Assignment 4

Michael Cai

March 7, 2016

1. Let C_t be consumption and X_t be a predictor of consumption. Suppose you have quarterly data on C and X. Let D_{1t}, D_{2t}, D_{3t}, and D_{4t} be dummy variables such that D_{1t} takes the value 1 in quarter 1 and 0 otherwise, etc. Which of the following, if any suffer from perfect multicollinearity and why?

a)
Q1: $C_t = \alpha + \beta X_t + \gamma_1 X_t + u_t$
Q2: $C_t = \alpha + \beta X_t + \gamma_2 X_t + u_t$
Q3: $C_t = \alpha + \beta X_t + \gamma_3 X_t + u_t$
Q4: $C_t = \alpha + \beta X_t + \gamma_4 X_t + u_t$

Yes, this model suffers from perfect multicolllinearity because you have dummy variables for all four quarters. Thus the linear combination of Q1 + Q2 + Q3 + Q4 equals the "dummy" variable for the constant α.

b)
Q1: $C_t = \alpha + \beta X_t + \gamma_1 X_t + u_t$
Q2: $C_t = \alpha + \beta X_t + \gamma_2 X_t + u_t$
Q3: $C_t = \alpha + \beta X_t + \gamma_3 X_t + u_t$
Q4: $C_t = \alpha + \beta X_t + \gamma_4 X_t + u_t$

Yes. Although the model looks slightly different than the first model, they are mathematically equivalent and thus this model also suffers from perfect multicollinearity.

c)
Q1: $C_t = \alpha + \delta_1 + \gamma_1 X_t + u_t$
Q2: $C_t = \alpha + \delta_2 + \gamma_2 X_t + u_t$
Q3: $C_t = \alpha + \gamma_3 X_t + u_t$
Q4: $C_t = \alpha + \gamma_4 X_t + u_t$

Yes, this model also suffers from perfect multicollinearity since there are again 4 dummy variables for 4 possible outcomes (if there are m possible outcomes, there must be m-1 dummy variables if you are to include an intercept, or else there will be perfect multicollinearity with the intercept).

d)
Q1: $C_t = \alpha + \delta_1 + \beta X_t + \gamma_1 X_t + u_t$
Q2: $C_t = \alpha + \delta_2 + \beta X_t + \gamma_2 X_t + u_t$
Q3: $C_t = \alpha + \beta X_t + \gamma_3 X_t + u_t$
Q4: $C_t = \alpha + \beta X_t + u_t$

No, this model does not suffer from perfect multicollinearity because the intercept term acts as the value of C_t when the outcome is Q4, thus the values of Q1-Q3 are in reference to the values in Q4 and thus there is no extraneous dummy variable.

2. a) In models (a)-(d) of question 1, what are the slope coefficients of X_t in each of the 4 quarters? b) Suppose you estimate model (c) and wrote down the estimated slope coefficients for X_t in each of the 4 quarters. You then estimate model (d) and write down the estimated

slope coefficients for X_t in each of the 4 quarters. Do your estimates change? Why or why not? a)

(a)
Q1: $(\beta + \gamma_1)$
Q2: $(\beta + \gamma_2)$
Q3: $(\beta + \gamma_3)$
Q4: $(\beta + \gamma_4)$
(b)
Q1: $(\beta + \gamma_1)$
Q2: $(\beta + \gamma_2)$
Q3: $(\beta + \gamma_3)$
Q4: $(\beta + \gamma_4)$
(c)
Q1: γ_1
Q2: γ_2
Q3: γ_3
Q4: γ_4
(d)
Q1: $(\beta + \gamma_1)$
Q2: $(\beta + \gamma_2)$
Q3: $(\beta + \gamma_3)$
Q4: β

b)

No, they do not change. In (d) the slope coefficients are all based on the slope coefficient for the 4th quarter, β, whereas the slope coefficients in (c) have are not; however, the use of an extra letter to represent the slope coefficients is arbitrary.

To make the two models symbolically equivalent, let's consider the gammas in model (c), γ, to instead be gamma primes, γ'.

Then you would just change the β in Q4 of model (d) to be γ_4', and then $\gamma_1' = \gamma_4' + \gamma_1$, $\gamma_2' = \gamma_4' + \gamma_2$, etc.

The estimates remain the same because structurally the two equations are equivalent there are only arbitrary differences in the labeling of the coefficients.

3. Determine whether the following are true or false and explain why: a) Adjusted R^2 can be negative. b) Adjusted R^2 can be larger than 1.

a)

The definition of adjusted R^2 is $\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = R^2 - (1 - R^2)\frac{p}{n-p-1}$, where p is the number of explanatory variables, and n is the sample size.

As you can see by the definition, if R^2 is close to zero, which means if the explanatory power of the first explanatory variable is close to zero then it is possible that the adjusted R^2 becomes negative.

True.

b)

Also from the definition of adjusted R^2, we see that the adjusted R^2 can never exceed one because the first equation, $\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$, shows that the value of adjusted R^2 has a maximum bound of 1 ($R^2 \leq 1$, and $\frac{n-1}{n-p-1}$ is strictly positive since you can't have a sample size of less than 0).

4. Estimate this regression model in R etc.

a)

$EARN_i = \beta_0 + \beta_1 GEN_i + \beta_2 ED_i + \beta_3 GEN_i ED_i + u_i$
$\hat{\beta_0} = -11325.3$ with $s.e.\hat{\beta_0} = 3848.0$
$\hat{\beta_1} = -2677.9$ with $s.e.\hat{\beta_1} = 5777.47$
$\hat{\beta_2} = 2203.3$ with $s.e.\hat{\beta_2} = 282.1$

$\hat{\beta}_3 = 1017.4$ with $s.e.\hat{\beta}_3 = 420.2$
Adjusted $R^2 = 0.1966$

b)
95% Confidence Interval for β_1
$[-14001.7412, 8645.9412]$
95% Confidence Interval for β_2
$[1650.384, 2756.216]$

c)
$H_0 : \beta_3 = 0$
$H_1 : \beta_3 > 0$
$t = \frac{\hat{\beta}_3 - \beta_{3,0}}{s.e.\hat{\beta}_3} = \frac{1017.4-0}{420.2} = 2.421228$
p-value $= 1 - \phi(2.421228) = 0.007734 < 0.05$
Therefore, we reject the null hypothesis and say that β_3 is significant.

d)
$H_0 : \beta_1 = \beta_3 = 0$
$H_1 : \beta_i \neq 0$
We are testing $EARN_i = \beta_0 + \beta_2 ED_i$ against the full model.

```
fm0 <- lm(EARN ~ ED, data=incomedata)
fm1 <- lm(EARN ~ GEN + ED + GENED, data=incomedata)
waldtest(fm0,fm1,vcov=vcovHC,test=''Chisq'')
```

Thus, the waldtest gives a p-value that is well below a 5% level of significance, and thus we can reject the null hypothesis and say that at at least one of the two βs is significant.

e)
$LOGINC_i = \beta_0 + \beta_1 GEN_i + \beta_2 ED_i + \beta_3 GEN_i ED_i + u_i$
$H_0 : \beta_3 = 0$
$H_1 : \beta_3 > 0$
$t = \frac{\hat{\beta}_3 - \beta_{3,0}}{s.e.\hat{\beta}_3} = \frac{-0.03858-0}{0.02006} = -1.923$
p-value $= 0.02724$, which is still less than 0.05, however had it been a two-tailed test, then the p-value would have been greater than 0.05 and thus we would not be able to reject the null hypothesis in that case.