Answers:

4.2. The sample size $n = 200$. The estimated regression equation is

$$\widehat{Weight} = (2.15) - 99.41 + (0.31)\,3.94\,Height. \qquad R^2 = 0.81, \quad SER = 10.2.$$

(a) Substituting $Height = 70,\ 65,\ \text{and}\ 74$ inches into the equation, the predicted weights are 176.39, 156.69, and 192.15 pounds.

(b) $\widehat{\Delta Weight} = 3.94 \times \Delta Height = 3.94 \times 1.5 = 5.91.$

(c) We have the following relations: $1in = 2.54\,cm$ and $1lb = 0.4536\,kg$. Suppose the regression equation in the centimeter-kilogram space is

$$\widehat{Weight} = \hat{\gamma}_0 + \hat{\gamma}_1\, Height$$

The coefficients are $\hat{\gamma}_0 = -99.41 \times 0.4536 = -45.092\ kg$; $\hat{\gamma}_1 = 3.94 \times \frac{0.4536}{2.54} = 0.7036\,kg$ per *cm*. The R^2 is unit free, so it remains at $R^2 = 0.81$. The standard error of the regression is $SER = 10.2 \times 0.4536 = 4.6267\,kg$.

5.4. (a) $-12.12 + 2.37 \times 16 = \$25.80$ per hour

(b) The wage is expected to increase by $2.37 \times 2 = \$4.74$ per hour.

(c) The increase in wages for college education is $\beta_1 \times 4$. Thus, the counselor's assertion is that $\beta_1 = 10/4 = 2.50$. The *t*-statistic for this null hypothesis is $t = \frac{2.37 - 2.50}{0.10} = -1.3$, which has a *p*-value of 0.19 Thus, the counselor's assertion cannot be rejected at the 10% significance level. A 95% confidence for $\beta_1 \times 4$ is $4 \times (2.37 \pm 1.96 \times 0.10)$ or $\$8.70 \le \text{Gain} \le \$10.36$.

6.2. (a) Workers with college degrees earn \$10.47/hour more, on average, than workers with only high school degrees.

(b) Men earn \$4.69/hour more, on average, than women.

6.3. (a) On average, a worker earns \$0.61/hour more for each year he ages.

(b) Sally's earnings prediction is $0.11 + 10.44 \times 1 - 4.56 \times 1 + 0.61 \times 29 = 23.68$ dollars per hour. Betsy's earnings prediction is $0.11 + 10.44 \times 1 - 4.56 \times 1 + 0.61 \times 34 = 26.73$ dollars per hour. The difference is 3.05 \$/hour (= $0.61 \times (34-29)$).

6.4. (a) Workers in the Northeast earn \$0.74 more per hour than workers in the West, on average, controlling for other variables in the regression. Workers in the Midwest earn \$1.54 less per hour than workers in the West, on average, controlling for other variables in the regression. Workers in the South earn \$0.44 less than workers in the West, controlling for other variables in the regression.

(b) The regressor *West* is omitted to avoid perfect multicollinearity. If *West* is included, then the intercept can be written as a perfect linear function of the four regional regressors.

(c) The expected difference in earnings between Juanita and Jennifer is
$-0.44 - (-1.54) = \$0.90$/hour.

7.2
(a) Yes, the college-high school earnings difference is statistically significant at the 5% level. The *t*-statistic is 10.47/0.29 = 36.1 which is larger in absolute value than 1.96, the 5% critical value. The 95% confidence interval is $10.47 \pm 1.96 \times 0.29 = [9.90, 11.04]$
(b) Yes, the female-male earnings difference is statistically significant at the 5% level. The *t*-statistic is -4.69/0.29 = -16.2 which is larger in absolute value than 1.96, the 5% critical value. The 95% confidence interval is $-4.69 \pm 1.96 \times 0.29 = [-5.26, -4.12]$

7.3.
(a) Yes, age is an important determinant of earnings. The *t*-statistic is 0.61/0.04 = 15.3, with a *p*-value less than .01; this implies that the coefficient on age is statistically significant at the 1% level. The 95% confidence interval is $0.61 \pm (1.96 \times 0.04) = [0.53, 0.69]$.
(b) $\Delta Age \times [\$0.53, \$0.69] = 5 \times [\$0.53, \$0.69] = [\$2.65, \$3.45]$.

7.4.
(a) The *F*-statistic testing the coefficients on the regional regressors are zero is 9.32. The 1% critical value (from the $F_{3,\infty}$ distribution) is 3.78. Because 9.32 > 3.78, the regional effects are significant at the 1% level.
(b) The expected difference between Juanita and Molly is $(X_{6,\text{Juanita}} - X_{6,\text{Molly}}) \times \beta_6 = \beta_6$. Thus a 95% confidence interval is $-0.44 \pm (1.96 \times 0.37) = [-\$1.17, \$0.29]$.
(c) The expected difference between Juanita and Jennifer is

$$(X_{5,\text{Juanita}} - X_{5,\text{Jennifer}}) \times \beta_5 + (X_{6,\text{Juanita}} - X_{6,\text{Jennifer}}) \times \beta_6 = -\beta_5 + \beta_6.$$

A 95% confidence interval could be constructed using the general methods discussed in Section 7.3. In this case, an easy way to do this is to omit *Midwest* from the regression and replace it with $X_5 = West$. In this new regression the coefficient on *South* measures the difference in wages between the *South* and the *Midwest*, and a 95% confidence interval can be computed directly.

9.1. As explained in the text, potential threats to external validity arise from differences between the population and setting studied and the population and setting of interest. The statistical results based on New York in the 1970's are likely to apply to Boston in the 1970's but not to Los Angeles in the 1970's. In 1970, New York and Boston had large and widely used public transportation systems. Attitudes about smoking were roughly the same in New York and Boston in the 1970s. In contrast, Los Angeles had a considerably smaller public transportation system in 1970. Most residents of Los Angeles relied on their cars to commute to work, school, and so forth. The results from New York in the 1970's are unlikely to apply to New York in 2018. Attitudes towards smoking changed significantly from 1970 to 2018.

9.3. The key is that the selected sample contains only employed women. Consider two women, Beth and Julie. Beth has no children; Julie has one child. Beth and Julie are otherwise identical. Both can earn $25,000 per year in the labor market. Each must compare the $25,000 benefit to the costs of working. For Beth, the cost of working is forgone leisure. For Julie, it is forgone leisure and the costs (pecuniary and other) of child care. If Beth is just on the margin between working in the labor market or not, then Julie, who has a higher opportunity cost, will decide not to work in the labor market. Instead, Julie will work in "home production," caring for children, and so forth. Thus, on average, women with children who decide to work are women who earn higher wages in the labor market.

Empirical Exercise 4.2

Calculations for this exercise are carried out in the STATA file **EE_4_2.do.**

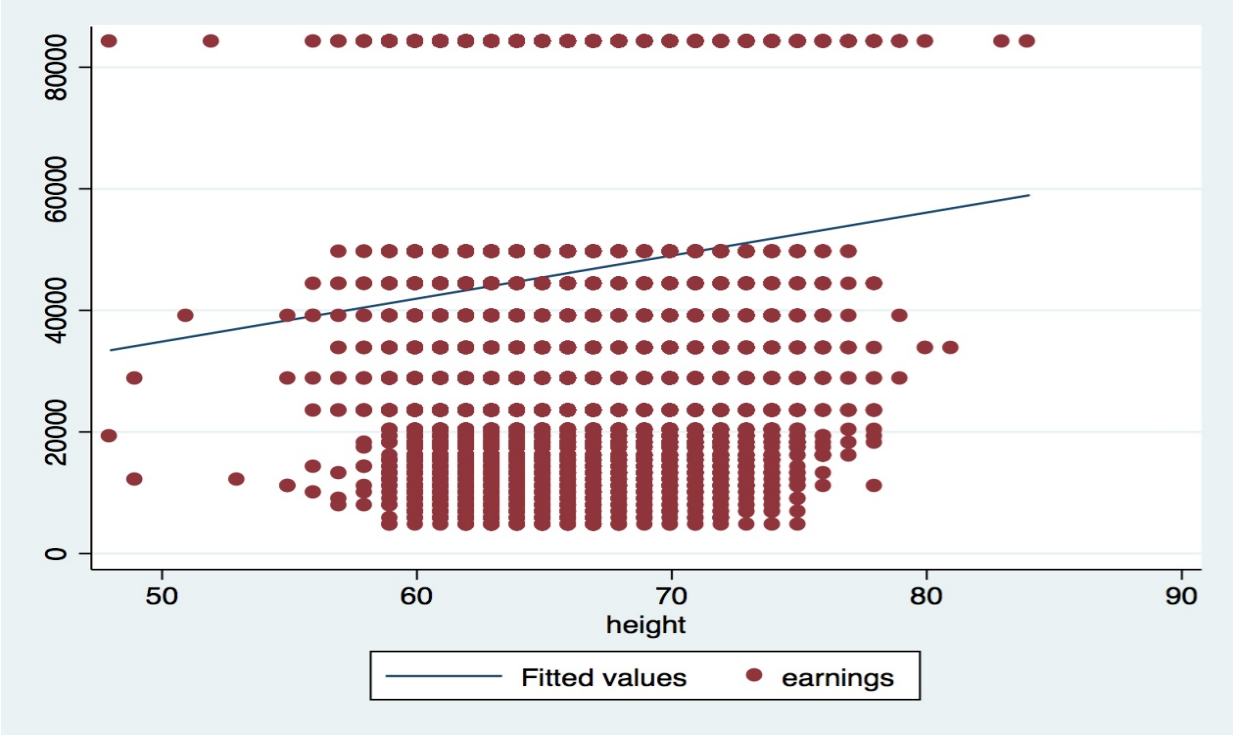
(a) The median height in the sample is 67 inches.

(b) The following table shows average annual earnings for taller and shorter workers

Height	Annual Earnings ($)		
	$\bar{x}$	SE($\bar{x}$)	95% Confidence Interval for mean
Height ≤ 67 inches	44,488	265.5	43,968 – 45,009
Height > 67 inches	49,988	305.4	49,389 – 50,587
Difference (Tall – Short)	5,499	404.7	4,706 – 6,293

The estimated average annual earnings for shorter workers is $44,488, is $49,988 for taller workers, for a difference of $5,499. The 95% confidence interval is $4,706 to $6,293. The difference is large (more than 10% of average earnings), precisely estimated (a standard error of $404) and statistically significantly different from zero.

(c) Scatterplot



The data documentation reports that individual earnings were reported in 23 brackets, and a single average value is reported for earnings in the same bracket. Thus, the dataset contains 23 distinct values of earnings.

(d) The estimated regression is

$$\widehat{Earnings} = -512.7 + 707.7 \times Height,\ R^2 = 0.011$$

The estimated slope is 707.7 (Dollars per year). The estimated earnings are

Height (in inches)	$\widehat{Earnings}$ **(in Dollars per year)**
65	45,486
67	46,901
70	49,204

(e) Recall that 1 cm = 0.394 inches. The estimated regression in (d), with units shown, is

$\widehat{Earnings}(\$) = -512.7(\$) + 707.7(\$/\text{inch}) \times Height(\text{inches})$,

R^2 (unit free) = 0.011, and SER = 26777(\$).

Note that

$$\begin{aligned} 707.7(\$/\text{inch}) \times Height(\text{inches}) &= 707.7(\$/\text{inch}) \times (0.394\text{inch/cm}) \times Height(\text{cm}) \\ &= 278.8(\$/\text{cm}) \times Height(\text{cm}) \end{aligned}$$

So the regression is

$\widehat{Earnings}(\$) = -512.7(\$) + 278.8(\$/\text{cm}) \times Height(\text{cm})$,

R^2 (unit free) = 0.011, and SER = 26777(\$)

(f) The regression for females is

$$\widehat{Earnings} = 12650 + 511.2 \times Height,\ R^2 = 0.002$$

A women who is one inch taller than average is predicted to have earnings that are \$511.2 per year higher than average.

(g) The regression for males is

$$\widehat{Earnings} = -43130 + 1306.9 \times Height,\ R^2 = 0.021$$

A man who is one inch taller than average is predicted to have earnings that are $1306.9 per year higher than average.

(h) Height may be correlated with other factors that cause earnings. For example, height may be correlated with "strength," and in some occupations, stronger workers may by more productive. There are many other potential factors that may be correlated with height and cause earnings and you will investigate of these in future exercises.

Empirical Exercise 5.3

Calculations for this exercise are carried out in the STATA file **EE_5_3.do.**

(a) Average birthweights, along with standard errors are shown in the table below. (Birthweight is measured in grams.)

	All Mothers	Non-smokers	Smokers
$\bar{X}$	3383	3432.1	3178.8
SE($\bar{X}$)	10.8	11.9	24.0
n	3000	2418	582

(b) The estimated difference is $\bar{X}_{\text{Smokers}} - \bar{X}_{\text{NonSmokers}} = -253.2$. The standard error of the difference is $SE\left(\bar{X}_{\text{Smokers}} - \bar{X}_{\text{NonSmokers}}\right) = \sqrt{SE(\bar{X}_{\text{Smokers}})^2 + SE(\bar{X}_{\text{NonSmokers}})^2} = 26.8$.

The 95% confidence for the difference is $-253.2 \pm 1.96 \times 26.8 = (-305.9, -200.6)$.

(c) The estimated regression is

$$\widehat{Birthweight} = \underset{(11.9)}{3432.1} \; \underset{(26.8)}{-253.2} Smoker$$

(i) The intercept is the average birthweight for non-smokers (*Smoker* = 0). The slope is the difference between average birthweights for smokers (*Smoker* = 1) and non-smokers (*Smoker* = 0).

(ii) They are the same.

(iii) This the same as the confidence interval in (b).

(d) Yes – and we'll investigate this more in future empirical exercises.

Empirical Exercise 6.1

Calculations for this exercise are carried out in the STATA file **EE_6_1.do.**

(a) The estimated regression is

$$\widehat{Birthweight} = 3432.1 - 253.2Smoker$$

The estimated effect of smoking on birthweight is –253.2 grams.

(b) The estimated regression is

$$\widehat{Birthweight} = 3051.2 - 217.6Smoker - 30.5Alcohol + 34.1Nprevist$$

(i) Smoking may be correlated with both alcohol and the number of pre-natal doctor visits, thus satisfying (1) in Key Concept 6.1. Moreover, both alcohol consumption and the number of doctor visits may have their own independent affects on birthweight, thus satisfying (2) in Key Concept 6.1.

(ii) The estimated is somewhat smaller: it has fallen to 217 grams from 253 grams, so the regression in (a) may suffer from omitted variable bias.

(iii) $\widehat{Birthweight} = 3051.2 - 217.6\times1 - 30.5\times0 + 34.1\times8 = 3106.4$

(iv) $R^2 = 0.0729$ and $\overline{R}^2 = 0.0719$. They are nearly identical because the sample size is very large ($n = 3000$).

(v) *Nprevist* is a control variable. It captures, for example, mother's access to healthcare and health. Because *Nprevist* is a control variable, its coefficient does not have a causal interpretation.

(c) The results from STATA are

```
. ** FW calculations;
. regress birthweight alcohol nprevist;

      Source |       SS       df       MS              Number of obs =    3000
-------------+------------------------------           F(  2,  2997) =   82.64
       Model |    54966381     2  27483190.5           Prob > F      =  0.0000
    Residual |   996653623  2997  332550.425           R-squared     =  0.0523
-------------+------------------------------           Adj R-squared =  0.0516
       Total |  1.0516e+09  2999  350656.887           Root MSE      =  576.67

------------------------------------------------------------------------------
 birthweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     alcohol |  -103.2781   76.53276    -1.35   0.177    -253.3402    46.78392
```

```
    nprevist |   36.49956   2.870272    12.72   0.000     30.87166    42.12746
       _cons |   2983.739   33.35198    89.46   0.000     2918.344    3049.134
------------------------------------------------------------------------------

. predict bw_res, r;

. regress smoker alcohol nprevist;

      Source |       SS       df       MS              Number of obs =    3000
-------------+------------------------------           F(  2,  2997) =   38.97
       Model |  11.8897961     2  5.94489803           Prob > F      =  0.0000
    Residual |  457.202204  2997  .152553288           R-squared     =  0.0253
-------------+------------------------------           Adj R-squared =  0.0247
       Total |     469.092  2999  .156416139           Root MSE      =  .39058

------------------------------------------------------------------------------
      smoker |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     alcohol |    .334529   .0518358     6.45   0.000     .2328917    .4361662
    nprevist |  -.0111667    .001944    -5.74   0.000    -.0149785   -.0073549
       _cons |   .3102729   .0225893    13.74   0.000     .2659807    .3545651
------------------------------------------------------------------------------

. predict smoker_res, r;

. regress bw_res smoker_res;

      Source |       SS       df       MS              Number of obs =    3000
-------------+------------------------------           F(  1,  2998) =   66.55
       Model |  21644450.1     1  21644450.1           Prob > F      =  0.0000
    Residual |   975009170  2998   325219.87           R-squared     =  0.0217
-------------+------------------------------           Adj R-squared =  0.0214
       Total |   996653621  2999   332328.65           Root MSE      =  570.28

------------------------------------------------------------------------------
      bw_res |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  smoker_res |  -217.5801    26.6707    -8.16   0.000    -269.8748   -165.2854
       _cons |  -2.75e-07   10.41185    -0.00   1.000    -20.41509    20.41509
------------------------------------------------------------------------------
```

(d) The estimated regression is

$$\widehat{Birthweight} = 3454.5 - 228.8Smoker - 15.1Alcohol$$
$$-698.0Tripre0 - 100.8\ Tripre2 - 137.0\ Tripre3$$

(i) *Tripre*1 is omitted to avoid perfect multicollinearity. (*Tripre*0+ *Tripre*1+ *Tripre*2+ *Tripre*3 = 1, the value of the "constant" regressor that determines the intercept). The regression would not run, or the software will report results from an arbitrary normalization if *Tripre*0, *Tripre*1, *Tripre*2, *Tripre*3, and the constant term all included in the regression.

(ii) Babies born to women who had no prenatal doctor visits (*Tripre*0 = 1) had birthweights that on average were 698.0 grams (≈ 1.5 lbs) lower than babies from others who saw a doctor during the first trimester (*Tripre*1 = 1).

(iii) Babies born to women whose first doctor visit was during the second trimester (*Tripre*2 = 1) had birthweights that on average were 100.8 grams (≈ 0.2 lbs) lower than babies from others who saw a doctor during the first trimester (*Tripre*1 = 1). Babies born to women whose first doctor visit was during the third trimester (*Tripre*3 = 1) had birthweights that on average were 137 grams (≈ 0.3 lbs) lower than babies from others who saw a doctor during the first trimester (*Tripre*1 = 1).

Empirical Exercise 7.1

Calculations for this exercise are carried out in the STATA file **EE_7_1.do.**

The following table summarizes some regressions

Dependent variable is *Birthweight*

Regressor	(1)	(2)	(3)	(4)
Smoker	-253.2 (26.8) [-305.8, -200.7]	-217.6 (26.1) [-268.8, -166.4]	-175.4 (26.8) [-228.0, -122.8]	-177.0 (27.3) [-230.5, -123.4]
Alcohol		-30.5 (72.6)	-21.1 (73.0)	-14.8 (72.9)
Nprevist		34.1 (3.6)	29.6 (3.6)	29.8 (3.6)
Unmarried			-187.1** (27.7)	-199.3 (31.0)
Age				-2.5 (2.4)
Years of education				-0.238 (5.53)
Intercept	3432.1 (11.9)	3051.2 (43.7)	3134.4 (44.1)	3199.4 (90.6)
SER	583.7	570.5	565.7	565.8
$\bar{R}^2$	0.028	0.072	0.087	0.087
n	3000	3000	3000	3000

Standard errors are shown in parentheses and 95% confidence interval for *Smoker* is shown in brackets

(a) See the table

(b) see table

(c) Yes it seems so. The coefficient changes falls by roughly 30% in magnitude when additional regressors are added to (1). This change is substantively large and large relative to the standard error in (1).

(d) Yes it seems so. The coefficient changes falls by roughly 20% in magnitude when *unmarried* is added as an additional regression. This change is substantively large and large relative to the standard error in (2).

(e) (i) -241.4 to -132.9

(ii) Yes. The 95% confidence interval does not include zero. Alternatively, the *t*-statistics Is -6.76 which is large in absolute value than the 5% crtical value of 1.96.

(iii) Yes. On average, birthweight is 187 grams lower for unmarried mothers.

(iv) As the question suggests, *unmarried* is a control variable that captures the effects of several factors that differ between married and unmarried mothers such as age, education, income, diet and other health factors, and so forth.

f. I have added on additional regression in the table that includes *Age* and *Educ* (years of education) in regression (4). The coefficient on *Smoker* is very similar to its value in regression (3).

Empirical Exercise 9.2

Calculations for this exercise are carried out in the STATA file **EE_9_2.do.**

The following regressions will be referenced in these answers.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Smoker	-175.4 (26.8)	-177.0 (27.3)	-178.2 (27.3)	-180.2 (27.3)	-177.8 (27.1)	275.6 (140.9)	-214.4 (29.8)
Alcohol	-21.1 (73.0)	-14.8 (72.9)	-10.9 (73.5)	-17.2 (73.7)	-14.5 (73.0)	2.6 (73.1)	-18.0 (72.8)
Nprevist	29.6 (3.6)	29.8 (3.6)	30.2 (3.6)	30.2 (3.6)	29.8 (3.6)	29.2 (3.6)	29.2 (3.6)
Unmarried	-187.1 (27.7)	-199.3 (31.0)	-204.8 (31.6)	-190.1 (31.6)	-196.1 (32.3)	-200.4 (30.6)	-176.1 (30.8)
Age		-2.5 (2.4)	-1.8 (2.5)		4.6 (18.1)	0.66 (2.42)	
Years of education		-0.238 (5.53)					
m_educ2 (Yrs ed = 12)			-52.0 (36.2)	-65.8 (37.2)			
m_educ3 (12 < Yrs ed < 16)			-34.9 (41.6)	-52.7 (42.6)			
m_educ4 (Yrs ed = 16)			-16.8 (44.2)	-38.4 (44.6)			
m_educ5 (Yrs ed > 16)			-70.0 (57.3)	-96.0 (55.7)			
Young = (Age ≤ 20)				-32.4 (38.1)			-73.8 (42.4)
Age^2					-0.13 (0.33)		
Smoker×Age						-17.7 (5.6)	
Smoker×Young							216.6 (64.6)
F-statistic and p-values on joint hypotheses							
mr_educ variables			0.89 (0.47)	1.14 (0.33)			
age and age²					0.64 (0.53)		
Smoker and interactions						23.87 (< 0.01)	25.9 (< 0.01)
SER	565.7	565.8	565.7	565.7	565.7	564.6	565.0
$\bar{R}^2$	0.087	0.087	0.087	0.087	0.087	0.91	0.90

Note: intercept included in all regressions.

(a) (i) The table shows various regressions. (1) and (2) were used in the answers to exercise 7.1. They suggested a 95% confidence interval of the effect of smoking on birthweight that ranged from (roughly) -230 to -120 grams. Regression (3) changes the education control and uses binary variables for a high school diploma (12 years of education), some college (12 < years of education < 16), a bachelor's degree (years of education = 16) and graduate work (years of education > 16), and where "years of education < 12" is the omitted category. Regression (4) additionally changes *Age* to the binary variable *Young* (*Age* ≤ 20). Regression (5) drops the education variables (which are not statistically significant in (3) and (4)) and adds Age^2 to check for nonlinear effect

of age on birthweight (which is insignificant). These modifications have little effect on the estimated effect of smoking on birthweight.

Regressions (6) and (7) investigate potential interaction effects of smoking and age. Both regressions suggest a significant interaction effect, with the effect of smoking on birthweight larger (that is, more negative) for older mothers. For example, from (6), the estimated effect of smoking on birthweight is 275.6-17.7×20 = -78.4 grams for a 20-year old mother, but is 275.6-17.7×30 = -255.5 grams for a 30-year old mother.

(ii) Omitted variables: There is the potential for omitted variable bias when a variable is excluded from the regression that (i) has an effect on birthweight and (ii) is correlated with smoking. There are several candidates. First, the dataset does not contain data on race and ethnicity and to the extent that these are related to birthweight and smoking, then they are potential omitted variables. There are other environmental variables such as mother's diet, exercise, and so forth that may affect birthweight and be correlated with smoking. These too are potential omitted variables. The size and significance of *unmarried* suggests that it is an important control variable, but it is undoubtedly an imperfect control.

Misspecification of the functional form: The regressions reported above suggest that an important nonlinearity arises from the interaction smoking and mother's age. Other nonlinearities do not seem to be important.

Errors-in-variables: All of the variables except birthweight are self-reported by the mother and may contain error. For example, some mothers may be reticent to respond that they smoked or drank during their pregnancy. How these kind of measurement errors affect the OLS estimates depends on the specifics of the measurement error, as discussed in section 9.2.

Sample selection: The data are a random sample of all babies born in Pennsylvania in 1989, and thus there is no sample-selection bias.

Simultaneous causality: This is a problem to the extent that women who are more likely to have low-birthweight children are more likely to stop smoking during pregnancy. This would induce a positive correlation between the regression error, u, and the binary variable *smoker*, which would result in an upward bias in the OLS coefficient.

Inconsistency of OLS standard errors: Heteroskedastic robust standard errors were used in the analysis, so that heteroskedasticity is not a concern. The data are collected using i.i.d. sampling from all babies born in Pennsylvania in 1989, so that correlation across the errors is unlikely to be a problem.

(b) To the extent that the OLS regression estimates the causal effect of smoking on birthweight, the results will be externally valid for these three populations. (Biology is the same in 1989 and 2015, and the same in Pennsylvania and Korea.) However, to the

extent that the OLS estimate is influenced by omitted variable bias associated, for example, with other environmental factors (mother's diet, exercise, etc.), then the results may be different in these populations because the correlation between smoking and these omitted factors may differ.