Final Exam – Suggested Solutions

Program: Professional Master in Economics
Instructor: Hedibert Freitas Lopes
Insper Institute of Education and Research

Course: Advanced Econometrics
Teaching Assistant: Guilherme Piantino
Date: December 13th 2025

Answers and comments in red. These are suggested solutions; alternative well-justified answers should also receive credit.

1 Bayesian updating

Suppose a doctor suggests to one of her patients to be tested for a particular Illness. Let I denote the event that the person has the illness. From a large population, the doctor knows by experience that the prevalence (prior probability) of the illness is estimated to be 0.3, i.e. $Pr(I) = 0.3$. He also knows that the test is imperfect. Let A denote the event that the test result is positive (i.e., Abnormal or Alert). The test's characteristics are as follows:

- If a person has the illness, the chance that the test will be positive is 0.8, i.e. $P(A|I) = 0.8$.
- If a person does not have the illness ($\bar{I}$), the chance that the test will be positive is 0.4, i.e. $P(A|\bar{I}) = 0.4$.

What is the chance (up to two decimal points) that the person actually has the illness (I), knowing for a fact that their test result came back positive (A)? In other words, what is the value of $P(I|A)$?

(a) 0.80 (b) 0.52 (c) 0.46 (d) 0.24 (e) None of the previous cases.

Hint: For events A and B, and $P(B) > 0$, it follows that

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Solution. First compute the marginal (predictive) probability of a positive test by the law of total probability:

$$P(A) = P(A|I)P(I) + P(A|\bar{I})P(\bar{I}) = 0.8 \times 0.3 + 0.4 \times 0.7 = 0.24 + 0.28 = 0.52.$$

Then, by Bayes' theorem,

$$P(I|A) = \frac{P(A|I)P(I)}{P(A)} = \frac{0.24}{0.52} = \frac{6}{13} \approx \mathbf{0.46}. \qquad \text{Answer: (c).}$$

Comments. The distractors correspond to common mistakes: 0.80 is $P(A|I)$ (confusing the inverse conditional), 0.52 is $P(A)$, and 0.24 is the joint $P(A \cap I)$. The positive test raises the probability of illness from 0.30 to 0.46; the update is modest because the false-positive rate (0.4) is high: the likelihood ratio is $P(A|I)/P(A|\bar{I}) = 2$, so the prior odds $0.3/0.7 = 0.43$ double to posterior odds 0.86, i.e. $0.86/1.86 = 0.46$.

2 Linear regression

Suppose that you have a data set of monthly temperatures in the city of São Paulo. Denote by y_t the temperature in month t, for $t = 1, \ldots, n$. You want to model y_t as a linear function of the temperatures in the previous two months, but you also want to take into account the monthly season effect. Write the mean function of such linear auto-regression with dummies for the months. Give as much details about your derivation as possible.

Solution. Dummies. For $j = 1, \ldots, 12$ define $D_{jt} = 1$ if observation t falls in month j (1 = January, ..., 12 = December) and $D_{jt} = 0$ otherwise. Since $\sum_{j=1}^{12} D_{jt} = 1$ for every t, we cannot include an intercept *and* all 12 dummies (perfect multicollinearity, the "dummy-variable trap"). Take January as the baseline and include $D_{2t}, \ldots, D_{12t}$.

Model (AR(2) with seasonal dummies). For $t = 3, \ldots, n$,

$$y_t = \beta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \sum_{j=2}^{12} \gamma_j D_{jt} + \varepsilon_t, \qquad \varepsilon_t \text{ iid } N(0, \sigma^2)$$

(or, more generally, $E(\varepsilon_t | y_{t-1}, y_{t-2}, D_t) = 0$).

Mean function.

$$E(y_t \mid y_{t-1}, y_{t-2}, D_t) = \beta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \sum_{j=2}^{12} \gamma_j D_{jt}.$$

Hence, if t is a January, $E(y_t|\cdot) = \beta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2}$; if t is month $j \neq 1$, $E(y_t|\cdot) = (\beta_0 + \gamma_j) + \phi_1 y_{t-1} + \phi_2 y_{t-2}$. That is, each month has its own intercept, while the autoregressive dynamics (ϕ_1, ϕ_2) are common to all months.

Interpretation.

- γ_j: expected temperature difference between month j and January, *holding the two previous temperatures fixed.*
- ϕ_1, ϕ_2: persistence of temperature shocks. Stationarity requires the roots of $1 - \phi_1 z - \phi_2 z^2 = 0$ to lie outside the unit circle (equivalently $\phi_1 + \phi_2 < 1$, $\phi_2 - \phi_1 < 1$, $|\phi_2| < 1$).
- Equivalent parametrization: drop β_0 and include all 12 dummies, $E(y_t|\cdot) = \sum_{j=1}^{12} \delta_j D_{jt} + \phi_1 y_{t-1} + \phi_2 y_{t-2}$, with $\delta_1 = \beta_0$ and $\delta_j = \beta_0 + \gamma_j$.

Matrix form and estimation. Stack $\mathbf{y} = (y_3, \ldots, y_n)'$ (the first two observations are lost as initial conditions, so the effective sample size is $n-2$) and let row t of X be $x_t' = (1,\ y_{t-1},\ y_{t-2},\ D_{2t}, \ldots, D_{12t})$, a $(n-2) \times 14$ matrix. Then $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with
$\boldsymbol{\beta} = (\beta_0, \phi_1, \phi_2, \gamma_2, \ldots, \gamma_{12})'$, and the (conditional) OLS/MLE estimator is $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$, with $\hat{\sigma}^2 = \|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|^2/(n - 2 - 14)$. (Under a flat prior, this is also the posterior mean of $\boldsymbol{\beta}$.)

Hint for grading: key elements are (i) two lags, (ii) 11 dummies + intercept (or 12 dummies without intercept), (iii) the conditional expectation written explicitly, (iv) interpretation of the coefficients.

3 Logistic regression

In class, we exemplify the estimation of a logistic regression using the Credit Card Default Data, with 10 thousand observations and four variables. The response variable was `default`, a factor with levels No and Yes indicating whether the customer defaulted on their debt and the explanatory/regressor variable is `student`, A factor with levels No and Yes indicating whether the customer is a student. Below is the R script which transforms the variables into 0/1 Bernoulli trials.

```
install.packages("ISLR")
library(ISLR)
data(Default)
Default[c(1:3,9998:10000),]     # First and last few observations
      default student   balance   income
1          No      No  729.5265 44361.63
2          No     Yes  817.1804 12106.13
3          No      No 1073.5492 31767.14
9998       No      No  845.4120 58636.16
9999       No      No 1569.0091 36669.11
10000      No     Yes  200.9222 16862.95
attach(Default)
n = nrow(Default)
y = rep(0,n)
y[default=="Yes"]=1
x1 = rep(0,n)
x1[student=="Yes"]=1
x2 = scale(income)
x3 = scale(balance)
```

- The variable y is binary with $y = 1$ if the customer defaulted and $y = 0$ otherwise.
- The estimation output for the logistic regression with x_1, x_2 and x_3 regressors is given below.
- The regressors are, respectively, dummy for student (1 if student, 0 if non student), income and balance.
- Both, income and balance are scaled, ie they are measured as standardized deviations from the mean income and mean balance, respectively.

Hint: The probability of default for a customer i in the logistic regression is given by

$$Pr(y_i = 1|x_{i1}, x_{i2}, x_{i3}) = \pi_i = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})\}}$$

```
glm.fit = glm(y~x1+x2+x3,family=binomial)
summary(glm.fit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.97524    0.19368 -30.851  < 2e-16 ***
x1          -0.64678    0.23625  -2.738  0.00619 **
x2           0.04046    0.10940   0.370  0.71152
```

```
x3           2.77483    0.11217  24.738  < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

Questions. Using the point estimates of the logistic regression parameters, compute the point estimate for the probability of default for a student customer with an income of 1 standard deviation above the mean and balance of 2 standard deviations above the mean. Compare with the probability of default for a non-student customer with the same characteristics. Comment your findings, including your understanding about the importance of x_1, x_2 and x_3 to explain y.

Solution. Set $x_2 = 1$ and $x_3 = 2$.

Student ($x_1 = 1$): linear predictor

$$\eta_S = -5.97524 - 0.64678 + 0.04046(1) + 2.77483(2) = -1.0319,$$
$$\hat{\pi}_S = \frac{1}{1 + e^{1.0319}} = \frac{1}{1 + 2.806} \approx \mathbf{0.263}.$$

Non-student ($x_1 = 0$):

$$\eta_N = -5.97524 + 0.04046 + 5.54966 = -0.3851,$$
$$\hat{\pi}_N = \frac{1}{1 + e^{0.3851}} = \frac{1}{1 + 1.470} \approx \mathbf{0.405}.$$

Comparison. For the same income and balance, the estimated default probability of a student (26.3%) is about 14 percentage points lower than that of a non-student (40.5%). In terms of odds, $e^{\hat{\beta}_1} = e^{-0.64678} \approx 0.52$: being a student roughly halves the odds of default, *ceteris paribus*. (Odds: $e^{-1.0319} = 0.356$ vs. $e^{-0.3851} = 0.680$; ratio $= 0.52$.)

Importance of the regressors.

- x_3 (balance) is by far the most important: $z = 24.7$, $p < 2 \times 10^{-16}$. One s.d. increase in balance multiplies the odds of default by $e^{2.775} \approx 16$. Both customers above are high-risk mainly because their balance is 2 s.d. above the mean (at mean income and balance, a non-student has $\hat{\pi} = 1/(1 + e^{5.975}) \approx 0.25\%$).
- x_1 (student) is significant at 1% ($z = -2.74$, $p = 0.006$) with a negative effect *conditional on balance and income*.
- x_2 (income) is not significant ($z = 0.37$, $p = 0.71$): once balance and student status are accounted for, income adds essentially no information (an odds ratio of $e^{0.040} \approx 1.04$ per s.d.); it could be dropped.
- The large drop in deviance (2920.6 $\to$ 1571.5 for 3 d.f.) shows the regressors jointly explain default very well.

Hint (confounding / Simpson-type reversal): a logistic regression of y on x_1 alone gives a *positive* student coefficient (students default more, unconditionally). The sign flips here because students tend to carry higher balances (and have lower incomes; student and income are strongly correlated). For a *given* balance, students are less risky. This is a good illustration of why the marginal and conditional effects of a regressor can differ.

4 Causality

The following is taken from the Uber page:

> "While we run a lot of experiments at Uber, we do not want to negatively impact the customer experience with our research. To estimate the causal impact of some variable that hasn't been randomized, we can leverage observational data rather than data obtained via experimentation.
>
> As an example, we might want to know how experiencing an event like a delay in food delivery (treatment) can influence Uber Eats customers' future engagement with the platform (outcome). However, since delaying food deliveries would impact the customer experience negatively, we need to use alternative approaches to estimate the impact of such an event.
>
> Examining observational data consisting of users who happened to experience a delayed delivery and users who didn't, we can begin to understand the impact without resorting to experimentation. However, simply calculating the difference in future customer engagement between these two groups would not likely give us a meaningful answer. The users who have experienced delays may be importantly different from those who haven't." https://www.uber.com/en-BR/blog/causal-inference-at-uber/

Questions:

a) Explain their statement *simply calculating the difference in future customer engagement between these two groups would not likely give us a meaningful answer.*

b) Give an example of a potential observable or unobservable confounder that may affect both treatment and outcome.

c) Draw a direct acyclic graph (DAG) representation.

d) Which econometric method is the best for identifying the causal effect?

Solution. Notation: $T_i = 1$ if user i experienced a delay, Y_i = future engagement, $Y_i(1), Y_i(0)$ the potential outcomes, X_i (observed) and U_i (unobserved) confounders.

a) Delays were not randomly assigned, so the two groups are not comparable. The naïve difference decomposes as

$$\underbrace{E[Y|T=1] - E[Y|T=0]}_{\text{naïve difference}} = \underbrace{E[Y(1) - Y(0)|T=1]}_{\text{ATT (causal)}} + \underbrace{E[Y(0)|T=1] - E[Y(0)|T=0]}_{\text{selection bias}}.$$

The selection bias is zero only if, absent the delay, both groups would have had the same engagement, which is implausible: the users who get delays order at different times, from different places and restaurants, and with different frequency. The naïve difference therefore mixes the causal effect of the delay with pre-existing differences between the groups (confounding/selection bias), and can even have the wrong sign.

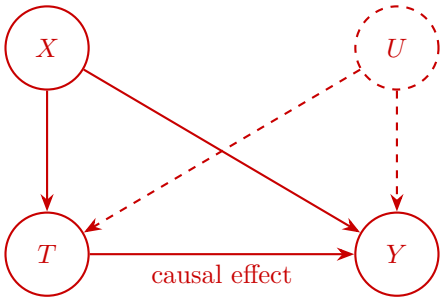
b) Examples (any one, well justified, is enough):

- *Observable*: past order frequency/tenure (heavy users order more often, so are mechanically more likely to experience at least one delay, and are also more engaged in the future: bias *understates* the harm of delays); order time (peak hours, Friday night); weather/rain; distance between restaurant and customer; neighborhood (traffic, urban vs. peripheral); restaurant popularity/busyness; order size.

- *Unobservable*: customer patience or loyalty/attachment to the platform; availability of competing apps or restaurants nearby; the customer's income/time constraints.

For instance, *rain* increases the probability of delays and also changes ordering behavior (people order more delivery when it rains), affecting both T and Y.

c) DAG:



T: delay (treatment); Y: future engagement (outcome); X: observed confounders (past usage, time, weather, distance, region); U: unobserved confounders (patience, loyalty). The paths $T \leftarrow X \rightarrow Y$ and $T \leftarrow U \rightarrow Y$ are *back-door paths* that must be blocked to identify $T \rightarrow Y$.

d) There is no single "best" method; it depends on the identifying assumption one is willing to defend:

- If all confounders are observed (*unconfoundedness*: $\{Y(0), Y(1)\} \perp T \,|\, X$, plus overlap $0 < P(T = 1|X) < 1$): control for X via regression adjustment, **propensity-score** methods (matching, inverse-probability weighting), or doubly-robust estimators; flexible versions such as BART / Bayesian Causal Forests or double machine learning allow heterogeneous effects (CATE) across users. This is the natural answer for the Uber setting, which has a rich set of logged covariates, and is what the Uber blog describes.
- If important confounders U are unobserved: **instrumental variables** (e.g., an exogenous shock that affects delays but not engagement directly, such as courier shortages or traffic incidents), **difference-in-differences** with user panel data (compare engagement before/after a delay, removing time-invariant U), or **regression discontinuity** if delays are defined by a cutoff (e.g., arrival just after vs. just before the promised time).

Hint for grading: accept propensity score / matching / IPW / regression adjustment (under unconfoundedness) or IV / DiD / RDD (with unobserved confounders), as long as the student states the identifying assumption.

5 Time Series

Assume that the temporally observed data $\{y_1, \ldots, y_n\}$ follows an stationary AR(1) model:

$$y_t = \alpha + \beta y_{t-1} + \varepsilon_t \qquad \varepsilon_t \text{ iid } N(0, \sigma^2).$$

where $\theta = (\alpha, \beta, \sigma^2)$. Derive the following mean and variance forecasting functions: $E(y_{n+3}|y_{1:n}, \theta)$ and $V(y_{n+3}|y_{1:n}, \theta)$. These are the mean and variance of the time series 3-steps ahead after observing $y_1, \ldots, y_n$.

Hint 1: It might be easier to first derive y_{n+3} as a function of $y_{1:n}$, θ, ε_{n+1}, ε_{n+2} and ε_{n+3}, and then compute the mean and the variance forecasting functions.
Hint 2: $E(a + bZ + cW) = a + bE(Z) + cE(W)$.
Hint 3: $V(a + bZ + cW) = b^2V(Z) + c^2E(W)$, when Z and W are uncorrelated random variables. [Typo: should read $c^2V(W)$.]

Solution. Step 1 (recursive substitution).

$$\begin{aligned}
y_{n+1} &= \alpha + \beta y_n + \varepsilon_{n+1}, \\
y_{n+2} &= \alpha + \beta y_{n+1} + \varepsilon_{n+2} = \alpha(1+\beta) + \beta^2 y_n + \beta\varepsilon_{n+1} + \varepsilon_{n+2}, \\
y_{n+3} &= \alpha + \beta y_{n+2} + \varepsilon_{n+3} = \alpha(1+\beta+\beta^2) + \beta^3 y_n + \beta^2\varepsilon_{n+1} + \beta\varepsilon_{n+2} + \varepsilon_{n+3}.
\end{aligned}$$

Given $y_{1:n}$ and θ, the first two terms are constants and the future shocks $\varepsilon_{n+1}, \varepsilon_{n+2}, \varepsilon_{n+3}$ are iid $N(0, \sigma^2)$, independent of $y_{1:n}$. Note that only y_n matters (Markov property of the AR(1)).

Step 2 (mean). Using Hint 2 and $E(\varepsilon_{n+h}) = 0$:

$$\boxed{E(y_{n+3}|y_{1:n}, \theta) = \alpha(1+\beta+\beta^2) + \beta^3 y_n = \alpha\,\frac{1-\beta^3}{1-\beta} + \beta^3 y_n.}$$

Step 3 (variance). Using Hint 3 (corrected) and independence of the shocks:

$$\boxed{V(y_{n+3}|y_{1:n}, \theta) = \beta^4\sigma^2 + \beta^2\sigma^2 + \sigma^2 = \sigma^2(1+\beta^2+\beta^4) = \sigma^2\,\frac{1-\beta^6}{1-\beta^2}.}$$

Moreover, since it is a linear combination of normals, $y_{n+3}|y_{1:n}, \theta \sim N\big(\alpha(1+\beta+\beta^2) + \beta^3 y_n,\ \sigma^2(1+\beta^2+\beta^4)\big)$.

Hint (general h and intuition): for any horizon h,

$$E(y_{n+h}|y_{1:n}, \theta) = \alpha\frac{1-\beta^h}{1-\beta} + \beta^h y_n \xrightarrow{h\to\infty} \frac{\alpha}{1-\beta}, \qquad V(y_{n+h}|y_{1:n}, \theta) = \sigma^2\frac{1-\beta^{2h}}{1-\beta^2} \xrightarrow{h\to\infty} \frac{\sigma^2}{1-\beta^2},$$

since $|\beta| < 1$ (stationarity). As the horizon grows, the forecast reverts to the unconditional mean and the forecast variance increases toward the unconditional variance of the process: the information in y_n fades away.