proof:

$$\begin{aligned}\mathbb{E}^*\left[Y_i|X_i\right] &= X_i'\mathbb{E}\left[X_iX_i'\right]^{-1}\mathbb{E}\left[X_iY_i\right] = X_i'\mathbb{E}\left[X_iX_i'\right]^{-1}\mathbb{E}\left[X_i\mathbb{E}\left[Y_i|X_i, Z_i\right]\right] \\ &= \mathbb{E}^*\left[\mathbb{E}\left[Y_i|Z_i, X_i\right]|X_i\right]\end{aligned}$$

where $X_i'\mathbb{E}\left[X_iX_i'\right]^{-1}\mathbb{E}\left[X_iA\right] = \mathbb{E}^*\left[A|X_i\right]$ for all A, and in our case $A = \mathbb{E}\left[Y_i|Z_i, X_i\right]$.

Projections

- Projections also follow the *law of iterated projections* (LIP),

$$\mathbb{E}^*\left[Y_i|X_i\right] = \mathbb{E}^*\left[\mathbb{E}^*\left[Y_i|Z_i, X_i\right]|X_i\right]$$

- Consider the following special case, T_i is a binary variable (i.e., $T_i \in \{0, 1\}$). Show that,

$$\mathbb{E}\left[Y_i|T_i\right] = \mathbb{E}\left[\mathbb{E}^*\left[Y_i|X_i, T_i\right]|T_i\right]$$

- The proof is left as a homework assignment.

Regression as a structural model: The CEF is linear

- When $\mathbb{E}[Y_i|X_i] = X_i'\beta$, β has a *causal* interpretation. It is no longer a linear approximation based on correlations, but a structural parameter that describes a relationship between X_i and Y_i.
- The parameter β has two different possible interpretations:

(i) *Conditional mean interpretation*: β indicates the effect of increasing X on the conditional mean $\mathbb{E}[Y|X]$ in the model $\mathbb{E}[Y|X] = X\beta$.

(ii) *Unconditional mean interpretation*: By the LIE it follows that $\mathbb{E}[Y] = \mathbb{E}[X]\beta$.
Hence, β can be interpreted as the effect of increasing the mean value of X on the (unconditional) mean value of Y (i.e., $\mathbb{E}[Y]$).

Regression as a prediction algorithm

- Denote by $\mathbf{X}$ the design matrix. This is the matrix on which we will project $\mathbf{y}$.
- In other words we have an input matrix $\mathbf{X}$ with dimensions $n \times p$ and an output vector $\mathbf{y}$ with dimensions $n \times 1$.

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix}_{n \times p}$$

- The X_{ji} denotes the value of characteristic j for individual i.
- Example: If the researcher observes 2 covariates: education and age. A possible design matrix is,

$$\mathbf{X} = \begin{pmatrix} 1 & \text{education}_1 & \text{age}_1 & \text{age}_p^2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \text{education}_n & \text{age}_n & \text{age}_n^2 \end{pmatrix}_{n \times 4}$$

Regression as a prediction algorithm

- The linear regression algorithm has the following form:

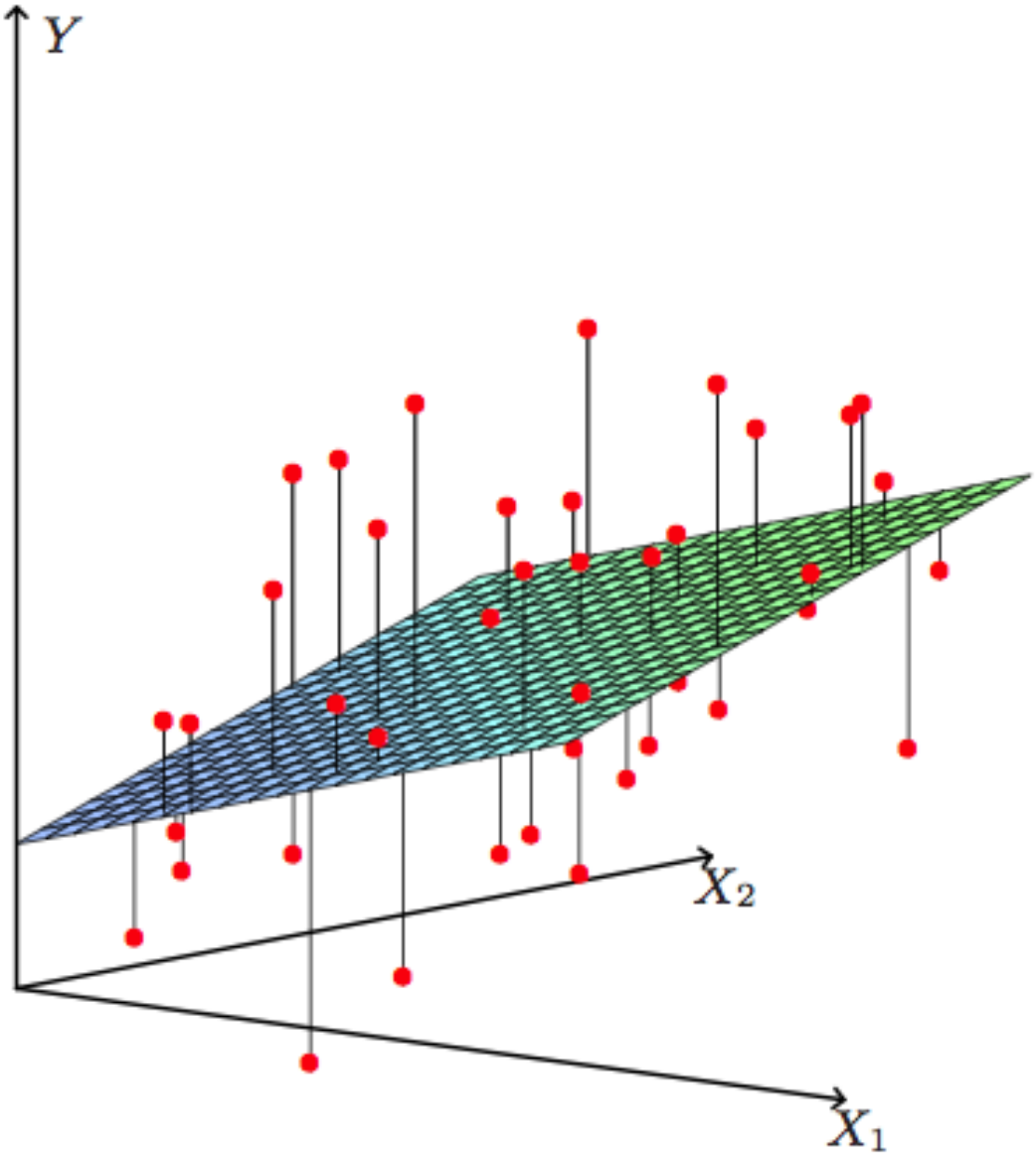
$$\hat{y}_i = f(X_i) = \beta_0 + \sum_{j=1}^{p} X_{ji}\beta_j$$

- We can estimate the coefficients β in a variety of ways but OLS is by far the most common, which minimizes the **residual sum of squares** (RSS):

$$RSS(\beta) = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{P} X_{ji}\beta_j)^2$$

- The estimation of β is according to a square loss function, i.e. $L(a, \hat{a}) = (a - \hat{a})^2$.

Visualization of OLS



The OLS estimator of β

- Write the RSS as:

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{x}\beta)$$

- Differentiate with respect to β:

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) \tag{1}$$

- Assume that $\mathbf{X}$ is full rank (no perfect collinearity among any of the independent variables) and set first derivative to 0:

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = 0$$

- Solve for β:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Rank condition for identifying β

- What happens if $\mathbf{X}$ is not full rank? There is an infinite number of ways to invert the matrix $\mathbf{X}'\mathbf{X}$, and the algorithm does not have a unique solution. There are many values of β that satisfy the F.O.C

- Therefore if $\mathbf{X}$ is not full rank, then $\hat{\beta}$ is not well defined!
- Note, that not all prediction algorithms require that the input matrix $\mathbf{X}$ be of full rank. For example Regression trees or more advanced methods such as Bagging or Boosting do not require a full rank condition.

Regression as a prediction: Making a Prediction

- The *hat matrix*, or *projection matrix* is,

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \text{ with } \tilde{\mathbf{H}} = \mathbf{I} - \mathbf{H}$$

- We use the hat matrix to find the fitted values:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

- The regression residuals are,

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

- $\mathbf{H}\mathbf{y}$ yields the part of $\mathbf{y}$ that can be explained by a linear combination of $\mathbf{X}$, and $\tilde{\mathbf{H}}\mathbf{y}$ is the part of $\mathbf{y}$ that is not explained using a linear combination of $\mathbf{X}$, which is the regression *residual*.
- Denote by $\mathbf{M}$ the regression residuals, $\mathbf{M} \equiv (I - \mathbf{H})\mathbf{y}$, hence $\mathbf{e} = \mathbf{M}\mathbf{y}$. What are the dimensions of $\mathbf{M}$ and $\mathbf{H}$? $n \times n$

Regression as a prediction: Making a Prediction

- Do we make any assumption on the distribution of $\mathbf{y}$? *No!*
- Can the dependent variable (the response), $\mathbf{y}$, be a binary variable, i.e $y \in \{0, 1\}$? *Yes!*
- Do we assume homoskedasticity, i.e that $Var(Y_i) = \sigma^2$, $\forall_i$? *No!*
- Is the residuals, $\mathbf{e}$, correlated with $\mathbf{y}$? Do we need to make any additional assumption in order for $corr(\mathbf{e}, \mathbf{X}) = 0$? *No!* The OLS algorithm will always yield residuals which are not correlated with the covariates, i.e., that are orthogonal to $\mathbf{X}$.
- The procedure we discussed so far is an algorithm, which solves an optimization problem (minimizing a square loss function). The algorithm requires an assumption of full rank in order to yield a unique solution, however it does not require any assumption on the distribution or the type of the response variable, $\mathbf{y}$.

Regression as a structural model: From algorithm to model

- Now we make stronger assumptions, most importantly we assume a data generating process (hence DGP), i.e we assume a functional form for the relationship between y and X

- Do we now assume y is a linear function of the covariates? *No*, we assume it is a linear function of β. An example of a response variable that is non-linear in β is:

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \exp(\beta_3 \cdot X_{2i} + \beta_4 \cdot X_{3i})$$

- What are the classic assumptions of the regression model? What is the role of the classic assumptions, why are we assuming them?

The classic assumptions of the regression model

1. The dependent variable is linearly related to the coefficients of the model and the model is correctly specified, $\mathbf{y} = \mathbf{X}\beta + \epsilon$
2. The independent variables, $\mathbf{X}$, are fixed, i.e are not random variables (this **should** be relaxed to the no *endogeneity* assumption $\mathbb{E}[\epsilon|X] = 0$, note that by LIE $\mathbb{E}[\epsilon|X] = 0 \Rightarrow \mathbb{E}[\epsilon X] = 0$).
3. The conditional mean of the error term is zero, $\mathbb{E}(\epsilon|X) = 0$
4. Homoscedasticity. The variance of the error term is independent of $\mathbf{X}$, i.e $\mathbb{V}(\epsilon_i) = \sigma^2$ (This assumption can easily be relaxed).
5. The error terms are uncorrelated with each other, $\mathbb{E}[\epsilon_i \cdot \epsilon_j] = 0$.
6. The design matrix, $\mathbf{X}$, has full rank (this is necessary for the OLS algorithm and the for identifying parameters of a regression structural model).
7. The error term is normally distributed,$\epsilon \sim N(0, \sigma^2)$ (This assumption can easily be relaxed. This assumption makes the OLS estimator also a maximum likelihood estimator (MLE)).

Notes on the classic assumptions of the regression model

- The assumption that $\mathbb{E}(\epsilon|X) = 0$ will always be satisfied when there is an intercept term in the model, i.e when the design matrix contains a constant term
- When $X \perp \epsilon$ it follows that $Cov(X, \epsilon) = 0$
- The normality assumption of ϵ_i is required for hypothesis testing on β in finite samples without asymptotic approximations.
  The assumption can be relaxed for sufficiently large sample sizes, as by the CLT, $\hat{\beta}_{OLS}$ converges to the normal distribution when $N \to \infty$.

Properties of the OLS estimators: Unbiased?

The OLS estimator of β is,

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(X\beta + \epsilon) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon \\ &= \beta + (X'X)^{-1}X'\epsilon\end{aligned}$$

An estimator is unbiased if $E(\hat{\beta}) = \beta$. There are 2 cases to consider: (1) X is fixed. (2) X is random.
We begin with the case of fixed X:

$$\begin{aligned}E(\hat{\beta}|X) &= E(\beta + (X'X)^{-1}X'\epsilon|X) \\ &= E(\beta|X) + E((X'X)^{-1}X'\epsilon|X) \\ &= \beta + (X'X)^{-1}E(\epsilon|X) \\ &\qquad \text{where } E(\epsilon|X) = E(\epsilon) = 0 \\ E(\hat{\beta}) &= \beta\end{aligned}$$

Note, fixed X implies that the expectations are conditional on X $\mathbb{E}[X'\epsilon|X]$.

Properties of the OLS estimators: Unbiased?

Next consider that case that X is a random variable,

$$\begin{aligned}\mathbb{E}\left[\beta + (X'X)^{-1}X'\epsilon\right] &= \beta + \mathbb{E}\left[(X'X)^{-1}X'\epsilon\right] \\ &\neq \beta + (\mathbb{E}\left[X'X\right])^{-1}\mathbb{E}\left[X'\epsilon\right] = \beta\end{aligned}$$

OLS is, in finite samples, generally biased for the population linear projection since the expectations operator cannot pass through a ratio.

Note, the OLS estimator can also be written in terms of sums,

$$\hat{\beta} = \left(\sum_i X_i X_i'\right)^{-1}\left(\sum_i X_i Y_i\right)$$

Although the OLS can be biased in finite samples, it is an asymptotically unbiased estimator (consistent) for the population linear projection.

The variance of $\hat{\beta}_{OLS}$ under fixed X and homoskedasticity

- Recall:

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(X\beta + \epsilon) \\ \Rightarrow \quad \hat{\beta} - \beta &= (X'X)^{-1}X'\epsilon \end{aligned}$$

- Plugging this into the covariance equation:

$$\begin{aligned} cov(\hat{\beta}|X) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X] \\ &= E\left[\left((X'X)^{-1}X'\epsilon\right)\left((X'X)^{-1}X'\epsilon\right)'|X\right] \\ &= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}|X] \\ &= (X'X)^{-1}X'E(\epsilon\epsilon'|X)X(X'X)^{-1} \\ &\qquad \text{where } E(\epsilon\epsilon'|X) = \sigma^2 I_{p\times p} \\ &= (X'X)^{-1}X'\sigma^2 I_{p\times p}X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

Estimating σ^2

We estimate σ^2 by dividing the residuals squared by the degrees of freedom because the e_i are generally smaller than the ϵ_i due to the fact that $\hat{\beta}$ was chosen to make the sum of square residuals as small as possible.

$$\hat{\sigma}^2_{OLS} = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2$$

Compare the above estimator to the classic variance estimator:

$$\hat{\sigma}^2_{classic} = \frac{1}{n-1}\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$$

Is one estimator always preferable over the other? If not when each estimator is preferable?

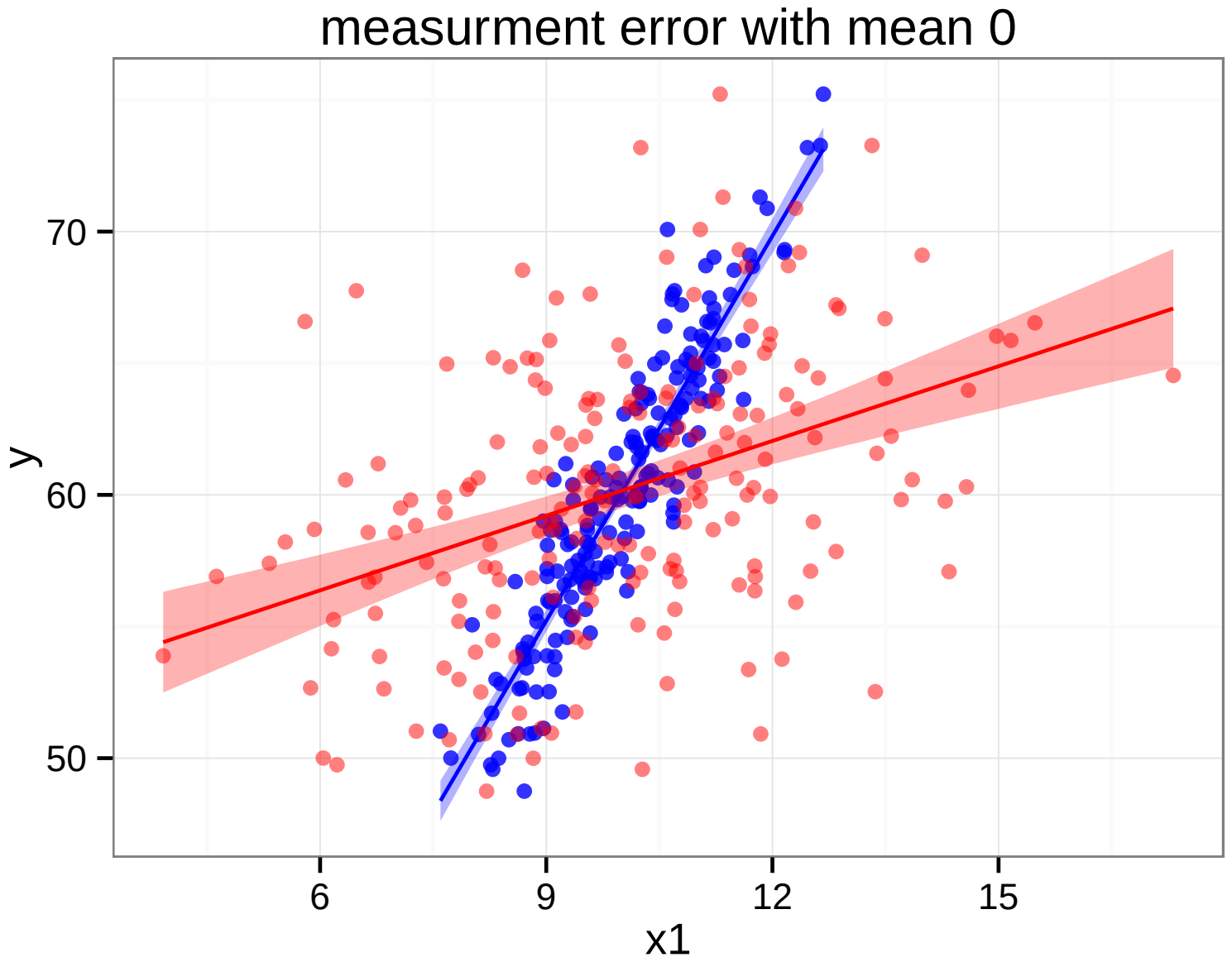
measurment error

Consider the following DGP (data generating process):

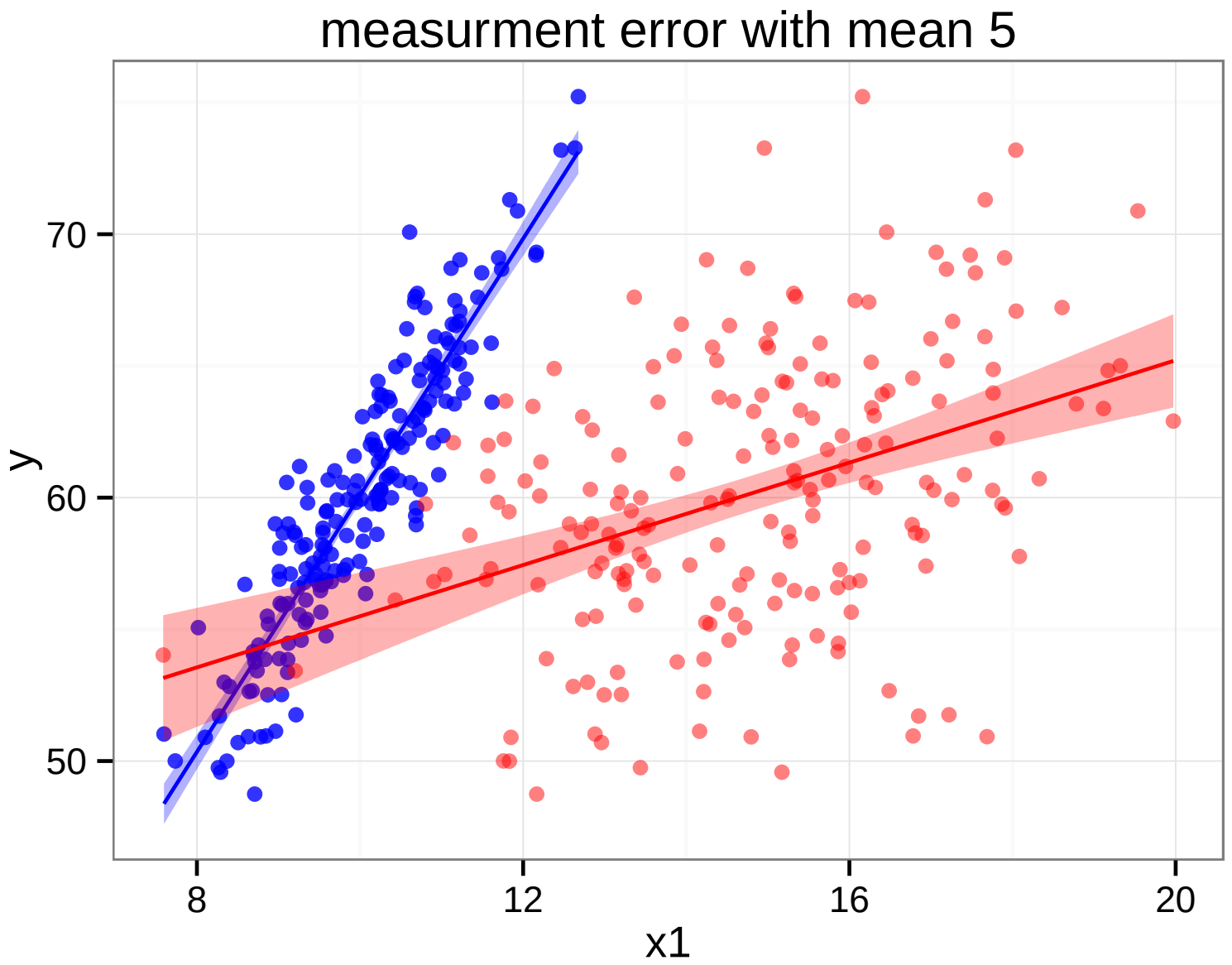
```
n=200
x1 = rnorm(n,mean=10,1)
epsilon = rnorm(n,0,2)
y = 10+5*x1+epsilon

### mesurment error:
noise = rnorm(n,0,2)
x1_noise = x1+noise
```

The true model has x_1, however we observe only x_1^{noise}. We will investigate the effect of the noise and the distribution of the noise on the OLS estimation of β_1. The true value of the parameter of interest is, $\beta_1 = 5$

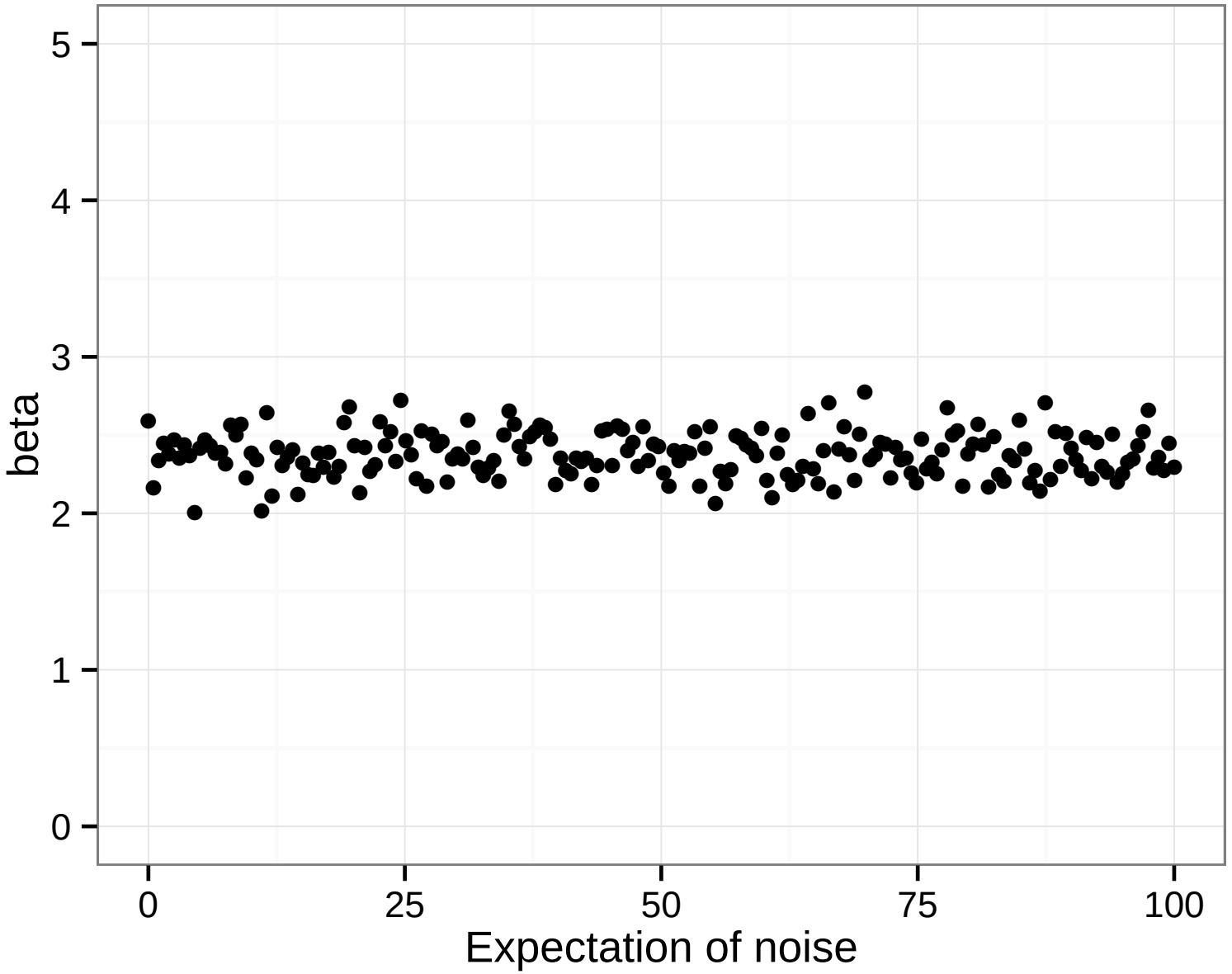
Measurement error: $noise \sim N(\mu = 0, \sigma = 2)$



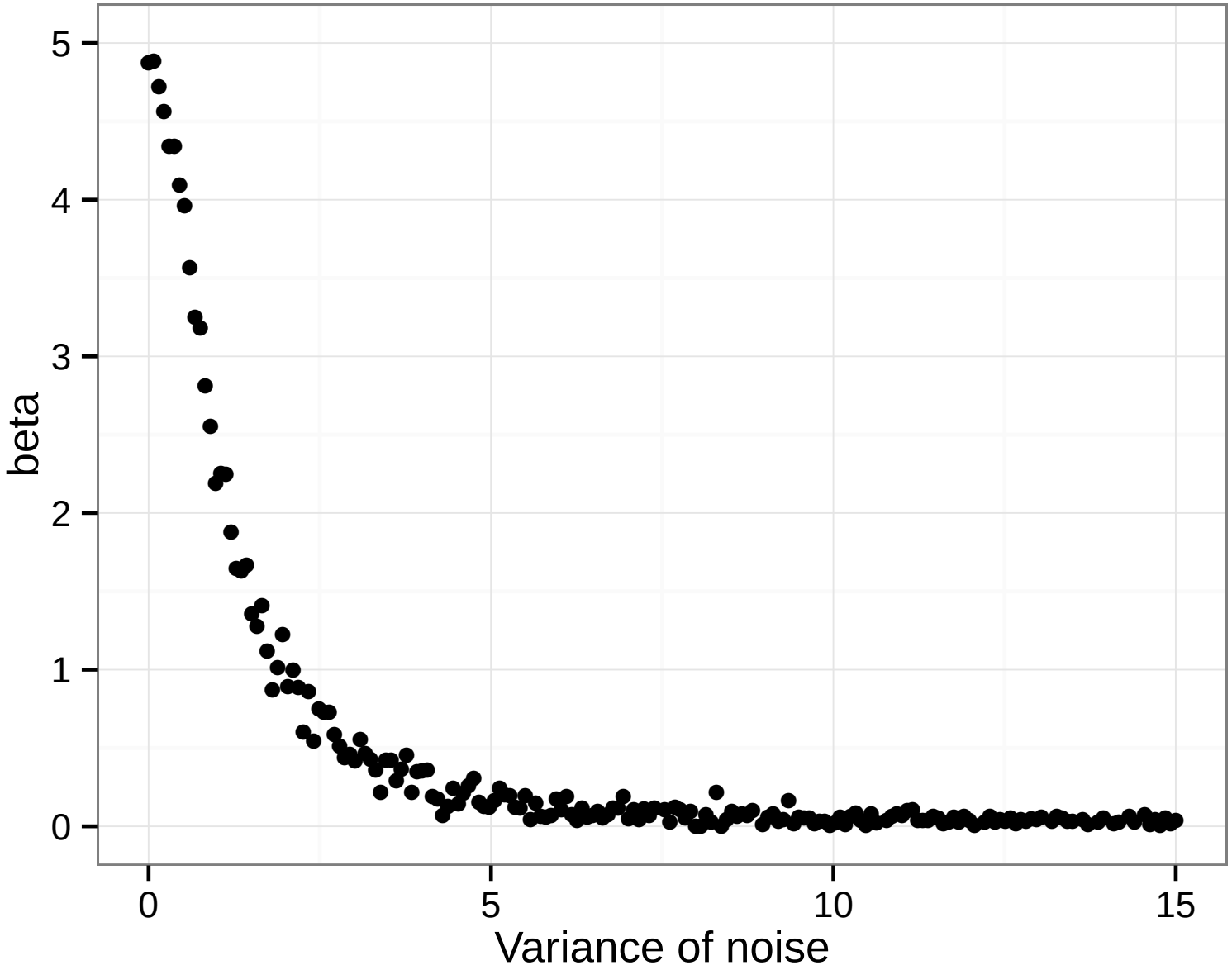
Measurement error: $noise \sim N(\mu = 5, \sigma = 2)$



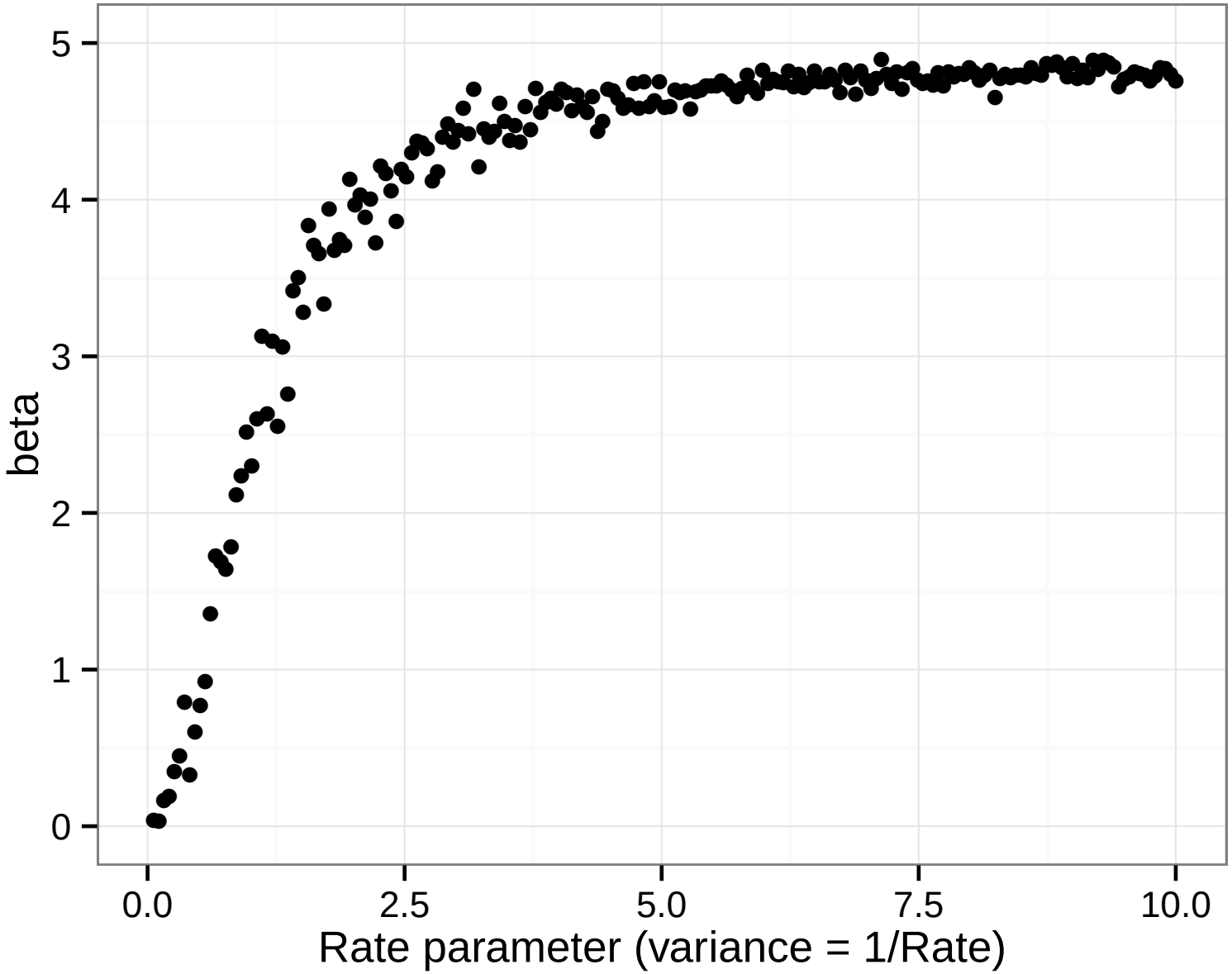
Measurement error: $noise \sim N(\mu = ?, \sigma = 2)$



Measurement error: $noise \sim N(\mu = 5, \sigma = ?)$



Measurement error: $noise \sim exp(\lambda = ?)$



Measurement error

- Could we reach the same conclusions as the simulations from analytical derivations? *Yes*
- As we saw before,

$$\mathbb{E}(\hat{\beta}_{OLS}) = \frac{Cov(y, x_1^{noise})}{\mathbb{V}(x_1^{noise})} = \frac{Cov(y, x_1 + noise)}{\mathbb{V}(x_1 + noise)}$$

$$= \frac{Cov(y, x_1)}{\mathbb{V}(x_1) + \mathbb{V}(noise)}$$

Therefore as $\mathbb{V}(noise) \to \infty$, the expectation of the OLS estimator of β will converge to zero,

$$\mathbb{V}(noise) \to \infty \Rightarrow \mathbb{E}(\hat{\beta}_{OLS}) = \frac{Cov(y, x_1)}{\mathbb{V}(x_1) + \mathbb{V}(noise)} \to 0$$

Measurement error in the dependent variable

- Consider the situation in which y_i is not observed, but y_i^{noise} is observed. There are no measurement error in x_1.
- The model (DGP) is,

$$y_i = 10 + 5 * x_{1i} + \epsilon_i$$

$$y_i^{noise} = y_i + noise_i$$

- Will the OLS estimator of β_1 be unbiased? Yes

$$\mathbb{E}(\hat{\beta}_{OLS}) = \frac{Cov(y^{noise}, x_1)}{\mathbb{V}(x_1)} = \frac{Cov(y + noise, x_1)}{\mathbb{V}(x_1)}$$

$$= \frac{Cov(y, x_1)}{\mathbb{V}(x_1)} = \beta_1$$

- This model is equivalent to the model, $y_i = 10 + 5 * x_{1i} + (\epsilon_i + noise_i)$, where y_i is observed.

Measurement error in the dependent variable

- Will the OLS estimator be unbiased if the measurement error was multiplicative instead of additive? Formally, if the DGP was:

$$y_i = 10 + 5 * x_{1i} + \epsilon_i$$

$$y_i^{noise} = y_i \cdot noise_i$$

- Analytic derivations:

$$\mathbb{E}(\hat{\beta}_{OLS}) = \frac{Cov(y^{noise}, x_1)}{\mathbb{V}(x_1)} = \frac{Cov(y \cdot noise, x_1)}{\mathbb{V}(x_1)}$$

$$Cov(y \cdot noise, x_1) = \mathbb{E}(y \cdot noise \cdot x_1) - \mathbb{E}(y \cdot noise) \cdot \mathbb{E}(x_1)$$
$$= \frac{\mathbb{E}(noise) \cdot Cov(y, x_1)}{\mathbb{V}(x_1)} = \mathbb{E}(noise) \cdot \beta_1$$

- When there is a multiplicative noise the bias of $\hat{\beta}$ is influenced by $\mathbb{E}(noise)$, not from $\mathbb{V}(noise)$.

Measurement error: Division bias

- Borjas (1980) introduced the problem of division bias in the context of labor supply estimation with respect to hourly wage.
- The problme arises when the researcher does not observe hourly wages, but observe:
  - Hours worked h - measured with error. The true hours of work are h^* and the observed hours of work are $h = \eta \cdot h^*$.
  - Earnings - measured without error. Earnings are defined as hourly wage times hours worked, $e = w^* \cdot h^*$.
  - The hourly wage is calculated as, $w = \frac{e}{h} = \frac{e}{\eta \cdot h^*}$. The true hourly wage is $w^* = \frac{e}{h^*}$.
  - If we observed h^* it would have been simple to extract w^* from e, $w^* = e/h^*$.
- By taking log() of w and h we can re-write the measurement error as,

$$\log(w) = \log(e) - \log(h^*) - \log(\eta) = \log(e/h^*) - \log(\eta) = \log(w^*) - \log(\eta)$$
$$\log(h) = \log(h^*) - \log(\eta)$$

Measurement error: Division bias

- The researcher is interested in estimating the following regression,

$$\log(h^*) = \alpha + \beta \cdot \log(w^*) + \epsilon_i \tag{2}$$

- What is the bias as a result of estimating instead,

$$\log(h) = a + b \cdot \log(w) + e_i \tag{3}$$

- Borjas (1980) showed that,

$$\text{plim } \hat{b} = \beta \cdot \frac{\sigma^2_{w^*}}{\sigma^2_{w^*} + \sigma^2_{\eta}} - 1 \cdot \frac{\sigma^2_{\eta}}{\sigma^2_{w^*} + \sigma^2_{\eta}}$$

- Does division bias always bias the OLS estimator downwards? No, if $\beta = -5$, the bias will be upwards.
- Division bias leads to an estimator that is biased towards -1, as $\hat{b}$ is a weighted average between β and -1.
- Unlike standard measurement error division bias can change the sign of the estimated OLS coefficient.

Measurement error: Division bias

- Until now we considered the affect of division bias when estimating a log-log regression. What is the effect when estimating:
  1. log-linear: $\log(h^*) = \alpha + \beta \cdot w^* + \epsilon_i$
  2. linear-log: $h^* = \alpha + \beta \cdot \log(w^*) + \epsilon_i$
  3. linear-linear: $h^* = \alpha + \beta \cdot w^* + \epsilon_i$
- In all this cases it is difficult (i.e., there is analytical solution) for the bias of the OLS coefficient.
- Simulations show the bias is downwards, and might change the sign of the coefficient similar to the log-log case.

Division bias: simulation

```
rm(list = ls())
set.seed(12345)
n=10000
elasticity=5
w.vec = seq(20,50,length=1000)
w=sample(w.vec,size=n,replace=TRUE)
h=w^elasticity
e = w*h
sigma.vec = seq(0.1,1,length=20)
noise = rnorm(n,mean=0,sd=0.7)+1
hs = h * noise
ws = e/hs
if(min(hs,ws)<0){
  index = (hs<0 | ws<0)
  hs[which(index)]=NA
  ws[which(index)]=NA
  cat("Fraction of missing observations: ",sum(index)/n,"\n")
}
```

Why do we need the *if* () statement? When the noise is negative it will make the observed hours of work a negative number, and $\log(\cdot)$ is not undefined over negative numbers.

The regression results without division bias

	Dependent variable:		
	log(h)		h
	(1)	(2)	(3)
log(w)	5.000***		
	(0.000)		
w		0.149***	9,163,099.000***
		(0.0002)	(36,657.090)
Constant	−0.000	12.415***	−234,614,765.000***
	(0.000)	(0.006)	(1,322,068.000)
Observations	10,000	10,000	10,000
R^2	1.000	0.986	0.862
Note:			*p<0.1; **p<0.05; ***p<0.01

The regression results with division bias

	Dependent variable:		
	log(hs)		hs
	(1)	(2)	(3)
log(ws)	-0.486^{***}		
	(0.018)		
ws		-0.0001^{***}	$-1{,}282.862^{**}$
		(0.00001)	(525.461)
Constant	19.267^{***}	17.504^{***}	$96{,}180{,}729.000^{***}$
	(0.067)	(0.016)	(1,239,380.000)
Observations	9,224	9,224	9,224
R^2	0.074	0.011	0.001
Note:			*p<0.1; **p<0.05; ***p<0.01

Gauss-Markov theorem: BLUE

- The regression estimator is a linear estimator, $\hat{\beta} = C\mathbf{y}$, where $C = (X'X)^{-1}X'$. A linear estimator is any $\hat{\beta}_j$ such that $\hat{\beta}_j = c_1 y_1 + c_2 y_2 + \cdots + c_p y_p$.

- The Gauss-Markov theorem: If assumptions following assumptions hold,
  - **Homoskedasticity** (without this the MMSE is GLS, which is not a feasible estimator as we do not know the covariance matrix).
  - **X is full rank** (necessary for the OLS algorithm to have a unique solution).
  - $\mathbb{E}[\epsilon_i | X_i] = 0$ (necessary for the estimator to be unbiased).
  - Y_i is a linear function X_i (linear in the parameters β).
  - The Gauss-Markov theorem is for fixed regressors, i.e. $\mathbb{E}\left[\hat{\beta}|X\right] = 0$.

  The OLS estimator is the best linear unbiased estimator (BLUE), in terms of MSE (Mean Squared Error).

What will be the critic on the Gauss-Markov theorem from a machine learning computer scientist, that is interested in making prediction?

Frisch-Waugh-Lovell: Regression Anatomy

- In the simple bivariate case:

$$\beta_1 = \frac{\mathrm{Cov}(Y_i, X_i)}{\mathrm{Var}(X_i)}$$

- In the multivariate case, β_j is:

$$\beta_j = \frac{\mathrm{Cov}(Y_i, \tilde{X}_{ij})}{\mathrm{Var}(\tilde{X}_{ij})} = \frac{\mathrm{Cov}(\tilde{Y}_i, \tilde{X}_{ij})}{\mathrm{Var}(\tilde{X}_{ij})}$$

  where $\tilde{A}_{ij}$ ($\tilde{Y}_i$ or $\tilde{X}_j$) are the residuals from the regression of A_{ij} on all other covariates.

- The multiple regression coefficient $\hat{\beta}_j$ represents the additional contribution of x_j on y, after x_j has been adjusted for $1, x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$
- What happens when x_j is highly correlated with some of the other x_k's?

Frisch-Waugh-Lovell: Regression Anatomy

- Claim: $\beta_j = \frac{\mathrm{Cov}(\tilde{Y}_i, \tilde{X}_{ij})}{\mathrm{Var}(\tilde{X}_{ij})}$, i.e $Cov(Y_i, \tilde{X}_{ij}) = Cov(\tilde{Y}_i, \tilde{X}_{ij})$
- Proof:
  Let $\tilde{Y}_i$ be the residuals of a regression of all the covariates except X_{ji} on Y_i, i.e

$$X_{ji} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_2 + \cdots + \beta_P X_{Pi} + f_i$$

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_2 + \cdots + \alpha_P X_{Pi} + e_i$$

  Then, $\hat{e}_i = \tilde{Y}_i$, and $\hat{f}_i = \tilde{X}_{ji}$
- It follows from the OLS algorithm that $Cov(x_{ki}, \tilde{X}_{ji}) = 0$, $\forall_{k \neq j}$. As the residuals of a regression are not correlated with any of the covariates

$$Cov(\tilde{Y}_i, \tilde{X}_{ij}) = Cov(Y_i - \hat{\alpha}_0 - \hat{\alpha}_1 X_{1i} - \hat{\alpha}_2 X_2 - \cdots + \hat{\alpha}_P X_{Pi}, \tilde{X}_{ij})$$

$$= Cov(Y_i, \tilde{X}_{ij})$$

Asymptotics of OLS

- Is the OLS estimator of β consistent? Yes
- Proof:
- Denote the observed characteristics of observation i by X_i. What is the dimensions of X_i? $p \times 1$
- $X_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{pmatrix}$
- $X_i X_i' = \begin{pmatrix} X_{i1}^2 & X_{i1}X_{i2} & \ldots & X_{i1}X_{ip} \\ X_{i2}X_{i1} & X_{i2}^2 & \ldots & X_{i2}X_{ip} \\ \vdots & \vdots & \vdots & \\ X_{ip}X_{i1} & X_{ip}X_{i2} & \ldots & X_{ip}^2 \end{pmatrix}$

Asymptotics of OLS

- Verify at home that,

$$X'X = \begin{pmatrix} \sum_{i=1}^n X_{i1}^2 & \sum_{i=1}^n X_{i1}X_{i2} & \cdots & \sum_{i=1}^n X_{1i}X_{ip} \\ \sum_{i=1}^n X_{i2}X_{i1} & \sum_{i=1}^n X_{i2}^2 & \cdots & \sum_{i=1}^n X_{i2}X_{ip} \\ \vdots & \vdots & \vdots & \\ \sum_{i=1}^n X_{ip}X_{i1} & \sum_{i=1}^n X_{ip}X_{i2} & \cdots & \sum_{i=1}^n X_{ip}^2 \end{pmatrix}_{(n\times p)}$$

- Hence, $X'X = \sum_{i=1}^n X_i'X_i$ or $X'X = \sum_{i=1}^n X_iX_i'$?
  $X'X = \sum_{i=1}^n X_iX_i'$
- Note (and verify at home),

$$X'y = \begin{pmatrix} \sum_{i=1}^n X_{i1}Y_i \\ \sum_{i=1}^n X_{i2}Y_i \\ \vdots \\ \sum_{i=1}^n X_{ip}Y_i \end{pmatrix} = \sum_{i=1}^n X_i'Y_i$$

Asymptotics of OLS

- The OLS estimator is, $\beta = (X'X)^{-1} X'y$
- Recall $(X \cdot k)^{-1} = k^{-1} \cdot (X)^{-1}$
- Multiplying and dividing by $\frac{1}{n}$ and $\frac{1}{\sqrt{n}}$ yields,

$$\beta = \left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{\sqrt{n}}X'y\right) = \left(\frac{1}{n}\sum_{i=1}^{n}X_i'X_i\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i'y_i\right)$$

$$\rightarrow \mathbb{E}\left(X_i'X_i\right)^{-1} \cdot \mathbb{E}\left(X_i Y_i\right) = \mathbb{E}\left(X_i X_i'\right)^{-1} \cdot \mathbb{E}\left(X_i\left(X_i'\beta + \epsilon_i\right)\right)$$

- The converges follows from the central limit theorem (CLT).

$$= \mathbb{E}\left(X_i X_i'\right)^{-1} \cdot \mathbb{E}\left(X_i X_i'\right)\beta + \mathbb{E}\left(X_i X_i'\right)^{-1} \cdot \mathbb{E}\left(X_i \epsilon_i\right) = \beta$$

A random coefficient model

- Assume the following DGP,

$$Y_i = \alpha + \beta \cdot T_i + \epsilon_i$$

  where β_i is the treatment effect that can change between individuals and is a random variable. ϵ_i is an error term independent of β_i and T_i.
- Assume the treatment T_i is assigned at random.
- Claim: The following regression provides an asymptotically unbiased estimate of a meaningful parameter of interest,

$$Y_i = a + b \cdot T_i + e_i$$

  is this claim correct?
- Asymptotically what is the OLS estimator (OLS algorithm) estimating? $b = \frac{Cov(Y_i, T_i)}{\mathbb{V}[T_i]}$
- What is $b = \frac{Cov(Y_i, T_i)}{\mathbb{V}[T_i]} =?$ $b = \mathbb{E}[\beta]$.

Regression in Causal Analysis

- Imagine we are analyzing a *randomized* experiment with a regression using the following model:

$$Y_i = \alpha + \beta_1 \cdot T_i + X_i' \cdot \beta_2 + \epsilon_i$$

  where T_i is an indicator variable for treatment status and X_i is a vector of *pre-treatment characteristics*.
- Under this model, what is random? We need to decide whether to assume X_i is fixed or a random variable.
- Note: If we consider X_i as a random variable we are interested in the average treatment effect, if we consider X_i as fixed we are interested in the conditional average treatment effect (conditional of this X characteristics).
- How do we interpret the coefficient β_1? What is the OLS estimator of the coefficient β_1?

Revisiting the regression example from lecture

- Lets recall the example from class:
  - Let Y be a IID standard normal random variable, indexed by t.
  - Define $\Delta Y_t = Y_t - Y_{t-1}$.
  - Lets estimate via OLS,

$$\Delta Y_t = \alpha + \beta_1 \Delta Y_{t-1} + \beta_2 \Delta Y_{t-2} + \beta_3 \Delta Y_{t-3}$$

  - Question: What are the values of the betas as $n \to \infty$?
- Next we consider a simplified regression specification and analytically calculate β_1 when $n \to \infty$,

$$\Delta Y_t = \alpha + \beta_1 \Delta Y_{t-1} + \beta_2 \Delta Y_{t-2}$$

- Using which theorem or method can we calculate β_1?

Revisiting the regression example from lecture

- Using which theorem or method can we calculate β_1?
  1. We know that as $n \to \infty$ the OLS estimator converges to,

     $$\beta = (\mathbb{E}[X'X])^{-1}\mathbb{E}[X'y]$$

     we can calculate $\mathbb{E}[X'X]$ and $\mathbb{E}[X'y]$ and than invert $\mathbb{E}[X'X]$. This might be a bit algebraic as $\mathbb{E}[X'X]$ is of dimensions 3×3.
  2. Another option is to use FWL theorem. What are the steps to calculate β_1 using FWL?
- Calculating β_1 using FWL, step-by-step:
  1. Regress ΔY_{t-1} on ΔY_{t-2} and an intercept. Denote the residuals as $\Delta \tilde{Y}_{t-1}$.
  2. Regress ΔY_t on $\Delta \tilde{Y}_{t-1}$ and an intercept.
  3. By FWL, $\beta_1 = \frac{Cov(\Delta Y_t, \Delta \tilde{Y}_{t-1})}{\mathbb{V}[\tilde{Y}_{t-1}]}$.
- Calculating β_1 using FWL might take a few steps and algebra, but is a powerful tool for extracting a single coefficient from a multivariate regression.

Calculating β_1 using FWL

Step 1: Regress ΔY_{t-1} on ΔY_{t-2} and an intercept:

$$\Delta Y_{t-1} = a + b\Delta Y_{t-2} + \Delta \tilde{Y}_{t-1}$$

$$\Rightarrow b = \frac{Cov(\Delta Y_{t-1}, \Delta Y_{t-2})}{\mathbb{V}[\Delta Y_{t-2}]} = \frac{-\sigma_Y^2}{\sigma_Y^2 + \sigma_Y^2} = \frac{-1}{1+1} = -\frac{1}{2}$$

we can calclulate a by using the fact that the regression line passes through the mean point,

$$\mathbb{E}[\Delta Y_{t-1}] = a + b\mathbb{E}[\Delta Y_{t-2}] \Rightarrow 0 = a + b \cdot 0 \Rightarrow a = 0$$

Hence, $\Delta \tilde{Y}_{t-1} = \Delta Y_{t-1} + \frac{1}{2}\Delta Y_{t-2}$ Step 2: Regress ΔY_t on $\Delta \tilde{Y}_{t-1}$ and an intercept:

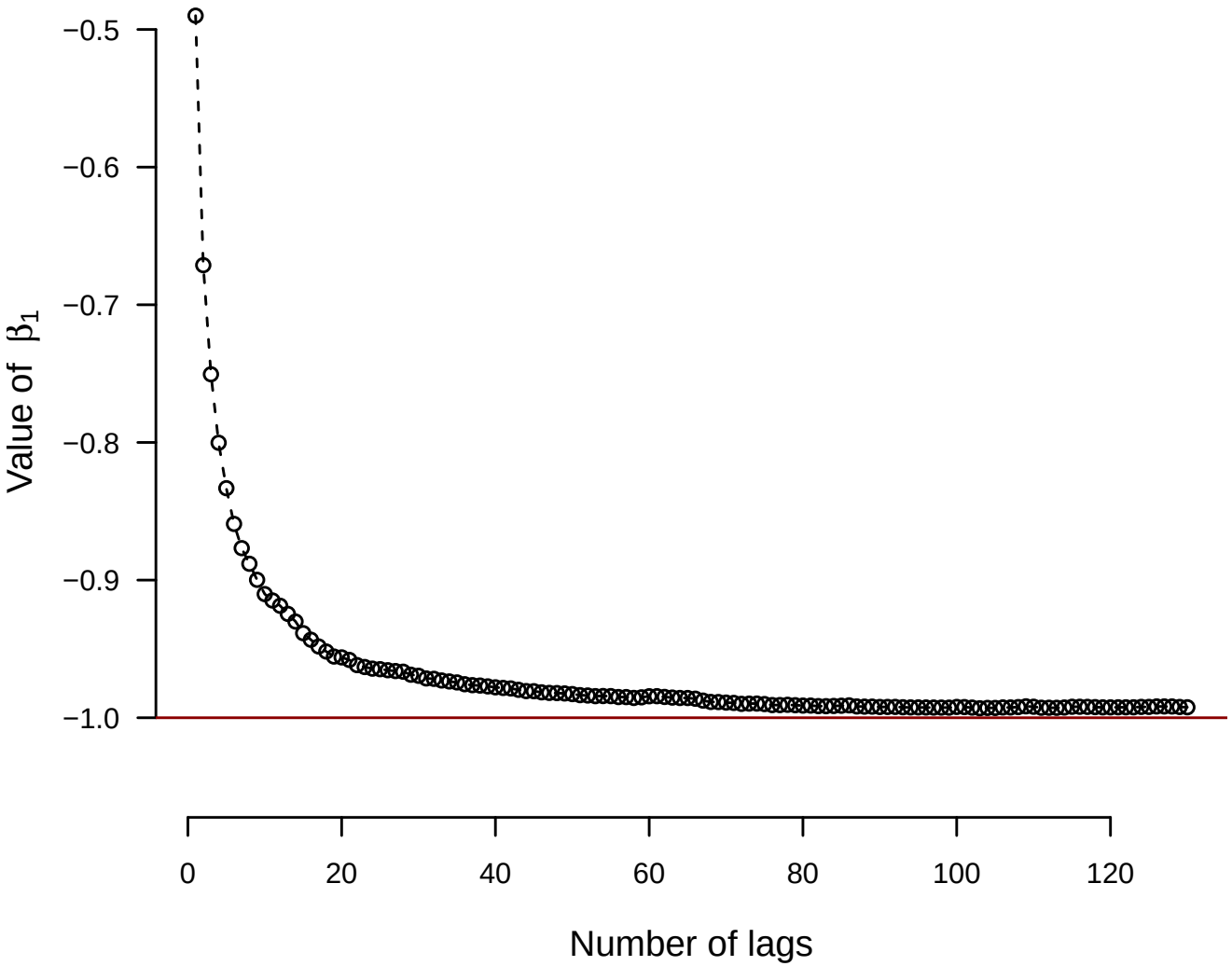
$$\beta = \frac{Cov(\Delta Y_t, \Delta \tilde{Y}_{t-1})}{\mathbb{V}\left[\Delta \tilde{Y}_{t-1}\right]} = \cdots = \frac{-\sigma_Y^2}{\sigma_Y^2 + \frac{1}{4}\sigma_Y^2 + \frac{1}{4}\sigma_Y^2} = -\frac{2}{3}$$

note that, $\sigma_Y^2 = 1$ as Y is distributed $i.i.d$ standard normal.

Revisiting the regression example from lecture: Coding exercise

- Next we use **R** to test the claim that as the number of Δlags increases and $n \to \infty$, the value of $\hat{\beta}_1$ converges.
- Write a simulation in **R** that:
  - Generates $\Delta Y_t, \Delta Y_{t-1}, \ldots, \Delta Y_{t-p}$ variables, where $p = 150$.
  - Write a loop that calculates $\hat{\beta}_1$ for each for each number of lags between 1 and 130.
  - Plot your results in a figure.

Simulation results



```
rm(list=ls())
set.seed(12345)

n=5000
p=150

Y = matrix(rnorm(n*p),ncol=p,nrow=n,byrow=TRUE)
for (l in c(1:(p-1))){
  Y[,l] = Y[,l]-Y[,l+1]
}
Y = Y[,-p]
data = data.frame(y=Y[,1],Y[,-1])

pp=130
num.lags = c(1:pp)
beta1 = rep(NA,pp)
for (k in c(1:pp)){
  beta1[k] = coef(lm(y~(.),data=data[,c(1:(num.lags[k]+1))]))[2]
}
```

```
par(cex=0.7,cex.lab=1.2)
plot(num.lags,beta1,
las=1,
frame=FALSE,
ylim=c(min(min(beta1,-1.05)),max(beta1)),
    xlab="Number of lags",
    ylab=expression(paste("Value of  ",beta[1])))

lines(num.lags,beta1,lty=2)
abline(h=-1,col="red4")
```

Proof: $\mathbb{E}^*[Y_i|X_i] = \mathbb{E}[Y_i|X_i]$ if $\mathbb{E}[Y_i|X_i]$ is linear

$$\begin{aligned}\mathbb{E}^*[Y_i|X_i] &= X_i'\mathbb{E}\left[X_iX_i'\right]^{-1}\mathbb{E}[X_iY_i] = X_i'\mathbb{E}\left[X_iX_i'\right]^{-1}\mathbb{E}[X_i\mathbb{E}[Y_i|X_i]] \\ &= X_i'\mathbb{E}\left[X_iX_i'\right]^{-1}\mathbb{E}\left[X_iX_i'\beta\right] = X_i'\mathbb{E}\left[X_iX_i'\right]^{-1}\mathbb{E}\left[X_iX_i'\right]\beta \\ &= X_i'\beta \qquad \square\end{aligned}$$

Back