Problem Set Answer Template

Your Name

July 17, 2025

Question 1

(placehoder for analytical question)

Answer:

We need to prove the additive property of expectations: if X and Y are random variables, then $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

By the definition of expectation, we have:

- The expectation of the sum $X+Y$ is given by:

$$\mathbb{E}[X+Y] = \int_{-\infty}^{\infty} (x+y) f_{X,Y}(x,y)\, dx\, dy, \tag{1}$$

- Using the linearity of integration, we can separate the terms inside the integral in Equation 1:

$$\mathbb{E}[X+Y] = \int_{-\infty}^{\infty} x f_X(x)\, dx + \int_{-\infty}^{\infty} y f_Y(y)\, dy.$$

- Simplifying, we obtain:

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]. \quad \blacksquare$$

Question 2

(placeholder for question involving data analyses in R)

Answer:

We use Angrist and Pischke (2009) in the answer. Table 1 describes the variables we will simulate.

Variable	Description
X_1	Random variable drawn from a normal distribution with mean 5 and standard deviation 2.

Variable	Description
X_2	Categorical variable with levels `"A"`, `"B"`, and `"C"`, sampled randomly with replacement.
X_3	Binary variable drawn from a Bernoulli distribution with probability 0.5.
Y	Dependent variable calculated as `3 + 2 * x1 + 1.5 * as.numeric(x2) + 0.8 * x3` plus some random noise.

Table 1: Variable description.

Code below simulates a dataset with 100 observations according to the description in Table 1 and adds some random missingness to variables X_2 and X_3. Table 2 shows the top 6 rows of this dataset.

```
# Generate a sample dataset
data <- dplyr::tibble(
   x1 = rnorm(100, mean = 5, sd = 2),
   x2 = factor(sample(c("A", "B", "C"), 100, replace = TRUE)),
   x3 = rbinom(100, 1, prob = 0.5),
   y = 3 + 2 * x1 + 1.5 * as.numeric(x2) + 0.8 * x3 + rnorm(100)
)

# Introduce some NAs into x2 and x3
data$x2[sample(1:100, 10)] <- NA
data$x3[sample(1:100, 10)] <- NA

dplyr::sample_n(data, size = 6) |>
   knitr::kable(
      format = "latex", digits = 3,
      align = "c", booktabs = TRUE, linesep = ""
   )
```

x1	x2	x3	y
3.851	C	0	15.340
4.119	A	1	13.455
9.832	C	0	26.608
9.141	A	NA	20.527
4.049	C	1	15.979
3.553	A	1	14.349

Table 2: Simulated data.

Code below produces Table 3 that shows point and uncertainty estimates for the model $Y = \alpha_1 + \beta_1 X_1 + \varepsilon$ (column 1) and $Y = \alpha_2 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + \varepsilon$ (column 2).

```
modelsummary::modelsummary(
   list(mod1, mod2), stars = TRUE,
   gof_omit = "BIC|AIC|RMSE",
```

```
    coef_omit = "(Intercept)", output = "latex")
```

	(1)	(2)
x1	1.995***	1.950***
	(0.093)	(0.057)
x2B		1.201*
		(0.493)
x2C		2.967***
		(0.391)
x3		1.272**
		(0.395)
x2B × x3		−0.330
		(0.643)
x2C × x3		−1.058+
		(0.558)
Num.Obs.	100	82
R2	0.878	0.949
R2 Adj.	0.876	0.945

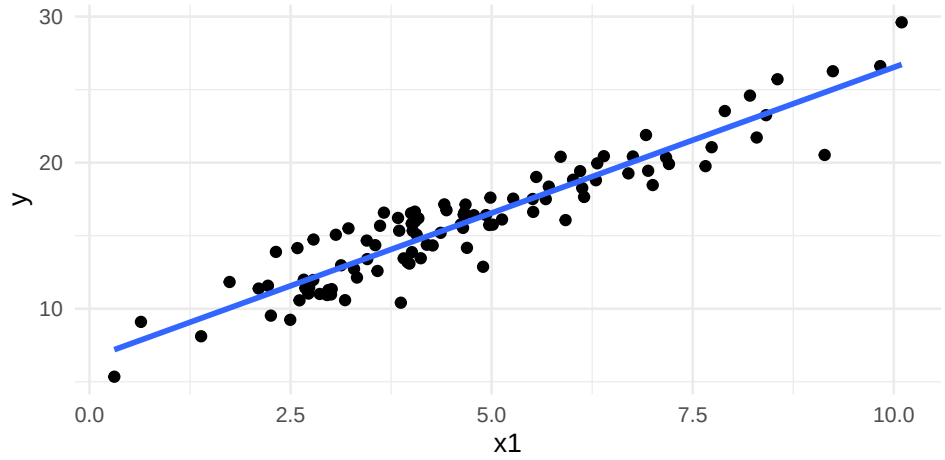
+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Table 3: Regression output.

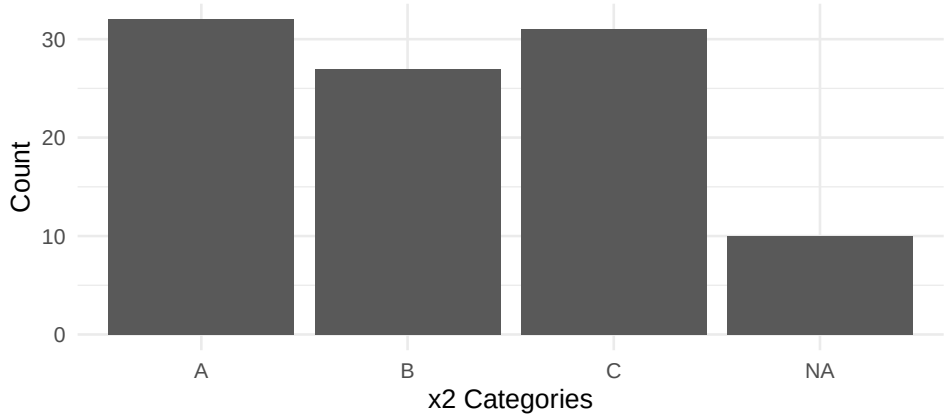
The code below produces Figure 1 with two panels. Figure 1a shows scatter plot of X_1 against Y with linear trend, while Figure 1b shows histogram of categorical variable X_2.

```
ggplot(data, aes(x = x1, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal(base_size = 10)

ggplot(data, aes(x = x2)) +
   geom_bar() +
   theme_minimal(base_size = 10) +
   labs(x = "x2 Categories", y = "Count", title = "")
```



(a) Scatter plot of X_1 vs Y variable and linear trend.



(b) Histogram of X_2 variable.

Figure 1: Example figures.

References

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Wconometrics: An Wmpiricist's Companion*. Princeton University Press.