Section 1 : Regression Review

Yotam Shem-Tov

Fall 2014

Contact information

- Yotam Shem-Tov, PhD student in economics
- E-mail: shemtov@berkeley.edu
- Office hours: Wednesday 2-4

There are two general approaches to regression

1. Regression as a model: a data generating process (DGP)
2. Regression as an algorithm, i.e as a predictive model

This two approaches are different, and make different assumptions

Regression as a prediction

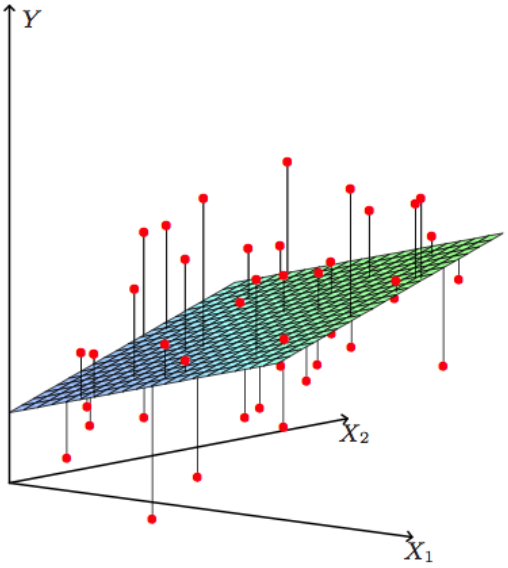
- We have an input vector $X^T = (X_1, X_2, \ldots, X_p)$ with dimensions of $n \times p$ and an output vector Y with dimensions $n \times 1$.
- The linear regression model has the form:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

- We can pick the coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ in a variety of ways but OLS is by far the most common, which minimizes the **residual sum of squares** (RSS):

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} x_{ij} \beta_j)^2 \end{aligned}$$

Regression as a prediction



Regression as a prediction: Deriving the Algorithm

- Denote $\mathbf{X}$ the $N \times (p+1)$ matrix with each row an input vector (with a 1 in the first position) and $\mathbf{y}$ is the output vector.
- Write the RSS as:

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

- Differentiate with respect to β:

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

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

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

- Solve for β:

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

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

- The matrix X is also referred as the design matrix

Regression as a prediction: Making a Prediction

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

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

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

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

- We can now write

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

- If $\mathbf{HY}$ yields part of $\mathbf{Y}$ that projects into $\mathbf{X}$, this means that $\tilde{\mathbf{H}}\mathbf{Y}$ is the part of $\mathbf{Y}$ that does not project into $\mathbf{X}$, which is the *residual* part of $\mathbf{Y}$. Therefore, $\tilde{\mathbf{H}}\mathbf{Y}$ makes the residuals
- $\mathbf{e}$ is the part of $\mathbf{Y}$ which is not a linear combination of X

Regression as a prediction: Deriving the Algorithm

- 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 that 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
- 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 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
- Is Y a linear function of the covariates? *No*, it is a linear function of β
- What are the classic assumptions of the regression model?

Regression as a model: 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, $Y = X\beta + \epsilon$
2. The independent variables, X, are fixed, i.e are not random variables (this can be relaxed to $Cov(X, \epsilon) = 0$)
3. The conditional mean of the error term is zero, $\mathbb{E}(\epsilon|X) = 0$
4. Homoscedasticity. The error term has a constant variance, i.e $\mathbb{V}(\epsilon_i) = \sigma^2$
5. The error terms are uncorrelated with each other, $Cov(\epsilon_i, \epsilon_j) = 0$
6. The design matrix, X, has full rank
7. The error term is normally distributed, i.e $\epsilon \sim N(0, \sigma^2)$ (the mean and variance follows from (3) and (4))

Discussion of 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
  The assumption can be relaxed for sufficiently large sample sizes, as by the CLT, $\hat{\beta}_{OLS}$ converges to a normal distribution when $N \rightarrow \infty$. What is a sufficiently large sample size?

Properties of the OLS estimators: Unbiased estimator

The OLS estimator of β is,

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

We know that $\hat{\beta}$ is unbiased if $E(\hat{\beta}) = \beta$

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

- What assumptions are used for the proof that $\hat{\beta}_{OLS}$ is an unbiased estimator?
  Assumption (1), the model is correct.
  Assumption (2), the covariates are independent of the error term

Properties of the OLS estimators: The variance of $\hat{\beta}_{OLS}$

- Recall:

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

- Plugging this into the covariance equation:

$$\begin{aligned} cov(\hat{\beta}|X) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X] \\ &= E\left[\left((X^TX)^{-1}X^T\epsilon\right)\left((X^TX)^{-1}X^T\epsilon\right)'|X\right] \\ &= E[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}|X] \\ &= (X^TX)^{-1}X^TE(\epsilon\epsilon^T|X)X(X^TX)^{-1} \\ &\qquad \text{where } E(\epsilon\epsilon^T|X) = \sigma^2 I_{p\times p} \\ &= (X^TX)^{-1}X^T\sigma^2 I_{p\times p}X(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1} \end{aligned}$$

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:

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

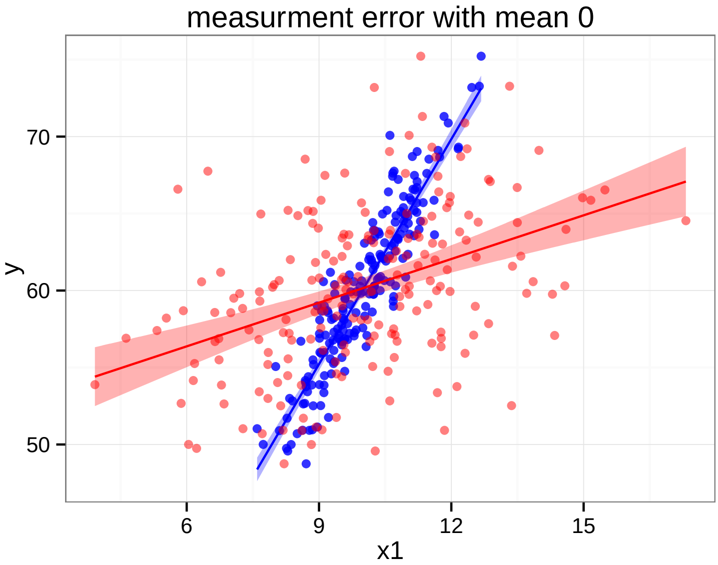
Is one estimator always preferable over the other? If not when each estimator is preferable?

Consider the following DGP (data generating process):

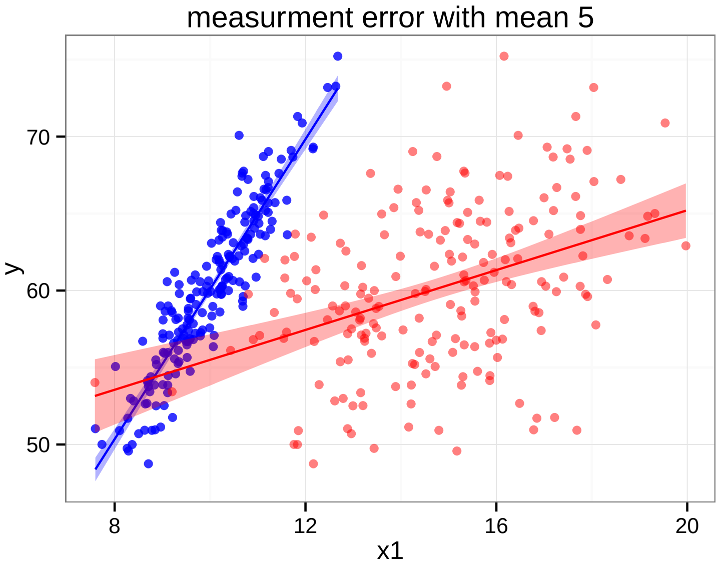
```
n=200
x1 = rnorm(n,mean=10,1)
epsilon = rnorm(n,0,2)
y = 10+5*x1+epsilon

### 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, $\beta_1 = 5$

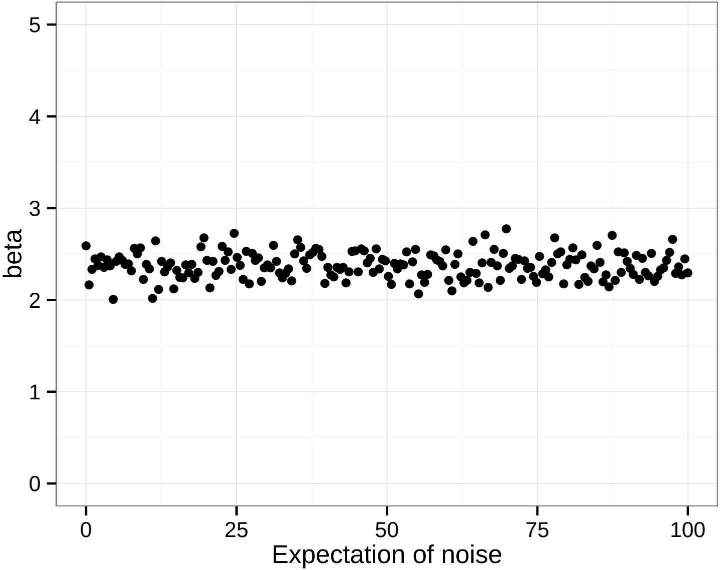
Measurement error: $noise \sim N(\mu = 0, \sigma = 2)$



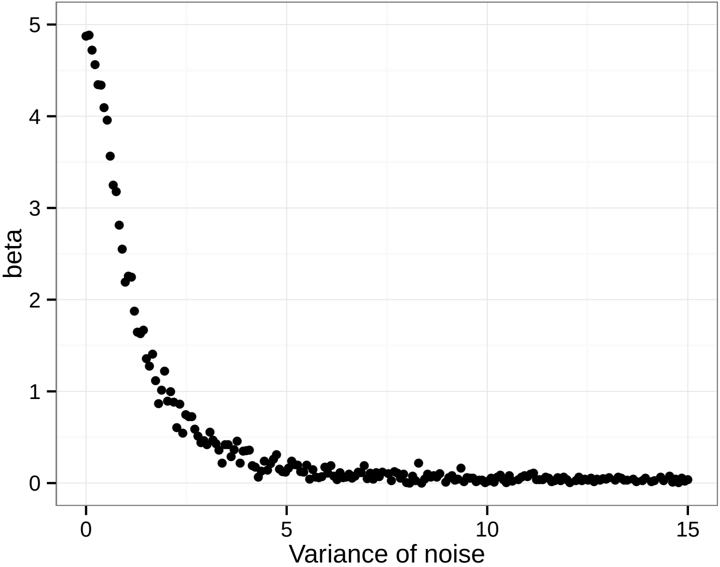
Measurement error: $noise \sim N(\mu = 5, \sigma = 2)$



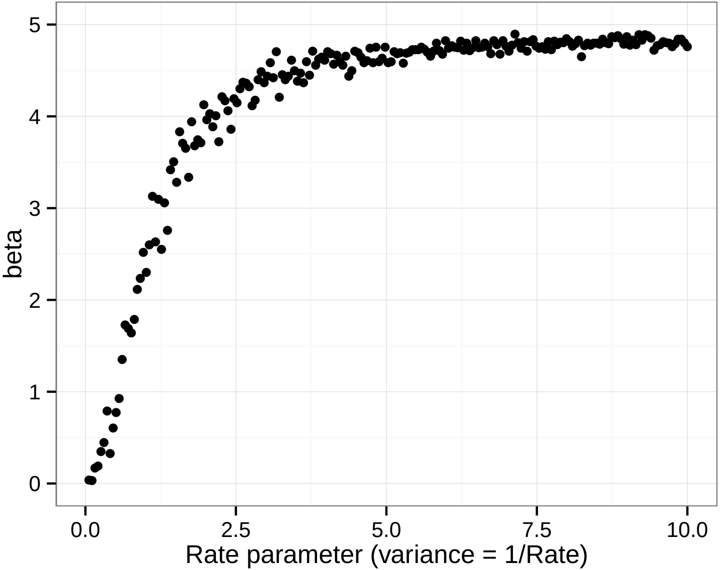
Measurement error: *noise* $\sim N(\mu = ?, \sigma = 2)$



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.

Measurment 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:

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

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

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

Measurment error in the dependent variable

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

Gauss-Markov theorem: BLUE

- The regression estimator is a linear estimator, $\hat{\beta} = Cy$, where $C = (X^T X)^{-1} X^T$. 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: (2),(3),(4),(5) hold. The regression estimator is the best linear unbiased estimator (BLUE), in terms of MSE (Mean Squared Error)

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:

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

  where $\tilde{X}_{ij}$ is the residual from the regression of X_{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?

- 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? $1 \times p$
- $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $x_i^T = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$
- $x_i^T x_i = \begin{pmatrix} x_{i1}^2 & x_{i1}x_{i2} & \ldots & x_{i1}x_{ip} \\ x_{i2}x_{i1} & x_{i2}^2 & \ldots & x_{i2}x_{ip} \\ \vdots & \vdots & \vdots & \\ x_{ip}x_{i1} & x_{ip}x_{i2} & \ldots & x_{ip}^2 \end{pmatrix}$

- Verify at home that,

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

- Hence, $X^TX = \sum_{i=1}^n x_i^T x_i$
- Note (and verify at home),

$$X^Ty = \begin{pmatrix} \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}y_i \\ \vdots \\ \sum_{i=1}^n x_{ip}y_i \end{pmatrix} = \sum_{i=1}^n x_i^T y_i$$

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

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

$$\rightarrow \mathbb{E}\left(x_i^T x_i\right)^{-1} \cdot \mathbb{E}\left(x_i^T y_i\right) = \mathbb{E}\left(x_i^T x_i\right)^{-1} \cdot \mathbb{E}\left(x_i^T \left(x_i \beta + \epsilon_i\right)\right)$$

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

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

Regression in Causal Analysis

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

$$Y_i = \alpha + \beta_1 \cdot T_i + \mathbf{X}_i^T \cdot \beta_2 + \epsilon_i$$

  where T_i is an indicator variable for treatment status and $\mathbf{X}_i$ is a vector of *pre-treatment characteristics*
- Under this model, what is random?
- How do we interpret the coefficient β_1?