ECON 4003 Econometrics I

Empirical Exercise 5

By Duong Trinh

Picture the Scenario

- **Objective:** Investigate the effect of smoking on baby's birth weight.
- **Dataset:** `birthweight_smoking.dta`
  - □ a random sample of 3,000 babies born in Pennsylvania in 1989.
- **Key variables:**
  - □ `birthweight`: birth weight of infant (in grams)
  - □ `smoker`: if the mother smoked during pregnancy or not.
  - □ `alcohol`: if the mother drank alcohol during pregnancy or not.
  - □ `nprevist`: total number of prenatal visits.
  - □ other various characteristics of the mother.

Question 1

Estimate the following two regression models:

- $birthweight_i = \beta_0 + \beta_1 smoker_i + u_i \quad (1)$
- $birthweight_i = \gamma_0 + \gamma_1 smoker_i + \gamma_2 alcohol_i + \gamma_3 nprevist_i + e_i \quad (2)$

Question (1a) Interpretation

	Dependent Variable: Birth Weight	
	(1)	(2)
Smoker	-253.23***	-217.58***
	(26.81)	(26.11)
Alcohol		-30.49
		(72.60)
Nprevist		34.07***
		(2.855)
Constant	3,432.06***	3,051.25***
	(11.89)	(43.71)
n	3,000	3,000
R^2	0.029	0.073
$\overline{R}^2$	0.028	0.072

Note: Standard errors are in parenthesis.
*p$<$0.1; **p$<$0.05; ***p$<$0.01

Question (1a) Interpretation

Model 1:

$$\underset{(se)}{\widehat{birthweight}} = \underset{(11.89)}{3432.06} - \underset{(26.81)}{253.23} \cdot smoker$$

- The average birth weight of babies born to non-smokers (smoker $=0$) is 3,432.06 grams.
- Babies born to smokers had birth weights that on average were 253.23 grams lower than babies born to non-smokers.

Question (1a) Interpretation (cont.)

Model 2:

$$\underset{(se)}{\widehat{birthweight}} = \underset{(43.71)}{3051.25} - \underset{(26.11)}{217.58} \cdot smoker - \underset{(72.60)}{30.49} \cdot alcohol + \underset{(2.855)}{34.07} \cdot nprevist$$

- The expected birth weight of babies born to women who didn't smoke during her pregnancy, did not drink alcohol, and didn't visit prenatal care is 3051.25 grams.
- Babies born to smokers had birth weights that were 217.58 grams lower than babies born to non-smokers on average, holding other factors constant.
- Babies born to women who drank alcohol had birth weights that were 30.49 grams lower than babies born to women who drank no alcohol on average, holding other factors constant.
- An extra prenatal care visit is associated with an increase in birth weight by 34.07 grams on average, holding other factors constant.

Question (1b) Omitted Variable Bias

Explain why the exclusion of `alcohol` and `nprevist` could lead to omitted variable bias in regression model (1).

OVB

Question (1b) Omitted Variable Bias

Explain why the exclusion of `alcohol` and `nprevist` could lead to omitted variable bias in regression model (1).

OVB

Check 2 conditions:

- Both *alcohol consumption* and *the number of prenatal doctor visits* may have direct effects on *birth weight*.
- *Smoking* may be correlated with both *alcohol consumption* and *the number of prenatal doctor visits*.

Question (1b) Omitted Variable Bias (cont.)

- Both *alcohol consumption* and *the number of prenatal doctor visits* may have direct effects on *birth weight*.

  Estimation results of Regression model (2):

  - □ $\hat{\gamma}_2 = -30.49 \Rightarrow \gamma_2$ is likely to be negative.
  - □ $\hat{\gamma}_3 = 34.07 \Rightarrow \gamma_3$ is likely to be positive.

- *Smoking* may be correlated with both *alcohol consumption* and *the number of prenatal doctor visits*.

  Regressing `smoker` on `alcohol` and `nprevist`:

  - □ `smoker` is positively correlated with `alcohol`
  - □ `smoker` is negatively correlated with `nprevist`

  The estimated coefficients are also statistically significant.

Question (1c) Interpretation - Omitted Variable Bias

Is the estimated coefficient of smoking on birth weight in the model (2) different from model (1)? Does regression model (1) seem to suffer from omitted variable bias?

Question (1c) Interpretation - Omitted Variable Bias

Is the estimated coefficient of smoking on birth weight in the model (2) different from model (1)? Does regression model (1) seem to suffer from omitted variable bias?

- **Model 1**: $\hat{\beta}_1 = -253.23$
- **Model 2**: $\hat{\gamma}_1 = -217.58$

$$\Longrightarrow \hat{\beta}_1 < \hat{\gamma}_1$$

The simple regression seems suffer from omitted variable bias (biased downward).

Question (1d) Prediction

A mother smoked during her pregnancy, did not drink alcohol, and had 8 prenatal care visits. Use the regression to predict the birth weight of the mother's child.

Question (1d) Prediction

A mother smoked during her pregnancy, did not drink alcohol, and had 8 prenatal care visits. Use the regression to predict the birth weight of the mother's child.

$$\underset{(se)}{\widehat{birthweight}} = \underset{(43.71)}{3051.25} - \underset{(26.11)}{217.58} \cdot smoker - \underset{(72.60)}{30.49} \cdot alcohol + \underset{(2.855)}{34.07} \cdot nprevist$$

$$\underset{(se)}{\widehat{birthweight}} = \underset{(43.71)}{3051.25} - \underset{(26.11)}{217.58} \times 1 - \underset{(72.60)}{30.49} \times 0 + \underset{(2.855)}{34.07} \times 8 = 3106.23$$

Question (1e) R^2 and adjusted R^2

Why are the R^2 and $\bar{R}^2$ in model (2) so similar? R2

	Dependent Variable: Birth Weight	
	(1)	(2)
Smoker	-253.23***	-217.58***
	(26.81)	(26.11)
Alcohol		-30.49
		(72.60)
Nprevist		34.07***
		(2.855)
Constant	3,432.06***	3,051.25***
	(11.89)	(43.71)
n	3,000	3,000
R^2	0.029	0.073
$\overline{R}^2$	0.028	0.072

Question (1e) R^2 and adjusted R^2

Why are the R^2 and $\bar{R}^2$ in model (2) so similar?

- They are nearly identical because the sample size is very large ($n = 3000$) and the number of regressors is small ($k = 3$)

Question 2

An alternative way to control for prenatal visits is to use the binary variables `tripre0` through `tripre3`. Regress `birthweight` on `smoker`, `alcohol`, `tripre0` , `tripre2`, and `tripre3`.

□ `nprevist`: total number of prenatal visits.

□ `tripre0`: indicator = 1 if no prenatal visit.

□ `tripre1`: indicator = 1 if the first prenatal visit in 1^{st} trimester.

□ `tripre2`: indicator = 1 if the first prenatal visit in 2^{nd} trimester.

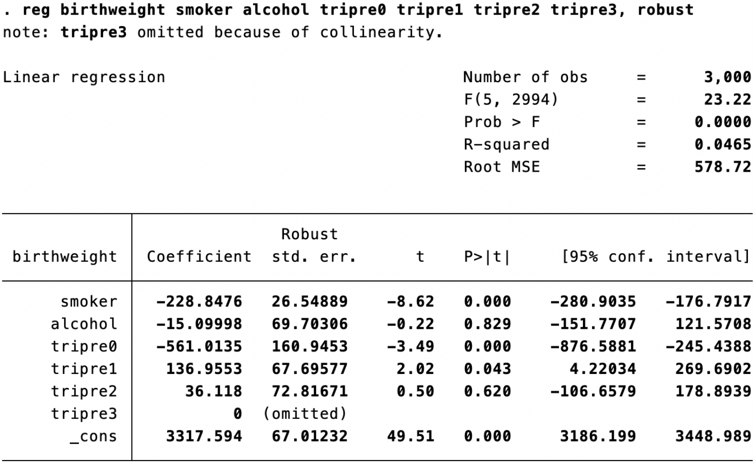
□ `tripre3`: indicator = 1 if the first prenatal visit in 3^{rd} trimester.

Question (2a) Perfect Multicollinearity

Why is `tripre1` excluded from the regression? What would happen if you included it in the regression?

Question (2a) Perfect Multicollinearity

Why is `tripre1` excluded from the regression? What would happen if you included it in the regression?



```
. reg birthweight smoker alcohol tripre0 tripre1 tripre2 tripre3, robust
note: tripre3 omitted because of collinearity.

Linear regression                               Number of obs     =      3,000
                                                F(5, 2994)        =      23.22
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0465
                                                Root MSE          =     578.72

------------------------------------------------------------------------------
             |               Robust
 birthweight | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      smoker |  -228.8476   26.54889    -8.62   0.000    -280.9035   -176.7917
     alcohol |  -15.09998   69.70306    -0.22   0.829    -151.7707    121.5708
     tripre0 |  -561.0135   160.9453    -3.49   0.000    -876.5881   -245.4388
     tripre1 |   136.9553   67.69577     2.02   0.043     4.22034    269.6902
     tripre2 |     36.118   72.81671     0.50   0.620    -106.6579    178.8939
     tripre3 |          0  (omitted)
       _cons |   3317.594   67.01232    49.51   0.000     3186.199    3448.989
------------------------------------------------------------------------------
```

Question (2a) Perfect Multicollinearity

Why is `tripre1` excluded from the regression? What would happen if you included it in the regression?

Let's ask a question...
When was the first prenatal visit of the woman *i*?

	`tripre1`	`tripre2`	`tripre3`	`tripre0`
In 1^{st} trimester	**1**	0	0	0
In 2^{nd} trimester	0	**1**	0	0
In 3^{rd} trimester	0	0	**1**	0
She had no visit	0	0	0	**1**

$\Longrightarrow$ Always: $\texttt{tripre0} + \texttt{tripre1} + \texttt{tripre2} + \texttt{tripre3} = 1$

Question (2a) Perfect Multicollinearity

Why is `tripre1` excluded from the regression? What would happen if you included it in the regression?

- `tripre1` is omitted to avoid perfect multicollinearity: This is because `tripre0` + `tripre1` + `tripre2` + `tripre3` = 1, which equals the value of the 'constant' regressor that determines the intercept
  → one regressor is an exact linear function of the other regressors.
  → Assumption (MR.3) is violated, the OLS estimator is not defined.
- Stata will drop one of the dummy variables if `tripre0`, `tripre1`, `tripre2`, and `tripre3`, and the constant term all included in the regression.
- If there are G dummy variables and each observation falls into one and only one category, you will include only G-1 of them as regressors to avoid **dummy variable trap** .

Question (2b,c) Interpretation

$$birthweight_i = \eta_0 + \eta_1 smoker_i + \eta_2 alcohol_i + \alpha_0 tripre0_i + \alpha_2 tripre2_i + \alpha_3 tripre3_i + e_i$$

$$\begin{aligned}
\mathbb{E}[birthweight|tripre1 = 1, smk, alc] &= \eta_0 + \eta_1 smk + \eta_2 alc \\
\mathbb{E}[birthweight|tripre0 = 1, smk, alc] &= \eta_0 + \eta_1 smk + \eta_2 alc + \alpha_0 \\
\mathbb{E}[birthweight|tripre2 = 1, smk, alc] &= \eta_0 + \eta_1 smk + \eta_2 alc + \alpha_2 \\
\mathbb{E}[birthweight|tripre3 = 1, smk, alc] &= \eta_0 + \eta_1 smk + \eta_2 alc + \alpha_3
\end{aligned}$$

- The coefficients on the **included dummy variables** represent the **incremental effect** of being in that category, relative to the base case of the **omitted category**, holding constant the other regressors.

Question (2b,c) Interpretation

	Dependent Variable: Birth Weight
	(1)
Smoker	-228.85 (26.55) ***
Alcohol	-15.10 (69.70)
Tripre0	-697.97 (146.58)***
Tripre2	-100.84 (31.55)***
Tripre3	-136.96 (67.70)**
Constant	3,454.55 (12.48)***
n	3,000
R^2	0.046
$\overline{R}^2$	0.045

Note: Standard errors are in parenthesis.
*p<0.1; **p<0.05; ***p<0.01

Question (2b,c) Interpretation

	Dependent Variable: Birth Weight
	(1)
Smoker	-228.85 (26.55) ***
Alcohol	-15.10 (69.70)
Tripre0	-697.97 (146.58)***
Tripre2	-100.84 (31.55)***
Tripre3	-136.96 (67.70)**
Constant	3,454.55 (12.48)***

Note: Standard errors are in parenthesis.
*p$<$0.1; **p$<$0.05; ***p$<$0.01

Babies born to women who had no prenatal doctor visits (tripre0 $= 1$) had birth weights that were on average 697.97 grams lower than babies from others who saw a doctor during the first trimester (tripre1 $= 1$).

Question (2b,c) Interpretation

	Dependent Variable: Birth Weight
	(1)
Smoker	-228.85 (26.55) ***
Alcohol	-15.10 (69.70)
Tripre0	-697.97 (146.58)***
Tripre2	-100.84 (31.55)***
Tripre3	-136.96 (67.70)**
Constant	3,454.55 (12.48)***

Note: Standard errors are in parenthesis.
*p$<$0.1; **p$<$0.05; ***p$<$0.01

Babies born to women whose first doctor visit was during the second trimester (tripre2 $= 1$) had birth weights that on average were 100.84 grams lower than babies from others who saw a doctor during the first trimester (tripre1 $= 1$).

Question (2b,c) Interpretation

	Dependent Variable: Birth Weight
	(1)
Smoker	-228.85 (26.55) ***
Alcohol	-15.10 (69.70)
Tripre0	-697.97 (146.58)***
Tripre2	-100.84 (31.55)***
Tripre3	-136.96 (67.70)**
Constant	3,454.55 (12.48)***

Note: Standard errors are in parenthesis.
$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Babies born to women whose first doctor visit was during the third trimester (tripre3 $= 1$) had birth weights that on average were 136.96 grams lower than babies from others who saw a doctor during the first trimester (tripre1 $= 1$).

Question (2d)

Does the regression model in Q2 explain a larger fraction of the variance in birth weight than the second regression model in Q1?

Question (2d)

Does the regression model in Q2 explain a larger fraction of the variance in birth weight than the second regression model in Q1?

- **Q2 - Model:**

  $birthweight_i =$

  $\eta_0 + \eta_1 smoker_i + \eta_2 alcohol_i + \alpha_0 tripre0_i + \alpha_2 tripre2_i + \alpha_3 tripre3_i + e_i$

  Estimation results: $R^2 = 0.046$, $\overline{R}^2 = 0.045$
- **Q1 - Model 2:**

  $birthweight_i = \gamma_0 + \gamma_1 smoker_i + \gamma_2 alcohol_i + \gamma_3 nprevist_i + e_i$

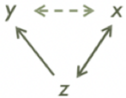
  Estimation results: $R^2 = 0.073$, $\overline{R}^2 = 0.072$

$\Longrightarrow$ Both R^2 and $\overline{R}^2$ are lower in Q2.

Appendix: Omitted Variable Bias

For omitted variable bias occur, the omitted variable Z must satisfy **both** conditions:

- Z is a determinant of Y; and
- Z is correlated with the regressor X



there is another factor z that causes y and is correlated with x, which makes x and y be associated

Appendix: Omitted Variable Bias (cont.)

$$plim\hat{\beta}_1 = \beta_1 + \beta_2 \frac{Cov(X,Z)}{Var(X)}$$

- $\hat{\beta}_1$ is biased upward $\Leftrightarrow plim\hat{\beta}_1 > \beta_1$
- $\hat{\beta}_1$ is biased downward $\Leftrightarrow plim\hat{\beta}_1 < \beta_1$

	$Cov(X,Z) > 0$	$Cov(X,Z) < 0$
$\beta_2 > 0$	$plim\hat{\beta}_1 > \beta_1$	$plim\hat{\beta}_1 < \beta_1$
$\beta_2 < 0$	$plim\hat{\beta}_1 < \beta_1$	$plim\hat{\beta}_1 > \beta_1$

Back

Appendix: Omitted Variable Bias (cont.)

$$plim\hat{\beta}_1 = \beta_1 + \beta_2 \frac{Cov(X,Z)}{Var(X)} + \beta_3 \frac{Cov(X,W)}{Var(X)}$$

- $\hat{\beta}_1$ is biased upward $\Leftrightarrow$ $plim\hat{\beta}_1 > \beta_1$
- $\hat{\beta}_1$ is biased downward $\Leftrightarrow$ $plim\hat{\beta}_1 < \beta_1$

Appendix: R^2 and adjusted R^2

- R^2 is the fraction of the sample variance of Y explained by X

$$R^2 = \frac{\sum_{i=1}^{n} \left(\widehat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2} = \frac{\text{Explained sum of squares (ESS)}}{\text{Total sum of squares (TSS)}}$$

- Adjusted R^2 (or $\bar{R}^2$) takes R^2 and penalise for additional regressors

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

  - □ $\frac{n-1}{n-k-1}$ is greater than 1 and grows with k
  - □ $\bar{R}^2 < R^2$, however two will be very close if n is large, k is small, or $R^2 = 0$ (which is very unlikely)

Back