ECMA31000: Introduction to Empirical Analysis
Hypothesis Testing; Linear Regression I

Joe Hardwick

University of Chicago

Autumn 2021

Outline

- Last time:
  - Confidence Intervals, Hypothesis Testing.

- Today:
  - Non-testable hypotheses
  - Linear Regression

Hypothesis Testing: Definitions

- The <u>null hypothesis</u> is a subset $\Theta_0 \subset \Theta$ of hypothesised values for θ_0, written as:

$$H_0 : \theta_0 \in \Theta_0.$$

- If Θ_0 is a singleton, the null hypothesis is called a <u>simple hypothesis</u>. If not, the hypothesis is <u>composite</u>.

- e.g. If $X \sim \mathcal{N}(\mu, 1)$, $H_0 : \mu = 0$ is a simple null, $H_0 : \mu \leq 0$ is a composite null.

- A test of H_0 is therefore a test of whether the data were generated by some F_θ such that $\theta \in \Theta_0$.

- The <u>alternative hypothesis</u> is $\Theta \backslash \Theta_0$.

Hypothesis Testing: Definitions

- A <u>test</u> of H_0 is a function $\phi_n(X_1, \ldots, X_n) \to \{0, 1\}$.
- We reject H_0 at sample size n iff $\phi_n = 1$.
- e.g. If $T_n(X_1, \ldots, X_n)$ is a sequence of statistics, and c_n a sequence of real numbers,

$$\phi_n(X_1, \ldots, X_n) = \mathbf{1}\left(T_n > c_n\right)$$

  is a test which rejects H_0 iff $T_n > c_n$.

Two undesirable outcomes

- Suppose $\theta_0 \in \Theta_0$ but $\phi_n = 1$. This is called a Type I Error.
- Suppose $\theta_0 \notin \Theta_0$ but $\phi_n = 0$. This is called a Type II Error.
- It is customary to control the probability of a Type I Error first, and then minimize the probability of a Type II Error subject to this constraint.
- The power function associated with ϕ_n is the function

$$\beta_n(\theta) = \mathrm{P}_\theta(\phi_n(X_1, \ldots, X_n) = 1),$$

which is the probability that ϕ_n rejects H_0 if the true parameter is θ.

Properties of tests

- Given a significance level α, we select a test ϕ_n such that

$$\beta_n(\theta) \leq \alpha \quad \text{for all } \theta \in \Theta_0,$$

or, alternatively

$$\sup_{\theta \in \Theta_0} \beta_n(\theta) \leq \alpha.$$

- Similarly, given a test ϕ_n, the size of ϕ_n with power function β_n is

$$\alpha := \sup_{\theta \in \Theta_0} \beta_n(\theta).$$

- The probability that ϕ_n rejects when H_0 is true cannot exceed α.

Example: Normal Distribution

- Suppose $\{X_i\}_{i=1}^n$ is an iid sample from $\mathcal{N}(\mu, 1)$. We wish to test $H_0 : \mu \leq 0$ vs. $H_1 : \mu > 0$, so

$$\Theta_0 = \left\{\mu \in (-\infty, 0]; \sigma^2 \text{ known}\right\}.$$

- .

- Let $c > 0$ and consider the test

$$\bar{X}_n > c$$
$$\Leftrightarrow \quad \sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma} > \frac{\sqrt{n}(c - \mu)}{\sigma}$$

$$\phi_n^{norm}(X_1, \ldots, X_n) = \mathbf{1}\left(\bar{X}_n > c\right).$$

$$\sim \mathcal{N}(0, 1)$$

- The power function is

$$\begin{aligned}
\beta_n^{norm}(\mu) &= \mathrm{P}_\mu\left(\bar{X}_n > c\right) \\
&= \mathrm{P}_\mu\left(\sqrt{n}\left(\bar{X}_n - \mu\right)/\sigma > \sqrt{n}\left(c - \mu\right)/\sigma\right) \\
&= 1 - \Phi\left(\sqrt{n}\left(c - \mu\right)/\sigma\right).
\end{aligned}$$

Example: Normal Distribution

- The uniformly (in μ) most powerful test of size α takes the form
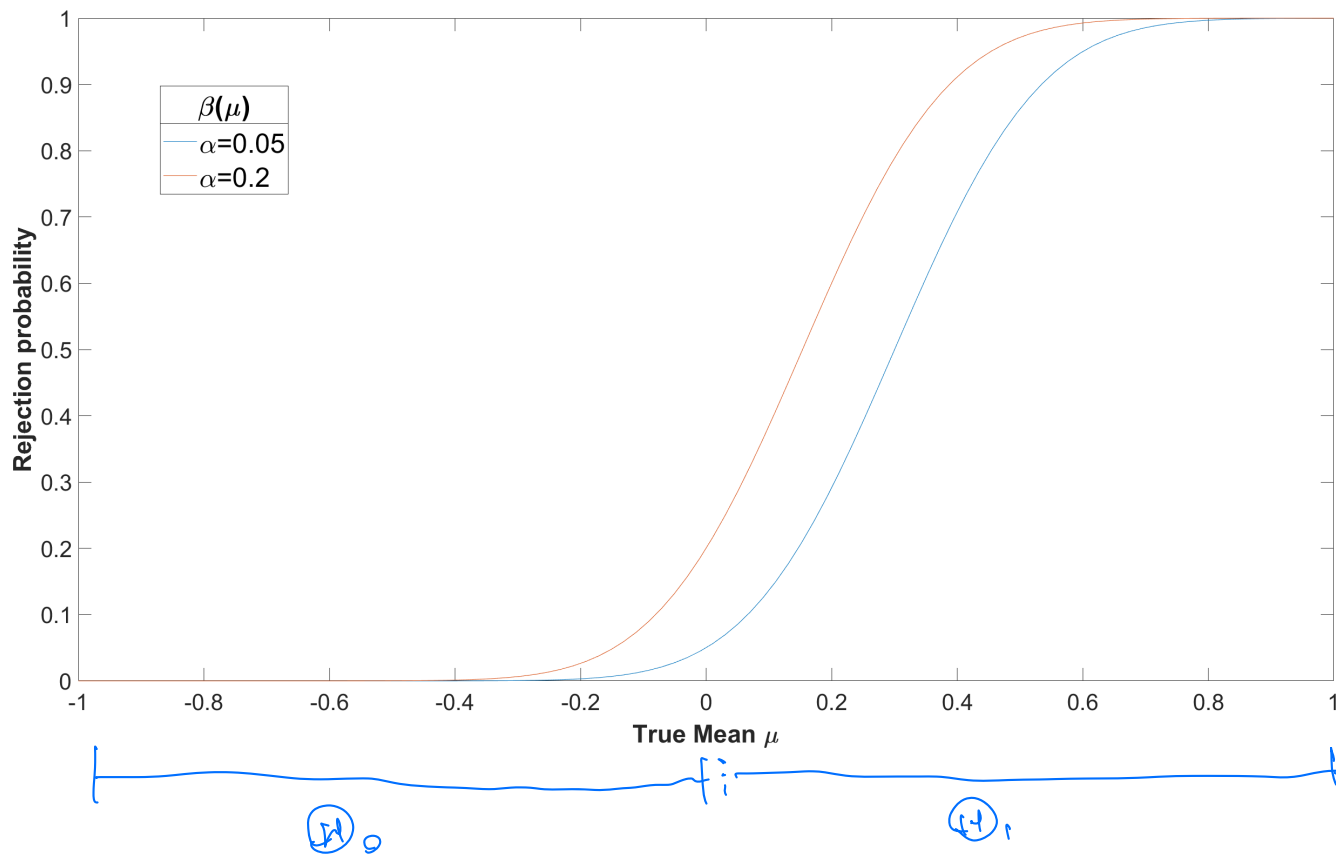
$$\phi_n^{norm}(X_1, \ldots, X_n) = \mathbf{1}\left(\bar{X}_n > c^*(\alpha)\right),$$

where $c^*(\alpha) = z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$. Its power function is:

$$\beta_n^{norm}(\mu) = 1 - \Phi\left(\sqrt{n}(c^* - \mu)/\sigma\right)$$

$$= 1 - \Phi\left(z_{1-\alpha} - \frac{\mu\sqrt{n}}{\sigma}\right).$$

Example: Normal Distribution

- Probability of rejecting H_0 is known for any true value of μ if, e.g. $X \sim \mathcal{N}\left(\mu, 1\right)$. This is the power curve:

Nuisance Parameters

- Suppose $\{X_i\}_{i=1}^n$ is an iid sample from $\mathcal{N}\left(\mu, \sigma^2\right)$. We wish to test $H_0 : \mu \leq 0$ vs. $H_1 : \mu > 0$, so

$$\Theta_0 = \left\{\mu \in (-\infty, 0]; \sigma^2 > 0\right\}, \quad \Theta_1 = \left\{\mu \in (0, \infty); \sigma^2 > 0\right\}.$$

- σ^2 is unobserved, and is called a <u>nuisance parameter</u> because it isn't of immediate interest but changing it may alter the distribution of our test statistic under the null and alternative hypotheses.

- One solution: Choose a test statistic whose distribution under H_0 controls size by accounting for changing σ^2 (e.g. the t-test).

- Problem: We've now added a restriction to the class of tests, so the most powerful one will be less powerful.

Nuisance Parameters

- Let $c > 0$ and consider the t-test:

$$\phi_n^t(X_1, \ldots, X_n) = \mathbf{1}\left(\frac{\sqrt{n}\left(\bar{X}_n - 0\right)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}} > c\right).$$

- If $\mu \in \mathbb{R}$, $Z \sim \mathcal{N}(0,1)$, $V \sim \chi_{n-1}^2$ and Z is independent of V, then

$$\frac{Z + \mu}{\sqrt{V/(n-1)}} \sim t_{\mu, n-1}$$

has a non-central $t-$distribution with $n-1$ degrees of freedom and non-centrality parameter μ, where $t_{\mu, n-1}$ denotes its CDF.

Power of the t-test

- Let

$$T_n = \frac{\sqrt{n}\left(\bar{X}_n - 0\right)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}} = \frac{\sqrt{n}\left(\bar{X}_n - 0\right)/\sigma}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2/\sigma^2}}.$$

- Therefore:

$$T_n = \frac{\sqrt{n}\left(\bar{X}_n - \mu\right)/\sigma + \sqrt{n}\mu/\sigma}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2/\sigma^2}} \sim t_{\frac{\sqrt{n}\mu}{\sigma},n-1}.$$

- The power function is

$$P(T_n > c)$$
$$= 1 - P(T_n \leq c).$$

$$\beta_n^t(\mu) = P_{\mu,\sigma}(T_n > c)$$
$$= 1 - t_{\frac{\sqrt{n}\mu}{\sigma},n-1}(c).$$

Power of the t-test

- When $\alpha \geq 0.5$, the uniformly (in (μ, σ)) most powerful test of size α takes the form

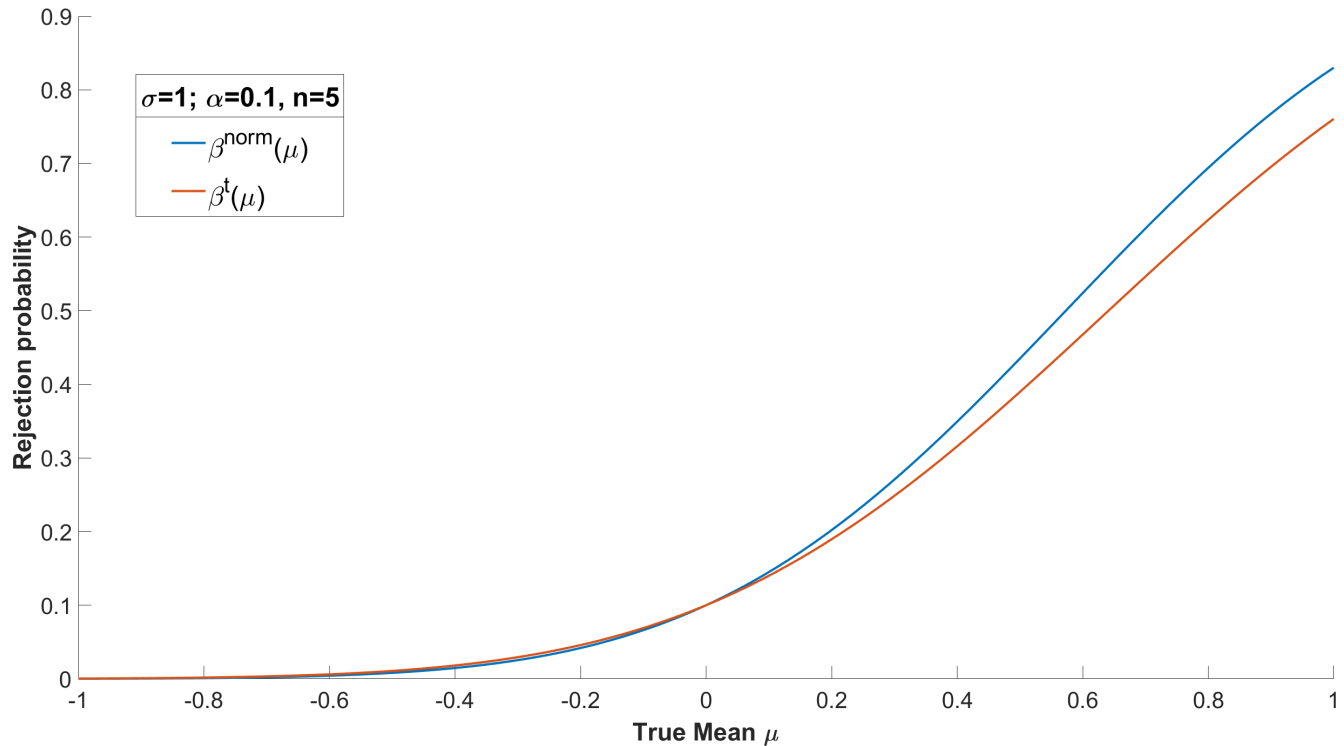$$\phi_n^t (X_1, \ldots, X_n) = \mathbf{1} \left(T_n > t_{n-1,1-\alpha}^* \right),$$

  where $t_{n-1,1-\alpha}^*$ is the $1 - \alpha$ quantile of the (central) t_{n-1} distribution. Its power function is:

$$\beta_n^t (\mu) = 1 - t_{\frac{\sqrt{n}\mu}{\sigma}, n-1} \left(t_{n-1,1-\alpha}^* \right).$$

- (When $\alpha < 0.5$, the most powerful test at the point (μ, σ) varies with (μ, σ), so there is no UMP test).

Power of the t-test

- Power curves for the t-test and UMP test when it is known that $\sigma = 1$:

Power of the t-test

- Expanding the statistical model to allow $\sigma^2 \neq 1$ made it harder to satisfy
$$\sup_{\theta \in \Theta_0} \beta_n(\theta) \leq \alpha.$$

- The t-test is less powerful than $\phi_n^{norm}(X_1, \ldots, X_n)$ when σ is known, but has significance level α for any value of σ when it is unknown.

- Moreover, $\phi_n^{norm}(X_1, \ldots, X_n)$ requires us to specify a value $\sigma = \sigma_0$, which, if incorrect, may cause the test to "over-reject" when H_0 is true:

$$\beta_n^{norm}(\mu) = 1 - \Phi\left(z_{1-\alpha} \cdot \frac{\sigma_0}{\sigma} - \frac{\mu\sqrt{n}}{\sigma}\right),$$

which roughly equals 0.5 if $\mu = 0$ and $\sigma_0/\sigma \approx 0$.

Testing the mean in general

- Now suppose we drop the restriction to normal distributions. The statistical model is greatly expanded:

$$\mathcal{F} = \left\{ F \text{ is a CDF and } 0 < \sigma^2(F) < \infty \right\}.$$

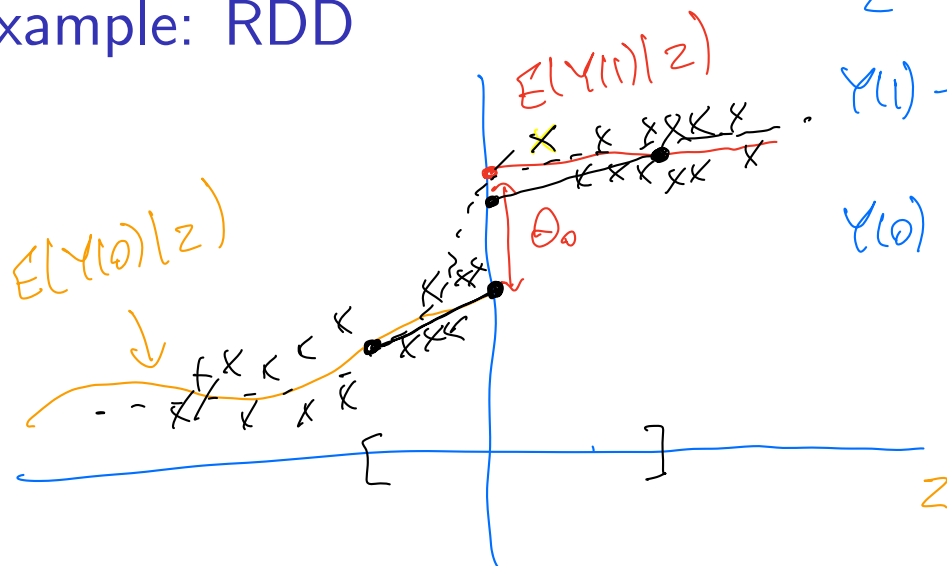- This is a 'non-parametric' class of distributions.

Theorem
Let $\mathcal{F}_0 = \{F \in \mathcal{F} : \mu(F) = 0\}$ be the null hypothesis. Any test ϕ which has size α has power $\leq \alpha$ for any $F \in \mathcal{F} \backslash \mathcal{F}_0$. Any test which has power β against some alternative $F \in \mathcal{F} \backslash \mathcal{F}_0$ has size $\geq \beta$.

Example: RDD

$E(Y(1)|z)$

$E(Y(0)|z)$

θ_0

Z — Standardized test Score

$Y(1)$ — Wage if attend a selective school.

$Y(0)$ — Wage if don't attend a selective school.

$D = \mathbb{1}(Z \geq 0)$

z

Suppose ----- $E(Y(0)|z)$

Linear Regression: Definitions

- Let (Y, X, U) be a random vector such that Y and U are scalar random variables and $X \in \mathbb{R}^{k+1}$.

- Assume the first component of X equals 1:

$$X = (X_0, X_1 \ldots, X_k),$$

where $X_0 = 1$.

- Let $\beta = (\beta_0, \ldots, \beta_k) \in \mathbb{R}^{k+1}$ be a constant vector of unknown parameters such that

$$Y = X'\beta + U.$$

- β_0 is the *intercept* and the remaining β_j are the *slope* parameters.

Interpretations

- Note that given (Y, X), and any β, we may construct U such that the representation

$$Y = X'\beta + U$$

  holds without loss of generality (simply take $U = Y - X'\beta$).
- We now consider additional restrictions on this relationship which lead us to interpret the quantity $X'\beta$ as a "best predictor" of Y, and also (uniquely) determine β.

Linear Conditional Expectation

- Consider
$$Y = X'\beta + U. \qquad \text{LIE}$$

- Suppose $\mathrm{E}(Y|X) = X'\beta$ and define $U = Y - \mathrm{E}(Y|X)$.
- Then $\mathrm{E}(U|X) = 0$ and so $\mathrm{E}(U) = \mathrm{E}(UX) = 0$, which means $Cov(U, X) = 0$. This model is often written equivalently as

$$Y = X'\beta + U; \qquad \mathrm{E}(U|X) = 0.$$

- The vector β is a feature of the joint distribution of (Y, X) which represents the conditional expectation.

$$E(U) = E\left(E(U|X)\right) = 0.$$

$$E(XX') \text{ has full rank}$$
$$\Rightarrow \beta \text{ is unique.}$$

Linear Conditional Expectation: Example I

- This model merely posits that if X_j increases by 1 unit, the best predictor of Y under square loss increases/decreases by β_j units.

- It does NOT claim that an increase in X_j *causes* a change in Y.

- For example, suppose we observe an iid sample of $\{Y_i, D_i\}_{i=1}^n$ where Y_i is the scrap rate of factory i, and

$$D_i = \begin{cases} 1 & \text{if factory } i \text{ receives job training grant,} \\ 0 & \text{otherwise.} \end{cases}$$

Linear Conditional Expectation: Example I

$$\mathrm{E}(Y|D=0) = \beta_0 \qquad \mathrm{E}(Y|D=1) = \beta_0 + \beta_1.$$

- Without loss of generality (see problem set 5):

$$\mathrm{E}(Y_i|D_i) = \beta_0 + \beta_1 D_i.$$

$$\beta_1 = \mathrm{E}(Y|D=1) - \mathrm{E}(Y|D=0).$$

- Hence, $\beta_0 = \mathrm{E}(Y|D=0)$ and
  $\beta_1 = \mathrm{E}(Y|D=1) - \mathrm{E}(Y|D=0)$.

- Suppose $\beta_1 < 0$, so the mean scrap rate of factories without a grant is higher than for those with a grant.

- This does NOT imply receiving a grant negatively impacts productivity, especially if grants are first come first served and the less efficient factories applied first.

- We lack a model of how Y is determined as a function of D and other variables (like prior scrap rate).

Linear Conditional Expectation: Example I

- A more complete model would consider *potential outcomes*:

$$Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0},$$

  where Y_{i1} is the scrap rate of firm i if they receive the grant and Y_{i0} is their scrap rate without. Note that

$$\beta_1 = E(Y(D=1) - E(Y(D=0)).$$

$$\begin{aligned}
\beta_1 &= \mathrm{E}(Y_1 | D = 1) - \mathrm{E}(Y_0 | D = 0) \\
&= \mathrm{E}(Y_1 | D = 1) - \mathrm{E}(Y_0 | D = 1) \Big\} \ \text{ATT} \\
&+ \mathrm{E}(Y_0 | D = 1) - \mathrm{E}(Y_0 | D = 0). \Big\} \ \text{Selection Bias.}
\end{aligned}$$

- It may also hold that $\beta_1 = 0$ but offering training grants is highly effective, if the baseline productivity of firms that don't enrol is much higher.

Linear Conditional Expectation: Example II

- If (Y, X) have a multivariate normal distribution, with $Y \in \mathbb{R}, X \in \mathbb{R}^k$, the assumption

$$\mathrm{E}\left(Y|X\right) = \beta_0 + X'\beta_1 = \left(1, X'\right)\beta$$

  holds for some vector $\beta = (\beta_0, \beta_1')' \in \mathbb{R}^{k+1}$.

- β is unique if X is non-degenerate (no component of X is a linear function of the others). *X is not perfectly multi-collinear.*

- If X is degenerate, β is not uniquely determined because multiple values of β lead to the same value of $(1, X')\beta$. To see this, suppose for some constants $a_1, \ldots, a_k$ that

$$1 = a_1 X_1 + \cdots + a_k X_k.$$

Linear Conditional Expectation: Example II

- Then:

$$E[Y|X] = (1, X')\, \beta = \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k$$

$$= (a_1 X_1 + \ldots a_k X_k)\, \beta_0 + \cdots + X_k\beta_k$$

$$= X_1 (a_1\beta_0 + \beta_1) + \cdots + X_k (a_k\beta_0 + \beta_k)$$

$$= (1, X')\, \tilde{\beta},$$

where

j-H-st component $\longleftarrow$ $\quad a_j \beta_0 + \beta_j$

$$\tilde{\beta} = (0, a_1\beta_0 + \beta_1, \ldots, a_k\beta_0 + \beta_k).$$

- Therefore, for any $m > 0$, $(1, X')\, \beta = (1, X')\, \beta_m^*$, where

$$\beta_m^* = \beta + m\left(\tilde{\beta} - \beta\right).$$

$$(1, X')\beta_m^* = (1, X')\beta + m\, (1, X')(\tilde{\beta} - \beta).$$

Questions?

$$\therefore \quad (1, X^1)\, \widehat{\beta} = (1, X^r)\, \beta.$$

Best Linear approximation to $\mathrm{E}(Y|X)$

- In the absence of joint normality, or if X can take more than two values, $\mathrm{E}(Y|X)$ is generally non-linear.

- Suppose $\mathrm{E}(Y^2) < \infty$ and $\mathrm{E}(X_j^2) < \infty$ for each $j = 1, \ldots, k$.

- Consider the problem

$$\min_{b \in \mathbb{R}^{k+1}} \mathrm{E}\left(\mathrm{E}(Y|X) - X'b\right)^2.$$

- The solution is the best linear approximation to $\mathrm{E}(Y|X)$ under square loss.

Best Linear predictor of Y

- Any solution to this problem is also represents a best linear predictor of Y, since

$$
\mathrm{E}\left[\left(\mathrm{E}\left(Y|X\right) - X'b\right)^2\right] = \mathrm{E}\left(\overbrace{\mathrm{E}\left(Y|X\right) - Y}^{a} + \overbrace{Y - X'b}^{b}\right)^2
$$

$$
= \mathrm{E}\left(Y - \mathrm{E}\left(Y|X\right)\right)^2
$$

$$
+ \mathrm{E}\left(Y - X'b\right)^2
$$

$$
- 2\mathrm{E}\left[\left(Y - \mathrm{E}\left(Y|X\right)\right)\left(Y - X'b\right)\right].
$$

- The law of iterated expectation implies

$$
\mathrm{E}\left[\left(Y - \mathrm{E}\left(Y|X\right)\right)\left(Y - X'b\right)\right] = \mathrm{E}\left(Y\left(Y - \mathrm{E}\left(Y|X\right)\right)\right).
$$

- Therefore,

$$
E[(Y - E(Y(X))]Y] \; - \; E[(Y - E(Y|X)] X'b)
$$

$$
\mathrm{E}\left[\left(\mathrm{E}\left(Y|X\right) - X'b\right)^2\right] = \mathrm{E}\left(Y - X'b\right)^2 + \textit{Constant.}
$$

$$
E\left(X'b \; E\left[\left(Y - E(Y(X)\right) | X\right]\right)
$$

28

Best Linear predictor of Y

- Since $\mathrm{E}\left(Y - X'b\right)^2$ is convex in b, any minimizer b^* is characterized by the FOC

$$\frac{\partial}{\partial b}\mathrm{E}\left(Y - X'b^*\right)^2 = 2\mathrm{E}\left(XX'\right)b^* - 2\mathrm{E}\left(XY\right) = 0.$$

- Hence b^* must satisfy

$$\mathrm{E}\left(X\left(Y - X'b^*\right)\right) = 0.$$

- Defining $U = Y - X'b^*$ yields

$$\mathrm{E}\left(XU\right) = 0.$$

Causal Model

- Suppose we <u>assume</u> that Y is determined by the equation

$$Y = X'\beta + U,$$

  where X is observed and U is not.
- The *ceteris paribus* effect of X_j on Y holding the other elements of X and U constant is β_j.
- We may assume $\mathrm{E}(U) = 0$ WLOG, by shifting β_0 accordingly, but $\mathrm{E}(XU), \mathrm{E}(U|X)$ etc. are not necessarily equal to 0.
- The statement $\mathrm{E}(UX) = 0$ is therefore an assumption about the joint distribution of (Y, X). If this assumption uniquely determines β, it implies that the best linear predictor of Y given X also represents the causal effect of X on Y.

Causal Model: Example I revisited

- We observe an iid sample of $\{Y_i, D_i\}$ where Y_i is the scrap rate of factory i, and

$$D_i = \begin{cases} 1 & \text{if factory } i \text{ receives a job training grant,} \\ 0 & \text{otherwise.} \end{cases}$$

- The observed outcome Y_i is a function of the potential outcomes Y_{i0}, Y_{i1} and treatment:

$$Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0}.$$

- WLOG we write

$$\mathrm{E}\left(Y_i | D_i\right) = \beta_0 + \beta_1 D_i.$$

Causal Model: Example I revisited

- We derived

$$\beta_1 = \underbrace{\mathrm{E}\left(Y_1|D=1\right) - \mathrm{E}\left(Y_0|D=1\right)}_{ATT} + \underbrace{\mathrm{E}\left(Y_0|D=1\right) - \mathrm{E}\left(Y_0|D=0\right)}_{\text{Selection Bias}}$$

- Suppose that assignment of grants is now independent of the factory's potential outcomes Y_0, Y_1. This means that factories are not applying for the grant based on Y_0. We say assignment to treatment is "randomized".

- The Selection Bias term vanishes, and

$$\beta_1 = \mathrm{E}\left(Y_1|D=1\right) - \mathrm{E}\left(Y_0|D=1\right) = \mathrm{E}\left(Y_1 - Y_0\right),$$

where the second equality holds by random assignment.

- β_1 now represents the mean effect (often called Average Treatement Effect or ATE) of receiving a grant in the population of firms.

Causal Model: Example I revisited

- It is still not the case that the model

$$Y_i = \beta_0 + \beta_1 D_i + U_i$$

has a causal interpretation, even though β_1 may be equal to a parameter we are interested in (namely the ATE).

- The reason is that the grants may not impact firms equally, and β_1 represents the average effect.

- If we go one step further, and assume

$$Y_{i0} = \beta_0 + U_i,$$
$$Y_{i1} = \beta_0 + \beta_1 + U_i,$$

then $\beta_1 = Y_{i1} - Y_{i0}$ represents an (homogeneous) additive treatment effect and our model has a causal interpretation.

Linear regression when $\mathrm{E}(XU) = 0$

- Let (Y, X, U) be a random vector such that Y and U are scalar random variables and $X \in \mathbb{R}^{k+1}$.

- Assume the first component of X equals 1:

$$X = (X_0, X_1 \ldots, X_k),$$

  where $X_0 = 1$.

- Let $\beta = (\beta_0, \ldots, \beta_k) \in \mathbb{R}^{k+1}$ be a constant vector of unknown parameters such that

$$Y = X'\beta + U.$$

- Suppose $\mathrm{E}(XU) = 0$, justified according to how we interpret the model.

- Suppose also that $\mathrm{E}\left(X_j^2\right) < \infty$ for $j \leq k$, so $\mathrm{E}(XX')$ exists.

Linear regression when $\mathrm{E}(XU) = 0$

- There is <u>perfect collinearity</u> in X if there exists a constant vector $a \neq 0$ such that

$$\mathrm{P}\left(a'X = 0\right) = 1.$$

- We assume there is no perfect collinearity in X. This assumption is equivalent to the condition that $\mathrm{E}(XX')$ is invertible.

- Since $\mathrm{E}(XX')$ is positive semidefinite, it is invertible iff it is positive definite.

Linear regression when $\mathrm{E}(XU) = 0$

Lemma
Suppose X is a $(K \times 1)$ random vector and $\mathrm{E}(XX')$ exists. Then $\mathrm{E}(XX')$ is invertible iff there is no perfect collinearity in X.

Proof.
If there is perfect collinearity X, then there exists a vector $a \neq 0$ such that $\mathrm{P}(X'a = 0) = 1$. For this vector,

$$\mathrm{E}(XX')\,a = \mathrm{E}(X(X'a)) = \mathrm{E}(X \cdot 0) = 0.$$

Therefore, $\mathrm{E}(XX')$ is not full column rank and so not invertible. Now suppose there is no perfect collinearity in X. For any vector $c \in \mathbb{R}^{k+1} \setminus \{0\}$,

$$c'\mathrm{E}(XX')\,c = \mathrm{E}\left((X'c)^2\right) > 0,$$

since the expectation equals 0 if and only if $\mathrm{P}(X'c = 0) = 1$ which is ruled out by assumption. $\square$

Linear regression when $\mathrm{E}(XU) = 0$

- Since $\mathrm{E}(XU) = \mathrm{E}(X(Y - X'\beta)) = 0$, we obtain

$$\mathrm{E}(XY) = \mathrm{E}(XX')\beta.$$

- Since there is no perfect collinearity in X, $\mathrm{E}(XX')$ is invertible, so we can solve for a unique β:

$$\beta = \mathrm{E}(XX')^{-1}\mathrm{E}(XY).$$

- In this case, β is point identified, since it is uniquely determined by $\mathrm{E}(XX'), \mathrm{E}(XY)$.

Linear regression when $\mathrm{E}\left(XU\right) = 0$

- If $\mathrm{E}\left(XX'\right)$ is not invertible, there are multiple solutions to

$$\mathrm{E}\left(XY\right) = \mathrm{E}\left(XX'\right)\beta.$$

- If $\tilde{\beta}$ satisfies $\mathrm{E}\left(XX'\right)\tilde{\beta} = \mathrm{E}\left(XY\right)$, then $\mathrm{E}\left(XX'\right)\left(\tilde{\beta} - \beta\right) = 0$, so

$$\left(\tilde{\beta} - \beta\right)'\mathrm{E}\left(XX'\right)\left(\tilde{\beta} - \beta\right) = \mathrm{E}\left(\left[X'\left(\tilde{\beta} - \beta\right)\right]^2\right) = 0,$$

which implies $\mathrm{P}\left(X'\tilde{\beta} = X'\beta\right) = 1$.

- If we interpret $X'\beta$ as a best linear predictor of Y, this says there are multiple best predictors.
- If the model is interpreted causally, however, different values of β imply different ceteris paribus effects, holding U and the other components of X fixed.

Questions?