Final BUEC 333

August 11, 2014, 12:00-15:00

Name: ______________________
Student ID: ______________________

Read carefully before starting

- Allowed on your desk: a pen, a ruler, your SFU ID, and something to eat+drink. **If we find anything else, you lose 20% 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.** This excludes math.
- Every subquestion is worth 1 point. **No partial marks** will be given for a subquestion: be complete and precise. Adding incorrect or irrelevant statements to an otherwise correct answer will result in 0 points.
- 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.

Questions

1. [Stock and Watson, 2.5] "In September, Seattle's daily high temperature has a mean of 59 degrees Fahrenheit. The standard deviation is 9 degrees Fahrenheit." Remember that, to convert from degrees Fahrenheit to degrees Celsius, we need to subtract 32 and then multiply by 5/9, so

$$T_C = \frac{5}{9} \times (T_F - 32)\,.$$

   For the questions that follow, indicate which formulas you are using.

   (a) What is the mean of Seattle's daily high temperature in September in degrees Celsius?
   (b) What is the standard deviation of Seattle's daily high temperature in September in degrees Celsius?
   (c) What is the variance of Seattle's daily high temperature in September in degrees Celsius?

2. Consider the joint probability distribution of the random variables X_1 and X_2 given in Table 1.

   (a) What is the probability that $X_1 = 0$ given that $X_2 = 1$? In other words, what is $P(X_1 = 0|X_2 = 1)$?
   (b) What is the probability that $X_1 = 0$ given that $X_2 = 3$?
   (c) What is $var(X_1|X_2 = 2)$?

3. Suppose you want to estimate the population mean 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 $\bar{Y} = (Y_1 + Y_2)/2$.

	$X_2 = 1$	$X_2 = 2$	$X_2 = 3$
$X_1 = 0$	0.2	0.1	0.1
$X_1 = 1$	0.3	0.3	0

Table 1: Joint probability distribution of X_1 and X_2.

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

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

(d) 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.)

4. Consider the following estimated regression equation that describes the relationship between a student's weight and height:

$$\widehat{WEIGHT} = 100 + 6.0\,HEIGHT$$

(a) A student has height 5. What is the regression's prediction for that student's weight?

(b) In the sample, the sample average of HEIGHT is 4. What can you say about the sample average for WEIGHT?

Now, an additional variable is included, is ID, a student's SFU ID. Obviously, this is a nonsensical variable to include: it is not in any way related to a student's weight. The new estimated regression equation is

$$\widehat{WEIGHT} = 101.5 + 5.98\,HEIGHT + 0.02\,ID$$

(c) Someone's weight has nothing to do with their SFU ID. Still, the R^2 went up from 0.74 to 0.75. How is this possible?

(d) If the post office box number is not related to a student's weight, should the estimated coefficient not be equal to 0? How could it be that it is 0.02?

5. Let D_i be a dummy variable. Consider the model that consists of the equation

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 D_i X_i + u_i$$

and the standard OLS assumptions

(a) Draw a graph to visualize this model.

(b) What is the interpretation of β_3?

6. [Based on SW 4.3]. 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) 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.

(d) 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.

7. [Based on 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

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

where the standard errors for the three regressors are 0.04, 0.004, and 0.003, respectively. The coefficient estimate for $Female$ has changed from -0.44 to -0.28. Why has it changed?

8. A mixed bag of questions:

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

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

   (c) What are the two conditions an instrumental variable must satisfy?

9. 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

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

   Alternatively, we can use fixed effects regression to estimate the effect fixed effects regression line is

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

   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? If YES: explain what causes the difference between 0.13 and -0.66. If NO: explain why the Least Square assumptions are unlikely to hold. Include in your answer: "tax on beer".

10. [Based on SW, 14.7] Suppose that Y_t follows the stationary AR(1) model

    $$Y_t = 3.9 + 0.5Y_{t-1} + u_t$$

    where u_t is i.i.d. with $E(u_i) = 0$ and $\text{var}(u_i) = 9$.

    (a) Compute the mean and variance of Y_t.

    (b) Compute the first autocovariance of Y_t.

    (c) Compute the second autocovariance of Y_t.

11. In the context of your second hand-in assignment, consider the following code and output. Why are there "NA"s in the row for "occ9"?

```
> unionData <- read.dta("http://www.sfu.ca/~cmuris/2014-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
```