Ann Marie Weideman

2017 Qual Section 2

(2)

2. (25 points) Suppose that the pair (X, Y) is distributed such that $X \sim$ normal$(0, 1)$ and $Y \sim$ Bernoulli$(\theta), 0 < \theta < 1$. For example, X could be the log of the level of a certain biomarker and Y an indicator of some disease. Assume that the value of θ is known and given (e.g. we know that the disease prevalence is 0.001).

Here we try to answer the following question: Given the above specifications, what is the largest possible correlation between X and Y?

Notation: Define $\rho = \text{corr}(X, Y)$. Use $\phi(\cdot)$ and $\Phi(\cdot)$ to denote the standard normal pdf and cdf, respectively. Define any new notation you use.

(a) (1 point) Is $\rho = 1$ possible? Explain.

(b) (6 points) Find the joint distribution for (X,Y) that maximizes corr(X,Y) subject to the model stated above. Show that it (that distribution) has the property that $E[X|Y = 1] = \theta^{-1}\phi(\Phi^{-1}(1 - \theta))$.

(c) (6 points) Obtain an explicit expression for $\rho^* = \text{corr}(X, Y)$ within the joint distribution found in the previous part. Compute the numerical value of ρ^* for the case $\theta = 0.001$.

(d) (6 points) Now suppose we are interested in various diseases with different prevalences (θ) ranging in $(0, 1)$. Find the value of θ that leads to the largest possible value of ρ^*, and compute that largest value, to be denoted ρ^{**} (compute its numerical value).

(e) (3 points) Note: This part is totally independent of the previous parts, even though the models have some similarity. You can reuse results obtained above if needed.
Suppose that the iid pairs $(X_i, Y_i), i = 1, \ldots, n$, are distributed such that $X_i \sim$ normal$(0, 1)$, $Y_i \sim$ Bernoulli$(\theta), 0 < \theta < 1$, and $\text{corr}(X_i, Y_i) = \rho$, where both θ and ρ are unknown parameters. Develop an estimating equation for (ρ, θ) based on the vectors $Z_i := (T_i, Y_i)^\top, i = 1, \ldots, n$, where $T_i := X_i Y_i$. Obtain the estimates $(\hat{\rho}, \hat{\theta})$ in explicit form.

(f) (3 points) Based on $\hat{\rho}$ from the previous part, develop a large-sample (as $n \to \infty$) 95% confidence interval for ρ. The interval should not depend on any unknown parameters. Describe and justify your procedure clearly.

Note: $\Phi(-3.09) \approx 0.001$, $\Phi(-2.33) \approx 0.01$, $\Phi(-1.96) \approx 0.025$, $\Phi(-1.64) \approx 0.05$, $\Phi(-1.28) \approx 0.1$

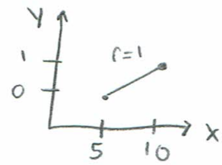
2017 Qual, Section 2 AMW

2a) Is $\rho = 1$ possible? Explain

⌈ Given $X \sim N(0,1)$ and $Y \sim Bern(\theta)$, $0 < \theta < 1$.

Take the following example:

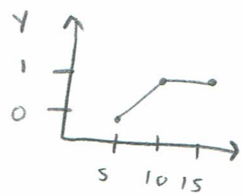
$X = (5, 10)$, $Y = (0, 1)$



For any two points with positive slope, you can connect these two points w/ a line of perfect fit and hence have $r = 1$.

However, for three or more points this is not possible b/c $Y \sim Bern(\theta)$.

Take, for example, $X = (5, 10, 15)$; $Y = (0, 1, 1)$.



As can be seen by this example,

Since $Y \sim Bern(\theta)$, there will never be a case where $r = 1$ b/c we can never create a linear fit between X and Y that perfectly fits all the data. ⌋

2017 Qual, Section 2 AMW

2b) Find the joint distn. for (X,Y) that maximizes $\mathrm{corr}(X,Y)$ subject to the model stated above. Show that the chosen distribution has the property that $E[X|Y=1] = \theta^{-1}\phi(\Phi^{-1}(1-\theta))$.

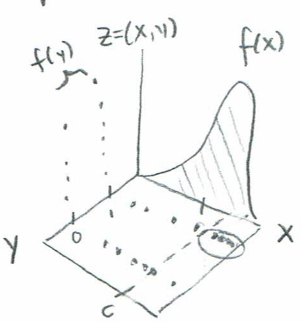
$$\mathrm{Corr}(X,Y) = \rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{E[XY] - \overbrace{E[X]}^{0}\cdot\overbrace{E[Y]}^{\theta}}{\sqrt{1\cdot\theta(1-\theta)}} = \frac{\overbrace{E[XY]}^{\text{want to maximize this}}}{\underbrace{\sqrt{\theta(1-\theta)}}_{\text{fixed}}}$$

$$\text{Maximize } E[XY] = E[XY|Y=0]\cdot P(Y=0) + E[XY|Y=1]\cdot P(Y=1)$$

$$= 0\cdot \overbrace{E[X|Y=0]}^{0}\cdot\underbrace{P(Y=0)}_{(1-\theta)} + 1\cdot E[X|Y=1]\cdot\underbrace{P(Y=1)}_{\theta}$$

$$= \theta\,\underbrace{E[X|Y=1]}_{\text{Need to maximize this}}$$

Notice that the joint pmf of (X,Y) is concentrated on the lines $Y=0$ & $Y=1$, like in this picture. Notice that, in order to maximize $E[X|Y=1]$ (the expected value of X given that $Y=1$),



there is some value of $X > c$ ∋ the positive density of $f_X(x)$ is maximized (so that the mean is greatest).

Thus, need a value of c such that we have $\{(x,y): x<c, y=0 \text{ \& } x>c, y=1\}$

Thus, c must satisfy $\underbrace{P(Y=1)}_{\theta} = P(\overbrace{X}^{\sim N(0,1)} > c)$

$\Rightarrow \theta = 1 - P(X \le c) \Rightarrow \theta = 1-\Phi(c) \Rightarrow \theta - 1 = -\Phi(c) \Rightarrow 1-\theta = \Phi(c)$

$\Rightarrow c = \Phi^{-1}(1-\theta)$.

Since $E[X|Y=1] = E[X|X>c] = \int_x x\cdot P(X|X>c)\,dx = \int_x x\cdot\frac{P(X, X>c)}{\underbrace{P(X>c)}_{P(Y=1)=\theta}}dx$

$= \int_x x\cdot\frac{P(x)\cdot I(x>c)}{\theta}dx = \int_c^\infty x\cdot\frac{\overbrace{P(x)}^{=\phi(x)}}{\underbrace{\theta}_{=1-\Phi^{-1}(c)}}dx = \int_c^\infty \frac{x\,\phi(x)\,dx}{1-\Phi^{-1}(c)}$

$= \frac{1}{1-\Phi^{-1}(c)}\int_c^\infty x\,\phi(x)\,dx$ (*)

Note that $\frac{d}{dx}(\phi(x)) = \frac{d}{dx}\left[\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\right] = \frac{-x}{\sqrt{2\pi}}e^{-x^2/2} = -x\,\phi(x)$

$\Rightarrow \frac{d\phi(x)}{dx} = -x\,\phi(x) \Rightarrow d\,\phi(x) = -x\,\phi(x)\,dx$, Sub into (*) to get:

$$\boxed{E[X|Y=1] = \frac{1}{1-\Phi^{-1}(c)}\int_{-\infty}^{c} d\phi(x)} = \frac{\phi(c)}{1-\Phi^{-1}(c)} = \boxed{\frac{\phi(\Phi^{-1}(1-\theta))}{\theta}}$$ ✓

where $\boxed{\mathrm{Corr}(X,Y) \text{ is maximized for } f(x,y) = \phi(x)\,I(x>\Phi^{-1}(1-\theta)) + \phi(x)\,I(x<\Phi^{-1}(1-\theta))}$

2c) Obtain an explicit expression for $\rho^* = \text{Corr}(X, Y)$ within the joint distribution found in the previous part. Compute the numerical value of ρ^* for $\theta = 0.001$.

From last part, know that the chosen distribution has the property that

$E[X|Y=1] = \frac{1}{\theta}\phi(\Phi^{-1}(1-\theta))$.

part b)

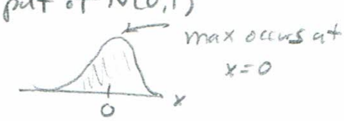
Then, since $\text{Corr}(X,Y) = \rho^* = \frac{\theta E[X|Y=1]}{\sqrt{\theta(1-\theta)}} = \frac{\cancel{\theta} \cdot \frac{1}{\cancel{\theta}}\phi(\Phi^{-1}(1-\theta))}{\sqrt{\theta(1-\theta)}}$

$\Rightarrow \boxed{\text{Corr}(X,Y) = \rho^* = \frac{\phi(\Phi^{-1}(1-\theta))}{\sqrt{\theta(1-\theta)}}}$

For $\theta = 0.001$, have $\text{Corr}(X,Y) = \rho^* = \frac{\phi(\Phi^{-1}(1-0.001))}{\sqrt{0.001(1-0.001)}} = \boxed{\frac{\phi(3.09)}{\sqrt{0.001(1-0.001)}}} = ?$ Can't do that shit in my head. I'm not a calculator!

2d) Now, suppose we are interested in various diseases w/ different prevalences ranging from $(0,1)$. Find the value of θ that leads to the largest possible value of ρ^* and label it ρ^{**} (compute numerical value).

From part c), $\rho^* = \frac{\phi(\Phi^{-1}(1-\theta))}{\sqrt{\theta(1-\theta)}} = \frac{\phi(c)}{\sqrt{\theta(1-\theta)}}$. Note that $\phi(x)$ is the pdf of $N(0,1)$



$\Rightarrow \phi(x)$ is maximized at $x=0$. $\Rightarrow c=0 \Rightarrow \Phi^{-1}(1-\theta)=0 \Rightarrow 1-\theta = \Phi(0) \Rightarrow \theta = 1-\underbrace{\Phi(0)}_{0.5}$

$\Rightarrow \theta = 0.5$

Thus, $\rho^{**} = \frac{\phi(0)}{\sqrt{0.5(1-0.5)}} = \frac{1}{2}\phi(0) = \frac{1}{2}\cdot\frac{1}{\sqrt{2\pi}}e^0 = \boxed{\sqrt{\frac{2}{\pi}}}$

AMW

2017 Qual, Section 2

2e) Suppose iid pairs (X_i, Y_i), $i=1,\dots,n$ $\ni$ $X_i \sim N(0,1)$ and $Y_i \sim \text{Bern}(\theta)$, $0<\theta<1$.

Given $\text{Corr}(X_i, Y_i) = \rho$, where both θ and ρ are unknown parameters.

Want to develop an estimating eqn. for (ρ, θ) based on $Z_i = (T_i, Y_i)^T = (X_i Y_i, Y_i)^T$.

Obtain estimates $(\hat{\rho}, \hat{\theta})$ in explicit form.

$$\Gamma \quad \rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{E[XY]-E[X]E[Y]}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}} \approx \frac{\frac{1}{n}\sum_i X_i Y_i}{\sqrt{\frac{1}{n}\sum_i Y_i\left(1-\frac{1}{n}\sum_i Y_i\right)}} = \frac{\bar{T}}{\sqrt{\bar{Y}(1-\bar{Y})}}$$

Thus, $\hat{\rho} = \dfrac{\bar{T}}{\sqrt{\bar{Y}(1-\bar{Y})}}$ where $\hat{\theta} = \bar{Y}$. $\lrcorner$

2017 Qual, Section 2 AMW

2f) Based on $\hat{\rho}$ from the previous part, develop a large sample (as $n \to \infty$) 95% CI for ρ.

NOTE: The interval should not depend on any unknown parameters.

$$95\%\ CI(\rho) = \left(\hat{\rho} - 1.96\sqrt{Var(\hat{\rho})},\ \hat{\rho} + 1.96\sqrt{Var(\hat{\rho})}\right)$$

From the previous part, know that $\hat{\rho} = \dfrac{\bar{T}}{\sqrt{\bar{Y}(1-\bar{Y})}}$. Will use delta method on $\left(\frac{\bar{T}}{\bar{Y}}\right)$ to derive asymptotic variance of $\hat{\rho}$.

Find dist. of $\bar{T}$: Know by CLT that $\sqrt{n}(\bar{T}_n - E[T_i]) \xrightarrow{d} N(0, Var[T_i])$

Find dist of $\bar{Y}$: Know by CLT that $\sqrt{n}(\bar{Y}_n - \underbrace{E[Y_i]}_{\theta}) \xrightarrow{d} N(0, \underbrace{Var[Y_i]}_{\theta(1-\theta)})$

Then, by multivariate CLT, have

$$\sqrt{n}\left(\begin{pmatrix}\bar{T}_n \\ \bar{Y}_n\end{pmatrix} - \begin{pmatrix}E[T_i] \\ \theta\end{pmatrix}\right) \xrightarrow{d} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \overbrace{\begin{pmatrix}Var(T_i) & \rho\sqrt{Var(T_i)\,\theta(1-\theta)} \\ '' \quad '' & \theta(1-\theta)\end{pmatrix}}^{\Sigma}\right)$$

since $\rho = \dfrac{Cov(T_i, Y_i)}{\sqrt{Var(T_i)\,Var(Y_i)}} \Rightarrow Cov(T_i, Y_i) = \ldots$

Then, by delta method, $\sqrt{n}\left(g(\bar{T}_n, \bar{Y}_n) - g(\underbrace{E[T_i]}_{a}, \underbrace{\theta}_{b})\right) \xrightarrow{d} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \nabla g^T \Sigma \nabla g\right)$

where $g(a,b) = a(b(1-b))^{-1/2} \Rightarrow \nabla g(a,b) = \begin{pmatrix}\frac{d}{da}\left[a(b(1-b))^{-1/2}\right] \\ \frac{d}{db}\left[a(\underbrace{b(1-b)}_{b-b^2})^{-1/2}\right]\end{pmatrix} = \begin{pmatrix}\frac{1}{\sqrt{b(1-b)}} \\ \frac{-a(1-2b)}{2[b(1-b)]^{3/2}}\end{pmatrix}$

$$= \begin{pmatrix}\frac{1}{\sqrt{\theta(1-\theta)}} \\ \frac{-E[T_i](1-2\theta)}{2[\theta(1-\theta)]^{3/2}}\end{pmatrix}$$

Then, $\sqrt{n}\left(\dfrac{\bar{T}}{\sqrt{\bar{Y}(1-\bar{Y})}} - \dfrac{E[T_i]}{\sqrt{\theta(1-\theta)}}\right) \xrightarrow{d} N\left(0, \underbrace{\underbrace{\begin{pmatrix}\frac{1}{\sqrt{\theta(1-\theta)}} & \frac{-E[T_i](1-2\theta)}{2[\theta(1-\theta)]^{3/2}}\end{pmatrix}}_{1\times2} \underbrace{\begin{pmatrix}Var(T_i) & \rho\sqrt{Var(T_i)\theta(1-\theta)} \\ \rho\sqrt{Var(T_i)\theta(1-\theta)} & \theta(1-\theta)\end{pmatrix}}_{2\times2} \underbrace{\begin{pmatrix}\frac{1}{\sqrt{\theta(1-\theta)}} \\ \frac{-E[T_i](1-2\theta)}{2[\theta(1-\theta)]^{3/2}}\end{pmatrix}}_{2\times1}}_{\tilde{\sigma}^2}\right)$

cont'd →

AMW

2f cont'd.

where $\tilde{\sigma}^2 =$

$$\left[\left(\frac{Var(T_i)}{\sqrt{\theta(1-\theta)}} - \frac{E[T_i](1-2\theta)\rho\sqrt{Var(T_i)\cancel{\theta(1-\theta)}}}{2[\theta(1-\theta)]^{3/2}}\right)\left(\frac{\rho\sqrt{Var(T_i)\cancel{\theta(1-\theta)}}}{\sqrt{\cancel{\theta(1-\theta)}}} \quad \frac{-E[T_i](1-2\theta)(\cancel{\theta(1-\theta)})}{2[\cancel{\theta(1-\theta)}]^{3/2}_{[\theta(1-\theta)]^{1/2}}}\right)\right]\begin{pmatrix}\frac{1}{\sqrt{\theta(1-\theta)}} \\ \frac{-E[T_i](1-2\theta)}{2[\theta(1-\theta)]^{3/2}}\end{pmatrix}$$

$$= \frac{Var(T_i)}{(\theta(1-\theta))} - \frac{E[T_i](1-2\theta)\rho\sqrt{Var(T_i)}}{2[\theta(1-\theta)]^{3/2}} - \frac{E[T_i](1-2\theta)\rho\sqrt{Var(T_i)}}{2[\theta(1-\theta)]^{3/2}} + \frac{E[T_i]^2(1-2\theta)^2}{4[\theta(1-\theta)]^2}$$

$$= \frac{4Var(T_i)[\theta(1-\theta)] + E[T_i]^2(1-2\theta)^2}{4(\theta(1-\theta))^2} - \frac{2E[T_i](1-2\theta)\rho\sqrt{Var(T_i)}}{2[\theta(1-\theta)]^{3/2}}$$

Replace θ w/ $\bar{Y}$, $E[T_i]$ w/ $\bar{T}$, $Var(T_i) = \underbrace{E[T_i^2] - E[T_i]^2}$ w/ $\underbrace{\frac{1}{n}\sum_i T_i^2 - \left(\frac{1}{n}\sum_i T_i\right)^2}_{\approx} = \frac{1}{n}\sum_i T_i^2 - \bar{T}^2$

and ρ w/ $\frac{\bar{T}}{\sqrt{\bar{Y}(1-\bar{Y})}}$, we get

$$\underbrace{\tilde{\sigma}^2}_{\text{asymptotic variance of } \hat{\rho}} = \frac{4\left[\frac{1}{n}\sum_i T_i^2 - \bar{T}^2\right][\bar{Y}(1-\bar{Y})] + \bar{T}^2(1-2\bar{Y})^2}{4(\bar{Y}(1-\bar{Y}))^2} - \frac{2\bar{T}(1-2\bar{Y})\frac{\bar{T}}{\sqrt{\bar{Y}(1-\bar{Y})}}\sqrt{\frac{1}{n}\sum_i T_i^2 - \bar{T}^2}}{2[\bar{Y}(1-\bar{Y})]^{3/2}}$$

Then, $95\%\ CI(\rho) = \left(\hat{\rho} - 1.96\sqrt{\tilde{\sigma}^2},\ \hat{\rho} + 1.96\sqrt{\tilde{\sigma}^2}\right) = \boxed{\left(\hat{\rho} - 1.96\cdot\tilde{\sigma},\ \hat{\rho} + 1.96\,\tilde{\sigma}\right)}$