Department of Economics
Yale University
Prof. Edward Vytlacil

Econ 136
Econometrics
Spring 2024

Problem Set 2A: Due Tuesday, February 20, at 2pm.

Assignments should be submitted through Canvas and include

1. a PDF document, and
2. a R or .Rmd code file.

The PDF document should include all your work (code, output, written typed answers, hand-written math answers). The problem set is due Tuesday, February 20, at 2pm. Late submission without professor's approval will not be accepted without a note from your residential Dean. You are allowed to work in groups of up to four students. However, each student must turn in their own problem set, and indicate on the problem set the other members of their group.

Part I: Theoretical Questions
The following questions are related to material covered in lecture through Thursday, February 8, and to Chapter 2 from the Hansen *Econometrics* textbook. It does not cover OLS or estimation, which will be covered on your next problem set

1. Consider the following model, $Y = X'\beta + e$, where $X = (1, X_1)'$, $\beta = (\beta_0, \beta_1)'$, and Y and X_1 are scalar random variables. Suppose $\mathbb{E}[Y^2] < \infty$, $\mathbb{E}[X_1^2] < \infty$ Suppose that we are defining $X'\beta$ as the best linear predictor of Y given X.
   (a) Why does $\mathbb{E}(Xe)$ necessarily equal zero in this example?
   (b) Why does $\mathbb{E}(Xe) = 0$ imply $\mathbb{E}(e) = 0$ and $\mathbb{E}(X_1 e) = 0$?
   (c) Show that $\mathbb{E}(Xe) = 0$ implies $\text{Cov}(X_1, e) = 0$.
   (d) Why is no perfect collinearity in X equivalent to $\text{Var}(X_1) > 0$?
   (e) Suppose that $\text{Var}(X_1) > 0$. Recall that $\mathbb{E}(XU) = 0$ implies that
   $$\beta = (E(XX'))^{-1} E(XY).$$
   Use this expression to solve for β_0 and β_1 in terms of $\mathbb{E}(X_1), \mathbb{E}(Y)$, $\text{Cov}(X_1, Y)$, and $\text{Var}(X_1)$.
2. Consider the following model, $Y = X'\beta + e$, where $X = (1, X_1)'$, $\beta = (\beta_0, \beta_1)'$, Y and X_1 are scalar random variables and suppose $\mathbb{E}[Y^2] < \infty$, $\mathbb{E}[X_1^2] < \infty$. Suppose that we are defining $X'\beta$ as the best linear predictor of Y given X. Suppose that a researcher is interested in the best linear predictor of X given Y, and concludes that it must be the case that
   $$X_1 = -\frac{\beta_0}{\beta_1} - \frac{1}{\beta_1} Y - \frac{1}{\beta_1} e,$$
   and thus that the best linear predictor of X_1 given Y must be $-\frac{\beta_0}{\beta_1} - \frac{1}{\beta_1} Y$. Will the BLP of X_1 given Y be $-\frac{\beta_0}{\beta_1} - \frac{1}{\beta_1} Y$? (hint: what is $\text{Cov}(Y, e)$?)

3. Consider the following model, $Y = X'\beta + e$, with Y a bernoulli random variable, $Y \in \{0, 1\}$ and X a $K+1$-dimensional random vector. Suppose $\mathbb{E}[e \mid X] = 0$. This model is called a *linear probability model*, and is frequently used to predict binary outcomes. Let $\mu_X = \mathbb{E}[X]$.

   (a) Find an expression for $\Pr[Y = 1]$ in terms of μ_X and β.

   (b) Find an expression for $\text{Var}[Y]$ in terms of μ_X and β.

   (c) Find an expression for $\Pr[Y = 1 \mid X]$ in terms of X and β.

   (d) Using your answer to part (c), find an expression for $\text{Var}[Y \mid X]$ in terms of X and β.

   (e) Using your answer to part (d), find an expression for $\text{Var}[e \mid X]$ in terms of X and β.

   (f) Under what conditions (if any) will e be heteroskedastic?

4. Let Y denote wage, F denote an indicator variable for being a woman, and let S denote years of schooling. Consider the wage regression model:

$$Y = \beta_0 + \beta_1 F + \beta_2 S + e$$

   and suppose $\mathbb{E}(e \mid F, S) = 0$. Suppose we are interested in β_1, the gender-wage gap in the model that includes years of schooling (which we might sometimes say is the gender-wage gap "controlling" for schooling). Consider also the projection of Y on F alone, omitting schooling:

$$Y = \gamma_0 + \gamma_1 F + v$$

   where, by construction, $\mathbb{E}[v] = \mathbb{E}[v \cdot F] = 0$. We will refer to γ_1 as the gender-wage gap not controlling for schooling.

   (a) Why does $\mathbb{E}[e \mid F, S] = 0$ imply $\mathbb{E}[e] = 0$?

   (b) Show that $\gamma_0 = \mathbb{E}[Y \mid F = 0]$, $\gamma_1 = \mathbb{E}[Y \mid F = 1] - \mathbb{E}[Y \mid F = 0]$.
   (hint: you may wish to look back at your answer to Q4 of PS1).

   (c) Show that $\gamma_1 = \beta_1 + \beta_2 \left(\mathbb{E}[S \mid F = 1] - \mathbb{E}[S \mid F = 0]\right)$, and interpret this result.

   (d) Historically in the United States, women tended to obtain less schooling than men. In that context, would you expect the gender-wage gap not controlling for schooling to be larger or smaller than the gap that does control for schooling? explain, using your answer to part (c).

   (e) Currently in the United States, women tended to obtain more schooling than men. In that context, would you expect the gender-wage gap not controlling for schooling to be larger or smaller than the gap that does control for schooling? explain, using your answer to part (c).

Part II: R Questions

The following questions involve coding with **R**. The content is closely related to the Example 4 of Handout 1. For this problem set, you will work with the data set `financeR.dta`, which you can download from https://edward-vytlacil.github.io/Data/financeR.dta. The dataset is in STATA `.dta` format, so you will need to use an appropriate package to load the data into **R**. I recommend using the `read_dta` function from the `haven` package. The main variables in the data set are the monthly returns of some U.S. stocks as well as the returns of the S&P 500 index, which is a measure of the returns of the U.S. stock market as a whole. In the data set, we denote the S&P 500 (index fund) returns as r_M, and the returns of SPDR Gold Shares (GLD) as r_A, Morgan Stanley (MS) shares as r_B, and Genworth Financial (GNW) shares as r_C.

1. Load the dataset fiananceR.dta. Note that the data set is in a STATA `.dta` format. To load the data, I recommend first downloading and installing the `haven` package, and then use the `read_dta` function to load the dataset as a data frame. Use `View` to examine the data (however, do not include `View` in your RMarkdown file or you will get a knitting error).

2. How many variables and observations are there? Use `dim`.

3. Use the `sapply` and `class` functions to learn about the data types in the data set. Which variable is not numeric?

4. Drop the variable from your data frame that is not numeric.

5. Suppose the risk free rate is $r_f = 0.0041$. Use `sapply` along with a function defined by you to create a new data frame in which all the variables are excess returns (returns minus r_f). For the rest of this problem set, use that new data frame containing excess returns.

6. Use the `summary()` command to compute some descriptive statistics of the variables in the dataset. Which asset had the highest mean return?

7. Use the `var()` function to compute the variances matrix of the variables.

   (a) Which asset was the most volatile, having the highest variance?

   (b) Which two assets had the highest covariance with each other?

   (c) Discuss the results in words. What is your intuition for why the asset in (a) above is the most volatile? What is your intuition for why the two assets in (b) above are highly positively dependent?

   (d) Use the `cor` function to compute the correlations of the numeric variables. Which two assets had the highest correlation? Is your answer the same as in part (b) above? Is there any contradiction between your result here and your result in part (b) above? discuss.

8. Use the data to create a figure similar to Figure 0.1 on slide 7 of the review slides. showing the risk and return tradeoff across the assets.

(a) Using the `c` (concatenate), `mean`, and `sapply` functions, create a vector containing the sample means of the asset return variables.

(b) Using the `c` (concatenate), `var`, and `sapply` functions, create a vector containing the sample variances of the asset return variables.

(c) Use the vectors you created in (a) and (b) above along with the `ggplot` function to create a figure similar to Figure 0.1 on slide 7 of the review slides, showing the risk and return tradeoff across the assets. Note that you will have to first install `ggplot2` (if you have not done so already) and then load the `ggplot2` library.

(d) Discuss the risk-return tradeoff shown in the figure.

9. Using `sapply` and a function defined by you, compute the Sharpe ratio for each asset. Which one has the best estimated risk-reward tradeoff according to the Sharpe ratio?

10. Consider creating a portfolio with weight w_A on SPDR Gold Shares (GLD) and weight $w_M = 1 - w_M$ on the S&P 500 index fund.

(a) Given your answers above and the rule for variance of a sum, why might it be appropriate to include both SPDR Gold Shares and the S&P 500 index shares in the same portfolio?

(b) Use the data to create a figure similar to Figure 0.3 on slide 18 of the review slides graphing the expected return and volatility of the portfolio as a function of w_A.

(c) How to evaluate the risk-return tradeoff for the portfolio vs. only SPDR Gold or only S&P 500 index? Given your estimates, would an investor ever wish to invest 100% in the S&P 500 index?