36-402 Homework 9

Eu Jing Chua
eujingc

9 April, 2019

Question 1

a)

$$\begin{aligned}
\text{Var}[Y] &= \text{Var}[\alpha X + \epsilon] \\
&= \alpha^2 \text{Var}[X] + \sigma^2 = \alpha^2 + \sigma^2 \\
\text{Var}[Z] &= \text{Var}[\beta_1 X + \beta_2 Y + \eta] \\
&= \text{Var}[\beta_1 X + \alpha\beta_2 X + \beta_2 \epsilon + \eta] \\
&= \text{Var}[(\beta_1 + \alpha\beta_2) X + \beta_2 \epsilon + \eta] \\
&= (\beta_1 + \alpha\beta_2)^2 + \beta_2^2 \sigma^2 + \sigma^2 \\
&= (\beta_1 + \alpha\beta_2)^2 + (\beta_2^2 + 1)\sigma^2
\end{aligned}$$

b)

$X \to Y$ is open when conditioned on nothing.
$X \to Z \leftarrow Y$ is open when conditioned on Z.

c)

$X \to Z$ and $X \to Y \to Z$ are open when conditioned on nothing.
$X \to Z$ is open when conditioned on Y.

d)

$$\begin{aligned}
\text{Cov}[X, Y] &= \text{Var}[X]\,\alpha = \alpha \\
\text{Cov}[X, Z] &= \text{Var}[X]\,\beta_1 + \text{Var}[X]\,\alpha\beta_2 = \beta_1 + \alpha\beta_2 \\
\text{Cov}[Y, Z] &= \text{Var}[X]\,\alpha\beta_1 + \text{Var}[Y]\,\beta_2 \\
&= \alpha\beta_1 + (\alpha^2 + \sigma^2)\beta_2
\end{aligned}$$

e)

The population coefficient for Y against X is:

$$\frac{\text{Cov}[X, Y]}{\text{Var}[X]} = \alpha$$

f)

The population coefficient for Z against X is:

$$\frac{\text{Cov}[X, Z]}{\text{Var}[X]} = \beta_1 + \alpha\beta_2$$

g) The population coefficient for Z against X and Y is:

$$
\begin{aligned}
c &= \begin{bmatrix} \beta_1 + \alpha\beta_2 \\ \alpha\beta_1 + (\alpha^2+\sigma^2)\beta_2 \end{bmatrix} \\
v &= \begin{bmatrix} 1 & \alpha \\ \alpha & (\alpha^2+\sigma^2) \end{bmatrix} \\
v^{-1} &= \frac{1}{\sigma^2}\begin{bmatrix} (\alpha^2+\sigma^2) & -\alpha \\ -\alpha & 1 \end{bmatrix} \\
v^{-1}c &= \frac{1}{\sigma^2}\begin{bmatrix} (\alpha^2+\sigma^2) & -\alpha \\ -\alpha & 1 \end{bmatrix}\begin{bmatrix} \beta_1 + \alpha\beta_2 \\ \alpha\beta_1 + (\alpha^2+\sigma^2)\beta_2 \end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix} (\alpha^2+\sigma^2)(\beta_1+\alpha\beta_2) - \alpha^2\beta_1 - (\alpha^2+\sigma^2)\alpha\beta_2 \\ -\alpha\beta_1 - \alpha^2\beta_2 + \alpha\beta_1 + (\alpha^2+\sigma^2)\beta_2 \end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix} (\alpha^2+\sigma^2)\beta_1 - \alpha^2\beta_1 \\ -\alpha^2\beta_2 + (\alpha^2+\sigma^2)\beta_2 \end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix} \sigma^2\beta_1 \\ \sigma^2\beta_2 \end{bmatrix} \\
&= \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}
\end{aligned}
$$

h)

$$
\begin{aligned}
c &= \begin{bmatrix} \alpha \\ \alpha\beta_1 + (\alpha^2+\sigma^2)\beta_2 \end{bmatrix} \\
v &= \begin{bmatrix} 1 & \beta_1 + \alpha\beta_2 \\ \beta_1 + \alpha\beta_2 & (\beta_1+\alpha\beta_2)^2 + (\beta_2^2+1)\sigma^2 \end{bmatrix} \\
v^{-1} &= \frac{1}{(\beta_2^2+1)\sigma^2}\begin{bmatrix} (\beta_1+\alpha\beta_2)^2 + (\beta_2^2+1)\sigma^2 & -\beta_1 - \alpha\beta_2 \\ -\beta_1 - \alpha\beta_2 & 1 \end{bmatrix}
\end{aligned}
$$

The population coefficient of X is then given by:

$$
\begin{aligned}
&\frac{1}{(\beta_2^2+1)\sigma^2}\begin{bmatrix} (\beta_1+\alpha\beta_2)^2 + (\beta_2^2+1)\sigma^2 & -\beta_1 - \alpha\beta_2 \end{bmatrix}\begin{bmatrix} \alpha \\ \alpha\beta_1 + (\alpha^2+\sigma^2)\beta_2 \end{bmatrix} \\
&= \frac{\alpha(\beta_1+\alpha\beta_2)^2 + \alpha\sigma^2(\beta_2^2+1) - \alpha\beta_1^2 - \alpha^2\beta_1\beta_2 - (\beta_1+\alpha\beta_2)(\alpha^2+\sigma^2)\beta_2}{(\beta_2^2+1)\sigma^2} \\
&= \frac{\alpha(\beta_1+\alpha\beta_2)^2 + \alpha\sigma^2(\beta_2^2+1) - \alpha\beta_1(\beta_1+\alpha\beta_2) - (\beta_1+\alpha\beta_2)(\alpha^2+\sigma^2)\beta_2}{(\beta_2^2+1)\sigma^2} \\
&= \frac{\alpha(\beta_1+\alpha\beta_2)^2 + \alpha\sigma^2(\beta_2^2+1) - (\alpha\beta_1 + (\alpha^2+\sigma^2)\beta_2)(\beta_1+\alpha\beta_2)}{(\beta_2^2+1)\sigma^2} \\
&= \frac{\alpha\sigma^2(\beta_2^2+1) + (\alpha\beta_1 + \alpha^2\beta_2 - \alpha\beta_1 - (\alpha^2+\sigma^2)\beta_2)(\beta_1+\alpha\beta_2)}{(\beta_2^2+1)\sigma^2} \\
&= \frac{\alpha\sigma^2(\beta_2^2+1) - \sigma^2\beta_2(\beta_1+\alpha\beta_2)}{(\beta_2^2+1)\sigma^2} \\
&= \frac{\alpha\beta_2^2 + \alpha - \beta_1\beta_2 - \alpha\beta_2^2}{(\beta_2^2+1)} \\
&= \frac{\alpha - \beta_1\beta_2}{(\beta_2^2+1)}
\end{aligned}
$$

Question 2

a) Open paths from X to Y.

Conditioning on nothing,

- $X \to R \to Y$
- $X \to R \to Q \to Y$
- $X \to Q \to Y$
- $X \leftarrow U \to Y$

Conditioning on U:

- $X \to R \to Y$
- $X \to R \to Q \to Y$
- $X \to Q \to Y$

Conditioning on R:

- $X \to Q \to Y$
- $X \leftarrow U \to Y$

Conditioning on Q,

- $X \to R \to Y$
- $X \leftarrow U \to Y$

Conditioning on R and Q,

- $X \leftarrow U \to Y$

Conditioning on U, R and Q, there are no open paths.

b) Open paths from R to Y.

Conditioning on nothing:

- $R \to Y$
- $R \to Q \to Y$
- $R \leftarrow X \to Q \to Y$
- $R \leftarrow X \leftarrow U \to Y$

Conditioning on X:

- $R \to Y$
- $R \to Q \to Y$

Conditioning on X and U:

- $R \to Y$
- $R \to Q \to Y$

Conditioning on X, U and Q:

- $R \to Y$

c) Open paths from X to Q.

Conditioning on nothing:

- $X \to Q$
- $X \to R \to Q$

Conditioning on Y:

- $X \to Q$
- $X \to R \to Q$
- $X \to R \to Y \leftarrow Q$
- $X \leftarrow U \to Y \leftarrow Q$

Conditioning on U:

- $X \to Q$
- $X \to R \to Q$

Conditioning on U and Y:

- $X \to Q$
- $X \to R \to Q$
- $X \to R \to Y \leftarrow Q$

Conditioning on R:

- $X \to Q$

Conditioning on R and Y:

- $X \to Q$
- $X \leftarrow U \to Y \leftarrow Q$

d) Population coefficient of X in linear regression of Y against X:

$$\begin{aligned}
\operatorname{Cov}[Y, X] &= \operatorname{Var}[X]\,\beta\delta_1 + \operatorname{Var}[X]\,\beta\gamma_2\delta_2 + \operatorname{Var}[X]\,\gamma_1\delta_2 + \operatorname{Var}[U]\,\alpha\delta_3 \\
&= (\alpha^2 + \sigma^2)(\beta\delta_1 + \beta\gamma_2\delta_2 + \gamma_1\delta_2) + \alpha\delta_3 \\
\frac{\operatorname{Cov}[Y, X]}{\operatorname{Var}[X]} &= \beta\delta_1 + \beta\gamma_2\delta_2 + \gamma_1\delta_2 + \frac{\alpha\delta_3}{\alpha^2 + \sigma^2}
\end{aligned}$$

e)

$$\begin{aligned}
\operatorname{Cov}[R, X] &= \operatorname{Var}[X]\,\beta + \operatorname{Var}[X]\,\gamma_1\gamma_2 \\
&= (\alpha^2 + \sigma^2)(\beta + \gamma_1\gamma_2) \\
\frac{\operatorname{Cov}[R]}{\operatorname{Var}[X]} &= \beta + \gamma_1\gamma_2
\end{aligned}$$

f) The coefficient for X is γ_1, and for R is γ_2.

g) The coefficient for X should not be 0, as given Q and R, there is still an open path from X to Y via $X \leftarrow U \to Y$, so X is not indepenedent on Y conditioned on Q and R.

Question 3

a) There is an open path from Amount of smoking → Amount of tar in lungs → Cellular damage → Cancer in the graph, as Cancer is a descendent of Amount of smoking.

b)

We could condition on the sets:

- {Amount of Tar in Lungs, Occupational Prestige}
- {Amount of Tar in Lungs, Asbestos Exposure}
- {Cellular Damage}

c) The only case in which conditioning on even more variables results in the possibility of an open path is when we condition on the middle variable of a collider.
However, the only collider relevant would be Amount of smoking → Yellowing of Teeth ← Access to Dental Care. Even if we open up this collider, all possible paths still have to pass through one of the sets of conditioned variables above, which block paths. Thus, this will not be possible for this graph.

d) We could condition on Occupational Prestige.
Then, additionally conditioning on Cellular Damage would make them dependent again.
Finally, we could additionally condition on Amount of smoking to make them independent again.

Question 4

a)

Table 1: Coefficient of smoking

	x
smoking	1.29

With every unit increase in smoking, we predict that the log-odds of smoking increases by 1.29.

b)

Table 2: Coefficient of smoking

	x
smoking	1.34

For a fixed level of yellowing of teeth, with every unit increase in smoking, we predict that the log-odds of smoking increases by 1.34.

c)

Table 3: Coefficient of smoking

	x
smoking	0.366

For a fixed value of asbestos exposure, with every unit increase in smoking, we predict that the log-odds of smoking increases by 0.0524.

d)

Table 4: Coefficient of smoking

	x
smoking	-0.0484

For a fixed value of all other covariates, with every unit increase in smoking, we predict that the log-odds of smoking decreases by 0.0484.

e) The regression of cancer against smoking, controlling for asbestos exposure.
By controlling for asbestos exposure, we block the backdoor path where asbestos exposure is confounding. After, there is only one direct path from smoking to cancer that is unblocked, satisfying the backdoor criterion, hence we can make causal inference from the regression.

f)

Table 5: CV errors of models

	x
Smoking	0.0940
Smoking and Asbestos	0.0786
Smoking and Teeth	0.0918
All	0.0800

As the model of cancer against smoking and asbestos exposure has the lowest cross-validation error, we expect it to have the least generalization error and thus would be the most ideal for an insurance company to predict how likely a customer is to get cancer.

Question 5

a) Amount of tar in lungs is statistically independent of Cancer conditioned on Amount of smoking in figure 4, but not figure 3.

b) There is no independence relation that holds in figure 3 but not figure 4. Assuming a independence relation between any two variables in figure 3, we cannot create dependence by removing edges only as in figure 4, as there are now less possible open paths.

c) We know that if the data came from figure 4, then Cancer is statistically independent of Tar in lungs controlling for Amount of smoking, which would result in a true logistic regression coefficient of 0. Otherwise if there is dependence, the coefficient is expected to be non-zero. We could test this with a bootstrapped confidence interval on the coefficient of Tar in lungs in the model to see if it contains 0.

Table 6: 95% C.I for tar

	lower	upper
tar	-14.4	32.3

We use case-resampling to bootstrap a 95% C.I for the coefficient of tar in lungs, and see that it does indeed contain 0. Hence, we can be 95% confident that the data came from figure 4.