PS 236: Causal Inference
Problem Set 1

UC Berkeley, Fall 2008

Due: Thursday, September 25

Your solutions must be submitted in hard copy to my mailbox in the Political Science main office prior to the office's closure on the due date (approximately 5pm).

Question 1

A fair coin is flipped twice. Let $H\cdot$ and $\cdot H$ represent heads on the first and second flip respectively, while the other flip may be heads or tails. For parts a–d, define the sample space and the set of desired outcomes and calculate the probabilities. Explain your responses to parts e and f.

a. $\Pr(\cdot H)$
Solution: The sample space is $\{HH, HT, TH, TT\}$. The desired outcomes are $\{HH, TH\}$. $\Pr(\cdot H) = \frac{1}{2}$.

b. $\Pr(\cdot H \cap H\cdot)$
Solution: The sample space is $\{HH, HT, TH, TT\}$. The desired outcome is $\{HH\}$. $\Pr(\cdot H \cap H\cdot) = \Pr(HH) = \frac{1}{4}$.

c. $\Pr(\cdot H \cup H\cdot)$
Solution: The sample space is $\{HH, HT, TH, TT\}$. The desired outcomes are $\{HH, HT, TH\}$. $\Pr(\cdot H \cup H\cdot) = \Pr(\text{at least one head}) = \frac{3}{4}$.

d. $\Pr(\cdot H | H\cdot)$
Solution: The sample space is $\{HH, HT\}$. *Notice that conditioning changes the sample space under consideration.* The desired outcome is $\{HH\}$.

$$\Pr(\cdot H | H\cdot) = \frac{\Pr(\cdot H \cap H\cdot)}{\Pr(H\cdot)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}.$$

e. Provide a pair of independent events in this problem.
Solution: If two events are independent, then $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$. Here, either outcome in the second flip is independent of either outcome in the first flip; for example, getting heads on the first flip is independent of getting heads on the second flip—$\Pr(H\cdot) \cdot \Pr(\cdot H) = \frac{1}{2}\frac{1}{2} = \frac{1}{4} = \Pr(HH) = \Pr(H \cdot \cap \cdot H)$.

f. Provide a pair of events that are not independent.
Solution: If $\Pr(A \cap B) \neq \Pr(A) \cdot \Pr(B)$, then A and B are not independent (*i.e.*, the contrapositive of the statement in part e). Two such events in this problem are getting two heads and getting two tails; $\Pr(HH) \cdot \Pr(TT) = \frac{1}{4}\frac{1}{4} = \frac{1}{16} \neq \Pr(HH \cap TT) = 0$.

Question 2

Assume that the true relationship between the random variables X and Y is $Y = X^2$. Let X be distributed symmetrically around 0; *i.e.*, $f(x) = f(-x)\ \forall\ x \in \mathbb{R}$. A researcher, unaware of the true model relating X and Y, estimates $Y_i = \alpha + \beta X_i + \epsilon_i$ using a random sample of X. The latter point posits that the empirical distribution of X is identical to its theoretical distribution.

a. What is the covariance between Y and X?
Solution: First, calculate the expectation of X:

$$\begin{aligned}
\mathbb{E}(X) &= \int x f(x)\,dx = \int_{-\infty}^{0} x f(x)\,dx + \int_{0}^{\infty} x f(x)\,dx \\
&= \int_{0}^{\infty} -x f(-x)\,dx + \int_{0}^{\infty} x f(x)\,dx \\
&= \int_{0}^{\infty} (-x + x) f(x)\,dx \\
&= 0.
\end{aligned}$$

The second line follows from the change of variables formula for integration and the third lines uses the fact that $f(x) = f(-x)$. Now, for the covariance:

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(XY) - 0 \cdot \mathbb{E}(Y) \\
&= \mathbb{E}(XY) = \mathbb{E}(X^3) \\
\mathbb{E}(X^3) &= \int x^3 f(x)\,dx = \int_{-\infty}^{0} x^3 f(x)\,dx + \int_{0}^{\infty} x^3 f(x)\,dx \\
&= \int_{0}^{\infty} (-x)^3 f(-x)\,dx + \int_{0}^{\infty} x^3 f(x)\,dx \\
&= \int_{0}^{\infty} (-x^3 + x^3) f(x)\,dx \\
&= 0.
\end{aligned}$$

b. Calculate $\hat{\beta}$.
Solution:

$$\hat{\beta} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} = 0.$$

c. Calculate $\hat{\alpha}$.
Solution:

$$\begin{aligned}
\hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X} = \bar{Y} \\
\bar{Y} &= \mathbb{E}(Y) = \mathbb{E}(X^2) = \mathrm{Var}(X) + [\mathbb{E}(X)]^2 = \mathrm{Var}(X) - 0 \\
&= \mathrm{Var}(X)
\end{aligned}$$

d. What might the researcher conclude about the relationship between X and Y?
Solution: Since $\hat{\beta}$ is 0, the researcher would be tempted to conclude that there is no relationship between X and Y. This result arises only because he specified his OLS model incorrectly; the relationship between the two variables is not linear. Covariances, the relationship that produces the coefficients in OLS models, only capture linear trends and thus fail to capture the true relationship between X and Y in this scenario.

Question 3

Using the home ownership data file that was used in section, perform the following analyses in `R`. Please provide the answers in two separate parts. In the first, simply give the answer to the question. In the second, provide "clean" code that generates the solutions that you have given, but without extraneous code or `R` printouts (carrots, calculations, *etc.*). The code should be printed as a single file, with comments used to define the code relevant to each question.

a. Do any variables have missing values? If so, how many?
Solution: No covariates have missing values *per se*, but notice that the ownership variable contains 2,353 `N/A` entries.

Create a new variable that is equal to 1 if the `poverty` variable, the ratio of an individual's income to the state poverty line, is less than 100 and 0 otherwise. This converts the poverty ratio to an indicator of poverty status.

People own their homes if the `ownershd` variable is equal to `Owned with mortgage or loan` or `Owned free and clear`. They rent their homes if this variable is equal to `With cash rent` or `No cash rent`.

b. How many observations do not fit into the homeowner *v.* renter dichotomy?
Solution: There are 2,353 entries that cannot be categorized because they are listed as `N/A`.

c. What fraction of homeowners is in poverty? What fraction of renters?
Solution:

	Poverty Rate
Renters	0.27
Owners	0.06

d. Are a larger fraction of homeowners in poverty in California or in Oregon?
Solution: In California, 5.2% of homeowners are in poverty, while 6.2% are in poverty in Oregon, giving the latter state a higher homeowner poverty rate.

e. What are the income means, standard deviations, and quartiles for homeowners and for renters?
Solution:

	Mean	St. Dev.	0%	25%	50%	75%	100%
Renters	43579.00	43701.91	−6539.00	17105.00	33406.00	56749.00	968963.00
Owners	86473.90	77938.54	−19287.00	40650.00	68421.00	105650.00	999998.00

f. What are the income means, standard deviations, and quartiles for homeowners and for renters in poverty?
Solution:

	Mean	St. Dev.	0%	25%	50%	75%	100%
Renters	14879.87	16899.22	−6539.00	6842.00	10666.00	18111.00	428638.00
Owners	21302.10	33110.08	−19287.00	7245.00	12074.00	22136.00	480960.00

g. Give the median income by state.
Solution: See following page.

State	Median Income
Alabama	50310.00
Alaska	61881.00
Arizona	60372.00
Arkansas	44322.50
California	69226.00
Colorado	68320.00
Connecticut	81502.00
Delaware	62021.50
District of Columbia	70433.00
Florida	57353.00
Georgia	57655.00
Hawaii	76974.00
Idaho	51618.00
Illinois	65327.00
Indiana	57152.00
Iowa	55592.00
Kansas	55391.00
Kentucky	49303.00
Louisiana	50471.00
Maine	52624.00
Maryland	82850.00
Massachusetts	77477.00
Michigan	56347.00
Minnesota	62283.00
Mississippi	41254.00
Missouri	52322.00
Montana	44474.00
Nebraska	56342.00
Nevada	62384.00
New Hampshire	75565.00
New Jersey	79539.50
New Mexico	50259.50
New York	65402.00
North Carolina	52624.00
North Dakota	50913.00
Ohio	58359.00
Oklahoma	45671.00
Oregon	58359.00
Pennsylvania	58963.00
Rhode island	70433.00
South Carolina	52377.50
South Dakota	55341.00
Tennessee	50310.00
Texas	54988.50
Utah	61176.00
Vermont	53187.00
Virginia	70232.00
Washington	63088.00
West Virginia	46466.00
Wisconsin	58611.00
Wyoming	66992.50