Section 2 : Regression Analysis

Andrew Bertoli

11 September 2013

Roadmap

1. Basics of Regression
2. Advanced Regression
3. Reading Questions
4. Applications
5. Homework Questions

Basics of Regression

Classic Frequentist Model

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$$

Estimate of Parameters

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y}$$

Standard Errors

$$\widehat{\mathbf{cov}} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} \text{ where } \hat{\sigma}^2 = \frac{1}{n-p}\sum_i e_i^2$$

Basics of Regression

Example

Son's Height$= \beta_0 + \beta_1 \cdot$ Father's Height $+ \epsilon$

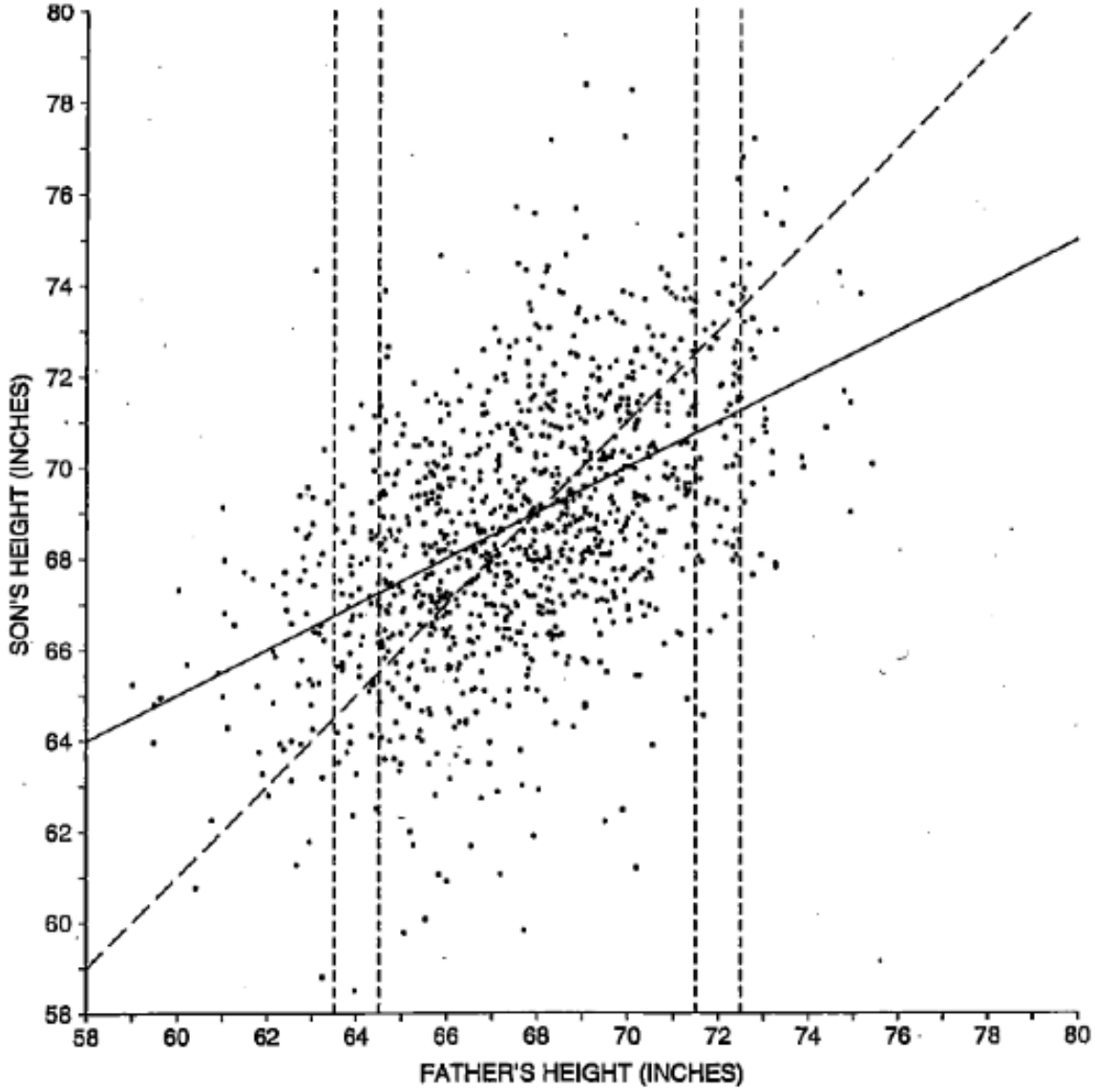
Nature draws ϵ from a Normal$(0, \sigma^2)$ distribution.

ϵ and the Son's Height are random. β_0, β_1, and the Father's Height are fixed.

This differs greatly from the experimental model (Neyman 1923) where only the treatment is random:

$Y_i = Y_{it} T_i + Y_{ic}(1 - T_i)$

Basics of Regression



Basics of Regression

The model with 1 independent variable

$$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \\ \mathbf{y_3} \\ . \\ . \\ . \\ . \\ . \\ \mathbf{y_n} \end{bmatrix} = \begin{bmatrix} 1 & x_{12} \\ 1 & x_{22} \\ 1 & x_{32} \\ . & . \\ . & . \\ . & . \\ . & . \\ . & . \\ 1 & x_{n2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon_1} \\ \boldsymbol{\epsilon_2} \\ \boldsymbol{\epsilon_3} \\ . \\ . \\ . \\ . \\ . \\ \boldsymbol{\epsilon_n} \end{bmatrix}$$

Basics of Regression

The real data with 1 independent variable

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} \\ 1 & x_{22} \\ 1 & x_{32} \\ . & . \\ . & . \\ . & . \\ . & . \\ . & . \\ 1 & x_{n2} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ . \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

Basics of Regression

The real data with $p-1$ independent variables

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & x_{13} & \dots & x_{1p} \\ 1 & x_{22} & x_{23} & \dots & x_{2p} \\ 1 & x_{32} & x_{33} & \dots & x_{3p} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ . \\ . \\ \hat{\beta}_{p-1} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ . \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

Basics of Regression

Model

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$$

Estimate of Parameters

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y}$$

Standard Errors

$$\widehat{\mathbf{cov}} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} \text{ where } \hat{\sigma}^2 = \frac{1}{n-p}\sum_i e_i^2$$

Basics of Regression

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ . \\ . \\ \hat{\beta}_{p-1} \end{bmatrix} = \begin{bmatrix} & & & \\ \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x'_{21} & x'_{22} & \cdots & x'_{2n} \\ x'_{31} & x'_{32} & \cdots & x'_{3n} \\ . & . & \cdots & . \\ x'_{p1} & x'_{n2} & \cdots & x'_{pn} \end{bmatrix} & & \\ & & & \end{bmatrix} \begin{bmatrix} 1 & x_{12} & \cdots & x_{1p} \\ 1 & x_{22} & \cdots & x_{2p} \\ 1 & x_{32} & \cdots & x_{3p} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots & . \\ 1 & x_{n2} & \cdots & x_{np} \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x'_{21} & x'_{22} & \cdots & x'_{2n} \\ x'_{31} & x'_{32} & \cdots & x'_{3n} \\ . & . & \cdots & . \\ x'_{p1} & x'_{n2} & \cdots & x'_{pn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ . \\ . \\ y_n \end{bmatrix}$$

Basics of Regression

$\hat{\boldsymbol{\beta}}$ minimizes the residual sum of squares. To prove this fact, consider some estimator $\dot{\boldsymbol{\beta}}$. It is easy to show that the RSS is minimized when $\dot{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$.

$RSS(\dot{\boldsymbol{\beta}}) = \mathbf{e}'\mathbf{e}$ $(= \sum_i e_i^2)$

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

$\frac{\partial RSS}{\partial \dot{\boldsymbol{\beta}}} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\dot{\boldsymbol{\beta}})$

Next, set the derivative at 0 to find the minimum.

$0 = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\dot{\boldsymbol{\beta}})$

$0 = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\dot{\boldsymbol{\beta}}$

So the RSS is minimized when

$\dot{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y}$

$\dot{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$

Basics of Regression

Model

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$$

Estimate of Parameters

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y}$$

Standard Errors

$$\widehat{\mathbf{cov}} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} \text{ where } \hat{\sigma}^2 = \frac{1}{n-p}\sum_i e_i^2$$

Basics of Regression

$$\begin{bmatrix} \widehat{var}(\hat{\beta}_0) & \widehat{cov}(\hat{\beta}_0, \hat{\beta}_1) & \dots & \widehat{cov}(\hat{\beta}_0, \hat{\beta}_{p-1}) \\ \widehat{cov}(\hat{\beta}_1, \hat{\beta}_0) & \widehat{var}(\hat{\beta}_1) & \dots & \widehat{cov}(\hat{\beta}_1, \hat{\beta}_{p-1}) \\ \widehat{cov}(\hat{\beta}_2, \hat{\beta}_0) & \widehat{cov}(\hat{\beta}_2, \hat{\beta}_1) & \dots & \widehat{cov}(\hat{\beta}_2, \hat{\beta}_{p-1}) \\ . & . & \dots & . \\ . & . & \dots & . \\ \widehat{cov}(\hat{\beta}_{p-1}, \hat{\beta}_0) & \widehat{cov}(\hat{\beta}_{p-1}, \hat{\beta}_1) & \dots & \widehat{var}(\hat{\beta}_{p-1}) \end{bmatrix} =$$

$$\hat{\sigma}^2 \begin{bmatrix} & & & \\ & \begin{bmatrix} 1 & 1 & \dots & 1 \\ x'_{21} & x'_{22} & \dots & x'_{2n} \\ x'_{31} & x'_{32} & \dots & x'_{3n} \\ . & . & \dots & . \\ x'_{p1} & x'_{n2} & \dots & x'_{pn} \end{bmatrix} & & \begin{bmatrix} 1 & x_{12} & \dots & x_{1p} \\ 1 & x_{22} & \dots & x_{2p} \\ 1 & x_{32} & \dots & x_{3p} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ 1 & x_{n2} & \dots & x_{np} \end{bmatrix} \\ & & & \end{bmatrix}^{-1}$$

Basics of Regression

$$\begin{bmatrix} \widehat{var}(\hat{\beta}_0) & \widehat{cov}(\hat{\beta}_0, \hat{\beta}_1) & \dots & \widehat{cov}(\hat{\beta}_0, \hat{\beta}_{p-1}) \\ \widehat{cov}(\hat{\beta}_1, \hat{\beta}_0) & \widehat{var}(\hat{\beta}_1) & \dots & \widehat{cov}(\hat{\beta}_1, \hat{\beta}_{p-1}) \\ \widehat{cov}(\hat{\beta}_2, \hat{\beta}_0) & \widehat{cov}(\hat{\beta}_2, \hat{\beta}_1) & \dots & \widehat{cov}(\hat{\beta}_2, \hat{\beta}_{p-1}) \\ . & . & \dots & . \\ . & . & \dots & . \\ \widehat{cov}(\hat{\beta}_{p-1}, \hat{\beta}_0) & \widehat{cov}(\hat{\beta}_{p-1}, \hat{\beta}_1) & \dots & \widehat{var}(\hat{\beta}_{p-1}) \end{bmatrix} =$$

$$\hat{\sigma}^2 \begin{bmatrix} & & & \\ & \begin{bmatrix} 1 & 1 & \dots & 1 \\ x'_{21} & x'_{22} & \dots & x'_{2n} \\ x'_{31} & x'_{32} & \dots & x'_{3n} \\ . & . & \dots & . \\ x'_{p1} & x'_{n2} & \dots & x'_{pn} \end{bmatrix} & & \begin{bmatrix} 1 & x_{12} & \dots & x_{1p} \\ 1 & x_{22} & \dots & x_{2p} \\ 1 & x_{32} & \dots & x_{3p} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ 1 & x_{n2} & \dots & x_{np} \end{bmatrix} \\ & & & \end{bmatrix}^{-1}$$

Basics of Regression

The Coefficient of Determination

R^2 is a measure of how well our data points fit the line.

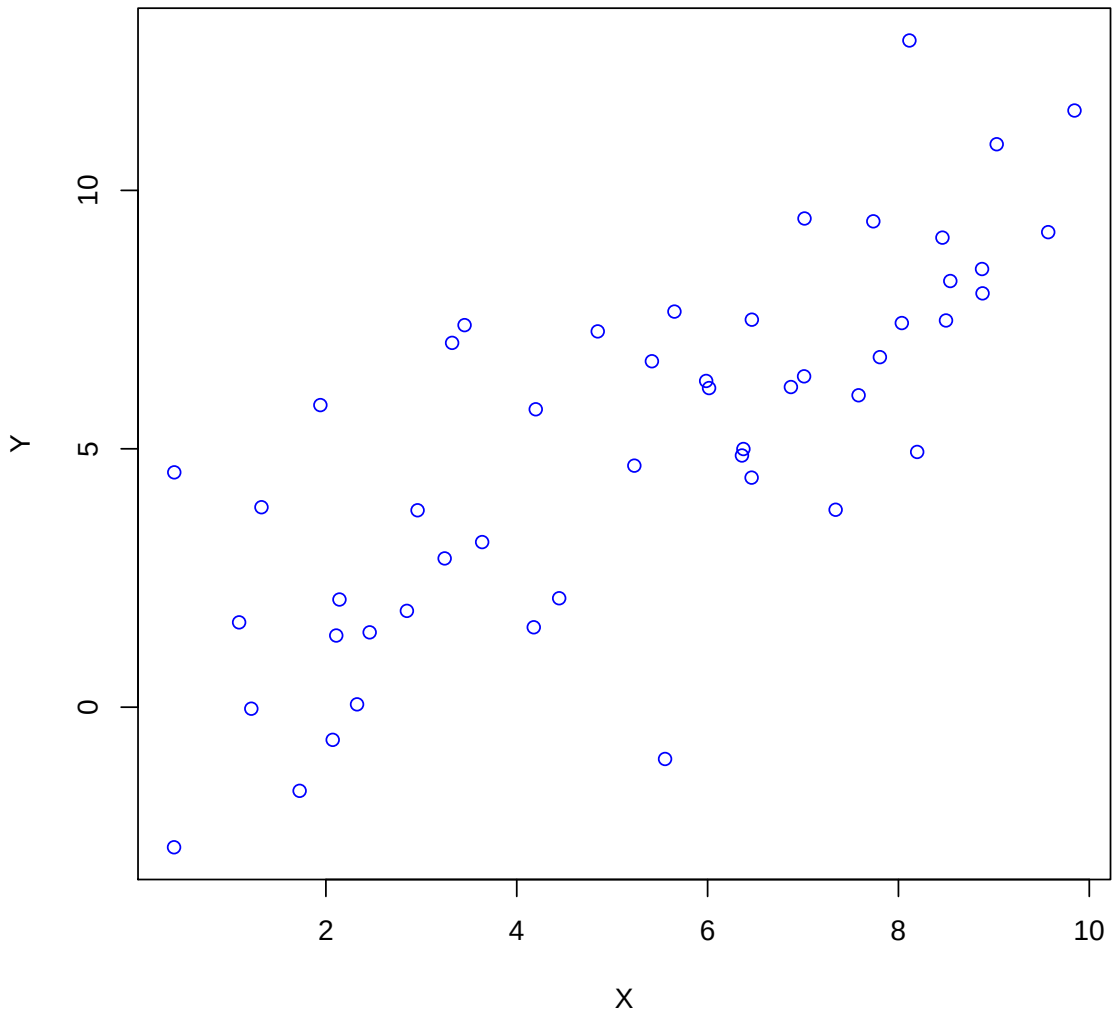
In univariate regression, R^2 is just the square of the correlation term

$R^2 = (\rho_{X,Y})^2 = (\frac{Cov(X,Y)}{\sigma_X \sigma_Y})^2$

In multiple regression, we calculate R^2 by using the formula

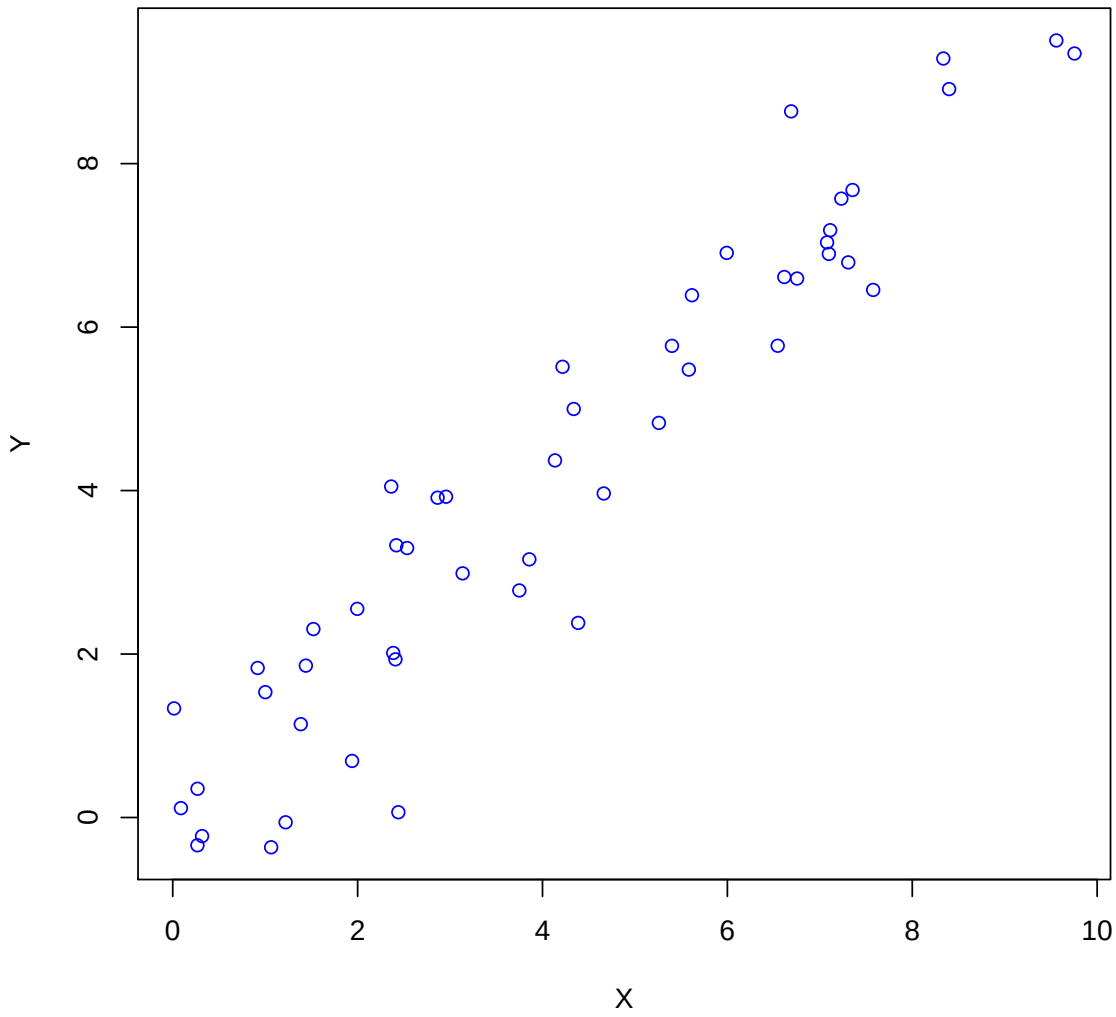
$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{\sum_i (y_i - \hat{y}_i)}{\sum_i (y_i - \bar{y})}$

Basics of Regression



$$R^2 = 0.58$$

Basics Regression



$$R^2 = 0.90$$

Basics of Regression

The Coefficient of Determination

A high R^2 is not proof of a causal relationship. For instance, between 1950 and 1999, the correlation between the purchasing power of the dollar and the death rate from lung cancer was -0.95. So $R^2 = (-0.95)^2 \approx 0.9$, which is extremely high. Why?

Basics of Regression

Pop Quiz (Continued)

Which of these assumptions are necessary for $\hat{\beta}$ to be unbiased?

1. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$
2. All independent and control variables are fixed (no measurement error)
3. There is no deterministic linear relationship between the X variables (no collinearity)
4. $E[\epsilon_i | X] = 0$ for all i
5. The ϵ_i are i.i.d. $N(0, \sigma^2)$ for all i

Basics of Regression

Pop Quiz (Continued)

Which of these assumptions are necessary for $\hat{\beta}$ to be unbiased?

1. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$

2. **All independent variables are fixed (no measurement error)**

3. **There is no deterministic linear relationship between the X variables (no collinearity)**

4. $E[\epsilon_i | X] = 0$ **for all** i

5. The ϵ_i are i.i.d. $N(0, \sigma^2)$ for all i

Basics of Regression

Unbiasedness

$E[\hat{\beta}|X] = E[(X'X)^{-1}X'Y|X]$ (Assumption 2: X is fixed)

$E[\hat{\beta}|X] = E[(X'X)^{-1}X'(X\beta + \epsilon)|X]$ (Assumption 1: Linear Additivity)

$E[\hat{\beta}|X] = E[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'X\epsilon)|X]$

$E[\hat{\beta}|X] = E[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'X\epsilon)|X]$

$E[\hat{\beta}|X] = E[(X'X)^{-1}X'X\beta|X] + E[(X'X)^{-1}X'X\epsilon)|X]$

$E[\hat{\beta}|X] = E[\beta] + E[(X'X)^{-1}X'\epsilon)|X]$ (Assumption 3: X is of full rank-No collinearity)

$E[\hat{\beta}|X] = \beta + (X'X)^{-1}E[X'\epsilon)|X]$

$E[\hat{\beta}|X] = \beta + (X'X)^{-1}[0]$ (Assumption 4: $E[\epsilon_i|X] = 0$-Exogeneity)

$E[\hat{\beta}|X] = \beta$

Basics of Regression

Pop Quiz (Continued)

Which of these assumptions are necessary for the standard error to be right?

1. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$
2. All independent and control variables are fixed (no measurement error)
3. There is no deterministic linear relationship between the X variables (no collinearity)
4. $E[\epsilon_i | X] = 0$ for all i
5. The ϵ_i are i.i.d. $N(0, \sigma^2)$ for all i

Basics of Regression

Deriving the Covariance Matrix

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

$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon)$ (Assumption 1: Linear Additivity)

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

$\mathbf{cov}(\hat{\beta}|\mathbf{X}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|\mathbf{X}]$ (Assumption 2: X is fixed)

$\mathbf{cov} = E[((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon)((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon)'|\mathbf{X}]$

$\mathbf{cov} = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon\epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}]$

$\mathbf{cov} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\epsilon\epsilon'|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$

So $E(\epsilon\epsilon'|\mathbf{X}) = \sigma^2 I_{pxp}$ (Assumptions 4 & 5: $\epsilon_i \sim N(0, \sigma^2)$ for all i)*

$\mathbf{cov} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2 I_{pxp}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$

$\mathbf{cov} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$

$\mathbf{cov} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ (Assumption 3: X is of full rank-No collinearity)

*Note: Normality is needed for the hypothesis tests to be accurate.

Basics of Regression

Deriving $\hat{\sigma}^2$

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

$\hat{\sigma}^2 = \frac{1}{n-p} \sum_i e_i^2$

Advanced Regression

$\hat{\beta}$ is BLUE (Best Linear Unbiased Estimator), meaning that it has the least variance of all unbiased linear estimators.

Formalization of Theorem:

Let $\gamma = \mathbf{c}'\boldsymbol{\beta}$, where $\mathbf{c}$ is p x 1. The parameter γ is a linear combination of the components of $\boldsymbol{\beta}$. Examples would include β_1 ($\mathbf{c} = (0, 1, 0, 0, 0, ...)$) or $\beta_2 - \beta_3$ ($\mathbf{c} = (0, 0, 1, -1, 0, ...)$).

The OLS estimator for γ is $\hat{\gamma} = \mathbf{c}'\hat{\boldsymbol{\beta}} = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. So $\hat{\gamma}$ is unbiased (since $E(\boldsymbol{\epsilon}|\mathbf{X}) = 0$ and $Cov(\boldsymbol{\epsilon}|\mathbf{X}) = \sigma^2\mathbf{I}$), and $Var(\hat{\gamma}) = \sigma^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}$.

Now let θ be another unbiased estimator of γ. Then $Var(\theta) \geq Var(\hat{\gamma})$, with equality holding only if $\theta = \hat{\gamma}$.

Advanced Regression

Pf:

Since θ is a linear estimator of Y, there is an n x 1 vector $\mathbf{d}$ with $\theta = \mathbf{d}'\mathbf{Y} = \mathbf{d}'\mathbf{X}\boldsymbol{\beta} + \mathbf{d}'\boldsymbol{\epsilon}$. Then $E(\theta) = \mathbf{d}'\mathbf{X}\boldsymbol{\beta}$. Since θ is unbiased, $\mathbf{d}'\mathbf{X}\boldsymbol{\beta} = \mathbf{c}'\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$. Therefore,

$$\mathbf{d}'\mathbf{X} = \mathbf{c}'$$

Let $\mathbf{q} = \mathbf{d} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}$, making $\mathbf{q}$ an n x 1 vector. So

$$\mathbf{q}' = \mathbf{d}' - \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Note that $\theta - \hat{\gamma} = \mathbf{q}'\mathbf{Y}$

So

$$\mathbf{q}'\mathbf{X} = \mathbf{d}'\mathbf{X} - \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}$$

$$\mathbf{q}'\mathbf{X} = \mathbf{d}'\mathbf{X} - \mathbf{c}'$$

$$\mathbf{q}'\mathbf{X} = \mathbf{0}_{1xp}$$

Advanced Regression

We also have

$\mathbf{d}' = \mathbf{q}' + \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

$Var(\theta) = Var(\mathbf{d}'\epsilon)$

$Var(\theta) = \sigma^2\mathbf{d}'\mathbf{d}$

$Var(\theta) = \sigma^2[\mathbf{q}' + \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{q} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}]$

$Var(\theta) = \sigma^2[\mathbf{q}'\mathbf{q} + \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}]$

The cross-product terms drop out because $\mathbf{q}'\mathbf{X} = \mathbf{0}_{1xp}$ and $\mathbf{X}'\mathbf{q} = \mathbf{0}_{px1}$. So we have

$Var(\theta) = \sigma^2\mathbf{q}'\mathbf{q} + Var(\hat{\gamma})$

So we have $Var(\tilde{\gamma}) \geq Var(\hat{\gamma})$, with equality holding only when $\mathbf{q} = \mathbf{0}_{nx1}$ (i.e. $\tilde{\gamma} = \hat{\gamma}$).

Advanced Regression

Pop Quiz (continued)

Question: You run an experiment and find the estimated treatment effect $\hat{\tau}$. You then fit the data to the model

$$y = \alpha + \beta x + \epsilon$$

and estimate β using OLS. What is the relationship between $\hat{\beta}$ and $\hat{\tau}$?

Advanced Regression

$$\hat{\tau} = ave(Y_i : X_i = 1) - ave(Y_i : X_i = 0)$$

$$\hat{\tau} = \frac{\sum X_i Y_i}{\sum X_i} - \frac{\sum (1-X_i) Y_i}{\sum (1-X_i)}$$

$$\hat{\tau} = \frac{ave(XY)}{ave(X)} - \frac{ave(Y)-ave(XY)}{1-ave(X)}$$

$$\hat{\tau} = \frac{ave(XY)-ave(X)ave(Y)}{ave(X)(1-ave(X))}$$

$$\hat{\tau} = \frac{ave(XY)-ave(X)ave(Y)}{p(1-p)}$$

where p is the proportion of subjects assigned to the treatment group. So

$$\hat{\tau} = \frac{Cov(X,Y)}{Var(X)}$$

$$\hat{\tau} = \hat{\beta}$$

Advanced Regression

Pop Quiz (Continued)

Question: After you verify that $\hat{\tau}$ and $\hat{\beta}$ are the same, you compute their standard errors. What is the relationship between $\widehat{SE}(\hat{\tau})$ and $\widehat{SE}(\hat{\beta})$?

They will be different.

$$\widehat{SE}(\hat{\tau}) = \sqrt{\frac{\hat{\sigma}_T^2}{m} + \frac{\hat{\sigma}_C^2}{n-m}}$$

where m is the number of units in the treatment group. This measure corrects for heteroscedasticity.

On the other hand, $\widehat{SE}(\hat{\beta})$ is the (2,2) element of the Covariance matrix $\hat{\sigma}(X'X)^{-1}$, which assumes homoscedasticity.

Advanced Regression

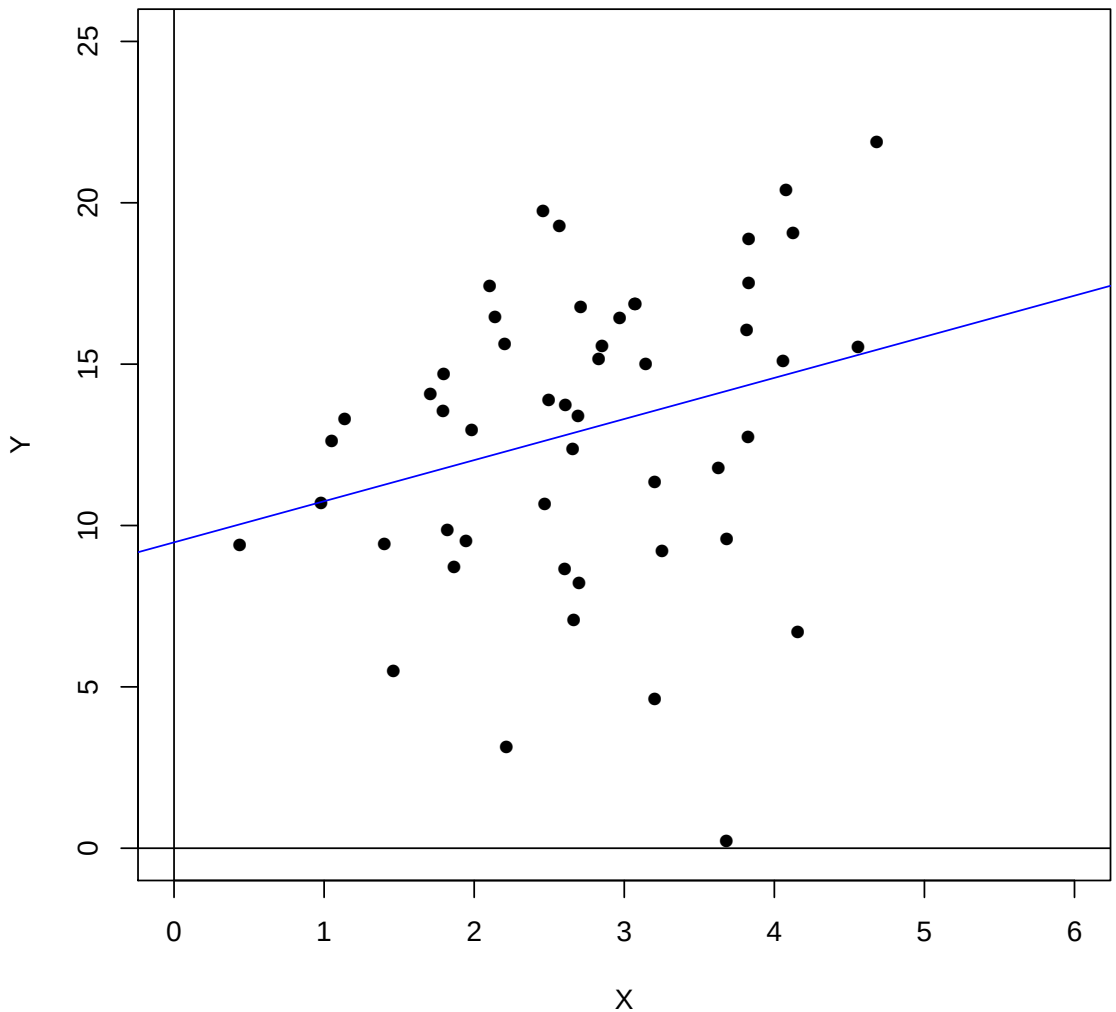
Pop Quiz (Continued)

Question: Assume that

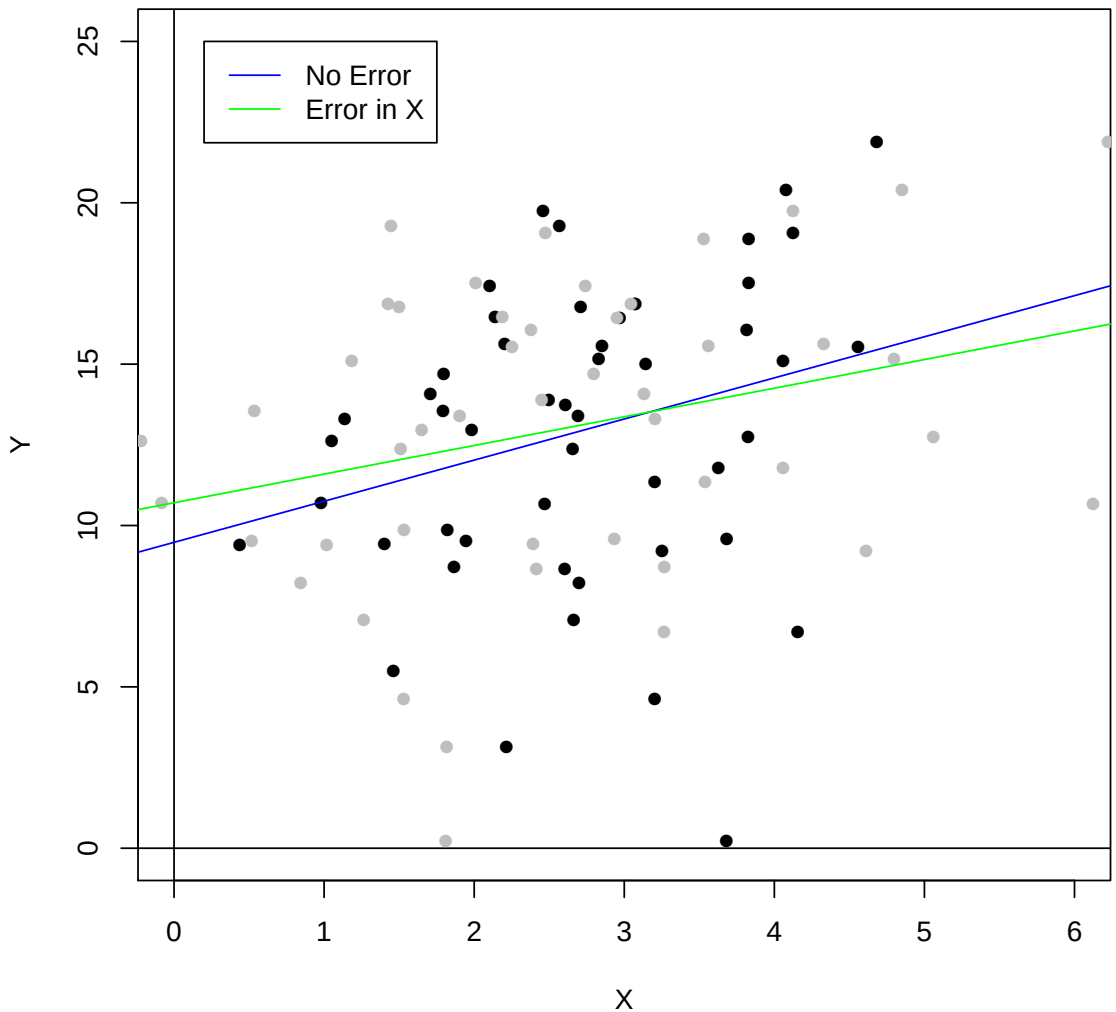
$$y = \alpha + \beta x + \epsilon$$

but the x's are measured with some error. Specifically, each x has the additional error term that is drawn from $N(0, \omega^2)$. Is $\hat{\beta}$ still unbiased? If not, will the bias cause you to underestimate or overestimate β?

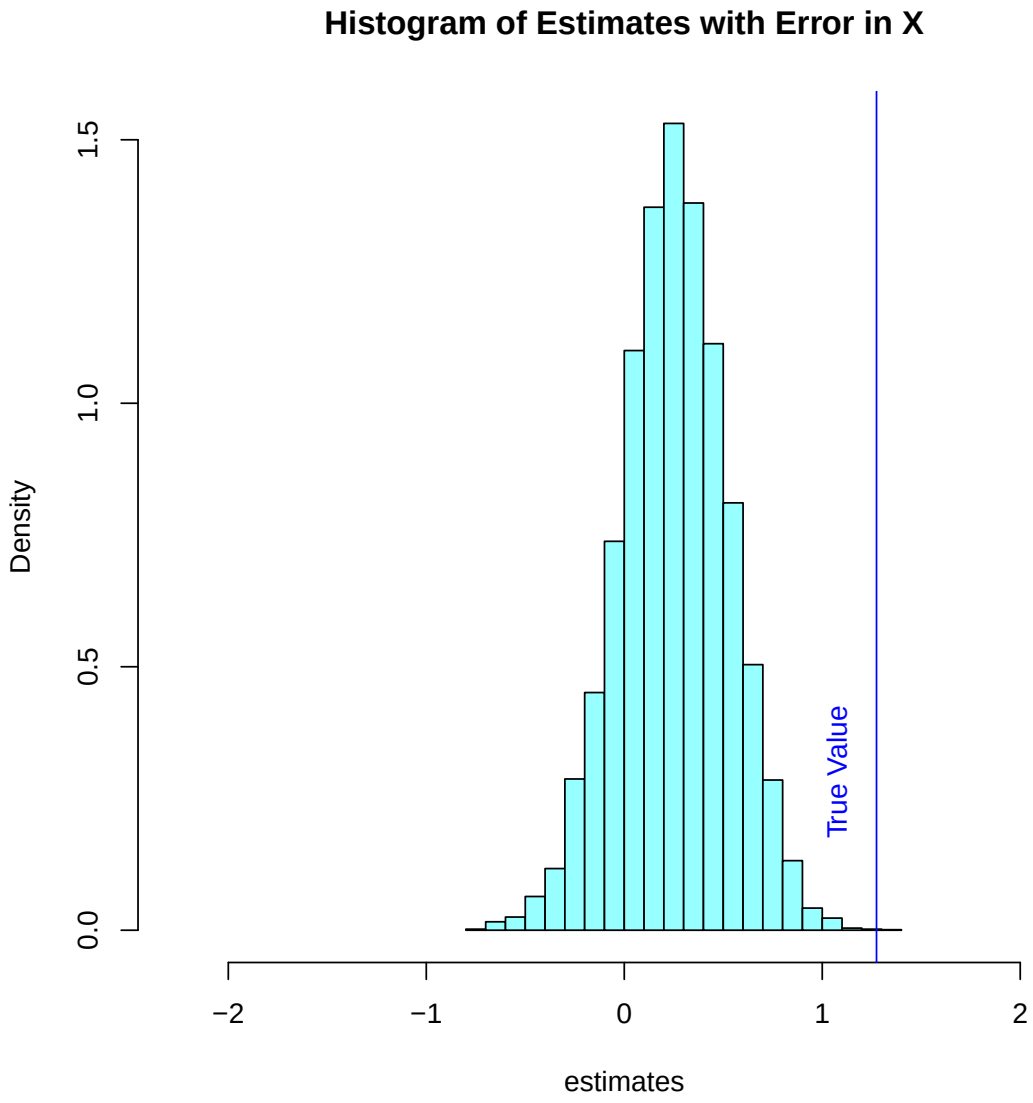
Advanced Regression



Regression



Advanced Regression



Advanced Regression

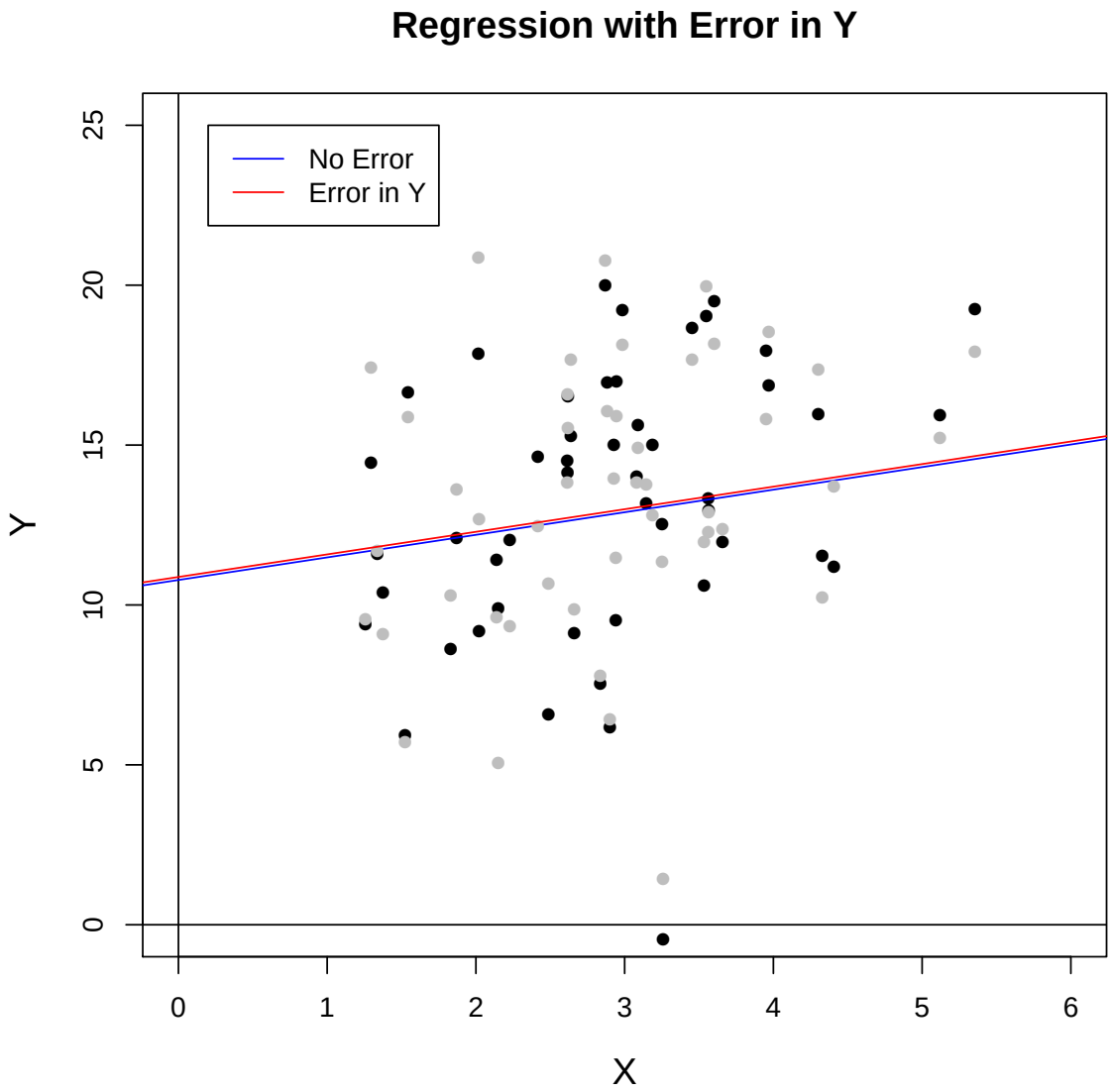
Pop Quiz (Continued)

Question: Now assume that

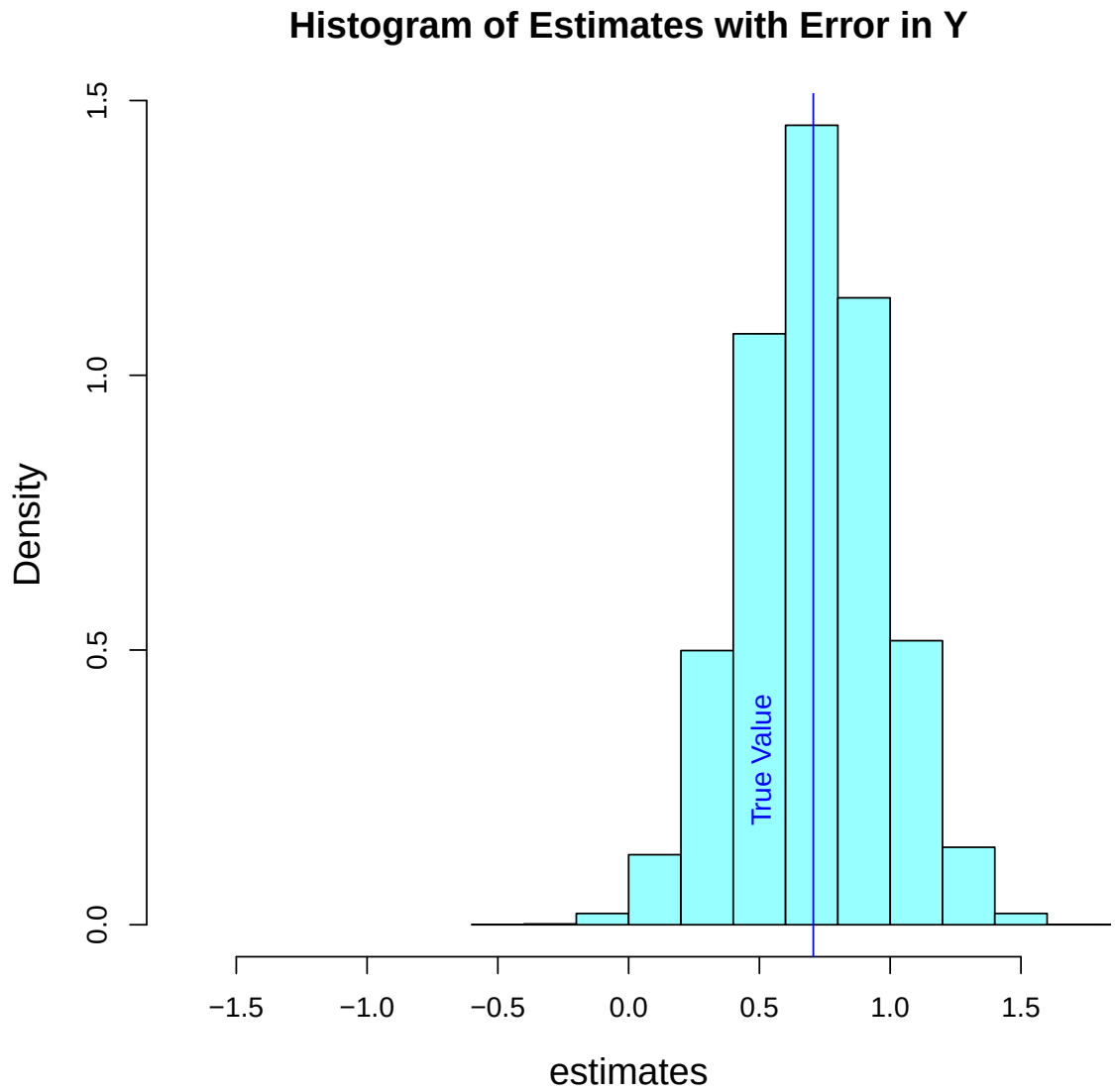
$$y = \alpha + \beta x + \epsilon$$

but the y's are measured with some error that is distributed $N(0, \omega^2)$. How will this error effect $\hat{\beta}$?

Advanced Regression



Advanced Regression



Advanced Regression

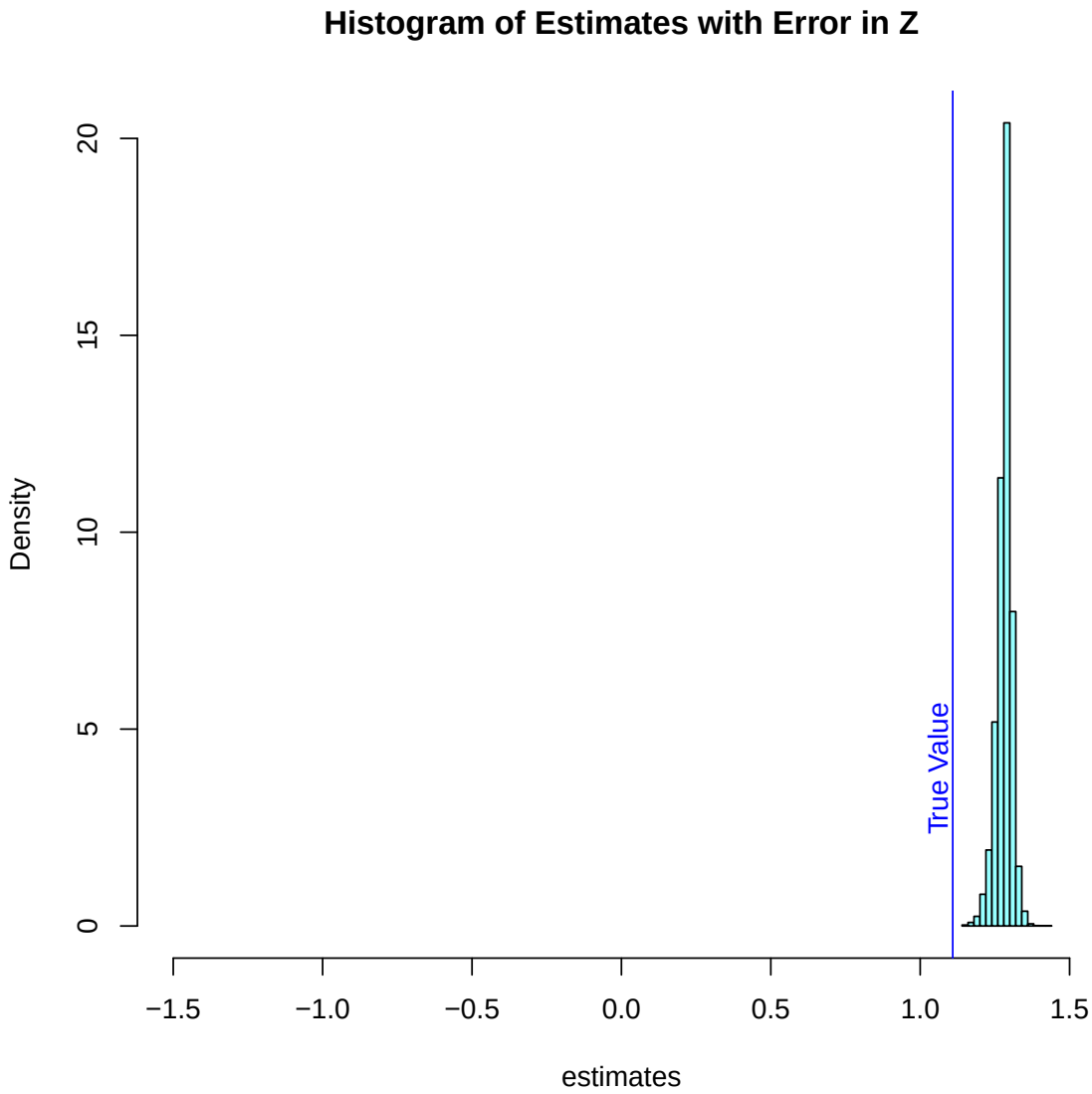
Pop Quiz (Continued)

Question: Now assume that we are using the model

$$y = \alpha + \beta_1 x + \beta_2 z + \epsilon$$

where z is some control variable. However, the z's are measured with some error that is distributed $N(0, \omega^2)$. How will this error effect $\hat{\beta}_1$?

Advanced Regression



Advanced Regression

The Hat Matrix

The hat matrix (or projection matrix) is

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

It is called that hat matrix because it puts a hat on Y

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

Note that

$\epsilon = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y} = \tilde{\mathbf{H}}\mathbf{Y}.$

So $\mathbf{HY} = \hat{\mathbf{Y}}$ gives us the part of $\mathbf{Y}$ that is explained by $\mathbf{X}$ and $\tilde{\mathbf{H}}\mathbf{Y} = \epsilon$ gives us the part of $\mathbf{Y}$ that is not explained by $\mathbf{X}$

Advanced Regression

Properties of the Hat Matrix

1. $\mathbf{H}$ is symmetric, and so is $\tilde{\mathbf{H}}$
2. $\mathbf{H}$ is idempotent ($\mathbf{H}^2 = \mathbf{H}$), and so is $\tilde{\mathbf{H}}$
3. $\mathbf{HX} = \mathbf{X}$
3. $\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{HY} \perp \mathbf{X}$

Reading Questions

The following equation is presented in the Winship and Morgan reading (p. 667)

$$\bar{Y}^{t}_{i\in T} - \bar{Y}^{c}_{i\in C} = \bar{\delta} + (\bar{Y}^{c}_{i\in T} - \bar{Y}^{c}_{i\in C}) + (1-\pi)(\bar{\delta}_{i\in T} - \bar{\delta}_{i\in C})$$

In Jas's Notation

$$E_S[Y_{it}|T_i = 1] - E_S[Y_{ic}|T_i = 0] = \bar{\tau} + (E_S[Y_{ic}|T_i = 1] - E_S[Y_ic|T_i = 0]) + \frac{\sum_i T_i}{n}(E_S[\tau_i|T_i = 1] - E_S[\tau_i|T_i = 0])$$

Reading Questions

The following equation is presented in the Winship and Morgan reading (p. 667)

$$\bar{Y}^t_{i\in T} - \bar{Y}^c_{i\in C} = \bar{\delta} + (\bar{Y}^c_{i\in T} - \bar{Y}^c_{i\in C}) + (1-\pi)(\bar{\delta}_{i\in T} - \bar{\delta}_{i\in C})$$

In actual words

$\bar{Y}^t_{i\in T} - \bar{Y}^c_{i\in C}$ = Observed Outcome for Treated minus Observed Outcome for Control

$\bar{\delta}$ = Average Treatment Effect

$\bar{Y}^t_{i\in T} - \bar{Y}^c_{i\in C}$ = The Selection Effect

$1-\pi$ = Proportion of Units Treated

$\bar{\delta}_{i\in T} - \bar{\delta}_{i\in C}$ = Difference in Average Treatment Effect between Treated and Controls

Reading Questions

Example: Say we want to estimate the effect of college education on mental ability. We compare a group of students who went to college to a group of students who did not, and we find that the group that went has a much higher mental ability. There are 3 potential sources of this difference:

1. The average treatment effect of going to college ($\bar{\delta}$)

2. The selection effect of smarter people going to college ($\bar{Y}^t_{i \in T} - \bar{Y}^c_{i \in C}$)

3. A difference between the average treatment effect across the sample (ATE) and the average treatment effect for the treated (ATT) ($\bar{Y}^t_{i \in T} - \bar{Y}^c_{i \in C}$)

Reading Questions

Imagine a world where there was no average treatment effect of going to college and people who went to college were on average just as smart as those who did not ($\bar{\delta} = 0$ and $\bar{Y}^t_{i \in T} - \bar{Y}^c_{i \in C} = 0$). We might still see a very large treatment effect. For example, imagine that

1. $n/2$ people in our sample would work hard in college and gain 10 points in their mental ability scores.

2. The other $n/2$ people would just party, and their mental ability scores would drop 10 points.

3. The group in (1) has the same average mental ability as the group in (2) if neither goes to college

4. Only group (1) goes to college

(1) and (2) imply that there is no average treatment effect. (3) means that there is no selection effect. But we will estimate a large average treatment effect (wrongly), simply because the $ATT \neq ATE$.

Reading Questions

Question: In Jas's article, he talks about how a treatment can change the conditional probability that an event will occur. This sounds inconsistent with the potential outcomes framework, where potential outcomes are considered fixed, at least for estimating the sample average treatment effect. This seems to imply a deterministic view of history, which may or may not be right. How would you rethink the potential outcomes framework if the potential outcomes are random variables?

Applications

In 1972, the Stanford Research Institute (SRI) was asked to evaluate schools in the Bay Area. There were 20 special classes set up for minority students. SRI was tasked with determining whether these new classes were fundamentally different than a set of control classes. Their evaluation was to be based on their standardized test scores.

After getting the data, SRI found that the new classes had a mean score of 78, with a standard deviation of 4.2. In comparison, the old classes had a mean score of 60.

They estimated the standard error of their estimator as $4.2/\sqrt{20} \approx 1$. Then they calculated their z-score as $(78 - 60)/1 = 18$. The resulting p-value is very close to 0.

Applications

Problem: They calculate a standard error, but where is the error in their model coming from?

1. It cannot be uncertainty in the experimental sense, since treatment was not randomized.

2. It cannot be sampling uncertainty, since we have all the data.

3. It cannot be uncertainty from measurement error. We have the correct class averages.

4. If they claim that the uncertainty comes from the fact that the tests scores are realizations of random variables, then what makes them think that these random variables are i.i.d. They will likely differ by school, location, and many other factors.

Bottom Line: Even if there is some uncertainty in the data generating process, you cannot determine the correct standard errors without an accurate chance model of the data.

Applications

A researcher wants to determine if there is discrimination against women in the workforce. He uses the model

$$\text{Salary} = a + b \cdot \text{Education} + c \cdot \text{Experience} + d \cdot \text{Man} + \epsilon$$

Where man is a dummy variable that is coded 1 if the person is a man and 0 if the person is a woman.

After getting the data, he estimates the coefficients as follows.

$$\text{Salary} = \$7,100 + \$1,300 \cdot \text{Edu} + \$2,200 \cdot \text{Exp} + \$700 \cdot \text{Man} + \epsilon$$

To test for discrimination, the researchers checks if the coefficient on Man is statistically significant.

Applications

Problems

1. There is no clear counterfactual, so it is hard to know what to control for. Should you control for education, or is education is part of the mechanism (because men more likely to go to college)?

2. For the standard errors to be right, the model needs to be right. This means

a) The effect of every extra year of education should be constant, regardless of whether it's from 2nd to 3rd grade or from 11th to 12th grade.

b) Likewise, the effect of every extra year of experience should be constant.

c) The ϵ_i must be drawn i.i.d. $N(0, \sigma^2)$

3. A better test for discrimination would be to randomize male and female names on resumes and see if the resumes with male names receive more interviews.