ECMA 31000: Solutions to Problem Set 5

Joe Hardwick

Due Nov 11 by 11:59PM

Question 1 Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$ and $\mu \in \mathbb{R}$ are unknown. We wish to test $H_0 : \mu \leq 1$ vs. $H_1 : \mu > 1$ based on an iid sample $\{X_i\}_{i=1}^n$.

a) What is the parameter space Θ indexing our statistical model? What is Θ_0?

ANS: $\Theta = \{(\mu, \sigma^2) \in \mathbb{R}^2 : \mu \in \mathbb{R}, \sigma^2 > 0\}$, and $\Theta_0 = \{(\mu, \sigma^2) \in \mathbb{R}^2 : \mu \leq 1, \sigma^2 > 0\}$.

b) Show that the test

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

with associated power function β_n satisfies

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

where $t^*_{n-1,1-\alpha}$ is the $1-\alpha$ quantile of the (central) t distribution with $n-1$ degrees of freedom.

Hint: Do $\mu = 1$ first, then do the other cases by arguing that T_n is distributed as a central t_{n-1} plus something negative.

ANS: Let

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

where the RHS is a non-central t distribution with parameter $\sqrt{n}(\mu - 1)$. When $\mu = 1$, T_n has a (central) t_{n-1} distribution, and so exceeds $t^*_{n-1,1-\alpha}$ with probability α. When $\mu < 1$,

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

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

Therefore, $\mathrm{P}\left(T_n > t^*_{n-1,1-\alpha}\right) < 1 - t_{n-1}\left(t^*_{n-1,1-\alpha}\right) = \alpha$. Therefore, the supremum of the power function is attained when $\mu = 1$, and:

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

c) Show that the test

$$\phi_n(X_1, \ldots, X_n) = \mathbf{1}\left(\frac{\sqrt{n}\left(\bar{X}_n - 1\right)}{\sigma_0} > c\right),$$

where σ_0 is chosen by the researcher, and with associated power function β_n, satisfies

$$\sup_{\theta\in\Theta_0} \beta_n(\theta) = \frac{1}{2}$$

for any fixed value of $c \geq 0$. What do you conclude about the use of such a test when σ is unknown?

ANS: Note that when $\mu \leq 1$,

$$\begin{aligned}
\beta_n(\theta) &= \mathrm{P}\left(\frac{\sqrt{n}\left(\bar{X}_n - 1\right)}{\sigma_0} > c\right) \\
&= \mathrm{P}\left(\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} > c\cdot\frac{\sigma_0}{\sigma} - \frac{\sqrt{n}(\mu-1)}{\sigma}\right) \\
&= 1 - \Phi\left(c\cdot\frac{\sigma_0}{\sigma} - \frac{\sqrt{n}(\mu-1)}{\sigma}\right) \leq 1 - \Phi\left(c\cdot\frac{\sigma_0}{\sigma}\right) \to 1 - \Phi(0).
\end{aligned}$$

as $\sigma \to \infty$. Since $1 - \Phi(0) = \frac{1}{2}$ is an upper bound of $1 - \Phi\left(c\cdot\frac{\sigma_0}{\sigma}\right)$, the above convergence implies

$$\sup_{\theta\in\Theta_0} \beta_n(\theta) = \frac{1}{2}.$$

This result shows that when σ^2 is unknown, misspecifying it can lead us to reject with probability close to $\frac{1}{2}$ no matter what the cutoff. So, for typical significance levels such as $\alpha = 0.05$, the test does not have size less than α.

Question 2 Consider the simple linear regression model

$$Y = \beta_0 + \beta_1 X + U,$$

where both Y and X are scalar random variables. Suppose (β_0, β_1) represents the best linear predictor of Y given X.

a) Derive β_0 and β_1 as features of the joint distribution of (Y, X).

ANS: Define $Q(a,b) := \mathrm{E}\left[(Y-a-bX)^2\right]$. The best linear predictor (a^*, b^*) solves

$$(a^*, b^*) \in \arg\min_{a,b} Q(a,b).$$

Expanding the square gives the equivalent problem

$$\min_{a,b}\left[\mathrm{E}\left(Y^2\right) - 2a\mathrm{E}(Y) - 2b\mathrm{E}(XY) + a^2 + 2ab\mathrm{E}(X) + b^2\mathrm{E}\left(X^2\right)\right],$$

which shows that Q is strictly convex in (a,b) provided $\mathrm{E}(X^2) \neq 0$. If $\mathrm{E}(X^2) = 0$, then $X = 0$ almost surely. In that case, the minimization problem reduces to

$$\min_{a,b} \mathrm{E}\left[(Y-a)^2\right],$$

which is solved by $a^* = \mathrm{E}(Y)$ and $b^* \in \mathbb{R}$. If $\mathrm{E}(X^2) \neq 0$, the objective is convex in (a,b), so the FOCs are sufficient for a minimizer, and yield

$$\frac{\partial Q}{\partial a}(a^*, b^*) = 2a^* - 2\mathrm{E}(Y) + 2b^*\mathrm{E}(X) = 0, \tag{1}$$

$$\frac{\partial Q}{\partial b}(a^*, b^*) = -2\mathrm{E}(XY) + 2a^*\mathrm{E}(X) + 2b^*\mathrm{E}\left(X^2\right) = 0. \tag{2}$$

Substituting $a^* = \mathrm{E}(Y) - b^*\mathrm{E}(X)$ from (1) into (2) yields

$$b^*\mathrm{E}\left(X^2\right) = \mathrm{E}(XY) - \mathrm{E}(X)\left[\mathrm{E}(Y) - b^*\mathrm{E}(X)\right]. \tag{3}$$

Solving gives

$$b^* = \frac{Cov(X,Y)}{Var(X)},$$

provided $Var(X) > 0$. This yields

$$a^* = \mathrm{E}(Y) - \frac{Cov(X,Y)\mathrm{E}(X)}{Var(X)}.$$

If $Var(X) = 0$, X is constant almost surely, so the constant 1 and X are perfectly collinear. Note that in this case $Cov(X,Y) = 0$, so (3) reduces to $0 = 0$ after substituting the information from (1) into (2). Therefore, the minimizers are characterized by (1), and:

$$(a^*, b^*) \in \left\{(a,b) \in \mathbb{R}^2 : a = \mathrm{E}(Y) - b\mathrm{E}(X)\right\}.$$

b) Suppose X is almost surely constant. Characterize the best linear predictors now.

ANS: See above for $X = 0$ and $X = c \neq 0$.

c) Suppose Y and X have the same (non-degenerate) distribution. Show that $|\beta_1| \leq 1$. When is equality attained?

ANS: By the Cauchy-Schwarz inequality:

$$\begin{aligned}|Cov\left(X,Y\right)| &\leq \sqrt{Var\left(X\right)Var\left(Y\right)}\\ &= Var\left(X\right),\end{aligned}$$

since X has the same distribution as Y. This yields

$$|\beta_1| = \frac{|Cov\left(X,Y\right)|}{Var\left(X\right)} \leq \frac{Var\left(X\right)}{Var\left(X\right)} = 1.$$

Equality is attained in the Cauchy-Schwarz inequality if and only if $Y = a + bX$ for some constants a, b. Since X and Y have the same distribution, $b = \pm 1$ must hold (equal variances), so $|\beta_1| = 1$ if and only if $Y = a \pm X$ for some a. Since $\mathrm{E}\left(Y\right) = \mathrm{E}\left(X\right)$ must also hold, when $b = 1$ we must have $a = 0$, and when $b = -1$ we must have $a = 2\mathrm{E}\left(X\right)$. This is not surprising, since if $Y = X$, the best linear predictor sets $a^* = 0, b^* = 1$ to yield zero loss. If $Y = 2\mathrm{E}\left(X\right) - X$, $a^* = 2\mathrm{E}\left(X\right), b^* = -1$ yields zero loss.

d) Now consider the "reverse regression"

$$X = \gamma_0 + \gamma_1 Y + V.$$

Suppose (γ_0, γ_1) represents the best linear predictor of X given Y. When is $\gamma_1 = \beta_1$?

ANS: We have

$$\gamma_1 = \frac{Cov\left(X,Y\right)}{Var\left(Y\right)},$$

provided $Var\left(Y\right) \neq 0$. If, in addition, $Var\left(X\right) \neq 0$, $\beta_1 = \gamma_1$ iff $Var\left(X\right) = Var\left(Y\right)$. If either $Var\left(X\right) = 0$ or $Var\left(Y\right) = 0$, then β_1 can be made to equal γ_1 by appropriate choices of γ_0, β_0.

Question 3 Let (Y, X, U) be a random vector such that Y and U are scalar random variables and $X \in \mathbb{R}^{k+1}$. (Y, X) are observable, but U is not. Assume the first component of X equals 1:

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

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

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

Assume that $(Y, X_1, X_2, \ldots, X_k)$ has a non-degenerate multivariate normal distribution. Hint: For this question, it will help to look up the properties of conditional distributions of multivariate normal vectors.

a) Does this model have a causal interpretation? Is it possible to find β such that $\mathrm{E}\left(U|X\right) = 0$?

ANS: It is a mathematical property of the multivariate normal distribution that there exists a

$\beta = (\beta_0, \ldots, \beta_k)$ such that

$$\mathrm{E}\left(Y|X_1, \ldots, X_k\right) = X'\beta.$$

It follows that

$$\begin{aligned} Y &= \mathrm{E}\left(Y|X_1, \ldots, X_k\right) + \left(Y - \mathrm{E}\left(Y|X_1, \ldots, X_k\right)\right) \\ &\equiv X'\beta + U, \end{aligned}$$

where $U := Y - \mathrm{E}\left(Y|X_1, \ldots, X_k\right)$ satisfies

$$\begin{aligned} \mathrm{E}\left(U|X\right) &= \mathrm{E}\left(Y|X\right) - \mathrm{E}\left(\mathrm{E}\left(Y|X_1, \ldots, X_k\right)|X\right) \\ &= X'\beta - X'\beta = 0. \end{aligned}$$

This model does not necessarily have a causal intepretaton. Without further information, all we can conclude is that $\mathrm{E}\left(Y|X\right)$ is linear in X. A change in X_j ceteris paribus only affects our best predictor of Y, and this may or may not represent a causal effect of changing X_j on Y.

b) Show that our modelling assumptions imply that $Var\left(U|X\right)$ does not depend on X.

ANS: We have

$$\begin{aligned} Var\left(U|X\right) &= Var\left(Y - X'\beta|X\right) \\ &= Var\left(Y|X\right), \end{aligned}$$

which can be shown not to depend on the value of X. It depends only on the variance covariance matrix of $(Y, X_1, \ldots, X_k)$.

c) You observe an iid sample $\{Y_i, X_i\}_{i=1}^n$. Write down the model for the i-th observation and again in matrix form. Write down the OLS estimator $\hat{\beta}_n$ of β. Find the conditional distribution of $\hat{\beta}_n|\mathbf{X}$, where $\mathbf{X}$ is a matrix/vector containing the entire sample of observations $\{X_i\}_{i=1}^n$.

ANS: We have

$$Y_i = X_i'\beta + U_i; \qquad \mathrm{E}\left(U_i|X_i\right) = 0.$$

Stacking the n observations on top of each other gives

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} - & X_1' & - \\ - & X_2' & - \\ - & \vdots & - \\ - & X_n' & - \end{pmatrix}}_{\mathbf{X}} \beta + \underbrace{\begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}}_{\mathbf{U}}; \qquad \mathrm{E}\left(\mathbf{U}|\mathbf{X}\right) = 0.$$

Note that $\mathrm{E}\left(\mathbf{U}|\mathbf{X}\right) = 0$ holds if and only if $\mathrm{E}\left(U_i|\mathbf{X}\right) = 0$ for all i. Because the sample is iid, however, (U_i, X_i) is independent of X_j, for $j \neq i$, so the conditional expectation of U_i given X_i does not depend on X_j for $j \neq i$. Therefore, for each i, $\mathrm{E}\left(U_i|\mathbf{X}\right) = \mathrm{E}\left(U_i|X_i\right) = 0$, and the representation follows. Next, note that

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

Now note that since $\mathrm{E}(Y|X) = X'\beta$ and, in addition, we may write $Var(Y|X) = \sigma^2 > 0$ for some σ^2 that does not depend on X,

$$Y_i|X_i \sim \mathcal{N}\left(X_i'\beta, \sigma^2\right)$$

for each $i = 1, \ldots, n$. Since each of the vectors (Y_i, X_i) is independent of all other (Y_j, X_j), the joint conditional distribution is also normal and constructed as

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}\left(\mathbf{X}\beta, \sigma^2 I_n\right),$$

where I_n is the $n \times n$ identity matrix. Using properties of the normal distribution and the fact we have conditioned on $\mathbf{X}$ yields

$$\begin{aligned}
(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}|\mathbf{X} &\sim (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathcal{N}\left(\mathbf{X}\beta, \sigma^2 I_n\right) \\
&\overset{d}{=} \mathcal{N}\left(\beta, (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\left[\sigma^2 I_n\right]\left((\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\right)'\right) \\
&\overset{d}{=} \mathcal{N}\left(\beta, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right).
\end{aligned}$$

d) Suppose Y is independent of X. What is the value of β?

ANS: If Y is independent of X, then it is mean independent of X, so $\mathrm{E}(Y|X) = \mathrm{E}(Y)$. Since the joint distribution of $(Y, X_1, \ldots, X_k)$ is non-degenerate, for

$$\mathrm{E}(Y|X) = X'\beta = \mathrm{E}(Y),$$

it must be the case that $\beta_0 = \mathrm{E}(Y)$ and $\beta_j = 0$ for $j = 1, \ldots, k$.

Question 4 Let A be an $N \times K$ matrix with $N \geq K$. Prove that A has full column rank iff $A'A$ has full column rank.

ANS: If A has full column rank, and $v \in \mathbb{R}^k \backslash \{0\}$ is any non-zero vector, then $Av \neq 0$, so $v'A'Av = (Av)' Av > 0$, since this is a sum of squares, not all of which are zero. This implies $A'Av \neq 0$, so $A'A$ is full column rank. On the other hand, if $A'A$ is full column rank, and $v \in \mathbb{R}^k \backslash \{0\}$, $A'Av \neq 0$, which implies $Av \neq 0$.

Question 5 In class we argued that if (Y, X) is a random vector such that Y is a scalar random variable and $X \in \mathbb{R}^{k+1}$ is a random vector, we can always write

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

for any β, by defining U as $U := Y - X'\beta$. Suppose $\mathrm{E}(Y^2) < \infty$ and $\mathrm{E}\left(X_j^2\right) < \infty$ for each j.

a) Given Y, X, can we <u>always</u> find β, U such that

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

If yes, are there restrictions β must satisfy for this to hold? If no, provide a counterexample.

ANS: Yes. The restriction on β is that it is defined to be the best linear predictor of Y given X. Now define $U := Y - X\beta$. In that case, the first order conditions characterizing the best linear predictor imply

$$\mathrm{E}(X(Y - X'\beta)) = 0 \iff \mathrm{E}(XY) = \mathrm{E}(XX')\beta.$$

We now verify that a solution to this equation always exists. If $\mathrm{E}(XX')$ is full rank, we know that

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

is a solution to this equation. Now suppose $\mathrm{E}(XX')$ has rank $n < k+1$. This implies that there exist $k+1-n$ non-zero and linearly independent vectors $\{a_j\}_{j=1}^{k+1-n}$ such that

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

for each $j = 1, \ldots, k+1-n$. Now WLOG, suppose the last $k+1-n$ elements of X are linear combinations of the first n elements. (If this doesn't hold, reconstruct X by putting the n linearly independent components of X at the top). Then using the same argument as in lecture 11, we can show that $X'\beta$ depends only on the first n elements of X by substituting out the the other components of X. Denote this sub-vector of X by $X_n := (X_1, \ldots, X_n)$. This vector is not linearly dependent, so $\mathrm{E}(X_n X_n')$ is full rank. Now it follows that for each $\beta \in \mathbb{R}^{k+1}$ there exists $\tilde{\beta} \in \mathbb{R}^n$ such that

$$\mathrm{E}(Y - X'\beta)^2 = \mathrm{E}\left(Y - X_n'\tilde{\beta}\right)^2.$$

We know the minimizer of $\mathrm{E}\left(Y - X_n'\tilde{\beta}\right)^2$ is given by

$$\tilde{\beta} = \mathrm{E}(X_n X_n')^{-1}\mathrm{E}(X_n Y),$$

and satisfies

$$\mathrm{E}\left(X_n\left(Y - X_n'\tilde{\beta}\right)\right) = 0.$$

Finally, this implies that the minimum of $\mathrm{E}(Y - X'\beta)^2$ is attained by setting

$$\beta = \left(\tilde{\beta}', \underbrace{0, \ldots, 0}_{k+1-n \text{ times.}}\right)',$$

from which it follows that

$$\mathrm{E}(X(Y - X'\beta)) = 0.$$

b) Given Y, X, can we always find β, U such that

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

If yes, does the additional assumption restrict β? If no, provide a counterexample.

ANS: No. Let $X \sim \mathcal{N}(0,1)$ and suppose $Y = X^2$. Then $\mathrm{E}(U|X) = \mathrm{E}\left(X^2 - X'\beta|X\right) = X^2 - X'\beta$. For any β, $\mathrm{P}(\mathrm{E}(U|X) = 0) = 0$.

c) Suppose we drop the assumption of linearity, and instead write the model as

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

Given Y, X, can we always find a function h and random variable U such that the above holds? If yes, what is the function h? If no, provide a counterexample.

ANS: Yes: let $h(X) = \mathrm{E}(Y|X)$ and $U = Y - \mathrm{E}(Y|X)$. Then $Y = h(X) + U$ and $\mathrm{E}(U|X) = \mathrm{E}(Y|X) - \mathrm{E}(\mathrm{E}(Y|X)|X) = 0$.

d) Suppose X only takes two values. Show that the best linear predictor of Y given X is also the best predictor of Y given X (under square loss).

ANS: Let $X \in \{x_0, x_1\}$, where x_0 and x_1 denote the two possible values of X. The best predictor of Y given X is defined by

$$g^* \in \underset{g \in L^2(X)}{\arg\min}\, \mathrm{E}(Y - g(X))^2.$$

Now note that since X takes only two values, any function $g(X)$ also takes at most 2 values. It follows that g^* takes the form:

$$g^*(X) = g^*(x_0) + (g^*(x_1) - g^*(x_0)) \frac{(X - x_0)}{x_1 - x_0},$$

which is a linear function of X.

Question 6 Suppose we observe an iid sample of $\{Y_i, D_i\}_{i=1}^n$ where Y_i is the scrap rate of firm i, and

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

Define the potential outcomes Y_{i0}, Y_{i1} as

$$\begin{aligned} Y_{i0} &= \text{scrap rate of firm } i \text{ without grant,} \\ Y_{i1} &= \text{scrap rate of firm } i \text{ with grant.} \end{aligned}$$

a) Show that we can write

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

where $\mathrm{E}(U_i|D_i) = 0$. Write down what β_0, β_1 and U_i are explicitly in terms of Y_{i0}, Y_{i1}, D_i.

ANS: From Question 5d) we know that the best predictor is also the best linear predictor since D_i takes only 2 values. We also know that the best predictor under square loss is the conditional expectation $\mathrm{E}(Y_i|D_i)$. Therefore, we choose $\beta_0 = \mathrm{E}(Y_i|D_i = 0) = \mathrm{E}(Y_{i0}|D_i = 0)$ and $\beta_1 = \mathrm{E}(Y_{i1}|D_i = 1) - \mathrm{E}(Y_{i0}|D_i = 0)$. This gives

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

It follows that

$$\begin{aligned} U_i &= Y_i - \beta_0 - \beta_1 D_i \\ &= Y_i - \mathrm{E}(Y_i|D_i) \end{aligned}$$

satisfies $\mathrm{E}(U_i|D_i = 0)$. Finally, substitute $Y_i = Y_{i0}(1 - D_i) + Y_{i1} D_i$.

b) How does your answer change if D_i is independent of Y_{i0} and Y_{i1}? Does the model have a causal interpretation?

ANS: β_0 becomes $\mathrm{E}(Y_{i0})$, and β_1 becomes $\mathrm{E}(Y_{i1} - Y_{i0})$. Now U_i still satisfies $\mathrm{E}(U_i|D_i) = 0$, but the model does not have a causal interpretation necessarily, because the effect of the grant on firm i is $Y_{i1} - Y_{i0}$, which is not necessarily equal to β_1, the average treatment effect in the population of firms.

c) Show that if $Y_{i1} - Y_{i0}$ is constant, in addition to the randomization assumption in part b), then we can write

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

where U_i is independent of D_i. Write down what β_0, β_1 and U_i are explicitly in terms of Y_{i0}, Y_{i1}, D_i. Is the model causal now?

ANS: β_0 remains equal to $\mathrm{E}(Y_{i0})$, but now $\beta_1 = \mathrm{E}(Y_{i1} - Y_0) = Y_{i1} - Y_{i0}$ because $Y_{i1} - Y_{i0}$ is constant. The model is causal, because β_1 represents the effect of the grant on each firm's scrap rate. If firm i is switched from not receiving a grant to receiving one, its effect on their scrap rate will equal β_1. Finally,

$$\begin{aligned} U_i &= Y_i - \beta_0 - \beta_1 D_i \\ &= Y_{i0}(1 - D_i) + Y_{i1} D_i - \mathrm{E}(Y_{i0}) - (Y_{i1} - Y_{i0}) D_i \\ &= Y_{i0} - \mathrm{E}(Y_{i0}). \end{aligned}$$

Since D_i is assumed to be independent of Y_{i0}, U_i is independent of D_i.

d) Suppose you interpret

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

causally to begin with. Is it the case that $\mathrm{E}(U_i) = \mathrm{E}(D_i U_i) = 0$? Hint: Is the best linear predictor always equal to the causal effect? See lecture 11.

ANS: No. $\mathrm{E}(U_i) = \mathrm{E}(D_i U_i) = 0$ if and only if (β_0, β_1) also represents a best linear predictor of Y given D. Suppose there is a constant positive treatment effect: $Y_{i1} - Y_{i0} = \beta_1 > 0$. In general, if (b_0, b_1) represents a best linear predictor,

$$\begin{aligned} b_1 &= \mathrm{E}(Y_{i1}|D_i = 1) - \mathrm{E}(Y_{i0}|D_i = 0) \\ &= \mathrm{E}(Y_{i1} - Y_{i0}|D_i = 1) \\ &+ \mathrm{E}(Y_{i0}|D_i = 1) - \mathrm{E}(Y_{i0}|D_i = 0) \\ &= \beta_1 + \underbrace{\mathrm{E}(Y_{i0}|D_i = 1) - \mathrm{E}(Y_{i0}|D_i = 0)}_{\text{Selection Bias}}. \end{aligned}$$

This calculation implies that $b_1 \neq \beta_1$ in general, unless $\mathrm{E}(Y_{i0}|D_i = 1) - \mathrm{E}(Y_{i0}|D_i = 0) = 0$. We discussed in Lecture 11 how differences in Y_0 across treated and control units may lead to a negative b_1, even though β_1, the causal effect, is positive.

Question 7 Let $\{X_i\}_{i\geq 1}$ be an iid sequence such that $X_i \sim \mathit{Bernoulli}(q)$. That is:

$$X_i = \begin{cases} 1 & \text{with probability } q, \\ 0 & \text{with probability } 1-q. \end{cases}$$

In Problem Set 3, we showed that

$$\frac{\sqrt{n}\left(\bar{X}_n - q\right)}{\sqrt{\bar{X}_n\left(1 - \bar{X}_n\right)}} \xrightarrow{d} \mathcal{N}(0,1), \tag{4}$$

and that

$$\mathrm{P}_q\left(\bar{X}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}} < q < \bar{X}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}}\right) \to 1-\alpha, \tag{5}$$

where $z_{1-\frac{\alpha}{2}}$ is the $1-\alpha$ quantile of the normal distribution.

a) Explain how to conduct a test of asymptotic size α of the null hypothesis $H_0 : q = 0.4$ vs. the alernative $H_1 : p \neq 0.4$ by using result (4) or by using result (5). Provide asymptotic justification.

ANS: To use (4), note that under H_0:

$$T_n = \frac{\sqrt{n}\left(\bar{X}_n - 0.4\right)}{\sqrt{\bar{X}_n\left(1 - \bar{X}_n\right)}} \xrightarrow{d} \mathcal{N}(0,1),$$

so we can construct a test of size α by defining

$$\phi_n(X_1, \ldots, X_n) = \mathbf{1}\left(|T_n| > z_{1-\frac{\alpha}{2}}\right).$$

We have under H_0:

$$\mathrm{P}\left(\phi_n\left(X_1,\ldots,X_n\right)=1\right)=\mathrm{P}\left(\left|\frac{\sqrt{n}\left(\bar{X}_n-0.4\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}\right|>z_{1-\frac{\alpha}{2}}\right)$$
$$\to\alpha,$$

because

$$\left|\frac{\sqrt{n}\left(\bar{X}_n-0.4\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}\right|\xrightarrow{d}\left|\mathcal{N}\left(0,1\right)\right|.$$

To use (5), simply check whether 0.4 lies in the confidence interval

$$\left[\bar{X}_n-z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}},\bar{X}_n+z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}}\right].$$

The probability that this occurs when H_0 is true is equal to α, since (4) gives:

$$\mathrm{P}\left(-z_{1-\alpha/2}<\frac{\sqrt{n}\left(\bar{X}_n-0.4\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}<z_{1-\frac{\alpha}{2}}\right)\to 1-\alpha.$$

b) Define the power function for this test, $\beta_n\left(q\right)$. (You are not required to explicitly compute it).

ANS:

$$\beta_n\left(q\right):=\mathrm{P}_q\left(\left|\frac{\sqrt{n}\left(\bar{X}_n-0.4\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}\right|>z_{1-\frac{\alpha}{2}}\right).$$

c) Show explicitly that as $n\to\infty$:

$$\beta_n\left(q\right)\to\begin{cases}\alpha & \text{if } q=0.4,\\ 1 & \text{if } q\neq 0.4.\end{cases}$$

ANS: If $q=0.4$, then

$$\left|\frac{\sqrt{n}\left(\bar{X}_n-0.4\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}\right|\xrightarrow{d}\left|\mathcal{N}\left(0,1\right)\right|,$$

from which it follows that

$$\beta_n\left(0.4\right)\to\alpha.$$

If $q \neq 0.4$, then

$$\frac{\sqrt{n}\left(\bar{X}_n - 0.4\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}} = \frac{\sqrt{n}\left(\bar{X}_n - q\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}} - \frac{\sqrt{n}\left(0.4 - q\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}.$$

It follows that

$$\begin{aligned}
\beta_n(q) &= \mathrm{P}_q\left(\left|\frac{\sqrt{n}\left(\bar{X}_n - q\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}} - \frac{\sqrt{n}\left(0.4 - q\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}\right| > z_{1-\frac{\alpha}{2}}\right) \\
&\geq \mathrm{P}_q\left(\left|\frac{\sqrt{n}\left(0.4 - q\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}\right| - \left|\frac{\sqrt{n}\left(\bar{X}_n - q\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}\right| > z_{1-\frac{\alpha}{2}}\right) \\
&= \mathrm{P}_q\left(\left|\frac{\left(\bar{X}_n - q\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}\right| < \left|\frac{\left(0.4 - q\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}\right| - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right),
\end{aligned}$$

where the inequality holds because $|a-b| \geq ||a|-|b||$. Since

$$\left|\frac{\left(\bar{X}_n - q\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}\right| \xrightarrow{d} 0,$$

$$\left|\frac{\left(0.4 - q\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}\right| - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \xrightarrow{a.s.} \frac{|0.4-q|}{\sqrt{q\left(1-q\right)}} > 0$$

it follows that $\beta_n(q) \to 1$, because $\frac{|0.4-q|}{\sqrt{q(1-q)}}$ is a continuity point of the distribution of 0.

Question 8 (Computational question) Consider the model

$$y = x'\beta + u; \qquad \mathrm{E}\left(xu\right) = 0$$

where y is a scalar, $x = (1, x_1, x_2, x_3) \in \mathbb{R}^4$ and u is an error term.

a) Suppose (y, x_1, x_2, x_3) are drawn from a multivariate normal distribution. Choose and report the parameters of this distribution. Simulate $n = 1000$ draws from this distribution and run a regression of y on x. Do not use special commands designed for linear regression. Construct the OLS estimator yourself and report it. Verify that $\sum_{i=1}^{n} x_i \hat{u}_i = 0$, where $\hat{u}_i$ is the i-th residual.

b) Now suppose $y \sim \mathcal{N}(1, 4)$, $x_1 \sim t_3$, $x_2 \sim \chi_2^2 - 2$ and $x_3 \sim U[-1, 1]$. What other information do you need to actually simulate draws from the joint distribution of (y, x_1, x_2, x_3)? Argue that it is possible if y, x_1, x_2, x_3 are independent of each other.

ANS: We know the marginal distribution of each component of $(y, 1, x_1, x_2, x_3)$ but not the joint distribution, which is required to simulate the data. If the components of the distribution are independent of each other, the joint distribution is given by the product of the marginals, which are known, so we can simulate it.

c) Suppose y, x_1, x_2, x_3 from part b) are all independent of each other. Now repeat part a). What do you notice about the coefficient estimates on (x_1, x_2, x_3)? Can you justify this theoretically?

ANS: Since y is independent of x,

$$\mathrm{E}\left(y|x\right) = \mathrm{E}\left(y\right) = \beta_0 + \mathrm{E}\left[(x_1, x_2, x_3)\right]\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\end{pmatrix} = \beta_0,$$

because each x_j has mean 0. So, $\beta_0 = \mathrm{E}\left(y\right)$ and $\beta_j = 0$ for $j = 1, 2, 3$, so we expect to see $\hat{\beta}_0 \approx 1, \hat{\beta}_j \approx 0$ for $j = 1, 2, 3$ in large samples. Note that

$$\mathrm{E}\left(u^2\right) = \mathrm{E}\left(y - \mathrm{E}\left(y\right)\right)^2 = 4.$$

The asymptotic variance is given by

$$\sigma^2 \mathrm{E}\left(xx'\right)^{-1} = \sigma^2 \begin{pmatrix}1 & 0 & 0 & 0\\ 0 & \mathrm{E}\left(x_1^2\right)^{-1} & 0 & 0\\ 0 & 0 & \mathrm{E}\left(x_2^2\right)^{-1} & 0\\ 0 & 0 & 0 & \mathrm{E}\left(x_3^2\right)^{-1}\end{pmatrix} = 4\begin{pmatrix}1 & 0 & 0 & 0\\ 0 & \frac{1}{3} & 0 & 0\\ 0 & 0 & \frac{1}{4} & 0\\ 0 & 0 & 0 & 3\end{pmatrix}.$$

A "back of the envelope calculation" gives:

$$\mathrm{P}\left(\left|\hat{\beta} - \beta\right| > \epsilon\right) = \mathrm{P}\left(\left|\sqrt{n}\left(\hat{\beta} - \beta\right)\right| > \epsilon\sqrt{n}\right) \leq \frac{\sum_{j=0}^{3} Var\left(\sqrt{n}\hat{\beta}_j\right)}{\epsilon^2 n} \approx 4\frac{1 + \frac{1}{3} + \frac{1}{4} + 3}{\epsilon^2 n} < \frac{1}{100\epsilon^2}.$$

So we estimate the probability that $\hat{\beta}$ is a distance further than 1 in absolute value from β is less than 0.01.