MLDS 401 – Homework 4
Q6–Q8: Full Statements and Detailed Solutions

Junbo Lian, Arielle Weinstein, John Bailey, Veronica Lin, Jiashu Huang

Q6

Let $\tilde{x}_1 = x_1 - \bar{x}_1$ and $\tilde{x}_2 = x_2 - \bar{x}_2$ denote centered variables, and let $\mathbf{1}$ be the $n \times 1$ vector of ones. The OLS slope in the misspecified regression of y on x_1 is

$$\hat{\alpha}_1 = \frac{\tilde{x}_1^\top y}{\tilde{x}_1^\top \tilde{x}_1}.$$

Substitute the true DGP:

$$y = \beta_0 \mathbf{1} + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

Using $\tilde{x}_1^\top \mathbf{1} = 0$, we have

$$\hat{\alpha}_1 = \frac{\tilde{x}_1^\top (\beta_1 x_1 + \beta_2 x_2 + \varepsilon)}{\tilde{x}_1^\top \tilde{x}_1} = \beta_1 + \beta_2 \frac{\tilde{x}_1^\top \tilde{x}_2}{\tilde{x}_1^\top \tilde{x}_1} + \frac{\tilde{x}_1^\top \varepsilon}{\tilde{x}_1^\top \tilde{x}_1}.$$

Taking the conditional expectation given the regressors (design),

$$\mathbb{E}[\hat{\alpha}_1 \mid X] = \beta_1 + \beta_2 \frac{\tilde{x}_1^\top \tilde{x}_2}{\tilde{x}_1^\top \tilde{x}_1}, \qquad \text{since } \mathbb{E}[\tilde{x}_1^\top \varepsilon \mid X] = 0.$$

Using sample moments,

$$\tilde{x}_1^\top \tilde{x}_2 = (n-1)\, r\, s_1 s_2, \qquad \tilde{x}_1^\top \tilde{x}_1 = (n-1)\, s_1^2,$$

hence

$$\mathbb{E}[\hat{\alpha}_1] = \beta_1 + \beta_2\, r\, \frac{s_2}{s_1}.$$

Therefore, the omitted variable bias (OVB) equals

$$\underbrace{\mathbb{E}[\hat{\alpha}_1] - \beta_1}_{\text{Bias}} \;=\; \beta_2\, r\, \frac{s_2}{s_1}.$$

Zero-bias conditions. The bias is zero if either (i) $r = 0$ (the included and omitted regressors are uncorrelated) or (ii) $\beta_2 = 0$ (the omitted regressor does not affect y).

Sign interpretation. The sign of the bias equals $\text{sign}(\beta_2) \cdot \text{sign}(r)$. For example, if the omitted regressor has a positive effect on y ($\beta_2 > 0$) and is positively correlated with x_1 ($r > 0$), then the estimated slope on x_1 will be upward biased, and vice versa.

Connection to JWHT 3.14 (Q5). In that simulation, $r = \mathrm{corr}(x_1, x_2)$ is large and $\beta_2 \neq 0$, so single-predictor regressions exhibit noticeable OVB, while the two-predictor regression shows instability (inflated standard errors) due to collinearity.

Q7

Let $Z = [z_1\ z_2]$ be the $n \times 2$ design matrix (ignoring the intercept for notational simplicity). Since the predictors are standardized,

$$Z^\top Z \;=\; n \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}.$$

Under homoskedastic OLS,

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) \;=\; \sigma^2 (Z^\top Z)^{-1}.$$

The inverse is

$$(Z^\top Z)^{-1} = \frac{1}{n(1-r^2)} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}.$$

Therefore, the variance of the first slope is simply the $(1,1)$ element:

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{n(1-r^2)}.$$

VIF connection. The VIF for regressor z_1 is defined as

$$\mathrm{VIF}_1 \;=\; \frac{1}{1-R_1^2},$$

where R_1^2 is from regressing z_1 on the *other* regressors. With one other regressor z_2, $R_1^2 = r^2$, hence

$$\mathrm{VIF}_1 = \frac{1}{1-r^2}, \qquad \mathrm{Var}(\hat{\beta}_1) \;=\; \frac{\sigma^2}{n}\,\mathrm{VIF}_1\,.$$

Thus collinearity inflates the sampling variance multiplicatively by the VIF. As $|r| \to 1$, the denominator $1 - r^2 \to 0$, the VIF explodes, and coefficient estimates become unstable.

Generalization. With k regressors,

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{n}\,\frac{1}{1-R_j^2},$$

where R_j^2 is from regressing the j-th regressor on the other $k-1$ regressors. Large R_j^2 (predictable regressor) $\Rightarrow$ large VIF $\Rightarrow$ wide standard errors.

Q8

Core Bike rental demand around a station reflects users' perceived safety along first/last-mile routes. Violent crimes generally depress perceived safety more than property crimes, especially during night or at poorly low-visibility locations. This channel reduces trip initiation and station choice probability.

Primary hypotheses.

(H1) **Violent crime effect (dominant negative):** Holding access and land-use constant, violent-crime intensity near a station is negatively associated with rental counts, with a larger (more negative) magnitude than property-crime effects.

(H2) **Temporal moderation (night):** The negative effect of violent crime is stronger at night or in winter months (short daylight), i.e., $\beta_{\text{violent}\times\text{night}} < 0$.

(H3) **Access moderation:** Proximity to high-quality bike lanes/lighting and presence of nearby transit hubs mitigates the crime penalty (interaction terms attenuate the negative slope).

(H4) **Heterogeneity by station type:** Effects are stronger for stations in low-footfall or low-visibility areas; weaker in dense commercial districts with passive surveillance ("eyes on the street").

(H5) **Lagged salience vs. raw counts:** Recent violent events have salience beyond their frequency; short-lag crime measures (last 1–4 weeks) better track perceived risk than annual aggregates.

Why some crime types may have weak/no effects. Certain categories may not directly threaten prospective riders or be highly localized/indoor (limited route exposure), yielding small elasticities. Conversely, theft may matter if it targets bikes at or near stations, but the mechanism is through expected loss rather than personal safety en route.