Problem Set 1

Group 6

Christian Birchler
Fenqi Guo
Mingrui Zhang
Wenjie Tu
Zunhan Zhang

Basic statistical concepts

1(a).

$$\mathbb{E}(Y) = \sum_y y \cdot p(y)$$

$$\begin{aligned}
\mathbb{V}(Y) &= \mathbb{E}(Y - \mathbb{E}(Y))^2 \\
&= \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 \\
&= \sum_y y^2 \cdot p(y) - \left[\sum_y y \cdot p(y)\right]^2
\end{aligned}$$

1(b).

- $\mathbb{E}(Y)$ tells us the average value of the probability distribution of Y.
- $\mathbb{V}(Y)$ tells us the spread around the average value.

2(a).

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_y y \cdot p(y) \\
&= \sum_y y \sum_x p(x, y) \\
&= \sum_x \sum_y y \cdot p(y|x) \cdot p(x) \\
&= \sum_x \left[\sum_y y \cdot p(y|x)\right] p(x) \\
&= \sum_x \mathbb{E}(Y|X = x) p(x) \\
&= \mathbb{E}_x[\mathbb{E}(Y|X = x)]
\end{aligned}$$

2(b).

The Law of Iterated Expectations is useful when the probability distribution of X and a conditional random variable $Y|X$ are known, and the probability distribution of Y is desired.

Example:

Y	$WAGE$	Wage per hour
X	$EDUC$	Years of education
$Y\|X$	$WAGE\|EDUC$	Wage per hour given specific years of education

In practical data analysis, we have easier access to the data of *years of education* and we are interested in the unconditional *wage per hour*. We can run a simple regression on *years of education* to get the *wage per hour* conditional on *years of education*. Then the unconditional *wage per hour* can be easily calculated by applying the *Law of Iterated Expectations*.

$$\begin{aligned} \mathbb{E}(EDUC) &= 11.5 \\ \mathbb{E}(WAGE|EDUC) &= 4 + 0.6EDUC \end{aligned}$$

$$\begin{aligned} \mathbb{E}(WAGE) &= \mathbb{E}(\mathbb{E}(WAGE|EDUC)) \\ &= \mathbb{E}(4 + 0.6EDUC) \\ &= 4 + 0.6\mathbb{E}(EDUC) \\ &= 4 + 0.6 \times 11.5 \\ &= 10.9 \end{aligned}$$

3(a).

$$
\begin{aligned}
Cov(y,x) &= \mathbb{E}[(y-\mathbb{E}(y))(x-\mathbb{E}(x))] \\
&= \mathbb{E}[y\cdot x - y\mathbb{E}(x) - x\mathbb{E}(y) + \mathbb{E}(y)\mathbb{E}(x)] \\
&= \mathbb{E}(y\cdot x) - \mathbb{E}(y)\mathbb{E}(x) - \mathbb{E}(y)\mathbb{E}(x) + \mathbb{E}(y)\mathbb{E}(x) \\
&= \mathbb{E}(y\cdot x) - \mathbb{E}(y)\mathbb{E}(x)
\end{aligned}
$$

3(b).

$$
\begin{aligned}
Cov(y,x) &= \mathbb{E}[(y-\mathbb{E}(y))(x-\mathbb{E}(x))] \\
&= \mathbb{E}[(y-\mathbb{E}(y))x] - \mathbb{E}[(y-\mathbb{E}(y))\mathbb{E}(x)] \\
&= \mathbb{E}[(y-\mathbb{E}(y))x] - \mathbb{E}[y\mathbb{E}(x) - \mathbb{E}(y)\mathbb{E}(x)] \\
&= \mathbb{E}[(y-\mathbb{E}(y))x]
\end{aligned}
$$

Similarly, we can get

$$Cov(y,x) = \mathbb{E}[(x-\mathbb{E}(x))y]$$

4.

It is true that $Corr(x,y) = Corr(y,x)$. The correlation measures the degree to which two random variables are linearly related and it is normalized between -1 and 1. Therefore, it has nothing to do with the order of how two random variables enter into the formula since $Cov(x,y) = Cov(y,x)$.

The linear regression model

5(a).

Linearity assumption is already built into the structural model.

5(b).

$$\begin{aligned}
\mathbb{E}(y_i|x_i) &= \mathbb{E}(\beta_0 + \beta_1 x_i + u_i|x_i) \\
&= \mathbb{E}(\beta_0|x_i) + \mathbb{E}(\beta_1 x_i|x_i) + \mathbb{E}(u_i|x_i) \\
&= \beta_0 + \beta_1 x_i + \mathbb{E}(u_i|x_i)
\end{aligned}$$

We still need to assume $\mathbb{E}(u_i|x_i) = 0$ in order to identify β_0 and β_1 in the model. With the conditional mean-zero-error assumption, we can derive:

$$\begin{aligned}
\mathbb{E}(u_i) &= \mathbb{E}_{x_i}\mathbb{E}(u_i|x_i) \\
&= \mathbb{E}_{x_i} \cdot 0 \\
&= 0
\end{aligned}$$

$$\begin{aligned}
Cov(x_i, u_i) &= \mathbb{E}(x_i u_i) - \mathbb{E}(x_i)\mathbb{E}(u_i) \\
&= \mathbb{E}_{x_i}\mathbb{E}(x_i u_i|x_i) - 0 \\
&= \mathbb{E}_{x_i}[x_i \mathbb{E}(u_i|x_i)] \\
&= 0
\end{aligned}$$

5(c).

In 5(b), we derived $Cov(x_i, u_i) = 0$

$$\begin{aligned}
Cov(y_i, x_i) &= Cov(\beta_0 + \beta_1 x_i + u_i, x_i) \\
&= Cov(\beta_0, x_i) + Cov(\beta_1 x_i, x_i) + \underbrace{Cov(u_i, x_i)}_{\text{zero}} \\
&= \beta_1 Var(x_i) \\
&\Downarrow \\
\hat{\beta}_1 &= \frac{Cov(y_i, x_i)}{Var(x_i)}
\end{aligned}$$

In matrix notation:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (X'X)^{-1}X'y$$

5(d).

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'y \\
&= (X'X)^{-1}X'(X\beta + \epsilon) \\
&= \beta + (X'X)^{-1}X'\epsilon
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}(\hat{\beta}) &= \mathbb{E}[\mathbb{V}(\hat{\beta}|X)] + \mathbb{V}[\mathbb{E}(\hat{\beta}|X)] \\
&= \mathbb{E}[\mathbb{V}(\beta + (X'X)^{-1}X'\epsilon|X)] + \mathbb{V}[\mathbb{E}(\beta + (X'X)^{-1}X'\epsilon|X)] \\
&= \mathbb{E}[(X'X)^{-1}X'\mathbb{V}(\epsilon|X)X(X'X)^{-1}] + \mathbb{V}(\beta) \\
&= \mathbb{E}[\sigma^2(X'X)^{-1}] \\
&= \sigma^2(X'X)^{-1} \qquad \Downarrow \\
\mathbb{V}\left(\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}\right) &= \sigma^2(X'X)^{-1}
\end{aligned}$$

5(e).

$$\begin{aligned}\mathbb{V}(\hat{\beta}) &= \sigma^2(X'X)^{-1} \\ &= \frac{1}{N-1}\sigma^2(\frac{1}{N-1}X'X)^{-1}\end{aligned}$$

$\frac{1}{N-1}X'X$ is the "sample variance-covariance matrix" of X.

5(f).

Two conditions for consistency:

- $$\lim_{N\to\infty}\mathbb{E}(\tilde{\beta}) = \beta$$
- $$\lim_{N\to\infty}\mathbb{V}(\tilde{\beta}) = 0$$

Let us focus on the second condition,

$$\begin{aligned}\mathbb{V}(\hat{\beta}) &= \sigma^2(X'X)^{-1} \\ &= \frac{1}{N}\sigma^2(\frac{1}{N}X'X)^{-1}\end{aligned}$$

In order to get a consistent estimator, we need to make $\mathbb{V}(\hat{\beta})$ closer to zero as sample size N goes larger. We therefore have to assume that $\frac{1}{N}X'X$ will converge to a constant as sample size goes larger.

Assumption: **Regular X's**

$$\lim_{N\to\infty}\frac{1}{N}X'X = \mathbb{E}(X'X) \equiv \Sigma_{XX}$$

5(g).

In order to test the hypothesis that $\beta_1 = 0$, we need to construct a *t-statistic*,

$$\textit{t-statistic} = \frac{\hat{\beta}_1}{\sqrt{\mathbb{V}(\hat{\beta}_1)}}$$

This is a two-side t test. If the absolute value of *t-statistic* is larger than 1.96, we can safely reject the null hypothesis that $\beta_1 = 0$. In other words, β_1 is statistically different from zero.

6(a).

$$\begin{aligned}
\mathbb{E}(\hat{\alpha}) &= \frac{Cov(x_i, y_i)}{\mathbb{V}(x_i)} \\
&= \frac{Cov(x_i, \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i)}{\mathbb{V}(x_i)} \\
&= \beta_1 + \beta_2 \underbrace{\frac{Cov(x_i, z_i)}{\mathbb{V}(x_i)}}_{\text{non-zero}} \\
&\neq \beta_1
\end{aligned}$$

Therefore, $\hat{\alpha}_1$ is a biased estimator for the target parameter β_1.

6(b).

From 6(a), we get $\mathbb{E}(\hat{\alpha}) = \beta_1 + \beta_2 \frac{Cov(x_i, z_i)}{\mathbb{V}(x_i)}$. When $Cov(x_i, z_i) = 0$, $\hat{\alpha_1}$ is an unbiased estimator for β_1.

6(c).

y_i	the GPA of the i^{th} student
x_i	how many hours spent on study per week
z_i	innate ability

In this case, $\hat{\alpha}_1$ would be a biased estimator for β_1 since $Cov(x_i, z_i) \neq 0$. One's innate ability is an unobservable and those with higher innate ability probabily would spend less on study but still get a higher GPA.

6(d).

Bias term: $\beta_2 \frac{Cov(x_i, z_i)}{\mathbb{V}(x_i)}$ As discussed in previous question, $\mathbb{V}(x_i)$ will converge to its true value (population variance) in large sample. It also applies to $Cov(x_i, z_i)$. As $n \to \infty$, $Cov(x_i, z_i)$ will become more stable and converge to a fixed value. In conclusion, the bias term will converge to a fixed value as sample size becomes larger.