San Francisco State University **Michael Bar**
ECON 312 **Fall 2018**

Problem set 5

Due Thursday, November 15, in class

Names of collaborators (type below):

Name 1 MINJIN PARK

Name 2 NISHANLANG KHONGLAH

Assignment Rules

1. Homework assignments must be typed. For instruction on how to type equations and math objects please see notes "Typing Math in MS Word".
2. Homework assignments must be prepared within this template. Save this file on your computer and type your answers following each question. Do not delete the questions.
3. Your assignments must be stapled.
4. No attachments are allowed. This means that all your work must be done within this word document and attaching graphs, questions or other material is prohibited.
5. Homework assignments must be submitted at the end of the lecture, in class, on the listed dates.
6. Late homework assignments will not be accepted under any circumstances, but the lowest homework score will be dropped.
7. The first homework assignment cannot be dropped.
8. You are encouraged to work on this homework assignment in groups of up to 3 people, and submit one assignment with up to 3 names typed on this page. Sharing the electronic version of your assignment with other teams is absolutely prohibited.
9. All the graphs should be fully labeled, i.e. with a title, labeled axis and labeled curves.
10. In all the questions that involve calculations, you are required to show all your work. That is, you need to write the steps that you made in order to get to the solution.
11. This page must be part of the submitted homework.

Perfect Multicollinearity

1. (20 points). For this exercise use the `wage21` data posted on the course website.

   a. Suppose that you want to study the determinants of schooling attendance, and estimate the following model:

   $$S = \beta_1 + \beta_2 SF + \beta_3 SM + \beta_4 ASVABC + u$$

   where S is individual's schooling level in years, SF and SM is schooling of father and mother respectively (also in years) and ASVABC is ability score composed of math and verbal skills. Estimate this model, and present the R regression output here.

   The regression output of the model above is given below.

```
lm(formula = S ~ SF + SM + ASVABC, data = wage)

Residuals:
Min          1Q           Median       3Q           Max
-5.8209      -1.3835      -0.2735      1.2776       6.1738

Coefficients:
             Estimate     Std. Error   t value      Pr(>|t|)
(Intercept)  5.51014      0.48818      11.287       < 2e-16 ***
SF           0.15081      0.03066      4.919        1.16e-06 ***
SM           0.08908      0.03771      2.362        0.0185 *
ASVABC       0.10486      0.00950      11.037       < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.91 on 536 degrees of freedom
Multiple R-squared:  0.3683,  Adjusted R-squared:  0.3648
F-statistic: 104.2 on 3 and 536 DF,  p-value: < 2.2e-16
```

   b. Suppose that you generate another variable, $SP = SM + SF$, which stands for “Schooling of Parents” and add this as a regressor to the above model. In other words, you are trying to estimate:

   $$S = \beta_1 + \beta_2 SF + \beta_3 SM + \beta_4 ASVABC + SP + u$$

   Present the R regression output here and explain clearly and briefly, why R was not able to estimate your model and it had to drop one variable.

   The regression output of the model above is given below.

```
lm(formula = S ~ SF + SM + ASVABC + SP, data = wage)

Residuals:
Min          1Q           Median       3Q           Max
-5.8209      -1.3835      -0.2735      1.2776       6.1738

Coefficients: (1 not defined because of singularities)
```

```
             Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)  5.51014     0.48818      11.287     < 2e-16 ***
SF           0.15081     0.03066      4.919      1.16e-06 ***
SM           0.08908     0.03771      2.362      0.0185 *
ASVABC       0.10486     0.00950      11.037     < 2e-16 ***
SP           NA          NA           NA         NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.91 on 536 degrees of freedom
Multiple R-squared:  0.3683,  Adjusted R-squared:  0.3648
F-statistic: 104.2 on 3 and 536 DF,  p-value: < 2.2e-16
```

R was unable to estimate the above model because the model exhibits perfect multicollinearity and one of the regressors, in this case SP, has been dropped by R because SP can be obtained as a linear combination of SM and SF.

Imperfect Multicollinearity

2. (30 points). For this question use `multi.csv` posted on the course website. The file contains hypothetical data on consumption c, income i, and wealth w.
   a. Suppose that you want to study the determinants of consumption expenditure. Estimate the following models:

$$[model1]: c = \beta_1 + \beta_2 i + u$$
$$[model2]: c = \beta_1 + \beta_3 w + u$$
$$[model3]: c = \beta_1 + \beta_2 i + \beta_3 w + u$$

   Using the stargazer package, present one table with a summary of all 3 models. Your table must be a copy of html file, and not a copy from RStudio console.

	Dependent variable:		
	Consumption		
	(1)	(2)	(3)
Income	0.600***		2.628
	(0.106)		(2.940)
Wealth		0.060***	-0.204
		(0.011)	(0.295)
Constant	24.600	23.748	29.593
	(18.938)	(19.581)	(20.882)
Observations	10	10	10
R^2	0.802	0.793	0.814
Adjusted R^2	0.777	0.767	0.761
Residual Std. Error	19.172 (df = 8)	19.581 (df = 8)	19.831 (df = 7)

F Statistic	32.322*** (df = 1; 8)	30.654*** (df = 1; 8)	15.342*** (df = 2; 7)
Note:			*p<0.1; **p<0.05; ***p<0.01

b. You suspect that model 3 suffers from (imperfect) multicollinearity. What clues from the above table are a sign of presence of multicollinearity?

In the first model, consumption is regressed only on income, and in the second model, consumption is regressed only on wealth. We can observe from the table that in model 1 and model 2, the regressors are highly significant; however, when consumption is regressed on both income and wealth, both regressors become insignificant and the sign on wealth changed from positive to negative.

c. To further explore your suspicion of multicollinearity, report the correlation table created by the command: `cor(multi$i,multi$w)`. Does the result confirm your suspicion of multicollinearity?

	i	w
i	1.0000000	0.9993104
w	0.9993104	1.0000000

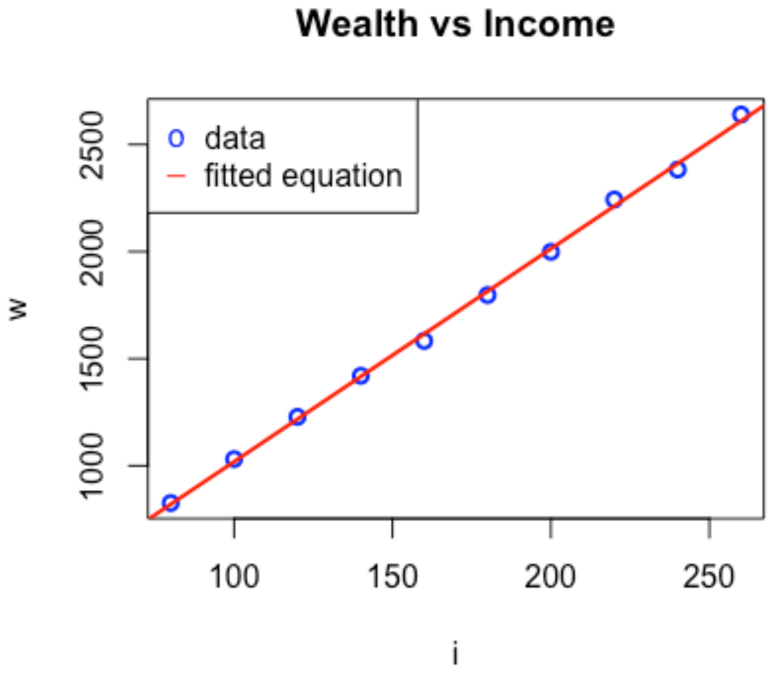
From the correlation table above, we can observe that the correlation between income and wealth is 0.9993104, which is close to 1, which confirms the suspicion of multicollinearity.

d. To further explore your suspicion of multicollinearity, regress wealth on income, and report the R regression output. Does the result confirm your suspicion of multicollinearity?

```
lm(formula = w ~ i, data = multi)

Residuals:
   Min      1Q  Median      3Q     Max
-32.558 -16.664   4.655  11.626  30.788

Coefficients:
             Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)  24.4788     23.4419      1.044      0.327
i            9.9442      0.1306       76.121     9.89e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.73 on 8 degrees of freedom
Multiple R-squared:  0.9986,  Adjusted R-squared:  0.9984
F-statistic:  5794 on 1 and 8 DF,  p-value: 9.886e-13
```

In the regression output presented above, the high R^2 indicates the presence of multicollinearity.

e. Visualize the result in the previous section by creating a graph that plots the actual data points of wealth (y-axis) against income (x-axis), with the fitted line from the regression in the last section. Does this graph confirm your suspicion of multicollinearity?



In the graph above, we can observe that there is almost an exact linear relationship between wealth and income, which therefore, confirms the presence of multicollinearity.

f. Suppose that economic theory tells you that both income and wealth are important for determining consumption, so model 3 is the correct one. However, because of multicollinearity, you were unable to estimate precisely the separate effects of income and wealth on consumption. What can you do in this case, to overcome the problem? (Hint: check the number of observations in the original sample.)

In this case, because economic theory suggests that both income and wealth are important for determining consumption, we can increase the sample size to include more observations to overcome the problem of multicollinearity.

Dummy Variables

3. (65 points). Use data at: http://online.sfsu.edu/mbar/ECON312_files/nlsw88.csv. Let the data frame containing this data be called `nlsw`. This is National Longitudinal Survey of employed Women (NLSW), so all the people in the sample are women. The purpose of

this project is to study the importance of `race` and `union` membership in predicting women's `wage` (measured in $ per hour). To verify that `race` and `union` are indeed factor variables, type `class(nlsw$race)` and `class(nlsw$union)`, and the result should be `factor`.

a. Present the table which gives the number of observations in each `race` category in the sample. The R commands:

```
table1 <- table(nlsw$race)
table1
          black     other     white
          583       26        1637
```

b. Present the table which gives the percentage of observations in each `race` category in the sample. The R command:

```
round(prop.table(table1)*100,2) #Rounding to 2 digits after .

          black     other     white
          25.96     1.16      72.89
```

c. What percentage of the sample are black women?

25.96% of the observations in the sample are black women.

d. Create a set of dummy variables for race and a dummy for union. Suppose W is the dummy for white. The R command that creates this dummy variable is:

```
nlsw$W <- ifelse(nlsw$race=="white", 1, 0) #Dummy for white
```

Write the commands that create dummy variables `B` and `O`, i.e. for “black” and “other” respectively, and a dummy `U` for “union”.

```
nlsw$B <- ifelse(nlsw$race=="black", 1, 0)
nlsw$O <- ifelse(nlsw$race=="other", 1, 0)
nlsw$U <- ifelse(nlsw$union=="union", 1, 0)
```

e. Estimate the model:

$$wage = \beta_1 + \beta_2 B + \beta_3 O + \beta_4 U + \beta_5 tenure + u$$

`tenure` is the number of years worked in the current job. Present the R regression output here.

```
lm(formula = wage ~ B + O + U + tenure, data = nlsw)

Residuals:
   Min          1Q           Median       3Q           Max
-14.566         -5.511       -2.023       3.649        70.125

Coefficients:
                Estimate     Std. Error   t value      Pr(>|t|)
```

```
(Intercept)    13.55716    0.32717    41.438    < 2e-16 ***
B              -2.74194    0.44352    -6.182    7.75e-10 ***
O              2.17856     1.73424    1.256     0.209
U              2.61025     0.45749    5.706     1.35e-08 ***
tenure         0.40540     0.03488    11.622    < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.415 on 1863 degrees of freedom
  (378 observations deleted due to missingness)
Multiple R-squared:  0.1042,  Adjusted R-squared:  0.1023
F-statistic: 54.18 on 4 and 1863 DF,  p-value: < 2.2e-16
```

f. Which is the reference category for race?

   The reference category for race is white.

g. Interpret the estimated coefficients on the race dummies (b_2, b_3).

   b_2 = -2.74 means that black women earn \$2.74/h lesser than white women with the same tenure and union membership.

   b_3 = 2.18 means that women of other races earn \$2.18/h more than white women with the same tenure and union membership.

h. What would happen if you tried to estimate:

   $$wage = \beta_1 + \beta_2 B + \beta_3 O + \beta_4 U + \beta_5 tenure + \beta_6 W + u?$$

   R would not be able to estimate the above model because the model exhibits perfect multicollinearity and one of the regressors, in this case W, will be dropped by R because W is the reference category for race.

i. Interpret the estimated coefficient on the union dummy (b_4).

   b_4 = 2.61 means that unionized women of the same race earn \$2.61/hour more than those with non-union membership with the same tenure.

j. Interpret the estimated coefficient on tenure (b_5).

   b_5 = 0.41 means that every additional year of tenure adds 41 cents to the hourly earnings of women of the same race and the same union membership.

k. Suppose you want to test whether union membership has the same effect on women of all races. The model you wish to estimate is:

$$wage = \beta_1 + \beta_2 B + \beta_3 O + \beta_4 U + \beta_5 tenure + \beta_6 (B \times U) + \beta_7 (O \times U) + u$$

Estimate this model, and present the R regression output.

```
lm(formula = wage ~ B + O + U + tenure + B * U + O * U, data = nlsw)

Residuals:
   Min          1Q           Median       3Q           Max
-15.360        -5.404        -2.075       3.537        68.384

Coefficients:
               Estimate      Std. Error   t value      Pr(>|t|)
(Intercept)    13.74069      0.33174      41.420       < 2e-16 ***
B              -3.74074      0.52042      -7.188       9.49e-13 ***
O              4.66250       2.11157      2.208        0.02736 *
U              1.63159       0.55250      2.953        0.00319 **
tenure         0.41088       0.03476      11.821       < 2e-16 ***
B:U            3.53436       0.98465      3.589        0.00034 ***
O:U            -7.11083      3.66964      -1.938       0.05281 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.379 on 1861 degrees of freedom
  (378 observations deleted due to missingness)
Multiple R-squared:  0.1127,  Adjusted R-squared:  0.1099
F-statistic: 39.41 on 6 and 1861 DF,  p-value: < 2.2e-16
```

l. Interpret the estimated coefficient on B:U (b_6).

   b_6 = 3.53 means that black union members earn $3.53/h more than non-union black workers of the same tenure.

m. Interpret the estimated coefficient on O:U (b_7).

   b_7 = -7.11 means that union members of other races earn $7.11/h lesser than non-union workers with the same tenure.