Question 1

You study the effect of media consumption on cognitive development (as measured by a standardized math score) for Australian teenagers. Media consumption (for example: watching TV, using social media, Playstation) may affect a child's learning. You investigate the research question: Is media consumption negatively associated with math scores?

You have available sample data on test score, media consumption, and other demographic characteristics for 8,764 Australian teenagers for the year 2018. The data are drawn from the Longitudinal Survey of Australian Children (LSAC).

Here a brief description of some of the variables:

- `mathscr`: respondent's score on a standardized math test (between 0 and 600)
- `mediahrs`: respondent's average daily media consumption in hours
- `male`: dummy equal 1 if respondent identifies as male

Use the following `Python` output (partially edited) to answer the questions below.

Python Code and Output

```
> formula = 'mathscr ~ mediahrs * male'
> reg = smf.ols(formula, data=df, missing='drop').fit(cov_type='HC1', use_t=False)
> print(reg.summary())

================================================================================
                    coef    std err          z      P>|z|     [0.025       0.975]
--------------------------------------------------------------------------------
Intercept       531.2458     11.868     44.763     <2e-16    507.985      554.507
mediahrs         -2.2392     xxxxxx     xxxxxx      0.050    xxxxxxx      xxxxxxx
male             -0.0232      0.008     -2.942      0.003     -0.039       -0.008
mediahrs:male    -1.2766      0.967     -1.324      0.187     -3.172        0.619
================================================================================
```

(a) [3 marks] Interpret and discuss the coefficient estimate for `male`. Is it statistically significant?

(b) [5 marks] Determine the t-statistic for `mediahrs`.

(c) [5 marks] Interpret and discuss the coefficient estimate for the interaction term.

(d) [7 marks] Consider your estimate of the effect of media consumption on math scores. Give an example of an omitted variable that could bias the above results. If this variable would be included, how would you expect your estimates to change? Explain.

(a)

sign: negative => being male associated with lower math scores, all else equal size: effect size is small when compared to intercept which captures the predicted math score for a non-male, non-media consuming person

significance: can reject $H_0: \beta_{male} = 0$ based on t-stat

(b) pvalue of 5% suggests $|tstat| = 1.96$
together with negative sign of coeff estimate
$\Rightarrow$ $tstat = -1.96$

(c) Interaction term allows for the effect of mediahrs on mathscr to differ by gender. The neg. sign of the coef. estimate for the ia term suggests that males have a more negative effect of mediahrs on mathscr.

(d) Example of omitted var: parents' ses
Hypothesis: Parents of lower ses may tend to allow their kids more mediahrs. At the same time their kids may have lower mathscr unrelated to their media consumption (worse schools, fewer resources).

Not controlling ses will result in an overstatement of the effect of mediators on mathscr. Mediators is a proxy for unobserved ses. I therefore would expect a smaller coeff estimate (in absolute value).

Question 2

Are the following statements true or false? Provide a brief and complete explanation. (Note: you will not receive any credit without providing a correct explanation.)

(a) [5 marks]

Statement:
In the multiple regression model, omitting an explanatory variable may not necessarily lead to omitted variables bias.

True.

If omitted variable is uncorrelated with included variable then there will be no ovb.

(b) [5 marks]

Statement:
In autoregressions, allowing for additional lags (that is, increasing p in the AR(p) model) increases R^2.

True.

In multiple regressions (such as AR(p)) adding regressors mechanically decreases the RSS and therefore increases R^2.

(c) [5 marks]

Statement:
Given a random sample $Y_1, \ldots, Y_{10}$, the following two estimators are equally good for estimating $\mathrm{E}(Y_7)$:

- the simple average of Y_1 and Y_{10};
- $2/3 \cdot Y_6 + 1/3 \cdot Y_9$.

False.

Consider expected values:

$$E((Y_1 + Y_{10})/2) = \frac{1}{2}(E(Y_1) + E(Y_{10})) = \frac{1}{2}(E(Y_7) + E(Y_7)) = E(Y_7)$$

$$E\left(\frac{2}{3}Y_6 + \frac{1}{3}Y_9\right) = \frac{2}{3}E(Y_6) + \frac{1}{3}E(Y_9) = \frac{2}{3}E(Y_7) + \frac{1}{3}E(Y_7) = E(Y_7)$$

$\Rightarrow$ both unbiased

Consider variance:

$$\begin{aligned} Var((Y_1 + Y_{10})/2) &= \frac{1}{4}(Var(Y_1) + Var(Y_{10})) \\ &= \frac{1}{4}(Var(Y_7) + Var(Y_7)) \\ &= \frac{1}{2} Var(Y_7) \end{aligned}$$

$$\begin{aligned} Var\left(\frac{2}{3}Y_6 + \frac{1}{3} Var\, Y_9\right) &= \frac{4}{9} Var(Y_6) + \frac{1}{9} Var(Y_9) \\ &= \frac{4}{9} Var(Y_7) + \frac{1}{9} Var(Y_7) \\ &= \frac{5}{9} Var(Y_7) \\ &> \frac{1}{2} Var(Y_7) \end{aligned}$$

$\Rightarrow$ both estimators are unbiased but the first one has a smaller variance and is therefore better

(d) [5 marks]

Statement:
In the simple regression model, if $\beta_1 = 0$, then $\hat{\beta}_0 = \bar{Y}$.

False, b/c

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

would be true if $\hat{\beta}_1 = 0$ which is not generally the case.

Question 3

Consider the linear regression model

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \qquad \beta_1 \neq 0.$$

(a) `[10 marks]`

Suppose that $\beta_2 = 0$. Define and derive the OLS estimator of β_1. Prove that this estimator is unbiased. Explicitly state any additional assumptions that you may need to establish unbiasedness.

(b) `[10 marks]`

Now suppose that $\beta_2 \neq 0$. Under which circumstances is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_{1i} Y_i}{\sum_{i=1}^{n} X_{1i}^2}?$$

Provide a complete proof.

(a) model becomes

$$y_i = \beta_1 X_{1i} + u_i$$

$$\hat{\beta}_1 := \arg\min_b \sum (y_i - b X_{1i})^2$$

first derivative

$$\frac{d \sum (y_i - b X_{1i})^2}{db} = 2 \sum (y_i - b X_{1i}) X_{1i}$$

<u>foc:</u>

$$\sum (y_i - \hat{\beta}_1 X_{1i}) X_{1i} = 0$$

solving gives: $\hat{\beta}_1 = \dfrac{\sum X_{1i} y_i}{\sum X_{1i}^2}$

for unbiasedness:

wts: $E(\hat{\beta}_1 | X_{1i}) = \beta_1$

let's look:

$$E(\hat{\beta}_1 | X_{1i}) = E\left(\frac{\sum X_{1i} Y_i}{\sum X_{1i}^2} \middle| X_{1i}\right)$$

$$= \frac{E(\sum X_{1i} Y_i | X_{1i})}{\sum X_{1i}^2}$$

$$= \frac{\sum E(X_{1i} Y_i | X_{1i})}{\sum X_{1i}^2}$$

$$= \frac{\sum X_{1i} E(Y_i | X_{1i})}{\sum X_{1i}^2}$$

$$= \frac{\sum X_{1i} E(X_{1i}\beta_1 + u_i | X_{1i})}{\sum X_{1i}^2}$$

$$= \frac{\sum X_{1i}^2 \beta_1 + E(u_i | X_{1i})}{\sum X_{1i}^2}$$

$$= \beta_1 + \underbrace{\frac{E(u_i | X_{1i})}{\sum X_{1i}^2}}_{= 0 \text{ under OLS assumption 1: } E(u_i | X_{1i}) = 0}$$

$$\Rightarrow \quad E(\hat{\beta}_1 \mid X_{1i}) = \beta_1$$

and therefore $\hat{\beta}_1$ is unbiased

(b)

$$\hat{\beta}_1 := \underset{b_1, b_2}{\operatorname{argmin}} \sum (y_i - b_1 X_{1i} - b_2 X_{2i})^2$$

first derivative:

$$\frac{\partial \sum (y_i - b_1 X_{1i} - b_2 X_{2i})^2}{\partial b_1} = -2 \sum (y_i - b_1 X_{1i} - b_2 X_{2i}) X_{1i} \quad (*)$$

$$\frac{\partial \sum (y_i - b_1 X_{1i} - b_2 X_{2i})^2}{\partial b_2} = -2 \sum (y_i - b_1 X_{1i} - b_2 X_{2i}) X_{2i} \quad (\#)$$

foc for $(*)$: $\sum (y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) X_{1i} = 0$

solving for $\hat{\beta}_2$:

$$\hat{\beta}_2 = \frac{\sum X_{1i} y_i}{\sum X_{1i} X_{2i}} - \hat{\beta}_1 \frac{\sum X_{1i}^2}{\sum X_{1i} X_{2i}}$$

foc for $\#$: $\sum (y_i - b_1 X_{1i} - b_2 X_{2i}) X_{2i} = 0$

also solving for $\hat{\beta}_2$:

$$\hat{\beta}_2 = \frac{\sum X_{2i} y_i}{\sum X_{2i}^2} - \hat{\beta}_1 \frac{\sum X_{1i} X_{2i}}{\sum X_{2i}^2} . \qquad \triangle$$

equating both $\circledcirc$ and $\triangle$:

$$\frac{\sum X_{1i} y_i}{\sum X_{1i} X_{2i}} - \hat{\beta}_1 \frac{\sum X_{1i}^2}{\sum X_{1i} X_{2i}} = \frac{\sum X_{2i} y_i}{\sum X_{2i}^2} - \hat{\beta}_1 \frac{\sum X_{1i} X_{2i}}{\sum X_{2i}^2} .$$

rearranging to move $\hat{\beta}_1$ - terms to LHS:

$$\hat{\beta}_1 \frac{\sum X_{1i}^2}{\sum X_{1i} X_{2i}} - \hat{\beta}_1 \frac{\sum X_{1i} X_{2i}}{\sum X_{2i}^2} . = \frac{\sum X_{1i} y_i}{\sum X_{1i} X_{2i}} - \frac{\sum X_{2i} y_i}{\sum X_{2i}^2} .$$

$\Leftrightarrow$

$$\hat{\beta}_1 \left(\frac{\sum X_{1i}^2}{\sum X_{1i} X_{2i}} - \frac{\sum X_{1i} X_{2i}}{\sum X_{2i}^2} . \right) = \frac{\sum X_{1i} y_i}{\sum X_{1i} X_{2i}} - \frac{\sum X_{2i} y_i}{\sum X_{2i}^2} .$$

$\Leftrightarrow$

. . .

$$\hat{\beta}_1 \left(\frac{(\Sigma X_{1i}^2)(\Sigma X_{2i})^2 - (\Sigma X_{1i} X_{2i})^2}{(\Sigma X_{1i} X_{2i})(\Sigma X_{2i}^2)} \right)$$

$$= \frac{(\Sigma X_{2i}^2)(\Sigma X_{1i} y_i) - (\Sigma X_{1i} X_{2i})(\Sigma X_{2i} y_i)}{(\Sigma X_{1i} X_{2i})(\Sigma X_{2i}^2)}$$

$\Leftrightarrow$

$$\hat{\beta}_1 = \frac{(\Sigma X_{2i}^2)(\Sigma X_{1i} y_i) - (\Sigma X_{1i} X_{2i})(\Sigma X_{2i} y_i)}{(\Sigma X_{1i}^2)(\Sigma X_{2i})^2 - (\Sigma X_{1i} X_{2i})^2}$$

Now, if $\Sigma X_{1i} X_{2i} = 0$ then RHS

collapses to $\dfrac{\Sigma X_{1i} y_i}{\Sigma X_{1i}^2}$

[This was discussed in the week 7 workshop]