Final BUEC 333

August 10, 2013, 12:00-15:00

Name: _____
Student ID: _____

Read carefully before starting

- Allowed on your desk: a non-graphical calculator, a pen, a ruler, your SFU ID, and something to eat+drink. **If we find anything else, you lose 10% of the max score.** This includes erasers, cases, ...
- On the **front page of this document** (the questions), write: (i) your name, and (ii) your student ID.
- On the **front page of your answer sheet,** write: (i) your name, and (ii) your student ID.
- Answer the questions in **chronological** order.
- For every subquestion (e.g. for 2 (c)), use **maximum 3 lines**
- If you finish this exam **before 14:30**, come forward and hand in the documents.
- If you finish it **after 14:30**, stay seated! Raise your hand, and we will come to collect your final.
- Each of the $5+5+4+6+4+2+4+3=33$ subquestions is worth 1 point. For each you receive either 0 or 1 point: no partial marks.
- Adding incorrect or irrelevant statements to an otherwise correct answer will result in 0 points for that subquestion.

Questions

1. **Three coin flips.** You are about to flip 3 coins. The random outcome is what the coins are showing. For example, you could throw HHT: first two coins show heads, third shows tails. Or HTH, TTT, etcetera. Consider the RV's:

   X_1**:** "the number of heads showing on the **first two** flips";

   X_2**:** "the number of heads showing **times** the number of tails showing".

   (a) What is the probability distribution for X_2? (Hint: You first need to determine the sample space.)

   (b) Compute $\mathrm{E}(X_2)$

   (c) Compute $\mathrm{Var}(X_1)$.

   (d) Compute $\mathrm{Cov}(X_1, X_2)$.

   (e) Compute $\mathrm{E}(X_2 | X_1 = 2)$.

2. **Estimation.** Suppose you want to know the mean value of Y, $\mathrm{E}(Y) = \mu_Y$. You have a random sample of size n, $\{Y_1, \cdots, Y_n\}$. For simplicity, assume that $n = 2$. Then, the sample average $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, simplifies to $\bar{Y} = (Y_1 + Y_2)/2$.

   (a) Is the sample average unbiased for μ_Y? Explain.

   (b) Now, consider $\tilde{Y} = \frac{1}{4}Y_1 + \frac{1}{4}Y_2$. What is the variance of $\tilde{Y}$?

   (c) On the basis of the variances, do you prefer $\bar{Y}$ or $\tilde{Y}$?

   (d) What is wrong with $\tilde{Y}$?

   (e) Now, consider the estimator $\check{Y} = \left(\mu_Y + \bar{Y}\right)/2$. Why is this not a good estimator? (Hint: the answer has nothing to do with efficiency, unbiasedness, or consistency.)

3. **Linear regression with one regressor.** This question tests your understanding of the concepts involved in regression analysis. For this question, consider the model with one regressor,

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

   (a) Write down the population regression function. Write down the sample regression function.

   (b) What does Least Squares Assumption 1, $\mathrm{E}(u_i|X_i) = 0$, mean? In your answer, use the words "other factors".

   (c) Describe the other two "Least Squares Assumptions"? You only need to use one sentence for each assumption.

   (d) Is β_1 a random variable? Is $\hat{\beta}_1$ an estimator? Is u_i a random variable? Is $\hat{u}_i$ a random variable? (First: 4x yes/no. Then: short explanations if you are not sure about your yes/no answers.)

4. **Regression and hypothesis testing.** This question is related to exercise 4.3 in Stock and Watson. A regression of average weekly earning (AWE, measured in dollars) on Age (in years) using a random sample of college educated full time wotkers aged 25-65 yields the following:

$$\widehat{AWE} = 696.7 + 9.6\,Age,\ R^2 = 0.023,\ SER = 604.3.$$

   (a) Interpret the value 9.6. Is that value, 9.6, an estimand, an estimator, or an estimate?

   (b) What are the units of the R^2? (Dollars? Years? Unit-free?)

   (c) What are the regression's predicted earnings for a 30-year-old worker? Will the regression give reliable predictions for a 99-year-old worker? Explain.

   (d) The average age in the sample is 41.6. What is the average value of AWE in the sample? Explain.

   (e) Assume that the standard error for the estimated regression coefficient of Age is 1.2. Construct a 95% confidence interval for β_1, the regression coefficient for Age. (Note: you should remember the critical value from all your CI practice.)

   (f) Consider the $p-$value associated with the two-sided test for $H_0 : \beta_1 = 0$. Do you expect this $p-$value to be smaller than, equal to, or greater than, 0.05? Explain.

5. **Nonlinear regression, SW Exercise 8.7.** This problem is inspired by the study of the gender gap in top corporate jobs in Bertrand and Hallock (2001). The study compares total compensation among top executives in a large set of U.S. public corporations in the 1990s.

   (a) Let $Female$ be an indicator variable that is equal to 1 for females and to 0 for males. A regression of the logarithm of earnings onto $Female$ yields

$$\log\left(\widehat{Earnings}\right) = 6.48 - 0.44 Female$$

   where the estimated regression coefficient -0.44 has a standard error of 0.05. Explain what the -0.44 means.

   (b) Does this regression suggest that there is gender discrimination? Explain.

   (c) Two new variables are added to the regression: $\log(MarketValue)$, where $MarketValue$ is a measure of firm size, in millions; and $Return$, the stock return, in percentage points. The resulting estimated regression line is

$$\log\left(\widehat{Earnings}\right) = 3.68 - 0.28 Female + 0.37 \log(MarketValue) + 0.004 Return$$

   where the standard errors for the three regressors are 0.04, 0.004, and 0.003, respectively. Explain what the 0.37 means, i.e. what is its interpretation?

   (d) The coefficient estimate for $Female$ has changed from -0.44 to -0.28. Why has it changed?

6. **Internal and external validity.**

   (a) Describe the difference between "internal validity" and "external validity".

(b) List three threats to internal validity. (The book lists five).

7. **Panel data.** Consider the example used in the chapter on panel data. We have a panel data set on $n = 48$ U.S. states during $T = 7$ periods, from 1982 up to and including 1988. The total number of observations is 336.

   (a) Is this a balanced panel? Explain.

   (b) For each state, in each time period, let Y_{it} denote the number of annual traffic deaths per 10000 in the population. Let X_{it} denote the beer tax in 1988 U.S. dollars. Temporarily ignore the data after 1982, so that we have a cross-section of 48 states. The estimated regression line gives

   $$\hat{Y}_{i,1982} = 2.01 + 0.13X_{i,1982}.$$

   If the Least Squares assumptions hold for this regression, how would you interpret the 0.13? Include in your answer: "tax on beer".

   (c) The estimated fixed effects regression line is

   $$\hat{Y}_{i,t} = \hat{\alpha}_i - 0.66X_{i,t}. \quad (0.29)$$

   How would you interpret the -0.66?

   (d) Consider the results for the fixed effects regression. Do you think that the Least Square assumptions hold, i.e. do you believe that the 0.13 in the first result comes from an unbiased estimator?

8. **A question about hand-in assignment 2.** In the context of your second hand-in assignment, consider the following code and output:

```
> unionData <- read.dta("http://www.sfu.ca/~cmuris/2013-Summer-333/wagepan.dta")
> summary(lm(lwage~union+hours+year+black+occ1+occ2+occ3+ ...
   ... + occ4+occ5+occ6+occ7+occ8+occ9+exper,data=unionData))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.422e+02  1.145e+01 -12.424  < 2e-16
union        2.376e-01  1.779e-02  13.355  < 2e-16
hours       -5.794e-05  1.373e-05  -4.219 2.50e-05
year         7.250e-02  5.784e-03  12.534  < 2e-16
black       -1.306e-01  2.354e-02  -5.549 3.05e-08
occ1         3.966e-01  3.218e-02  12.323  < 2e-16
occ2         3.542e-01  3.364e-02  10.529  < 2e-16
occ3         2.979e-01  3.922e-02   7.596 3.73e-14
occ4         1.778e-01  3.108e-02   5.721 1.13e-08
occ5         2.605e-01  2.738e-02   9.516  < 2e-16
occ6         1.831e-01  2.747e-02   6.667 2.94e-11
occ7         8.598e-02  3.278e-02   2.623  0.00874
occ8        -5.678e-02  6.579e-02  -0.863  0.38816
occ9                NA         NA      NA       NA
exper       -1.252e-02  4.667e-03  -2.682  0.00734

Residual standard error: 0.4888 on 4346 degrees of freedom
Multiple R-squared:  0.1601,    Adjusted R-squared:  0.1576
F-statistic: 63.74 on 13 and 4346 DF,  p-value: < 2.2e-16
```

   (a) Why are there "NA"s in the row for "occ9"?

   (b) The adjusted R-squared is lower than the R-squared. Are you surprised? Explain.

   (c) Use the reported p-value for exper to perform a two sided test for the null hypothesis that the regression coefficient of **exper** is equal to 0. The level of the test should be 1%.