BS2280 – Econometrics I
Homework 5: Multiple Regression Model I

1

When we interpret the coefficients of a multiple regression model, we always add "holding everything else constant". Using an example, explain what it means.

Holding all other variables constant: Other X variables do not change when X variable of interest is changing. If they all change at the same time, it would be difficult to assess the effect of a change in the X variable of interest on Y. A change in one X variable could increase Y, but a change in another X variable could decrease Y and so on. This would not be informative.
Example:

$$\widehat{officers}_i = \hat{\beta}_1 + \hat{\beta}_2 crimes_i + \hat{\beta}_3 population_i + \hat{\beta}_4 pcinc_i$$

A one unit increase in crimes would increase the number of police officers by $\hat{\beta}_2$ keeping everything else constant, i.e. population and per capita income.

2

Data on 935 individuals was collected to identify what factors can explain the variation in wage data. Firstly, a simple regression is run regressing wages (monthly earnings in USD) on years of education. Secondly, a multiple regression is run regressing wages on years of education and years of work experience.

Simple regression model:

```
Call:
lm(formula = wages2$wage ~ wages2$educ)

Residuals:
    Min      1Q  Median      3Q     Max
-877.38 -268.63  -38.38  207.05 2148.26

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  146.952     77.715   1.891   0.0589 .
wages2$educ   60.214      5.695  10.573   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 382.3 on 933 degrees of freedom
Multiple R-squared:  0.107,     Adjusted R-squared:  0.106
F-statistic: 111.8 on 1 and 933 DF,  p-value: < 2.2e-16
```

Multiple regression model:

```
Call:
lm(formula = wages2$wage ~ wages2$educ + wages2$exper)

Residuals:
    Min      1Q  Median      3Q     Max
-924.38 -252.74  -40.88  198.16 2165.70

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -272.528    107.263  -2.541   0.0112 *
wages2$educ    76.216      6.297  12.104  < 2e-16 ***
wages2$exper   17.638      3.162   5.578 3.18e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 376.3 on 932 degrees of freedom
Multiple R-squared:  0.1359,    Adjusted R-squared:  0.134
F-statistic: 73.26 on 2 and 932 DF,  p-value: < 2.2e-16
```

i Interpret the intercepts and coefficients of both regressions and make reference to their statistical significance.

Regression 1 - Simple regression model:
Intercept: estimate coefficient = 146.952
A person without education would have on average a monthly income of 146USD. The $p - value = 0.0589 > 5\%$ shows that we cannot reject the null hypothesis at the 5% significance level, therefore the intercept is statistically insignificant.

wages2$educ: estimate coefficient = 60.214
One more year of education will increase monthly earning on average by 60USD.

The $p - value < 2e - 16 < 1\%$. This value is highly statistically significant. We can reject the null hypothesis at the 1% significance level.

Regression 2 - Multiple regression model:
Intercept: estimate coefficient = -272.528
A person without education would have on average a negative monthly income of 272.53USD. The $p-value = 0.0112 < 5\%$ shows that we can reject the null hypothesis at the 5% significance level, therefore the intercept is statistically significant. Note that the intercept is an out-of-sample estimate, therefore be careful with its interpretation.

wages2$educ: estimate coefficient = 76.216
One more year of education will increase monthly earning on average by 76.22USD, ceteris paribus. The $p - value < 2e - 16 < 1\%$. This value is highly statistically significant. We can reject the null hypothesis at the 1% significance level.

wages2$exper: estimate coefficient = 17.638
One more year of work experience will increase monthly earning on average by 17.64USD, ceteris paribus. The $p - value = 3.18e - 08 < 1\%$. This value is highly statistically significant. We can reject the null hypothesis at the 1% significance level.

ii Explain why the omission of the work experience variable in the simple regression model led to an underestimation of the impact of education on wages.

Work experience and education both have a positive impact on wages. However, education and work experience are negatively correlated, i.e. individuals with more education have on average less work experience. By not controlling for work experience (not adding work experience to our model), the coefficient of education would also capture the work experience effect, therefore the impact of education would be underestimated.

iii Using the multiple regression model, predict the wage of someone who has 12 years of education and 1 year of work experience.
According to multiple regression model, we can write the fitted model as:

$$\widehat{wage}_i = -272.528 + 76.216 wages2\$educ_i + 17.638 wages2\$exper_i$$

Wage of someone who has 12 years of education and 1 year of work experience is:

$$\widehat{wage} = -272.528 + 76.216 \times 12 + 17.638 \times 1 = 659.702$$

3

The output below is the result of fitting an educational attainment function, regressing S on $ASVABC$, a measure of cognitive ability, SM, and SF, years of schooling (highest

grade completed) of the respondent's mother and father, respectively.

```
Call:
lm(formula = EAWE21$S ~ EAWE21$ASVABC + EAWE21$SM + EAWE21$SF)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9387 -1.6521  0.0186  1.5161  7.1553

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.59674    0.61428
EAWE21$ASVABC  1.24253    0.12359
EAWE21$SM      0.09135    0.04593
EAWE21$SF      0.20289    0.04251
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.254 on 496 degrees of freedom
Multiple R-squared:  0.329,     Adjusted R-squared:  0.3249
F-statistic: 81.06 on 3 and 496 DF,  p-value: < 2.2e-16
```

a. Give an interpretation of the regression coefficients.

Intercept: estimate coefficient = 10.59674
If all covariates were zero (zero ability score, no education of father and mother), the average years of education would be 10.6 years.

$ASVABC$: estimate coefficient = 1.24253
A one unit higher ability score will increase the number of years of education on average by 1.24 years, ceteris paribus.

SM: estimate coefficient = 0.09135
A year more of education of the mother will lead on average to a 0.09 years increase of years of education, ceteris paribus.

SF: estimate coefficient = 0.20289
A year more of education of the father will lead on average to a 0.20 years increase of education, ceteris paribus.

b. Undertake hypothesis tests to show whether the coefficients are statistically significant. The critical $t - value = 1.965$ (5% significance level).

To test whether any of the coefficients provided above are statistically significant, we need to follow these steps

(a) State the null and alternative hypotheses

(b) Select the significance level

(c) Select and calculate the test statistics

(d) Set the decision rule

(e) Make statistical decisions

Write hypothesis test for intercept: $H_0 : \beta_1 = 0; H_1 : \beta_1 \neq 0$
Significance level: 5%
Calculate t-statistic: $\hat{\beta}_1 = 10.59674$, s.e.$(\hat{\beta}_1) = 0.61428$

$$t = \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)} = \frac{10.59674 - 0}{0.61428} \approx 17.25$$

Compare t-statistic with critical t -value: $17.25 > 1.965$
Conclusion: We can reject the null hypothesis at the 5% significance level. The intercept is statistically significant.

Write hypothesis test for $ASVABC$: $H_0 : \beta_2 = 0; H_1 : \beta_2 \neq 0$
Significance level: 5%
Calculate t-statistic: $\hat{\beta}_2 = 1.24253$, s.e.$(\hat{\beta}_2) = 0.12359$

$$t = \frac{\hat{\beta}_2 - \beta_2^0}{s.e.(\hat{\beta}_2)} = \frac{1.24253 - 0}{0.12359} \approx 10.05$$

Compare t-statistic with critical t -value: $10.05 > 1.965$
Conclusion: We can reject the null hypothesis at the 5% significance level. The intercept is statistically significant.

Write hypothesis test for SM: $H_0 : \beta_3 = 0; H_1 : \beta_3 \neq 0$
Significance level: 5%
Calculate t-statistic: $\hat{\beta}_3 = 0.09135$, s.e.$(\hat{\beta}_3) = 0.04593$

$$t = \frac{\hat{\beta}_3 - \beta_3^0}{s.e.(\hat{\beta}_3)} = \frac{0.09135 - 0}{0.04593} \approx 1.98$$

Compare t-statistic with critical t -value: $1.98 > 1.965$
Conclusion: We can reject the null hypothesis at the 5% significance level. The intercept is statistically significant.

Write hypothesis test for SF: $H_0 : \beta_4 = 0; H_1 : \beta_4 \neq 0$
Significance level: 5%
Calculate t-statistic: $\hat{\beta}_4 = 0.20289$, s.e.$(\hat{\beta}_4) = 0.04251$

$$t = \frac{\hat{\beta}_4 - \beta_4^0}{s.e.(\hat{\beta}_4)} = \frac{0.20289 - 0}{0.04251} \approx 4.77$$

Compare t-statistic with critical t -value: $4.77 > 1.965$
Conclusion: We can reject the null hypothesis at the 5% significance level. The intercept is statistically significant.

c. Is the R^2 is statistically significant? The critical F value at the 5% significance level is 2.62. Interpret your result.

Write hypothesis test for R^2: $H_0 : R^2 = 0; H_1 : R^2 \neq 0$
Significance level: 5%
Calculate F-statistic:
You can use R^2 to calculate the F-statistic (or find it in the output table).
$R^2 = 0.329$

$$F(3, 496) = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} = \frac{0.329/(4-1)}{(1-0.329)/496} \approx 81.06$$

F-statistic = 81.06.
Compare test statistic with the critical F-value: $81.06 > 2.62$
The F statistic is greater than the critical F value, therefore we can reject the null hypothesis. The estimated model is statistically significant.

d. Calculate the 95% confidence interval for each coefficient.

The formula for calculating the confidence interval for the intercept is

$$\hat{\beta}_1 - s.e.(\hat{\beta}_1) \times t_{crit} \leq \beta_1 \leq \hat{\beta}_1 + s.e.(\hat{\beta}_1) \times t_{crit}$$

The regression output shows that the point estimate β_1 is 10.59674 and its standard error is 0.61428:

$$\hat{\beta}_1 - s.e.(\hat{\beta}_1) \times t_{crit} \leq \beta_1 \leq \hat{\beta}_1 + s.e.(\hat{\beta}_1) \times t_{crit}$$

$$10.59674 - 0.61428 \times 1.965 \leq \beta_1 \leq 10.59674 + 0.61428 \times 1.965$$

The 5% confidence level for intercept is

$$9.38 \leq \beta_1 \leq 11.80$$

Following the same logic, you can calculate the 5% confidence level for $ASVABC$, SM and SF, below are the results (slight differences are routing errors)

	2.5%	97.5%
$(Intercept)$	9.389834349	11.8036489
$ASVABC$	0.999708045	1.4853453
SM	0.001111899	0.1815941
SF	0.119365788	0.2864163

4

Explain the differences between R^2 and adjusted R^2 and calculate adjusted R^2 using the information from question 3. The formula of adjusted R^2 is

$$\bar{R^2} = R^2 - \frac{k-1}{n-k}(1-R^2)$$

R^2 is a measure of goodness of fit. However, the more X variables are on the right-hand side of a regression model the higher R^2 usually tends to be. Adding random variables could improve the model fit by chance, therefore a direct comparison of models with a different number of variables is misleading. Adjusted R^2 tries to mitigate the problem by introducing a negative adjustment factor for adding additional covariates. This allows us to compare "apples with apples".

$R^2 = 0.329$

$$\bar{R^2} = R^2 - \frac{k-1}{n-k}(1-R^2) = 0.329 - \frac{4-1}{496}(1-0.329) = 0.3249$$