Financial Econometrics
Shanghai Jiao Tong University
Summer School – 2025
Empirical Exercise 4A
With Solutions

Fatih Kansoy

Reference: Stock, J.H. and Watson, M.W., Introduction to Econometrics (4th Edition)

Data: `Birthweight_Smoking.csv`

Empirical Exercise 6.1. Use the `Birthweight_Smoking` data set introduced in Empirical Exercise E5.3 to answer the following questions.

(a) Regress *Birthweight* on *Smoker*. What is the estimated effect of smoking on birth weight?

(b) Regress *Birthweight* on *Smoker*, *Alcohol*, and *Nprevist*.

(i) Using the two conditions in Key Concept 6.1, explain why the exclusion of *Alcohol* and *Nprevist* could lead to omitted variable bias in the regression estimated in (a).

(ii) Is the estimated effect of smoking on birth weight substantially different from the regression that excludes *Alcohol* and *Nprevist*? Does the regression in (a) seem to suffer from omitted variable bias?

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

(iv) Compute R^2 and $\bar{R}^2$. Why are they so similar?

(v) How should you interpret the coefficient on *Nprevist*? Does the coefficient measure a causal effect of prenatal visits on birth weight? If not, what does it measure?

(c) Estimate the coefficient on *Smoking* for the multiple regression model in (b), using the three-step process in Appendix 6.3 (the Frisch–Waugh theorem). Verify that the three-step process yields the same estimated coefficient for *Smoking* as that obtained in (b).

(d) 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*.

(i) Why is *Tripre1* excluded from the regression? What would happen if you included it in the regression?

(ii) The estimated coefficient on *Tripre0* is large and negative. What does this coefficient measure? Interpret its value.

(iii) Interpret the value of the estimated coefficients on *Tripre2* and *Tripre3*.

(iv) Does the regression in (d) explain a larger fraction of the variance in birth weight than the regression in (b)?

Solution:

Key Concepts for Multiple Regression

Key Concept 6.1 – Omitted Variable Bias (OVB):

Omitted variable bias occurs when two conditions are met:

1. The omitted variable is correlated with an included regressor
2. The omitted variable is a determinant of the dependent variable

If both conditions hold, the OLS estimator of the included regressor's coefficient is biased.

Key Concept 6.3 – Dummy Variable Trap:

When using dummy variables for a categorical variable with J categories, include only $J - 1$ dummies. The omitted category becomes the "reference group," and coefficients measure effects relative to this group.

Data Overview

Variable	Description	Mean	SD	Min	Max
birthweight	Birth weight (grams)	3,382.9	592.2	425	5,755
smoker	Mother smoked (1=Yes)	0.194	0.395	0	1
alcohol	Mother drank (1=Yes)	0.019	0.138	0	1
nprevist	Prenatal care visits	10.99	3.67	0	35

Source: 1989 National Natality-Mortality Detail files (Pennsylvania). $n = 3{,}000$.

About 19.4% of mothers smoked during pregnancy, and 1.9% consumed alcohol.

(a) Simple Regression: Birthweight on Smoker

Model: $\text{Birthweight}_i = \beta_0 + \beta_1 \text{Smoker}_i + u_i$

Regression Output

	Coefficient	**Std. Error**	**t-statistic**	**p-value**
Smoker	−253.23	26.95	−9.40	0.000
Constant	3,432.06	11.87	289.12	0.000
$n = 3{,}000$,	$R^2 = 0.0286$,	Root MSE = 583.73		

Estimated Equation:

$$\widehat{\text{Birthweight}} = 3{,}432.1 - 253.2 \times \text{Smoker}$$

Answer: The estimated effect of smoking is $\hat{\beta}_1 = -253.2$ grams.
Interpretation: Babies born to mothers who smoked during pregnancy weigh, on average, **253 grams less** than babies born to non-smoking mothers. This effect is statistically significant ($t = -9.40$, $p < 0.001$).

(b) Multiple Regression: Adding Alcohol and Nprevist

Model: $\text{Birthweight}_i = \beta_0 + \beta_1 \text{Smoker}_i + \beta_2 \text{Alcohol}_i + \beta_3 \text{Nprevist}_i + u_i$

Regression Output

	Coefficient	**Std. Error**	**t-statistic**	**p-value**
Smoker	−217.58	26.68	−8.16	0.000
Alcohol	−30.49	76.23	−0.40	0.689
Nprevist	34.07	2.85	11.93	0.000
Constant	3,051.25	34.02	89.70	0.000
$n = 3{,}000$,	$R^2 = 0.0729$,	$\bar{R}^2 = 0.0719$,	Root MSE = 570.47	

(b)(i) OVB Analysis: Why Could Omitting Alcohol and Nprevist Cause Bias?

For OVB to occur, the omitted variable must satisfy **both** conditions:

Condition 1: Correlation with Smoker

	Smoker	Alcohol	Nprevist
Smoker	1.000		
Alcohol	**0.121**	1.000	
Nprevist	−0.109	−0.043	1.000

- $\text{Corr}(\text{Smoker}, \text{Alcohol}) = 0.121 > 0$: Smokers are more likely to drink alcohol
- $\text{Corr}(\text{Smoker}, \text{Nprevist}) = -0.109 < 0$: Smokers have fewer prenatal visits

Condition 2: Determinant of Birthweight
From the multiple regression:

- Alcohol: $\hat{\beta}_2 = -30.49$ (negative but not significant)
- Nprevist: $\hat{\beta}_3 = 34.07$ (positive and highly significant)

OVB Assessment:
Nprevist: Both conditions are satisfied. Nprevist is negatively correlated with Smoker and positively affects Birthweight. The expected bias direction:

$$\text{Bias} = \beta_{\text{Nprevist}} \times \frac{\text{Cov}(\text{Smoker}, \text{Nprevist})}{\text{Var}(\text{Smoker})} < 0$$

This suggests the simple regression *overstates* the negative effect of smoking.
Alcohol: Condition 1 is satisfied (positive correlation), but Condition 2 is weak (coefficient not significant). Alcohol likely contributes little to OVB.

(b)(ii) Comparison: Is There Evidence of OVB?

	Simple Regression (a)	**Multiple Regression (b)**
$\hat{\beta}_{\text{Smoker}}$	−253.23	−217.58
Std. Error	26.95	26.68
Difference	35.65 grams (14% change)	

Answer: The coefficient changes from −253 to −218 grams—a reduction of about **14%** in magnitude.
This is a modest but noticeable change. The simple regression appears to **slightly overstate** the negative effect of smoking because it does not control for the fact that smokers also have fewer prenatal visits (which independently reduces birthweight).
However, the effect of smoking remains **large and significant** in both specifications.

(b)(iii) Prediction for Jane

Jane's characteristics: Smoker = 1, Alcohol = 0, Nprevist = 8

$$\widehat{\text{Birthweight}} = 3{,}051.25 + (-217.58)(1) + (-30.49)(0) + (34.07)(8)$$

$$= 3{,}051.25 - 217.58 + 0 + 272.56 = 3{,}106.23$$

Jane's predicted birthweight: 3,106 grams

(b)(iv) R^2 vs. $\bar{R}^2$

$$R^2 = 0.0729 \qquad \bar{R}^2 = 0.0719$$

Answer: R^2 and $\bar{R}^2$ are nearly identical because the sample size is large ($n = 3{,}000$).
The adjustment formula is:

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

With $n = 3{,}000$ and $k = 3$, the adjustment factor $\frac{2999}{2996} \approx 1.001$ is very close to 1. The penalty for adding regressors is negligible when n is much larger than k.

(b)(v) Interpreting the Nprevist Coefficient

Answer: The coefficient $\hat{\beta}_{\text{Nprevist}} = 34.07$ means that each additional prenatal visit is associated with 34 grams higher birthweight, holding smoking and alcohol constant.
This is NOT a causal effect. The coefficient likely reflects **selection bias**:

- Mothers who expect healthier pregnancies may seek more prenatal care
- Mothers with pregnancy complications get more visits but may still have low-birthweight babies
- Unobserved factors (health consciousness, socioeconomic status) affect both visits and birthweight

The coefficient measures an **association**, not the causal impact of additional prenatal visits.

(c) Frisch-Waugh Theorem Verification

Frisch-Waugh-Lovell Theorem (Appendix 6.3):

The coefficient on X_1 in a multiple regression of Y on $X_1, X_2, \ldots, X_k$ equals the coefficient from regressing:

$$\text{residuals}(Y|X_2, \ldots, X_k) \quad \text{on} \quad \text{residuals}(X_1|X_2, \ldots, X_k)$$

Intuition: The residuals "partial out" the effects of other regressors, isolating the unique variation in X_1 that is uncorrelated with $X_2, \ldots, X_k$.

Three-Step Process

Step 1: Regress Smoker on Alcohol and Nprevist, obtain residuals $\tilde{S}_i$

$$\text{Smoker}_i = \gamma_0 + \gamma_1 \text{Alcohol}_i + \gamma_2 \text{Nprevist}_i + \tilde{S}_i$$

Step 2: Regress Birthweight on Alcohol and Nprevist, obtain residuals $\tilde{B}_i$

$$\text{Birthweight}_i = \delta_0 + \delta_1 \text{Alcohol}_i + \delta_2 \text{Nprevist}_i + \tilde{B}_i$$

Step 3: Regress $\tilde{B}_i$ on $\tilde{S}_i$ (no constant)

$$\tilde{B}_i = \beta_1^{FW} \tilde{S}_i + \text{error}$$

Result

	Coefficient	Std. Error	t-statistic	p-value
Smoker_resid	−217.58	26.67	−8.16	0.000

Verification:

- Frisch-Waugh coefficient: $\hat{\beta}_1^{FW} = -217.58$
- Multiple regression coefficient: $\hat{\beta}_{\text{Smoker}} = -217.58$

The coefficients are **identical**, confirming the Frisch-Waugh theorem. This demonstrates that multiple regression isolates the effect of each regressor by "controlling for" the other variables—mathematically equivalent to partialing out their influence.

(d) Regression with Trimester Dummies

Model: Replace continuous Nprevist with categorical timing of first prenatal visit.

$$\text{Birthweight}_i = \beta_0 + \beta_1 \text{Smoker}_i + \beta_2 \text{Alcohol}_i + \beta_3 \text{Tripre0}_i + \beta_4 \text{Tripre2}_i + \beta_5 \text{Tripre3}_i + u_i$$

Regression Output

	Coefficient	**Std. Error**	**t-statistic**	**p-value**
Smoker	−228.85	27.16	−8.42	0.000
Alcohol	−15.10	77.54	−0.19	0.846
Tripre0	−697.97	106.88	−6.53	0.000
Tripre2	−100.84	29.62	−3.40	0.001
Tripre3	−136.96	59.58	−2.30	0.022
Constant	3,454.55	12.65	273.08	0.000
$n = 3{,}000$,	$R^2 = 0.0465$,	$\bar{R}^2 = 0.0449$,	Root MSE = 578.72	

(d)(i) Why is Tripre1 Excluded?

The dummy variables satisfy:

$$\text{Tripre0}_i + \text{Tripre1}_i + \text{Tripre2}_i + \text{Tripre3}_i = 1 \quad \text{for all } i$$

Answer: Tripre1 is excluded to avoid the **dummy variable trap** (perfect multicollinearity).
If all four dummies were included with a constant, we would have:

$$\text{Constant} = \text{Tripre0} + \text{Tripre1} + \text{Tripre2} + \text{Tripre3}$$

This makes $(X'X)$ singular and $\hat{\beta}$ undefined. By omitting Tripre1, it becomes the **reference category**, and all coefficients are interpreted relative to mothers who began prenatal care in the first trimester.

(d)(ii) Interpreting Tripre0

Answer: $\hat{\beta}_{\text{Tripre0}} = -697.97$ grams.
Interpretation: Babies born to mothers who received **no prenatal care** (Tripre0 = 1) weigh, on average, **698 grams less** than babies born to mothers who began prenatal care in the **first trimester** (the reference group), holding smoking and alcohol constant.
This is a very large effect—nearly 700 grams (about 1.5 pounds) lower birthweight.

(d)(iii) Interpreting Tripre2 and Tripre3

Tripre2: $\hat{\beta}_{\text{Tripre2}} = -100.84$ grams
Mothers who began prenatal care in the **second trimester** have babies weighing 101 grams less than those who began in the first trimester.
Tripre3: $\hat{\beta}_{\text{Tripre3}} = -136.96$ grams
Mothers who began prenatal care in the **third trimester** have babies weighing 137 grams less than those who began in the first trimester.
Pattern: Later initiation of prenatal care is associated with lower birthweight, with no care (Tripre0) showing the largest negative effect.

(d)(iv) Comparing R^2

	Model (b) (Nprevist continuous)	**Model (d)** (Tripre dummies)
R^2	0.0729	0.0465
$\bar{R}^2$	0.0719	0.0449

Answer: No, model (d) explains **less** variance than model (b).
R^2 drops from 7.3% to 4.7%. The continuous variable Nprevist captures more variation in birthweight than the categorical timing variables. This suggests that **both** the timing and quantity of prenatal visits matter for birthweight.

Summary: Key Results

Model	$\hat{\beta}_{\text{Smoker}}$	**SE**	R^2
(a) Simple: Smoker only	−253.23	26.95	0.0286
(b) Multiple: + Alcohol, Nprevist	−217.58	26.68	0.0729
(d) Multiple: + Alcohol, Tripre dummies	−228.85	27.16	0.0465

The Big Picture

1. **Smoking Effect:** Across all specifications, smoking is associated with **200–250 grams lower birthweight**. The effect is robust to different control variables.

2. **Omitted Variable Bias:** Controlling for prenatal visits reduces the smoking coefficient by about 14%, suggesting mild OVB in the simple regression. Smokers have fewer prenatal visits, which independently lowers birthweight.

3. **Frisch-Waugh Theorem:** The three-step residualization process yields identical coefficients to multiple regression, demonstrating that "controlling for" variables is mathematically equivalent to partialing out their influence.

4. **Dummy Variable Trap:** When using categorical variables, omit one category to avoid perfect multicollinearity. Coefficients measure effects relative to the omitted reference group.

5. **Correlation vs. Causation:** The Nprevist coefficient likely reflects selection bias, not the causal effect of prenatal care.

Calculations performed using Stata. See accompanying `EE_4A_solution.do` *file.*