1) Recall the assumptions introduced with the linear model:

A0 The data $(Y_i, X_i)_{i=1}^N$ is iid

A1 The data has linear representation: $Y_i = X_i\beta + \epsilon_i$

A2 Strict exogeneity: $E[\epsilon_i|X_i] = 0$.

A3 Rank: if $\dim(X) = K$, then the data X has K linearly independent columns ($\text{Rank}(X) = K$)

A4 Spherical errors/Homoskedasticity: $\mathbb{V}[\epsilon_i|X_i] = \sigma^2$.

1. State the assumptions necessary to prove that the OLS estimator $\hat{\beta}$ is a consistent estimator of the relationship between X and Y in *conditional expectation* $\mathbb{E}[Y|X]$
2. State the assumptions necessary to prove that the OLS estimator $\hat{\beta}$ is a consistent estimator of the *causal effect* of each variable X on Y.
3. State which assumptions are necessary to derive the following expression for the asymptotic variance of $\hat{\beta}$, $\mathbb{V}(\hat{\beta})$:

$$\mathbb{V}[\hat{\beta}] = \mathbb{E}[X_i'X_i]^{-1}\frac{\sigma^2}{N}$$

2) Suppose you have iid data (C_i, I_i) where $C_i \in \{0, 1\}$ indicates whether an individual attends college, and I_i is a measure of parental income. You want to estimate the relationship:

$$\mathbb{E}[C|I] = \beta_0 + \beta_1 I$$

1. Describe your estimator $\hat{\beta}$ for $\beta = (\beta_0, \beta_1)$. State the formula you will use.
2. Describe the asymptotic distribution of $\hat{\beta}$.
3. Propose a way to estimate this asymptotic distribution, and use this to construct a $(1-\alpha)\times 100\%$ confidence interval for β_1.
4. Are you comfortable with concluding that your estimate of β_1 is also an estimate if the causal effect of parental income on college attendance? Why or why not?

3) Consider the model:

$$Y_i = \alpha + X_i\beta + \epsilon_i.$$

Notice that α is now the constant term, and so X_i does not contain a 1 in the first column. Further assume that assumptions A0-A4 still hold.

1. Let $\mu_X = \mathbb{E}[X]$ and $\hat{X}_i = X_i - \mu_X$. Write Y_i in terms of $\hat{X}_i$ and ϵ_i.
2. Does the equation you wrote above still satisfy A0-A4?
3. Based on this, do you expect any difference between estimating β using X compared to $\hat{X}$?

4) Consider the linear model:

$$Y_i = \beta_0 + X_{1,i}\beta_1 + X_{2,i}\beta_2 + \epsilon_i^*$$

Suppose that $\mathbb{E}[\epsilon_i^*|X_{1,i}, X_{2,i}] = 0$ and that $X_{1,i}$ and $X_{2,i}$ are independent. Let $\mathbb{E}[X_{1,i}] = \mu_1$ and $\mathbb{E}[X_{2,i}] = \mu_2$.

1. Calculate $\mathbb{E}[Y_i|X_{1,i}]$
2. Define $\epsilon_i = Y_i - \mathbb{E}[Y_i|X_{1,i}]$ and write Y_i in terms of $X_{1,i}$ and ϵ_i.
3. Use the above two steps to argue that if we run a regression of Y on $X_{1,i}$ without $X_{2,i}$, we still recover a consistent estimator of β_1.

5) Let:

$$Y_i = X_i\beta + Z_i\gamma + \epsilon_i$$

where X_i and Z_i are scalar variables, with $\mathbb{E}[X_i] = \mathbb{E}[Z_j] = 0$.[1] Suppose that $\mathbb{V}[X] = \sigma_X^2$, $\mathbb{V}[Z] = \sigma_Z^2$, and $\mathbb{C}(X, Z) = \sigma_{XZ}$.

1. Let $W_i = [X_i, Z_i]$. Write the matrix $\mathbb{E}[W_i'W_i]$ in terms of $\sigma_X^2, \sigma_Z^2, \sigma_{XZ}^2$.
2. Use the matrix inverse formula[2] to calculate $\mathbb{E}[W_i'W_i]^{-1}$.
3. Suppose that X_i and Z_i are *independent*, and calculate (a) the variance of the estimator $\hat{\beta}$ when Z_i is excluded from the regression; (b) the variance of the estimator $\tilde{\beta}$ when Z_i is included in the regression. Which estimator is more efficient? One hint: what is the value of σ_{XZ} when X and Z are independent.
4. Suppose that X_i and Z_i are *not* independent, but that $\gamma = 0$. Calculate (a) the variance of the estimator $\hat{\beta}$ when Z_i is excluded from the regression; (b) the variance of the estimator $\tilde{\beta}$ when Z_i is included in the regression. Which estimator is more efficient?

6) Consider the regression $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ when X_i is a single variable and A0-A4 are satisfied. Suppose that $\mathbb{V}[X_i] = 2$, $\sigma_\epsilon^2 = 1$, $N = 50$, and $\beta_1 = 0.5$. For the questions below, you will use some of the following facts about Z, a standard normal random variable.

$$P[Z > -3.36] = 0.9996,\ P[Z > -1.36] = 0.913,\ P[Z > 0.64] = 0.261,\ P[Z > |1.24] = 0.107$$

- Calculate $\mathbb{V}[\hat{\beta}_1]$ when $\hat{\beta}_1$ is estimated by OLS.
- Suppose you conduct a test of the Null hypothesis that $\beta_1 \leq 0$ with size 95% ($z_{0.05} = 1.64$). What is the power of this test?
- Suppose you conduct a test of the Null hypothesis that $\beta_1 \leq 0.2$. What is the power of this test?
- Suppose you conduct a test of the Null hypothesis that $\beta_1 \leq 0.4$. What is the power of this test?

[1] Note that based on question 3, you can always make this true by applying the logic of question (3).

[2] $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ab-cd}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

7) Consider the linear model:

$$Y_i = X_i\beta + \epsilon_i$$

where assumptions A0-A4 hold. Suppose that $\dim(\beta) = 4$.

1. Derive a test of the Null hypothesis that $\beta_1 + \beta_2 = 1$. Describe exactly how you would conduct the test with significance $\alpha \times 100\%$.

2. Derive a test of the joint Null hypotheses that $\beta_1 + \beta_2 = 1$ and $\beta_3 = \beta_4$. Describe exactly how you would conduct the test with significance $\alpha \times 100\%$.

8) For the below examples, write the correspoding R matrix and vector c in order to write each set of restrictions as $R\beta - c = 0$.

1. $\dim(\beta) = 4$, $\beta_1 = 0$, $\beta_2 - \beta_3 = 0$, $\beta_4 = 4$.

2. $\dim(\beta) = 5$, $\beta_1 = 1$, $\beta_3 = 4$.

3. $\dim(\beta) = 3$, $\beta_1 = 1.1$, $\beta_2 + 2\beta_3 = 1$.