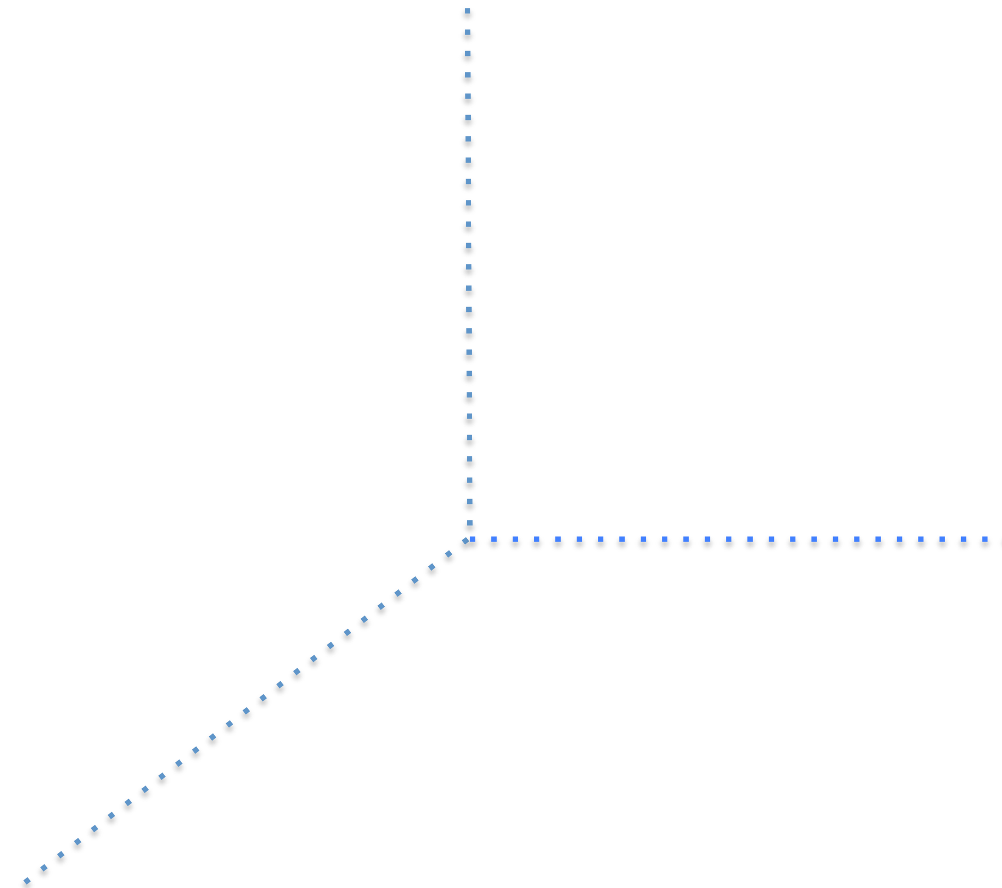
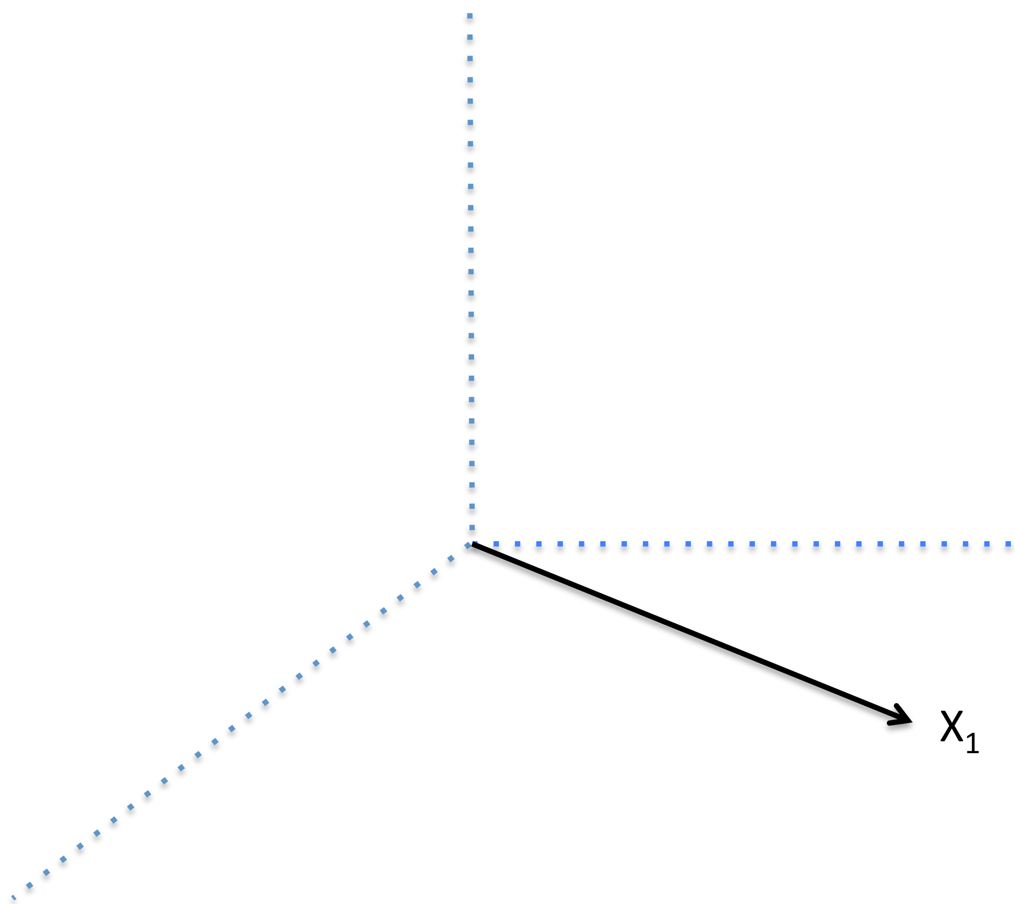
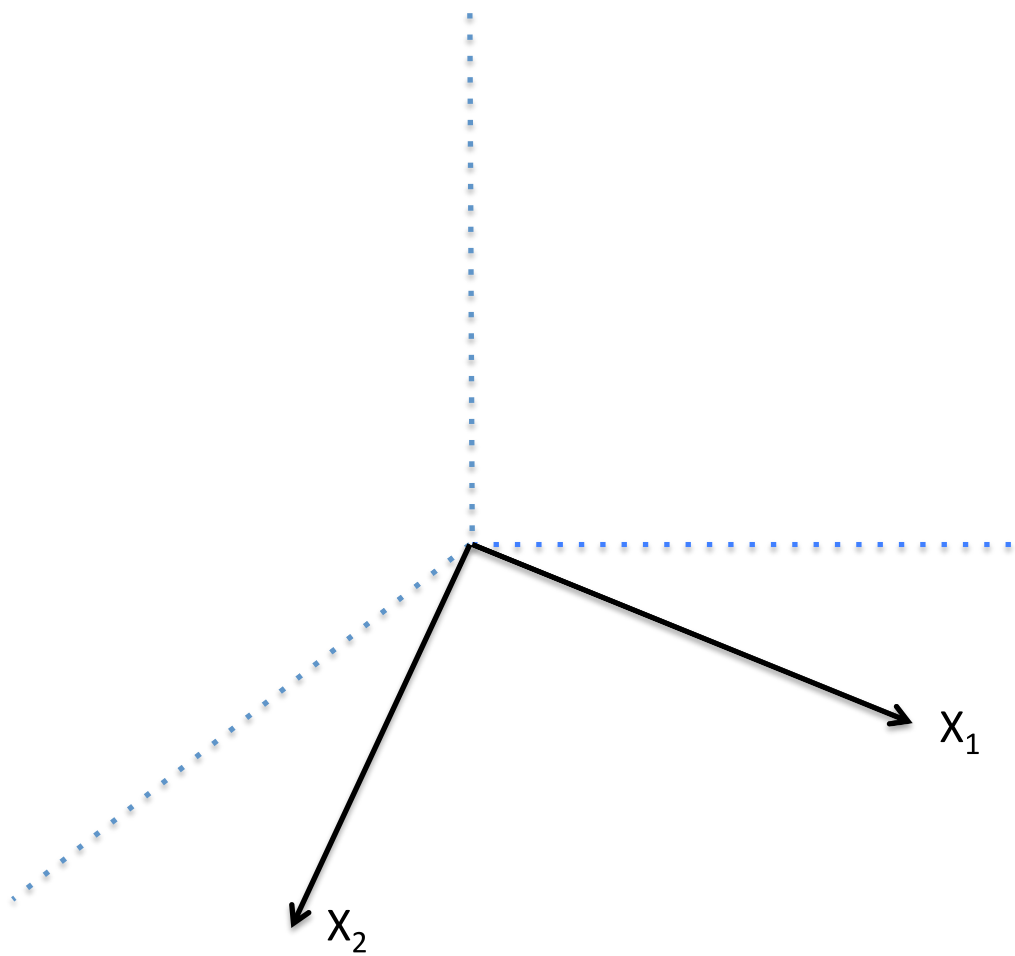
Review of Linear Least Squares

- Let's talk about regression as a deterministic mathematical operation
- Ingredients of a linear regression:
  - y regressand
  - $X = [x_1, ..., x_k]$ regressors
- y and $x_1, ..., x_k$ can be thought of as vectors in N dimensional Euclidean space E^N



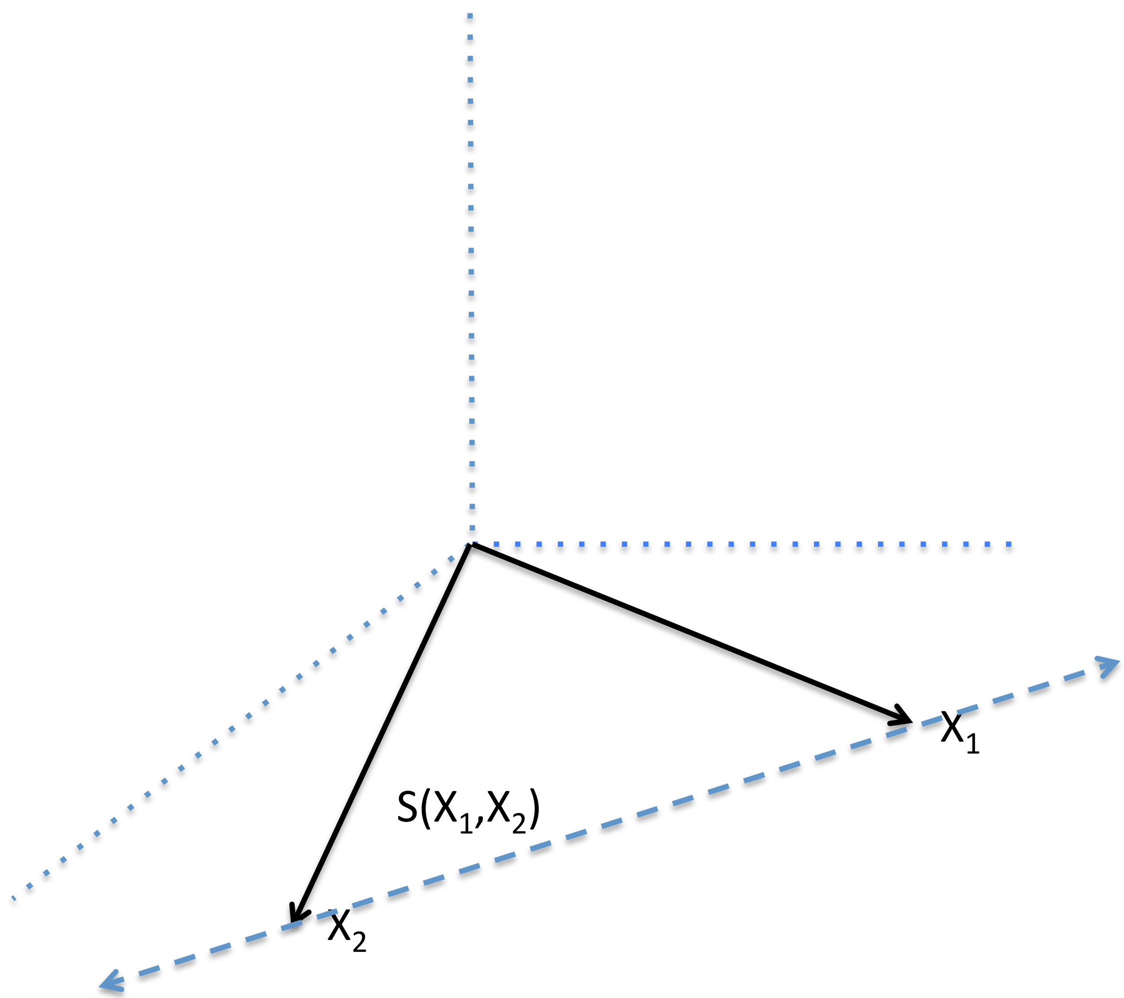

X_1



X_1
X_2

Review of Linear Least Squares

- Let's talk about regression as a deterministic mathematical operation
- Ingredients of a linear regression:
  - y regressand
  - $X = [x_1, ..., x_k]$ regressors
- y and $x_1, ..., x_k$ can be thought of as vectors in N dimensional Euclidean space E^N
- $S(X)$, or the span of X is the set of all vectors z in E^N such that $z = X\gamma$ for some γ.



X_1
$S(X_1,X_2)$
X_2

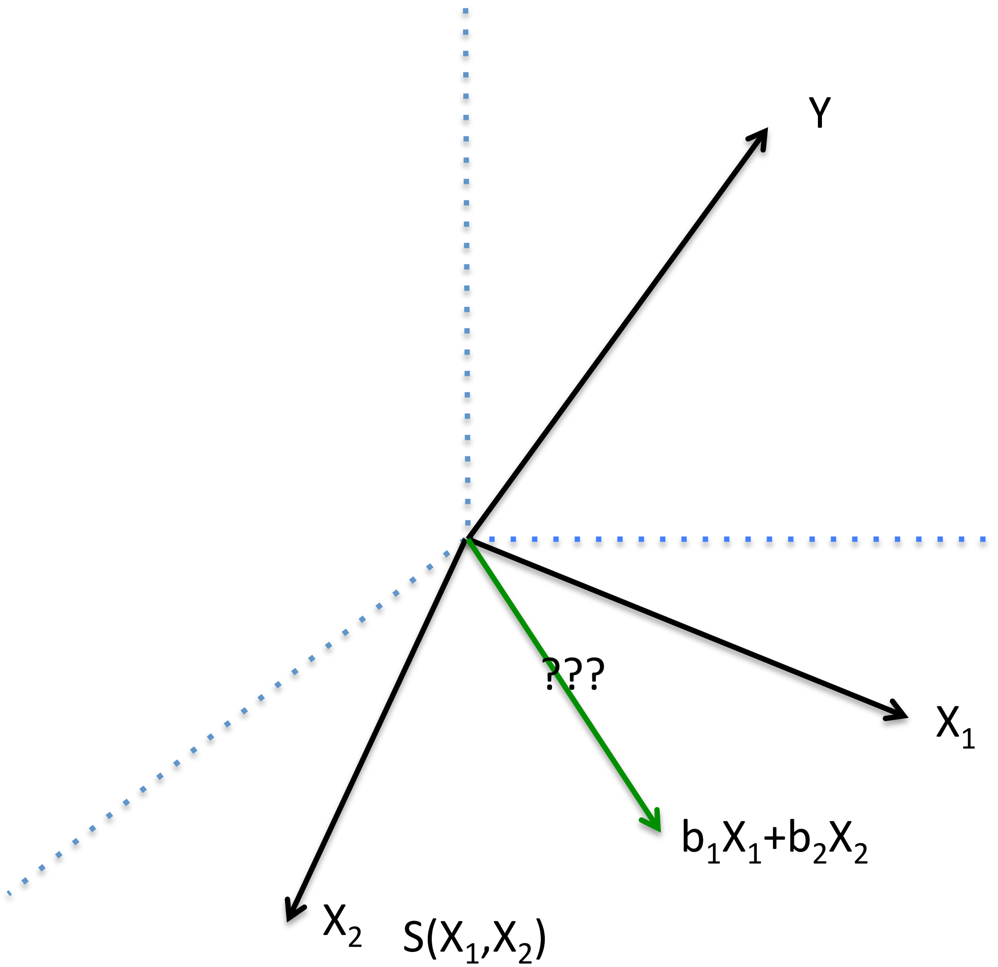
Review of Linear Least Squares

- Let's talk about regression as a deterministic mathematical operation
- Ingredients of a linear regression:
  - y regressand
  - $X = [x_1, ..., x_k]$ regressors
- y and $x_1, ..., x_k$ can be thought of as vectors in N dimensional Euclidean space E^N
- $S(X)$, or the span of X is the set of all vectors z in E^N such that $z = X\gamma$ for some γ.
- $dim(S(X)) = rank(X)$

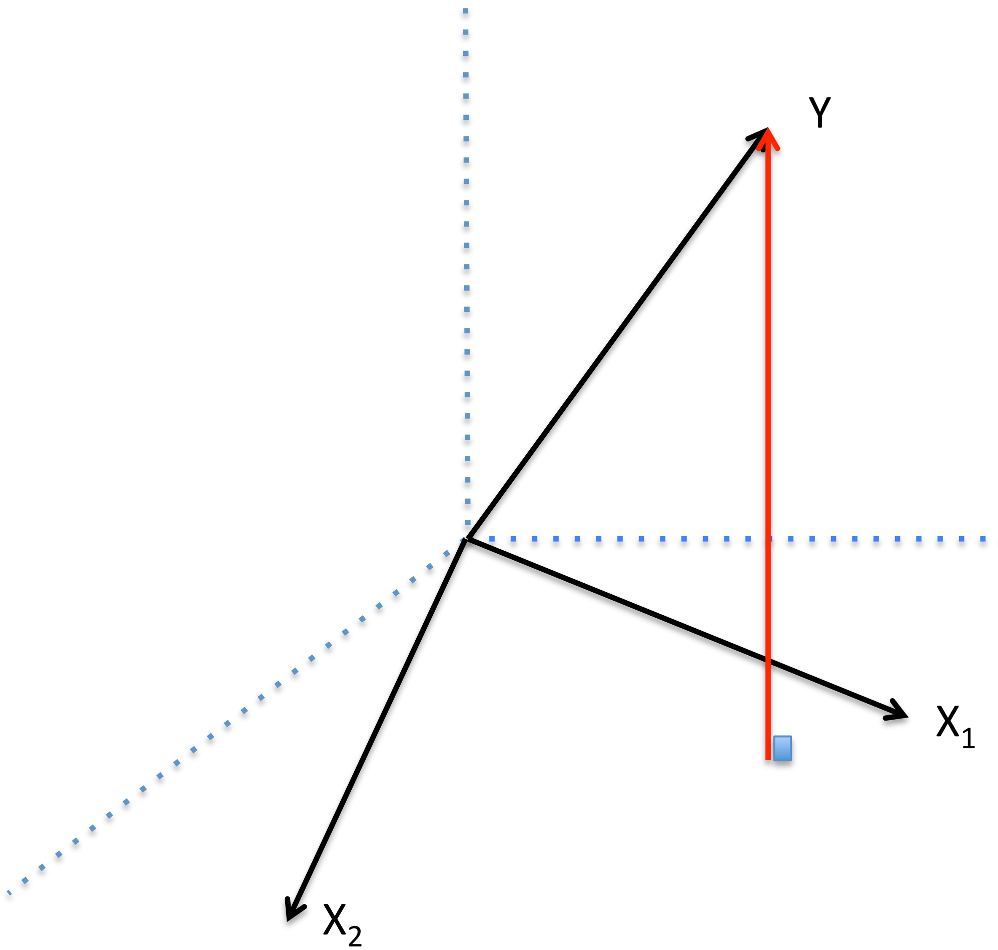
Review of Linear Least Squares

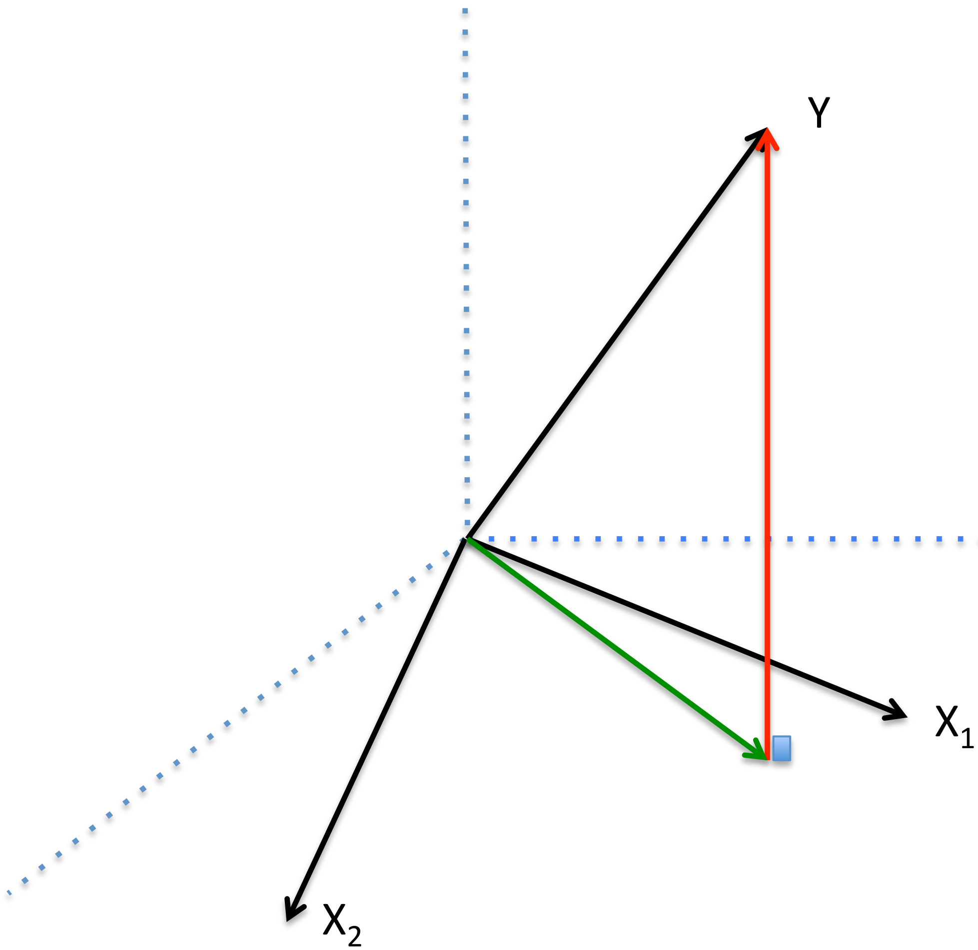
- Let's talk about regression as a deterministic mathematical operation
- Ingredients of a linear regression:
  - y regressand
  - $X = [x_1, ..., x_k]$ regressors
- y and $x_1, ..., x_k$ can be thought of as vectors in N dimensional Euclidean space E^N
- $S(X)$, or the span of X is the set of all vectors z in E^N such that $z = X\gamma$ for some γ.
- $dim(S(X)) = rank(X)$
- Define $S^o(X)$ as the orthogonal complement of $S(X)$, which is the set of all points w in E^N such that $w'z = 0$ for $z \in S(X)$.

- Problem: Given y, find the point/vector in $S(X)$ that is closest to y in the Euclidean norm.


Y
???
X$_1$
b$_1$X$_1$+b$_2$X$_2$
X$_2$
S(X$_1$,X$_2$)

- Problem: Given y, find the point/vector in $S(X)$ that is closest to y in the Euclidean norm.
- Equivalent to: $\min_b (y - Xb)'(y - Xb)$.


Y
X_1
X_2


Y
X_1
X_2

- Math Problem: $\min_b (y - Xb)'(y - Xb)$
- Let $\hat{b}$ be the minimizer. $y - X\hat{b} \in S^o(X)$.
- Which is equivalent to $X'(y - X\hat{b}) = 0$, first-order conditions of OLS.
- Solving, we get:

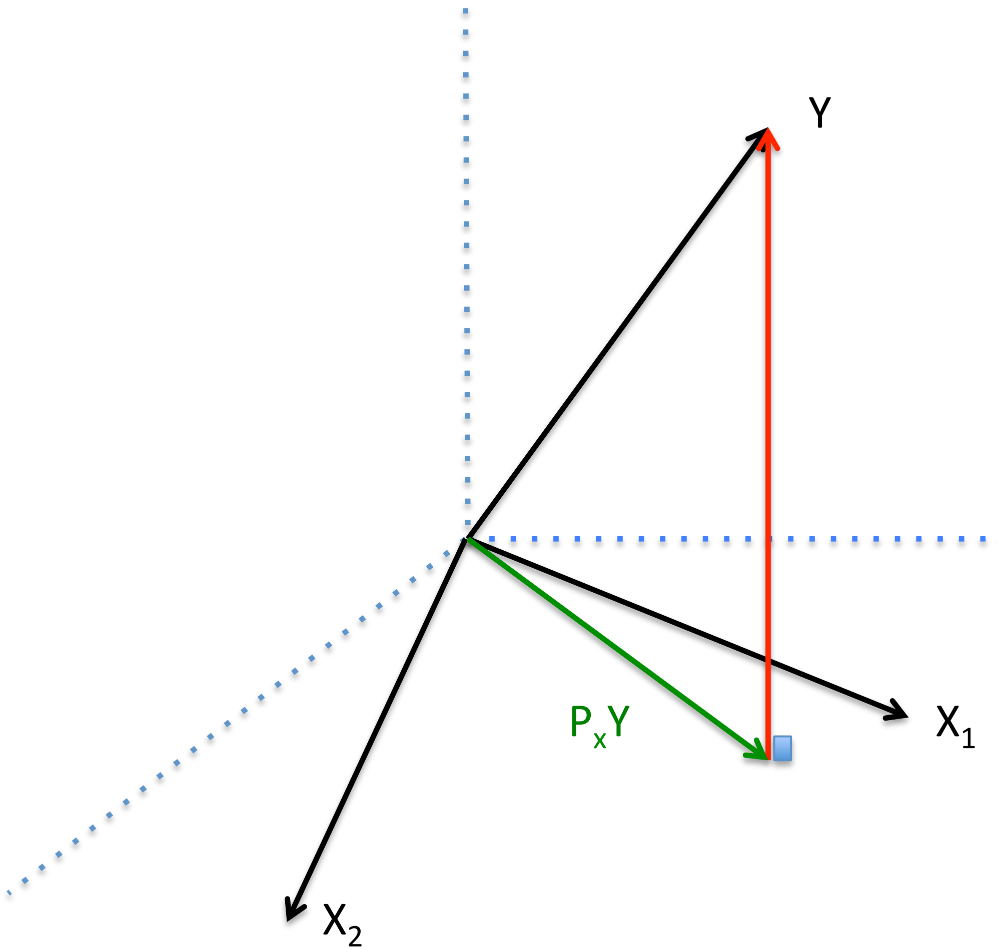
$$\hat{b} = (X'X)^{-1}X'y$$

- if $X'X$ is full rank (for $N > 3$).

- Look at the representation of y in $S(X)$, the "predicted" y
- $X\hat{b} = X(X'X)^{-1}X'y = P_X y$

$$P_X = X(X'X)^{-1}X'$$

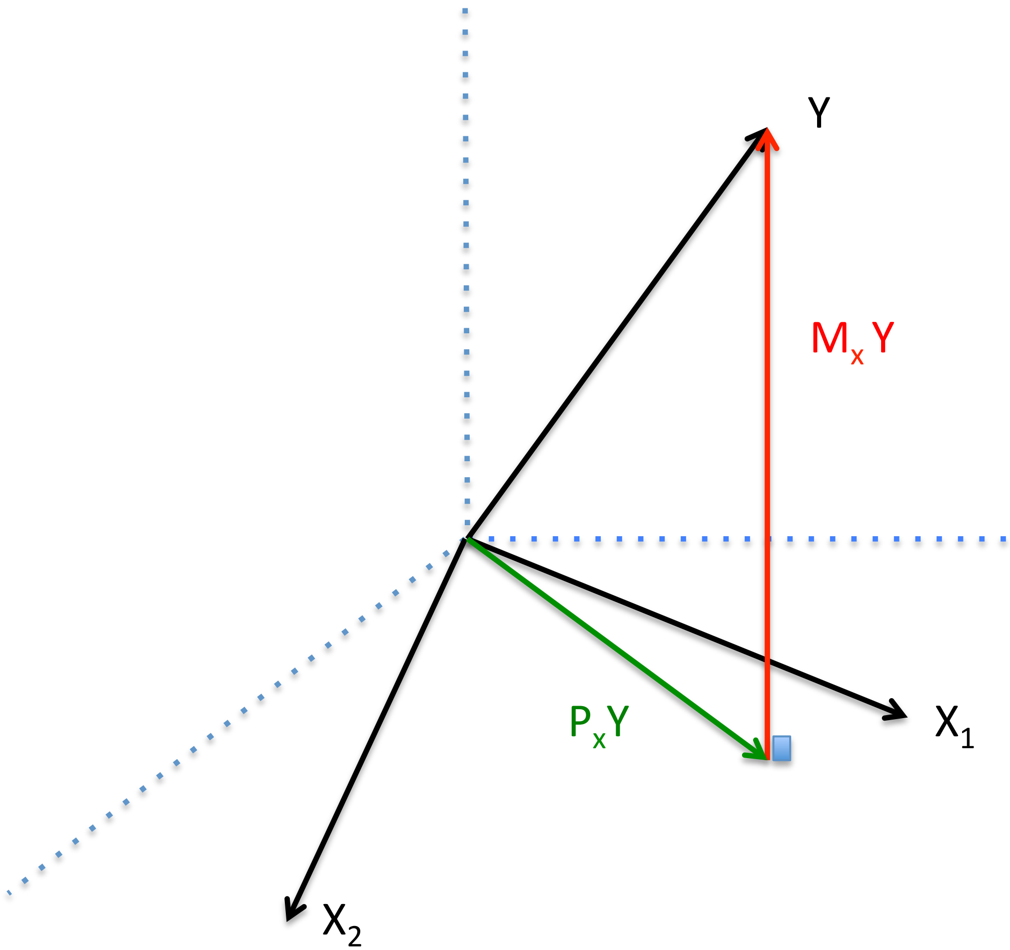
- This is the *projection* matrix that maps y onto $S(X)$.


Y
P_xY
X_1
X_2

- Now consider $y - X\hat{b}$, the "prediction error"
- $Y - X\hat{b} = (I - P_X)y$, with

$$M_X = I - P_X = I - X(X'X)^{-1}X'$$

- M_X projects y onto $S^o(X)$
- $y = M_X y + P_X y$, its *orthogonal decomposition*
- Note: $P_X P_X = P_X$, $M_X M_X = M_X$ (Why?)
- Also: $P_X M_X = 0$ (Why?)


Y
$M_x Y$
$P_x Y$
X_1
X_2

- In the regression context, $P_X y$ is the vector of fitted values from the regression
- $M_X y$ is the vector of regression residuals.
- Residuals are orthogonal to fitted values $(P_X y)'(M_X y) = 0$.

Long vs Short Regressions

- Suppose we run regression 1:

$$y = X_1\beta_1 + X_2\beta_2 + e_1$$

- Then we run regression 2:

$$M_1 y = M_1 X_2 \beta_2 + e_2$$

where $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$.

Frisch-Waugh-Lovell Theorem

- Claim 1: β_2 is numerically the same across the two regressions.
- Proof: Premultiply Reg 1 by $X_2'M_1$.
- This gives $X_2'M_1y = X_2'M_1X_2\beta_2$ (Why?)
- Solve for $\beta_2 = (X_2'M_1X_2)^{-1}(X_2'M_1)y$ which is the formula for β_2 in the second regression

Frisch-Waugh-Lovell Theorem II

- Claim 2: The residuals from Reg 1 and Reg 2 are identical
- Proof: Premultiply Reg 2 by M_1.
- This gives $M_1 y = M_1 X_2 \beta_2 + M_1 M_X y$
- Which is $M_1 y = M_1 X_2 \beta_2 + M_X y$ (What is $M_1 M_X$?)
- i.e. the residual term in Reg 2 is the same as in Reg 2.

Applications I

Regression on a constant:

- What does it mean to regress on a constant?
- i is a column of ones. $P_i y = ?$

Applications I

Regression on a constant:

- What does it mean to regress on a constant?
- i is a column of ones. $P_i y = i(i'i)^{-1}i'y$

Applications I

Regression on a constant:

- What does it mean to regress on a constant?
- i is a column of ones. $P_i y = i(i'i)^{-1}i'y = \bar{y}$
- What is $M_i y$?

Applications II

Regression on dummy variables:

- $D_n = 1$ if the n-th patient receives the drug, $D_n = 0$ if not.
- D is column of zeros and ones. $P_D y = ?$

Applications III

- Seasonality/group correction: suppose we run regression 1:

$$y = X\beta + D\gamma + e_1$$

  where D is a set of dummies for the various seasons/groups.
- Suppose instead of this, we calculate seasonal/group averages for X and y, and subtract these from y and X. We then run the regression for seasonally corrected variables (regression 2).
- How are the βs different across regressions 1 and 2?

Applications IV

- A formula for individual regression coefficients:

$$\beta_k = (\tilde{X}_k' \tilde{X}_k)^{-1} \tilde{X}_k' Y$$

where $\tilde{X}_k = M_{-k} X_k$ is the residual of the regression of X_k on all other regressors (X_{-k}).

Applications IV

- A formula for individual regression coefficients:

$$\beta_k = (\tilde{X}_k' \tilde{X}_k)^{-1} \tilde{X}_k' Y$$

  where $\tilde{X}_k = M_{-k} X_k$ is the residual of the regression of X_k on all other regressors (X_{-k}).
- Since $\tilde{X}_k$ is a vector, the formula boils down to (if the set of other regressors includes a constant)

$$\beta_k = \frac{Cov(\tilde{X}_k, Y)}{Var(\tilde{X}_k)}$$

Applications V

- Omitted variable bias: suppose our true model is

$$Y = \beta_0 + \beta_1 T + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Applications V

- Omitted variable bias: suppose our true model is
$$Y = \beta_0 + \beta_1 T + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$
- Suppose we do not have info on X_3, so we run
$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 T + \tilde{\beta}_2 X_2 + \eta$$

Applications V

- Omitted variable bias: suppose our true model is

$$Y = \beta_0 + \beta_1 T + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

- Suppose we do not have info on X_3, so we run

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 T + \tilde{\beta}_2 X_2 + \eta$$

- We know that

$$\tilde{\beta}_1 = \frac{Cov(Y, \tilde{T})}{Var(\tilde{T})}$$

where $\tilde{T}$ is the residual from the regression of T on X_2.

Applications V

- Omitted variable bias: suppose our true model is

$$Y = \beta_0 + \beta_1 T + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

- Suppose we do not have info on X_3, so we run

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 T + \tilde{\beta}_2 X_2 + \eta$$

- We know that

$$\tilde{\beta}_1 = \frac{Cov(Y, \tilde{T})}{Var(\tilde{T})}$$

  where $\tilde{T}$ is the residual from the regression of T on X_2.
- So

$$Cov(Y, \tilde{T}) = \beta_1 Var(\tilde{T}) + \beta_3 Cov(X_3, \tilde{T})$$

  (why?) and

$$\tilde{\beta}_1 = \beta_1 + \beta_3 \delta_{31}$$

  where δ_{31} is the coefficient on T when X_3 is regressed on T and X_2 (why?)

- Suppose Y is wages, T is years of schooling and X_3 is (unobserved) ability.
- Which way do we think the coefficient in our schooling regression (without ability) is biased?

OLS when Y and X are random

- Let Y be a random outcome variable, and X be a random vector of regressors.
- Let $\hat{Y}$, a *predictor* of Y, be a function of X.
- With $e = Y - \hat{Y}$, the prediction error, we want to minimize the expected loss

$$\min_{\hat{y}} E[L((Y - \hat{Y})|X]$$

  where $L(.)$ is the loss function.
- If $L(e) = e^2$, then the predictor minimizing the expected loss is

$$\hat{Y}(x) = E[Y|X = x]$$

  the conditional expectation function.

OLS and conditional expectation

- Ex: (Y, X) jointly normal

$$
\begin{aligned}
E(Y|X=x) &= \left(\mu_y - \left(\frac{\sigma_{xy}}{\sigma_x^2}\right)\mu_x\right) + \left(\frac{\sigma_{xy}}{\sigma_x^2}\right)x \\
&= \beta_0 + \beta_1 x
\end{aligned}
$$

- What is the connection with the OLS estimate?
  - In reality, you see N realizations, (y_i, x_i) from (Y, X). You do not know $\mu_y, \mu_x, \Sigma_{xy}$.
  - However, you can form an *estimate* of $E(Y|X=x)$ by regressing Y on X and a constant
  - In the *jointly normal* case, OLS will give you a consistent estimator of the function $E(Y|X=x)$

OLS as best linear predictor of $E(Y|X)$

- More generally, if the conditional expectation function is linear in X, i.e. $E[Y|X=x] = x'\beta$, then $x'\hat{\beta}_{OLS}$ is a consistent estimate of $E[Y|X=x]$.
- If $E[Y|X=x]$ is nonlinear, then OLS no longer represents the conditional expectation.
- Still, OLS is the *best linear predictor* for the conditional expectation function.
- Why? Take:

$$\begin{aligned}
\beta_{CEF} &= arg\min_b E[(E(Y|X) - X'b)^2] \\
\beta_{OLS} &= arg\min_b E[(Y - X'b)^2]
\end{aligned}$$

- Here, β_{CEF} is the best linear predictor for $E(Y|X)$ under square loss.

OLS as best linear predictor of $E(Y|X)$

- But:

$$\begin{aligned} E(Y - X'b)^2 &= E[Y - E(Y|X) + E(Y|X) - X'b]^2 \\ &= E[Y - E(Y|X)]^2 + E[E(Y|X) - X'b]^2 + \\ &\quad 2E[Y - E(Y|X)][E(Y|X) - X'b] \\ &= const + E[E(Y|X) - X'b]^2 \end{aligned}$$

so β_{CEF} and β_{OLS} are maximizing the same thing (plus a constant)

Review of Asymptotic Theory

- In most econometric analyses, we treat our data as the result of a random sampling process
- For example, survey data is of the form $(y_i, X_i, i = 1, .., N)$ where the outcome variable and covariates for respondent i is drawn from the joint distribution of (y, X) according to a sampling process
- Given the data, we form *statistics* of the data; i.e. *functions* of $(y_i, X_i, i = 1, .., N)$.
- Ex 1: $\bar{y}_N = \frac{1}{N}\sum_{i=1}^{N} y_i$, sample average of the outcome variable
- Ex 2: $\hat{\beta}_{OLS} = (X'X)^{-1}X'y$, the OLS coefficients
- These are both functions of the underlying random variables y and X.

- If we know (or assume) something about the distribution of y and X, we can also say something about the distribution of such statistics.
- Ex: if y_i are iid $N(0, \sigma^2)$, we know that $\sqrt{N}\bar{y}_N$ has the exact *finite sample* distribution $N(0, \sigma^2)$.
- Many times, however, statistics are complicated function of underlying random variables.
- This makes it difficult to derive exact *finite sample distributions* of the statistics
- However, under certain conditions, we can approximate the distribution of complicated statistics by using *laws of large numbers* and *central limit theorems* from probability theory.
- Ex: if y_i are iid with mean zero and finite variance σ^2, but not necessarily normal, as N grows large, the distribution of $\sqrt{N}\bar{y}_N$ is well-approximated by $N(0, \sigma^2)$.

Probability Limit

- Let θ_N be a statistic (function) of $(y_i, X_i, i = 1, .., N)$.
- We are interested in the behavior of the sequence θ_N, as N grows large.
- The first thing we look at is the *probability limit* of θ_N
- Definition: A sequence of random variables $\{\theta_N\}$ converges in probability to θ if, for any $\varepsilon > 0$ and $\delta > 0$, there exists $N^* = N^*(\varepsilon, \delta)$ such that for all $N > N^*$,

$$\Pr(|\theta_N - \theta| < \varepsilon) > 1 - \delta$$

- If such a limit exists, we write that $\operatorname{plim}_{N\to\infty} \theta_N = \theta$, or $\theta_N \xrightarrow{\mathrm{P}} \theta$.

Slutsky Theorem

- A really useful feature of plim's is that they are preserved under continuous transformations.
- Slutsky Theorem: Let $\boldsymbol{\theta_N}$ be a finite dimensional vector of random variables and let $g()$ be a real valued function continuous at $\boldsymbol{\theta}$.
- Then: $\operatorname{plim} \boldsymbol{\theta_N} = \boldsymbol{\theta}$ implies $\operatorname{plim} g(\boldsymbol{\theta_N}) = g(\boldsymbol{\theta})$
- Ex: $\hat{\beta}_{OLS} = \frac{N^{-1}\sum_{i=1}^{N} x_i y_i}{N^{-1}\sum_{i=1}^{N} x_i^2}$
- If we can calculate the plim's of the numerator and denominator, we can get the plim of the ratio.

Laws of Large Numbers

- *Kolmogorov LLN:* Let $\{X_i\}$ be iid.
- If and only if $E[X_i] = \mu$ and $E[|X_i|] < \infty$:

$$\text{plim}\, \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{N}\sum_{i=1}^{N} E[X_i] = \mu.$$

- *Markov LLN:* Let $\{X_i\}$ be independent, but not necessarily identical, with $E[X_i] = \mu_i$ and $Var(X_i) = \sigma_i^2$.
- If $\sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} < \infty$,

$$\text{plim}\left(\frac{1}{N}\sum_{i=1}^{N} X_i - \frac{1}{N}\sum_{i=1}^{N} E[X_i]\right) = 0$$

- I.e. in the non-identical iid case, variance can grow with i, but not too fast.

Convergence in Distribution

- Since N is finite in applications, specifying the probability limit of an estimator is not enough.
- We also need the distribution of θ_N. However, this might be difficult to derive exactly in most cases.
- We thus resort to central limit theorems to approximate the distribution of θ_N, as N grows large.
- A sequence of random variables, $\{\theta_N\}$ *converges in distribution* to a random variable θ, if

$$\lim_{N \to \infty} \Pr(\theta_N < x) = \Pr(\theta < x)$$

  for every x. (Here, the limit is in the deterministic sense.)
- Note that convergence in probability implies converges in distribution (but not the other way around).

- Convergence in distribution is also preserved under continuous transformations (*continuous mapping theorem*):

$$\theta_N \xrightarrow{\mathrm{D}} \theta \implies g(\theta_N) \xrightarrow{\mathrm{D}} g(\theta)$$

Central Limit Theorems

- Now we can state the most useful limit results for convergence in distribution
- Let
$$Z_N = \frac{\bar{X}_N - E[\bar{X}_N]}{\sqrt{Var(\bar{X}_N)}}$$
where $\bar{X}_N = N^{-1}\sum_{i=1}^{N} X_i$, the sample mean.
- *Central Limit Theorem I:* Let X_i be iid with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$, then $Z_N \xrightarrow{\text{D}} N(0,1)$.
- *Central Limit Theorem II:* Let X_i be independent with $E[X_i] = \mu_i$ and $Var(X_i) = \sigma_i^2$. If:
$$\lim_{N\to\infty} \frac{\sum_{i=1}^{N} E[|X_i - \mu_i|^{2+\delta}]}{\left(\sum_{i=1}^{N} \sigma_i^2\right)^{(2+\delta)/2}} = 0$$
for some $\delta > 0$, then $Z_N \xrightarrow{\text{D}} N(0,1)$.

Asymptotic Theory for OLS

- Given random variables $\{X_i\}$ and $\{u_i\}$, let y_i be generated as:

$$y = X\beta + u$$

- where y is the $(N \times 1)$ vector of y_i, u is $(N \times 1)$ vector of u_i, X is $(N \times K)$ matrix of X_i, β is $(K \times 1)$.
- We observe only (y, X), but not u.
- We form the OLS estimator of β, $\hat{\beta}_N$:

$$\begin{aligned}
\hat{\beta}_N &= (X'X)^{-1}X'(X\beta + u) \\
&= \beta + (X'X)^{-1}X'u \\
&= \beta + (\frac{1}{N}X'X)^{-1}\left(\frac{1}{N}X'u\right)
\end{aligned}$$

Consistency of OLS

- When is $\hat{\beta}_N \xrightarrow{\text{P}} \beta$?
- Need:

$$\frac{1}{N} X' u \xrightarrow{\text{P}} 0$$

- This is a $(K \times 1)$ vector, each element k of the form $\frac{1}{N}\sum_{i=1}^{N} X_i^{(k)} u_i$. Need each of these to have plim zero.
- If $E[X_i^{(k)} u_i] = E[w_i^{(k)}] = 0$, $w_i^{(k)}$ are independent draws, and $Var(w_i^{(k)})$ is not increasing too fast in i, we can use the Markov LLN to get the plim equal zero.
- Note: $E[X_i^{(k)} u_i] = Cov(X_i^{(k)}, u_i)$. $E[u_i | X_i^{(k)}] = 0$, $E[u_i] = 0$ sufficient for $Cov(X_i^{(k)}, u_i) = 0$.
- Practical question: What kinds of heteroskedasticity does this allow for?
- Also need $M_{XX} = \text{plim} \frac{1}{N} X' X$ to exist, and to be invertible.
- To use the Markov LLN for this, need restrictions on the variance of $X^{(i)} X^{(j)}$ (products of covariates) not growing too

Asymptotic distribution of OLS estimator

- Scale $\hat{\beta}_N$ by $\sqrt{N}$.

$$\begin{aligned} \sqrt{N}\,(\hat{\beta}_N - \beta) &= (\frac{1}{N}X'X)^{-1}\frac{1}{\sqrt{N}}X'u \\ (Why?) &\xrightarrow{\mathrm{D}} M_{XX}^{-1}\frac{1}{\sqrt{N}}X'u \end{aligned}$$

- But $\frac{1}{\sqrt{N}}X'u = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i u_i = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} q_i$.
- If $E[q_i] = 0$ and variance does not grow too fast, we can apply CLT II here.

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i u_i \xrightarrow{\mathrm{D}} N(0, M_{X\Omega X})$$

- where

$$\begin{aligned} M_{X\Omega X} &= \frac{1}{N}\sum_{i=1}^{N} Var(X_i u_i) \\ (u_i \text{ is scalar}) &= \frac{1}{N}\sum_{i=1}^{N} E(u_i^2 X_i X_i') \\ (\text{When?}) &= \text{plim}\,\frac{1}{N}\sum_{i=1}^{N} u_i^2 X_i X_i' \end{aligned}$$

- So:

$$\sqrt{N}\,(\hat{\beta}_N - \beta) \xrightarrow{\mathrm{D}} N(0, M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1})$$

where

$$\begin{aligned} M_{XX} &= \text{plim}\,\frac{1}{N} X'X \\ M_{X\Omega X} &= \text{plim}\,\frac{1}{N}\sum_{i=1}^{N} u_i^2 X_i X_i' \end{aligned}$$

Heteroskedasticity robust (White) standard errors

- Since we do not see M_{XX} and $M_{X\Omega X}$, we need to estimate them.
- White (1980) suggested:

$$
\begin{aligned}
\hat{M}_{XX} &= \frac{1}{N} X'X \\
\hat{M}_{X\Omega X} &= \frac{1}{N} \sum_{i=1}^{N} \hat{u}_i^2 X_i X_i'
\end{aligned}
$$

- and showed the restrictions under which

$$
\sqrt{N}\left(\hat{\beta}_N - \beta\right) \xrightarrow{\mathrm{D}} N(0, \hat{M}_{XX}^{-1} \hat{M}_{X\Omega X} \hat{M}_{XX}^{-1})
$$

- Basic idea of the proof is showing that

$$
\begin{aligned}
\text{plim}\, \hat{M}_{X\Omega X} &\overset{?}{=} M_{X\Omega X} \\
&= \text{plim}\, \frac{1}{N} \sum_{i=1}^{N} \left(y_i - X_i\hat{\beta}\right)^2 X_i X_i' \\
&= \text{plim}\, \frac{1}{N} \sum_{i=1}^{N} \left(u_i + X_i(\beta - \hat{\beta})\right)^2 X_i X_i' \\
&= \text{plim}\, \frac{1}{N} \sum_{i=1}^{N} u_i^2 X_i X_i' + \text{extra terms}
\end{aligned}
$$

White showed that the plim of extra terms is zero.

- Asymptotic theory can be used to show the consistency and derive limit distribution of the OLS estimator under quite general assumptions on u_i.
- Key statistical assumptions:

$$
\begin{aligned}
Cov(X_i, u_i) &= 0 \\
\operatorname{plim} \frac{1}{N} X'X &= M_{XX} \text{ exists and invertible}
\end{aligned}
$$

- White standard errors account for quite general forms of heteroskedasticity.
- You can use standard OLS to get consistent estimates of β, then use White standard error formula to do (asymptotically) correct inference.