THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

CREATE CHANGE

ECON2300 - Introductory Econometrics

Tutorial 7: Assessing Studies Based on Multiple Regression

Tutor: Francisco Tavares Garcia

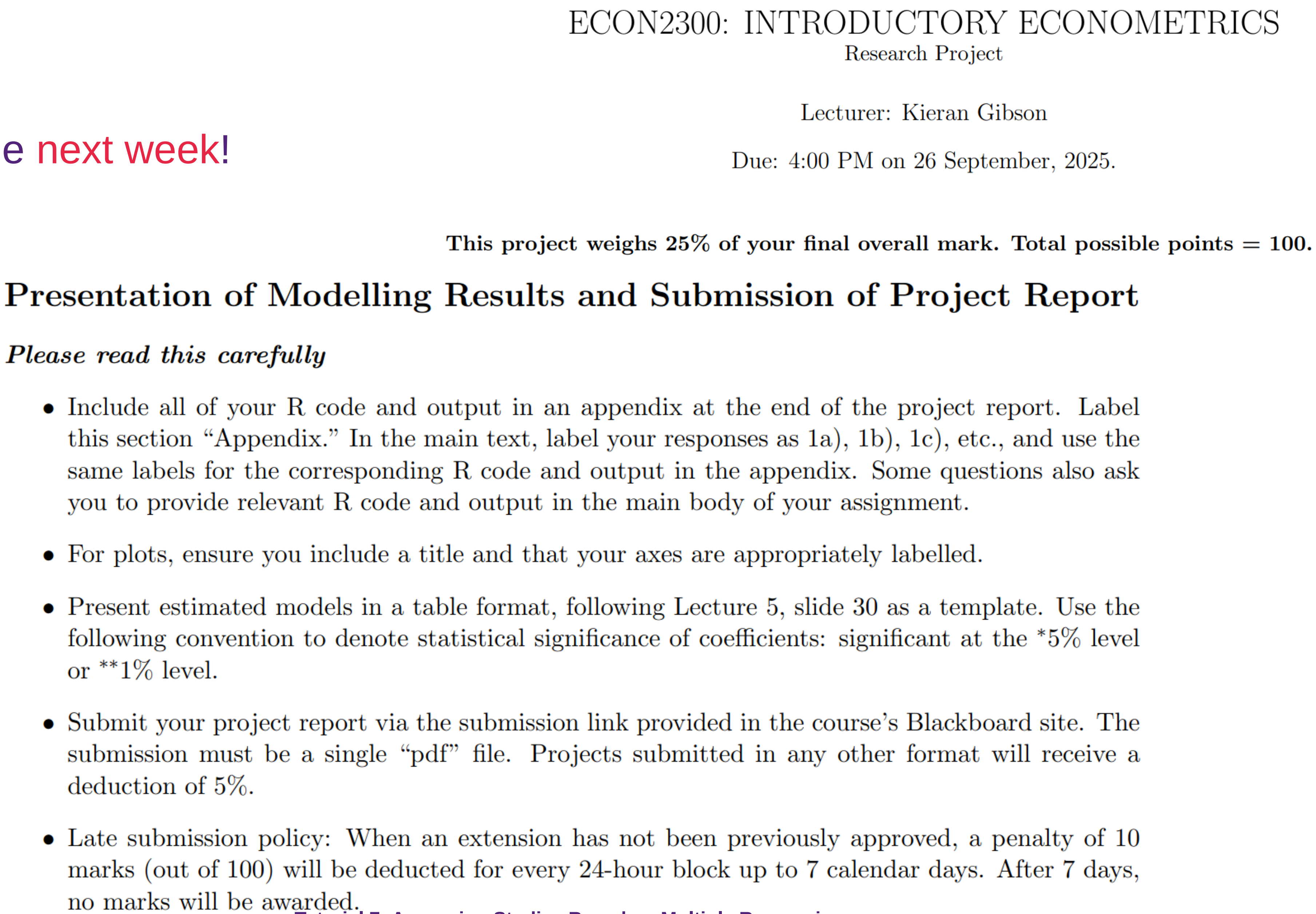
Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b

Quiz 6 is now available under the Assessment folder.

The due date is Thursday, 18th August, 16:00.

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b

Project 1 is due next week!



ECON2300: INTRODUCTORY ECONOMETRICS

Research Project

Lecturer: Kieran Gibson

Due: 4:00 PM on 26 September, 2025.

This project weighs 25% of your final overall mark. Total possible points = 100.

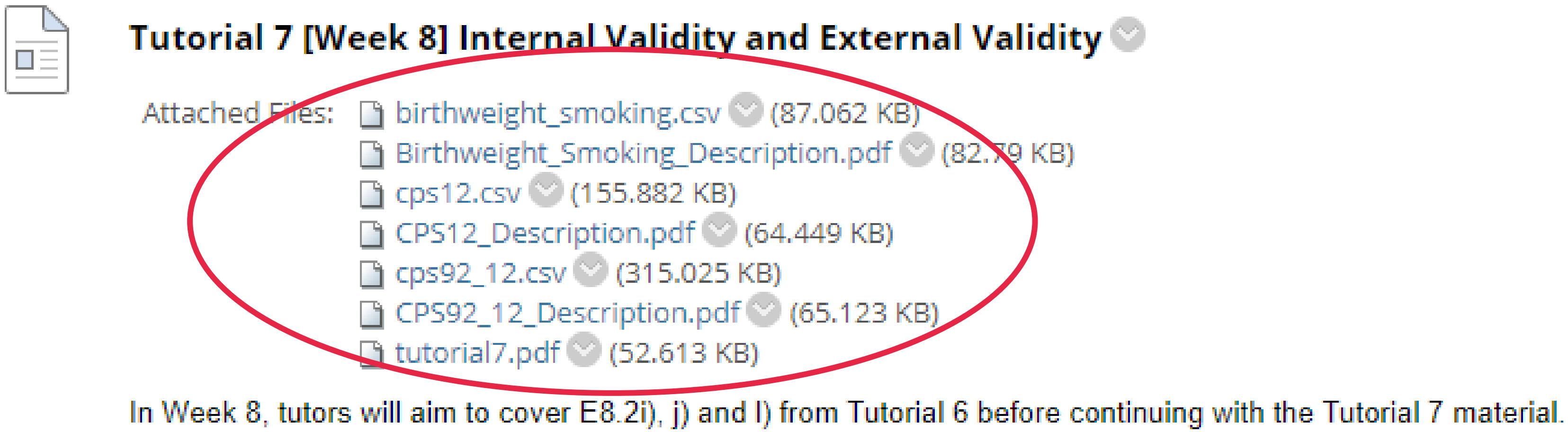
Presentation of Modelling Results and Submission of Project Report

Please read this carefully

- Include all of your R code and output in an appendix at the end of the project report. Label this section "Appendix." In the main text, label your responses as 1a), 1b), 1c), etc., and use the same labels for the corresponding R code and output in the appendix. Some questions also ask you to provide relevant R code and output in the main body of your assignment.
- For plots, ensure you include a title and that your axes are appropriately labelled.
- Present estimated models in a table format, following Lecture 5, slide 30 as a template. Use the following convention to denote statistical significance of coefficients: significant at the *5% level or **1% level.
- Submit your project report via the submission link provided in the course's Blackboard site. The submission must be a single "pdf" file. Projects submitted in any other format will receive a deduction of 5%.
- Late submission policy: When an extension has not been previously approved, a penalty of 10 marks (out of 100) will be deducted for every 24-hour block up to 7 calendar days. After 7 days, no marks will be awarded.

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b

- Download the files for tutorial 07 from Blackboard,
- save them into a folder for this tutorial.



Tutorial 7 [Week 8] Internal Validity and External Validity

Attached Files:
- birthweight_smoking.csv (87.062 KB)
- Birthweight_Smoking_Description.pdf (82.79 KB)
- cps12.csv (155.882 KB)
- CPS12_Description.pdf (64.449 KB)
- cps92_12.csv (315.025 KB)
- CPS92_12_Description.pdf (65.123 KB)
- tutorial7.pdf (52.613 KB)

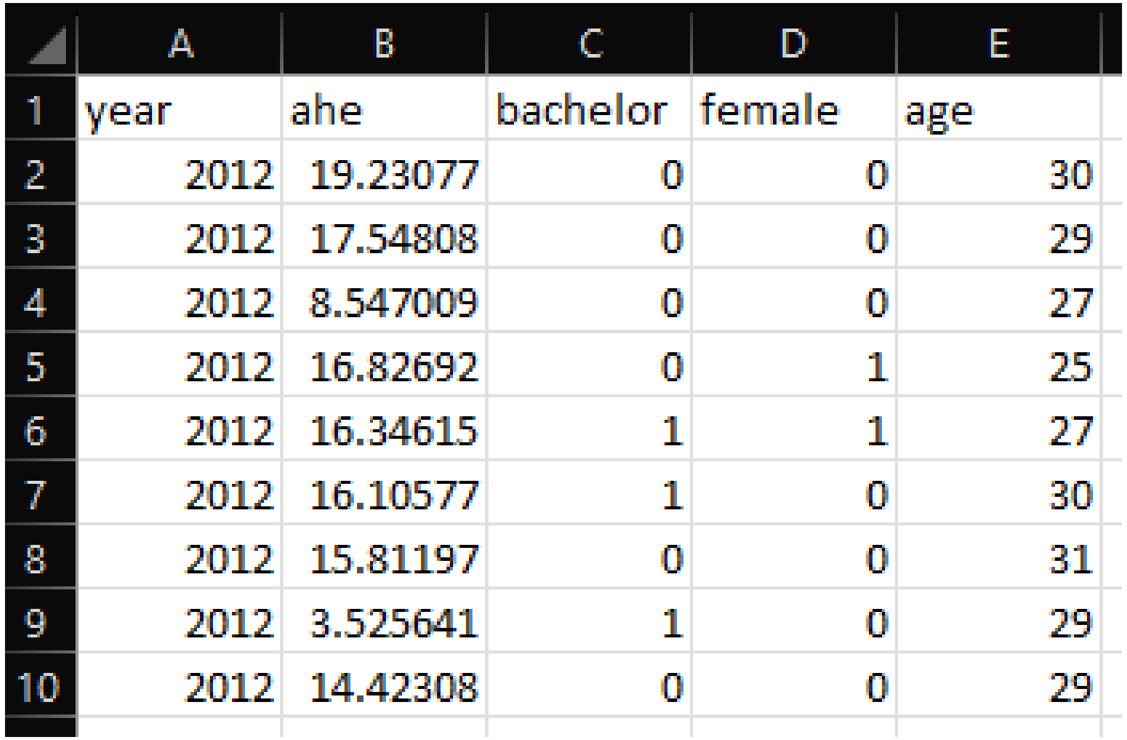
In Week 8, tutors will aim to cover E8.2i), j) and l) from Tutorial 6 before continuing with the Tutorial 7 material.

Now, let’s download the script for the tutorial.

- Copy the code from Github,
  - https://github.com/tavaresgarcia/teaching
- Save the scripts in the same folder as the data.

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b

E9.1 Use the data set cps12.csv, described in [E8.2] to answer the following questions.



	A	B	C	D	E
1	year	ahe	bachelor	female	age
2	2012	19.23077	0	0	30
3	2012	17.54808	0	0	29
4	2012	8.547009	0	0	27
5	2012	16.82692	0	1	25
6	2012	16.34615	1	1	27
7	2012	16.10577	1	0	30
8	2012	15.81197	0	0	31
9	2012	3.525641	1	0	29
10	2012	14.42308	0	0	29

Series in Data Set:

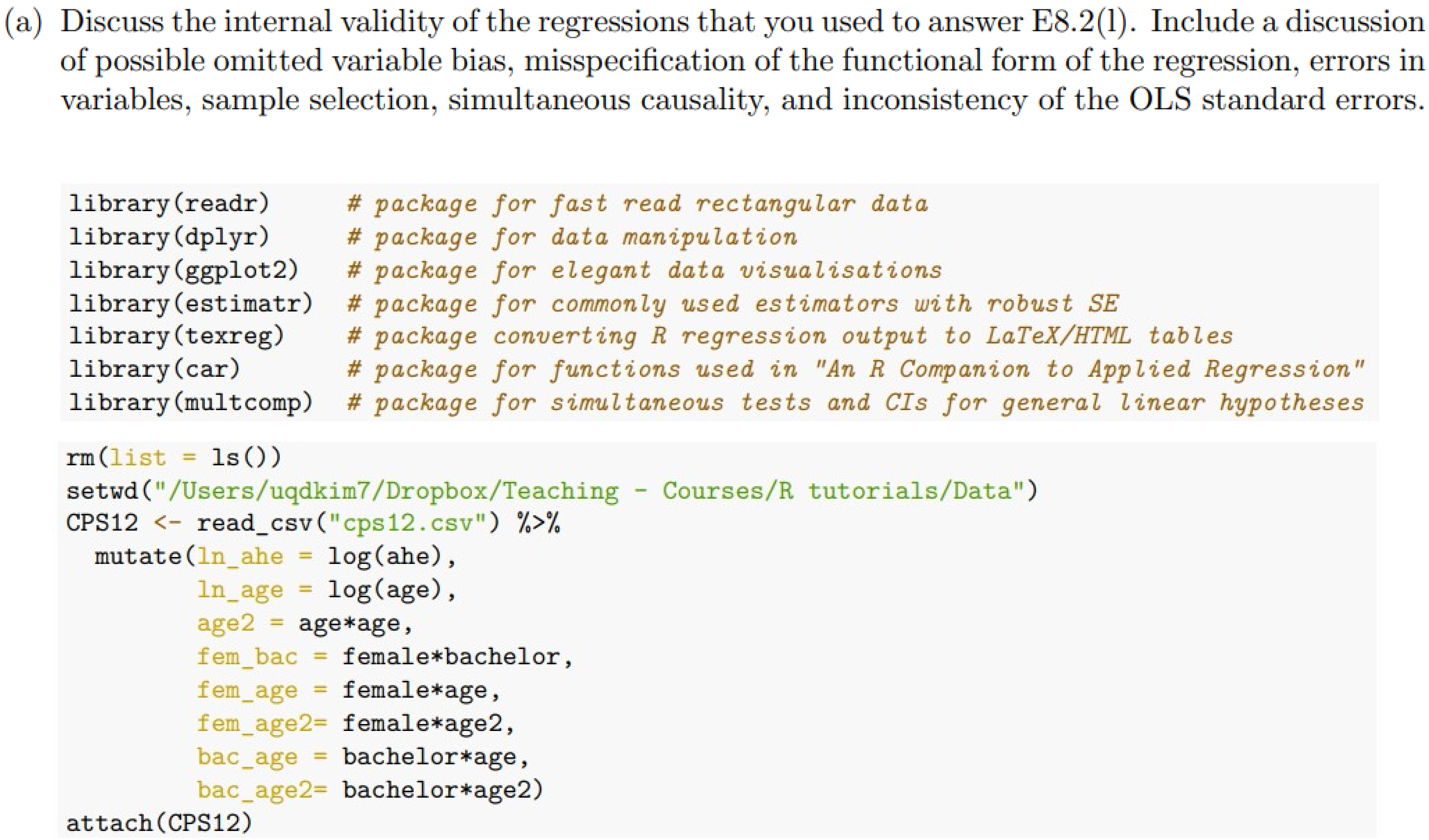
FEMALE: 1 if female; 0 if male

YEAR: Year

AHE : Average Hourly Earnings

BACHELOR: 1 if worker has a bachelor's degree;
0 if worker has a high school degree

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b

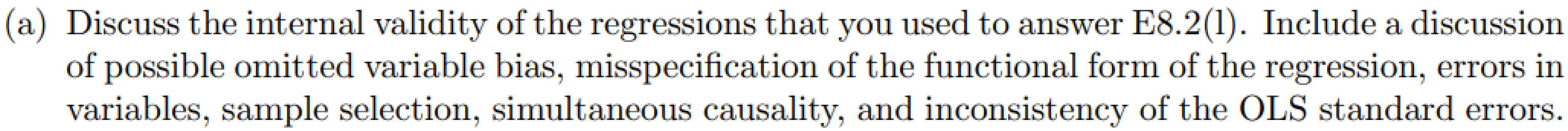


(a) Discuss the internal validity of the regressions that you used to answer E8.2(l). Include a discussion of possible omitted variable bias, misspecification of the functional form of the regression, errors in variables, sample selection, simultaneous causality, and inconsistency of the OLS standard errors.

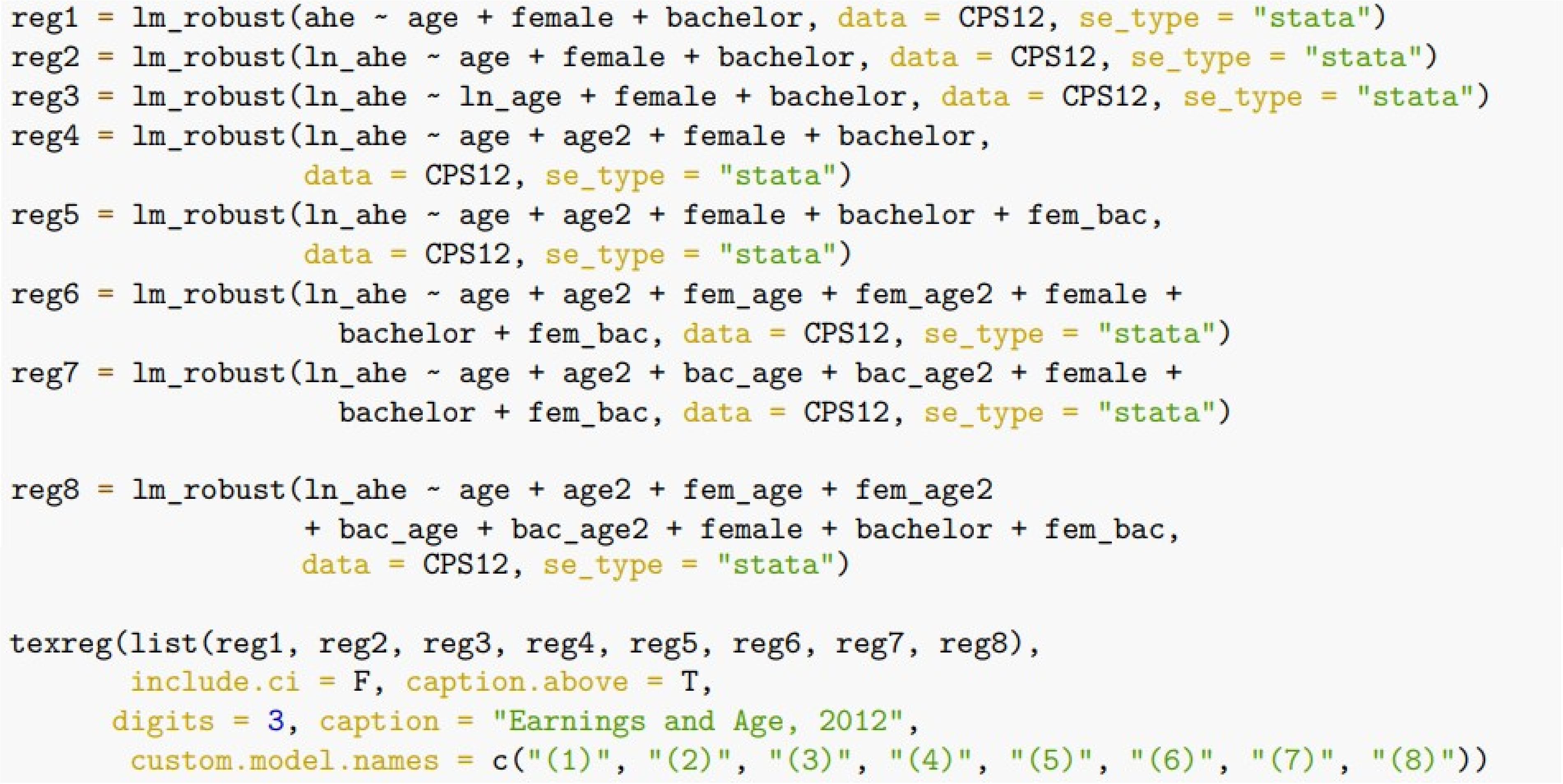
```r
library(readr)     # package for fast read rectangular data
library(dplyr)     # package for data manipulation
library(ggplot2)   # package for elegant data visualisations
library(estimatr)  # package for commonly used estimators with robust SE
library(texreg)    # package converting R regression output to LaTeX/HTML tables
library(car)       # package for functions used in "An R Companion to Applied Regression"
library(multcomp)  # package for simultaneous tests and CIs for general linear hypotheses
```

```r
rm(list = ls())
setwd("/Users/uqdkim7/Dropbox/Teaching - Courses/R tutorials/Data")
CPS12 <- read_csv("cps12.csv") %>%
  mutate(ln_ahe = log(ahe),
         ln_age = log(age),
         age2 = age*age,
         fem_bac = female*bachelor,
         fem_age = female*age,
         fem_age2= female*age2,
         bac_age = bachelor*age,
         bac_age2= bachelor*age2)
attach(CPS12)
```

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b



(a) Discuss the internal validity of the regressions that you used to answer E8.2(l). Include a discussion of possible omitted variable bias, misspecification of the functional form of the regression, errors in variables, sample selection, simultaneous causality, and inconsistency of the OLS standard errors.



```
reg1 = lm_robust(ahe ~ age + female + bachelor, data = CPS12, se_type = "stata")
reg2 = lm_robust(ln_ahe ~ age + female + bachelor, data = CPS12, se_type = "stata")
reg3 = lm_robust(ln_ahe ~ ln_age + female + bachelor, data = CPS12, se_type = "stata")
reg4 = lm_robust(ln_ahe ~ age + age2 + female + bachelor,
                 data = CPS12, se_type = "stata")
reg5 = lm_robust(ln_ahe ~ age + age2 + female + bachelor + fem_bac,
                 data = CPS12, se_type = "stata")
reg6 = lm_robust(ln_ahe ~ age + age2 + fem_age + fem_age2 + female +
                  bachelor + fem_bac, data = CPS12, se_type = "stata")
reg7 = lm_robust(ln_ahe ~ age + age2 + bac_age + bac_age2 + female +
                  bachelor + fem_bac, data = CPS12, se_type = "stata")

reg8 = lm_robust(ln_ahe ~ age + age2 + fem_age + fem_age2
                 + bac_age + bac_age2 + female + bachelor + fem_bac,
                 data = CPS12, se_type = "stata")

texreg(list(reg1, reg2, reg3, reg4, reg5, reg6, reg7, reg8),
       include.ci = F, caption.above = T,
      digits = 3, caption = "Earnings and Age, 2012",
       custom.model.names = c("(1)", "(2)", "(3)", "(4)", "(5)", "(6)", "(7)", "(8)"))
```

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b

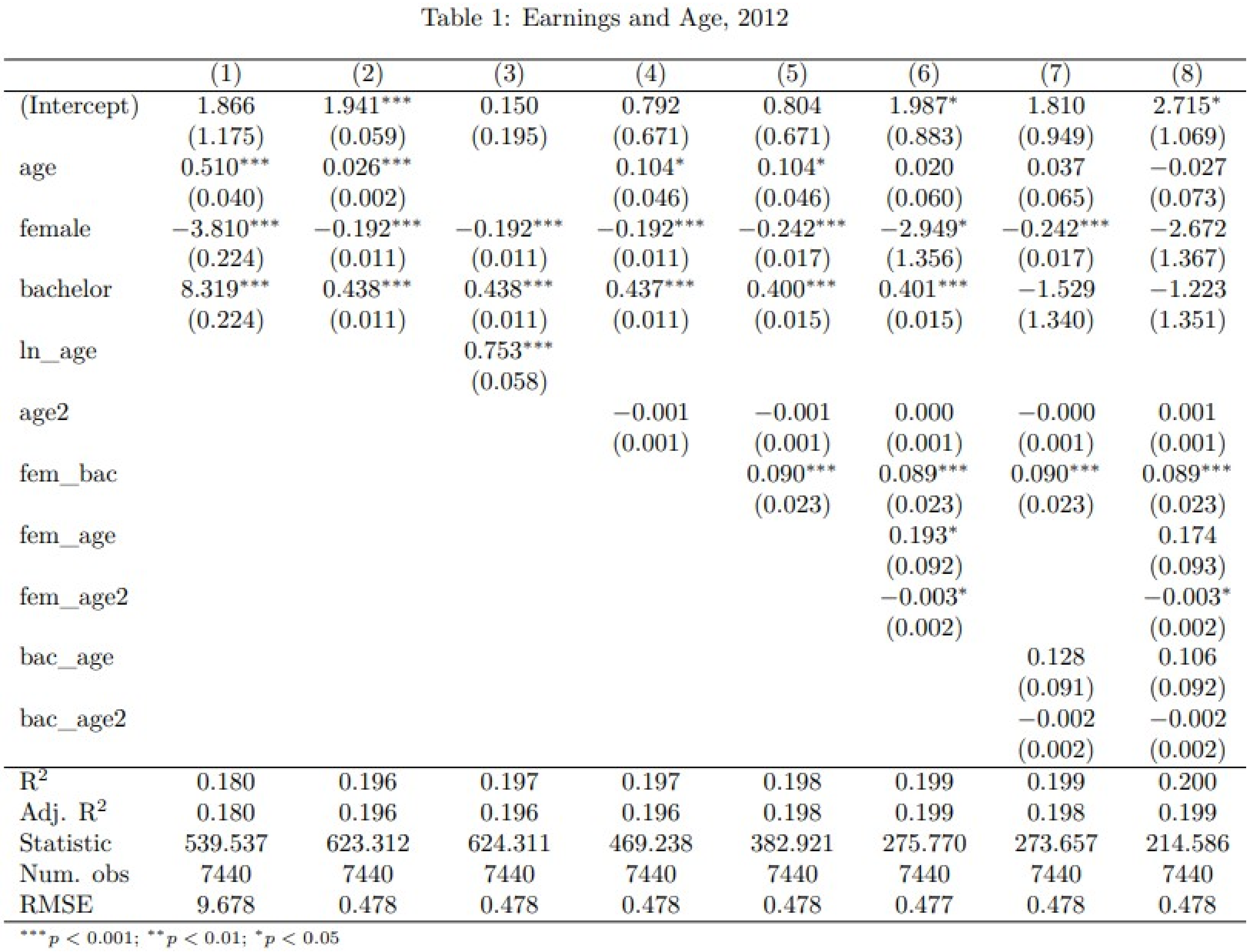
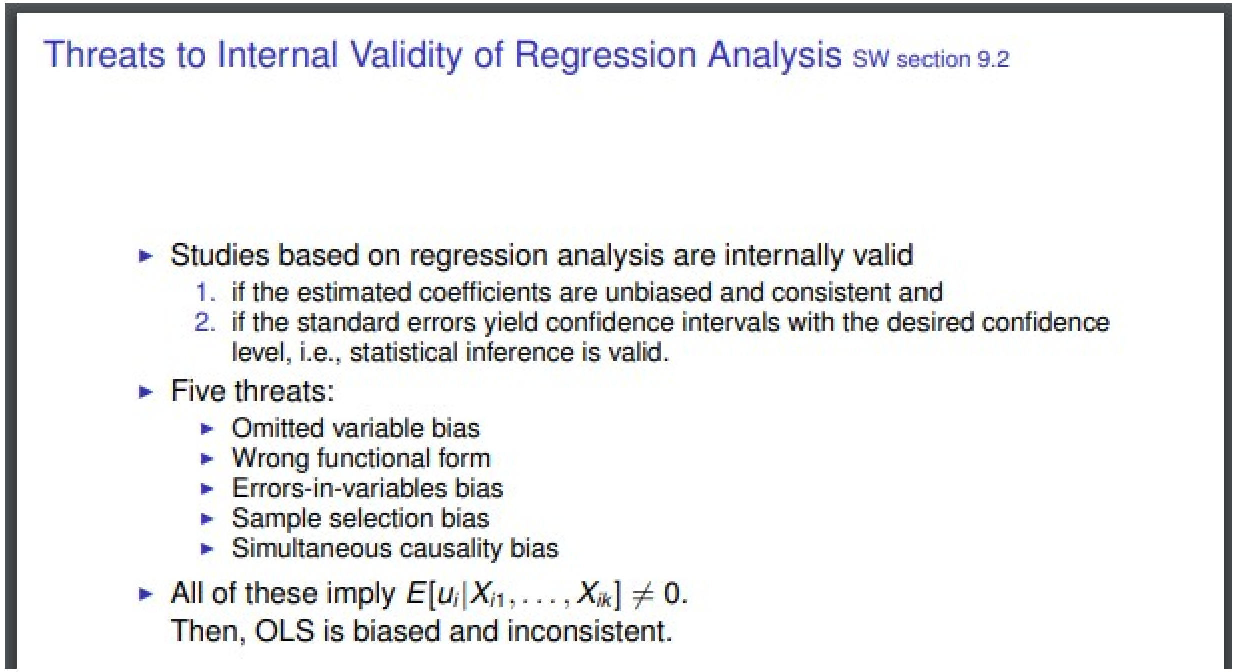


Table 1: Earnings and Age, 2012

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(Intercept)	1.866	1.941***	0.150	0.792	0.804	1.987*	1.810	2.715*
	(1.175)	(0.059)	(0.195)	(0.671)	(0.671)	(0.883)	(0.949)	(1.069)
age	0.510***	0.026***		0.104*	0.104*	0.020	0.037	−0.027
	(0.040)	(0.002)		(0.046)	(0.046)	(0.060)	(0.065)	(0.073)
female	−3.810***	−0.192***	−0.192***	−0.192***	−0.242***	−2.949*	−0.242***	−2.672
	(0.224)	(0.011)	(0.011)	(0.011)	(0.017)	(1.356)	(0.017)	(1.367)
bachelor	8.319***	0.438***	0.438***	0.437***	0.400***	0.401***	−1.529	−1.223
	(0.224)	(0.011)	(0.011)	(0.011)	(0.015)	(0.015)	(1.340)	(1.351)
ln_age			0.753***					
			(0.058)					
age2				−0.001	−0.001	0.000	−0.000	0.001
				(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
fem_bac					0.090***	0.089***	0.090***	0.089***
					(0.023)	(0.023)	(0.023)	(0.023)
fem_age						0.193*		0.174
						(0.092)		(0.093)
fem_age2						−0.003*		−0.003*
						(0.002)		(0.002)
bac_age							0.128	0.106
							(0.091)	(0.092)
bac_age2							−0.002	−0.002
							(0.002)	(0.002)
R^2	0.180	0.196	0.197	0.197	0.198	0.199	0.199	0.200
Adj. R^2	0.180	0.196	0.196	0.196	0.198	0.199	0.198	0.199
Statistic	539.537	623.312	624.311	469.238	382.921	275.770	273.657	214.586
Num. obs	7440	7440	7440	7440	7440	7440	7440	7440
RMSE	9.678	0.478	0.478	0.478	0.478	0.477	0.478	0.478

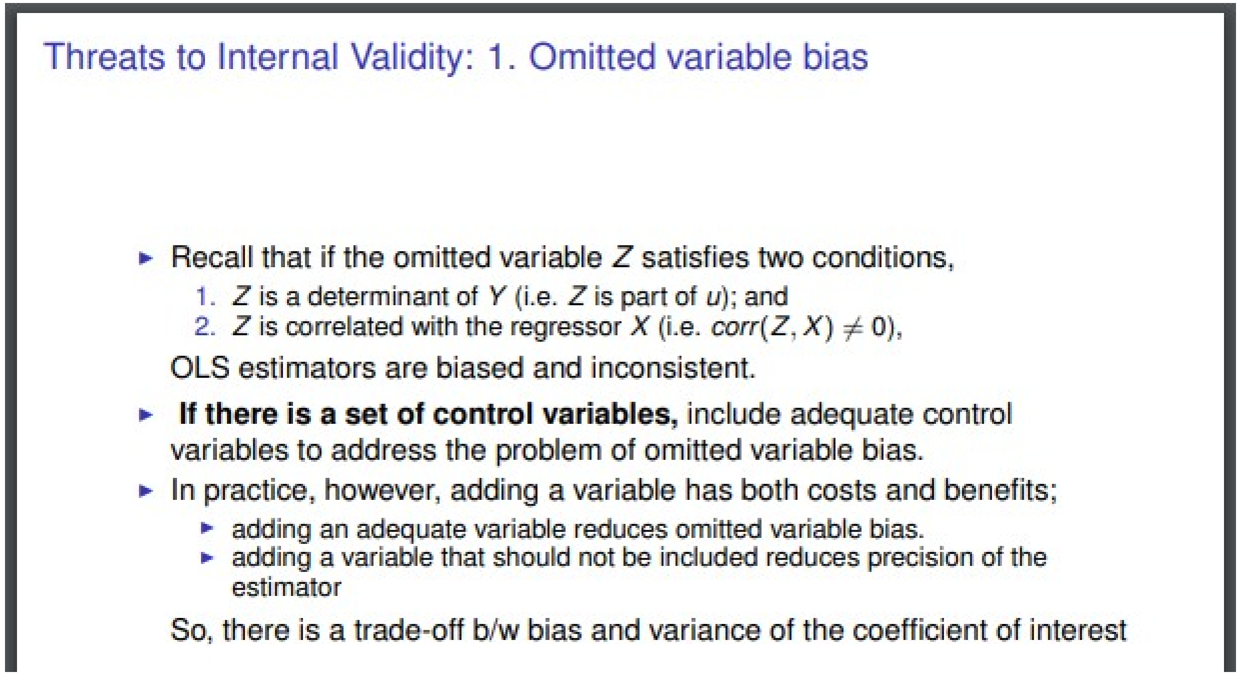
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b



Threats to Internal Validity of Regression Analysis SW section 9.2

- Studies based on regression analysis are internally valid
  1. if the estimated coefficients are unbiased and consistent and
  2. if the standard errors yield confidence intervals with the desired confidence level, i.e., statistical inference is valid.
- Five threats:
  - Omitted variable bias
  - Wrong functional form
  - Errors-in-variables bias
  - Sample selection bias
  - Simultaneous causality bias
- All of these imply $E[u_i|X_{i1}, \ldots, X_{ik}] \neq 0$.
  Then, OLS is biased and inconsistent.

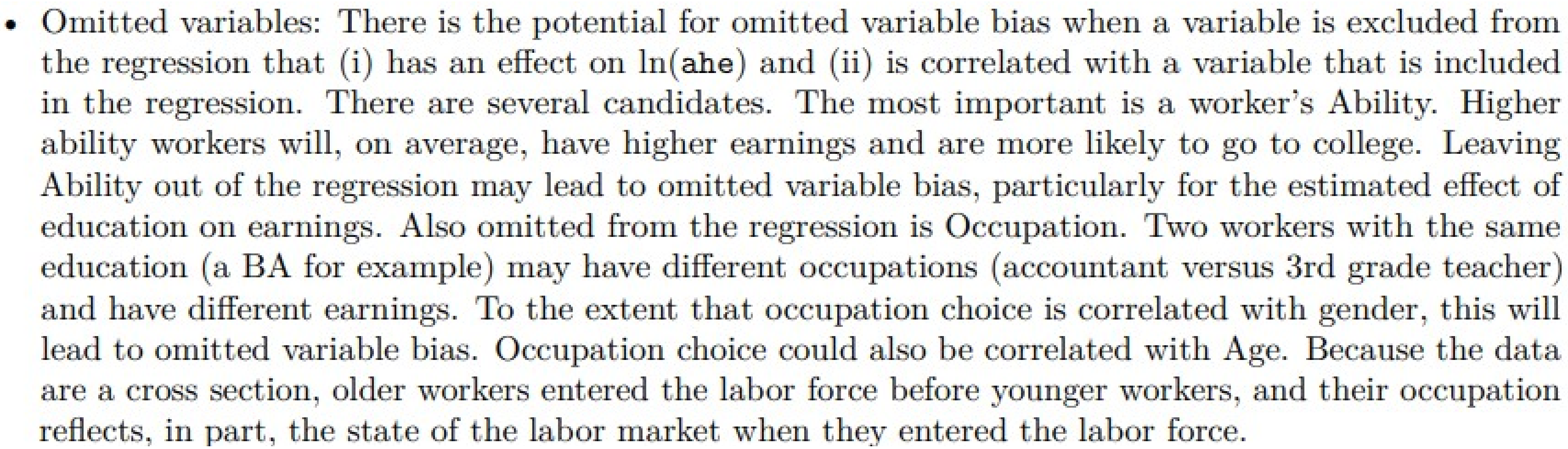


Threats to Internal Validity: 1. Omitted variable bias

- Recall that if the omitted variable Z satisfies two conditions,
  1. Z is a determinant of Y (i.e. Z is part of u); and
  2. Z is correlated with the regressor X (i.e. $corr(Z, X) \neq 0$),

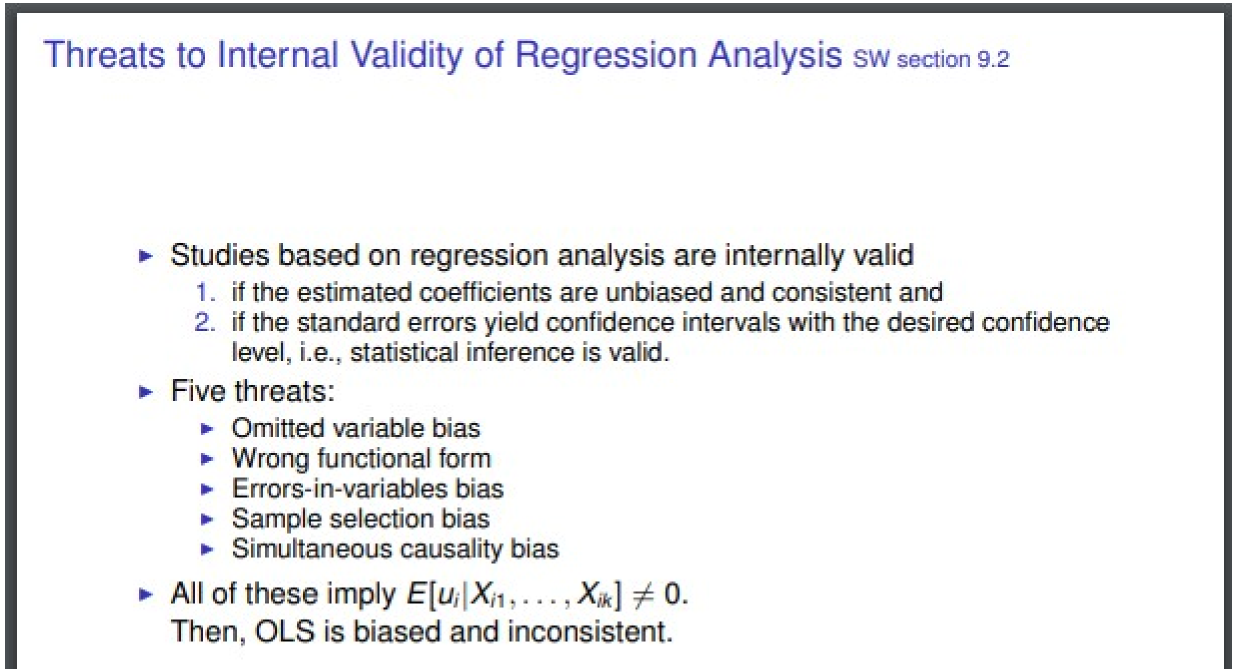
  OLS estimators are biased and inconsistent.
- **If there is a set of control variables,** include adequate control variables to address the problem of omitted variable bias.
- In practice, however, adding a variable has both costs and benefits;
  - adding an adequate variable reduces omitted variable bias.
  - adding a variable that should not be included reduces precision of the estimator

  So, there is a trade-off b/w bias and variance of the coefficient of interest



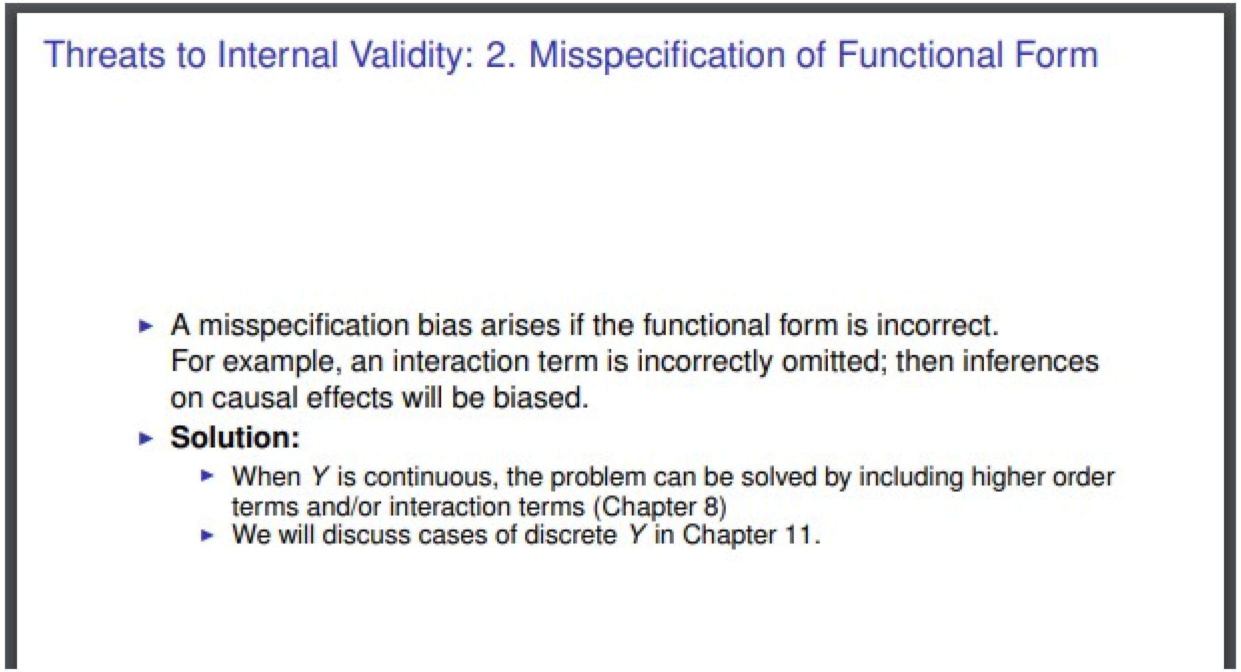
- Omitted variables: There is the potential for omitted variable bias when a variable is excluded from the regression that (i) has an effect on ln(ahe) and (ii) is correlated with a variable that is included in the regression. There are several candidates. The most important is a worker's Ability. Higher ability workers will, on average, have higher earnings and are more likely to go to college. Leaving Ability out of the regression may lead to omitted variable bias, particularly for the estimated effect of education on earnings. Also omitted from the regression is Occupation. Two workers with the same education (a BA for example) may have different occupations (accountant versus 3rd grade teacher) and have different earnings. To the extent that occupation choice is correlated with gender, this will lead to omitted variable bias. Occupation choice could also be correlated with Age. Because the data are a cross section, older workers entered the labor force before younger workers, and their occupation reflects, in part, the state of the labor market when they entered the labor force.

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b



Threats to Internal Validity of Regression Analysis SW section 9.2

- Studies based on regression analysis are internally valid
  1. if the estimated coefficients are unbiased and consistent and
  2. if the standard errors yield confidence intervals with the desired confidence level, i.e., statistical inference is valid.
- Five threats:
  - Omitted variable bias
  - Wrong functional form
  - Errors-in-variables bias
  - Sample selection bias
  - Simultaneous causality bias
- All of these imply $E[u_i|X_{i1},\ldots,X_{ik}] \neq 0$.
  Then, OLS is biased and inconsistent.

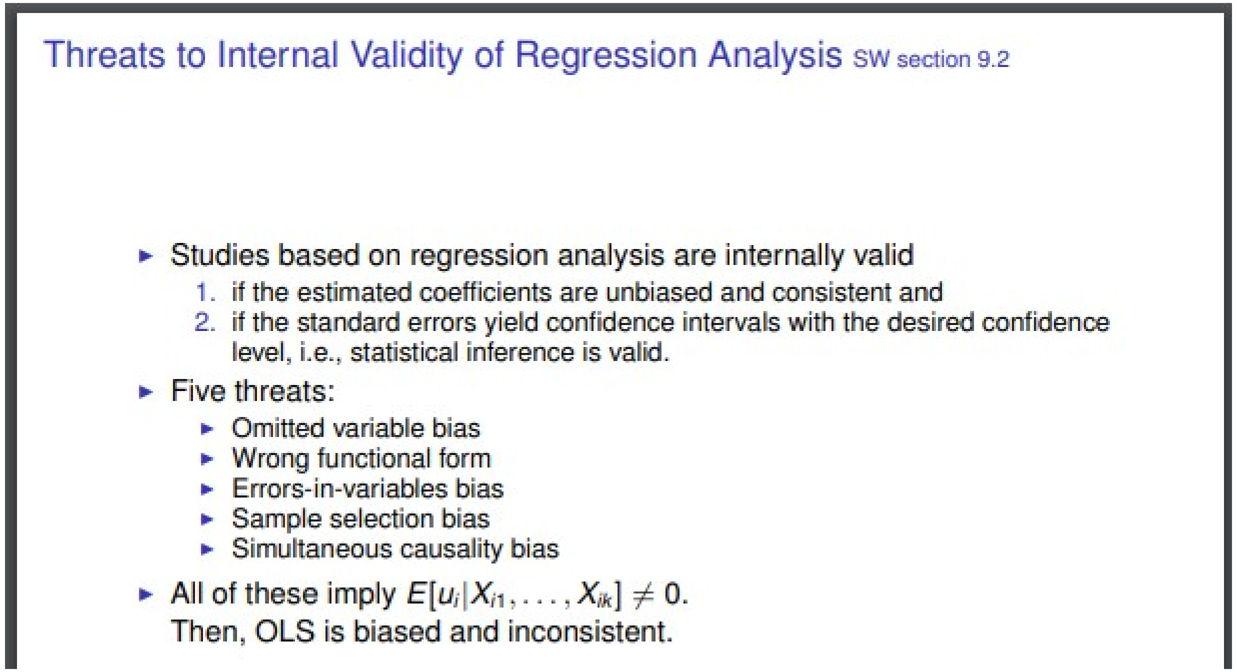


Threats to Internal Validity: 2. Misspecification of Functional Form

- A misspecification bias arises if the functional form is incorrect.
  For example, an interaction term is incorrectly omitted; then inferences on causal effects will be biased.
- **Solution:**
  - When Y is continuous, the problem can be solved by including higher order terms and/or interaction terms (Chapter 8)
  - We will discuss cases of discrete Y in Chapter 11.

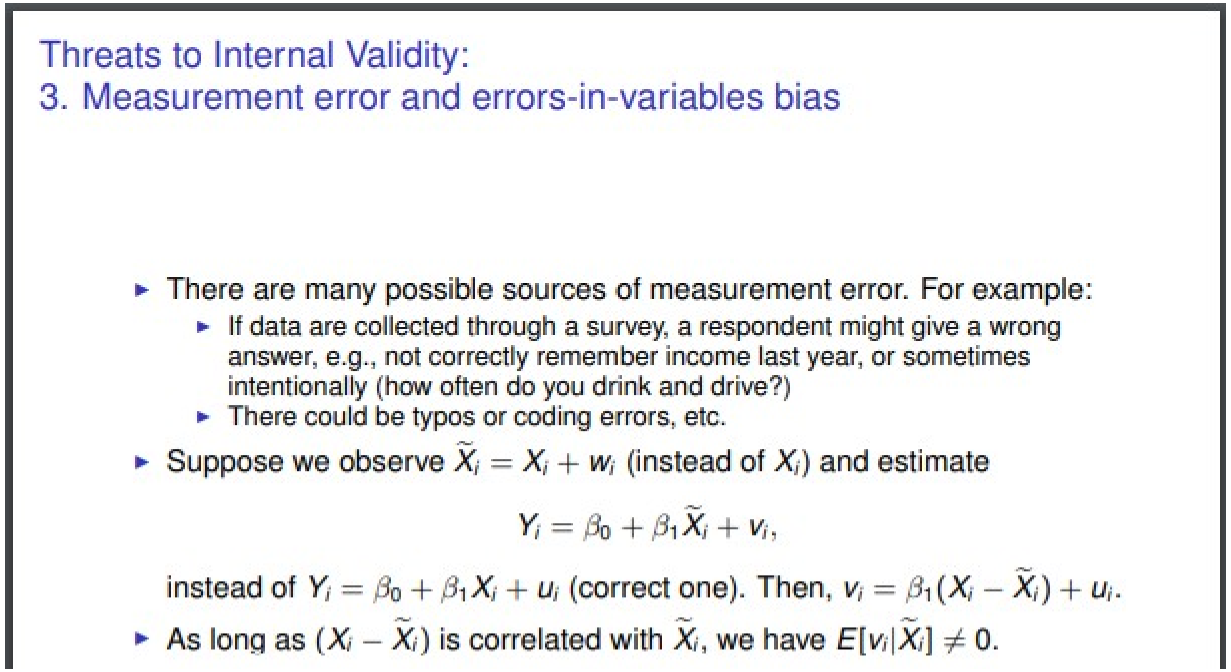
- Misspecification of the functional form: This was investigated carefully in (a)-(k). There does appear to be a nonlinear effect of Age on earnings, which is adequately captured by the polynomial regression with interaction terms.

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b



Threats to Internal Validity of Regression Analysis SW section 9.2

- Studies based on regression analysis are internally valid
  1. if the estimated coefficients are unbiased and consistent and
  2. if the standard errors yield confidence intervals with the desired confidence level, i.e., statistical inference is valid.
- Five threats:
  - Omitted variable bias
  - Wrong functional form
  - Errors-in-variables bias
  - Sample selection bias
  - Simultaneous causality bias
- All of these imply $E[u_i|X_{i1}, \ldots, X_{ik}] \neq 0$. Then, OLS is biased and inconsistent.



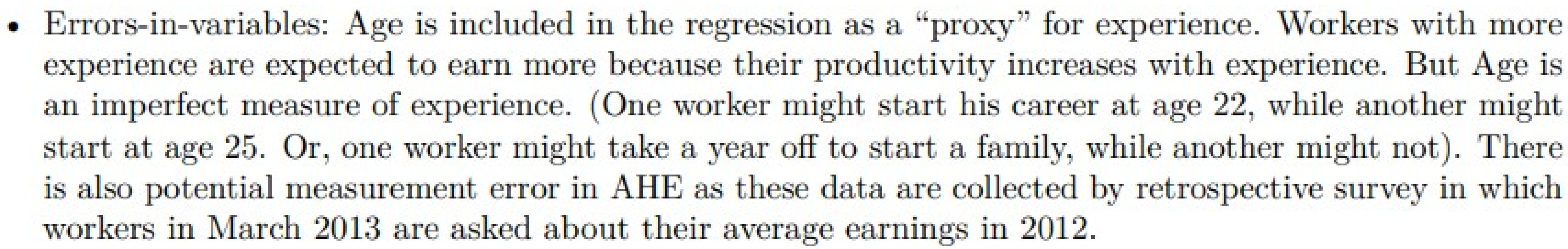
Threats to Internal Validity: 3. Measurement error and errors-in-variables bias

- There are many possible sources of measurement error. For example:
  - If data are collected through a survey, a respondent might give a wrong answer, e.g., not correctly remember income last year, or sometimes intentionally (how often do you drink and drive?)
  - There could be typos or coding errors, etc.
- Suppose we observe $\widetilde{X}_i = X_i + w_i$ (instead of X_i) and estimate

$$Y_i = \beta_0 + \beta_1 \widetilde{X}_i + v_i,$$

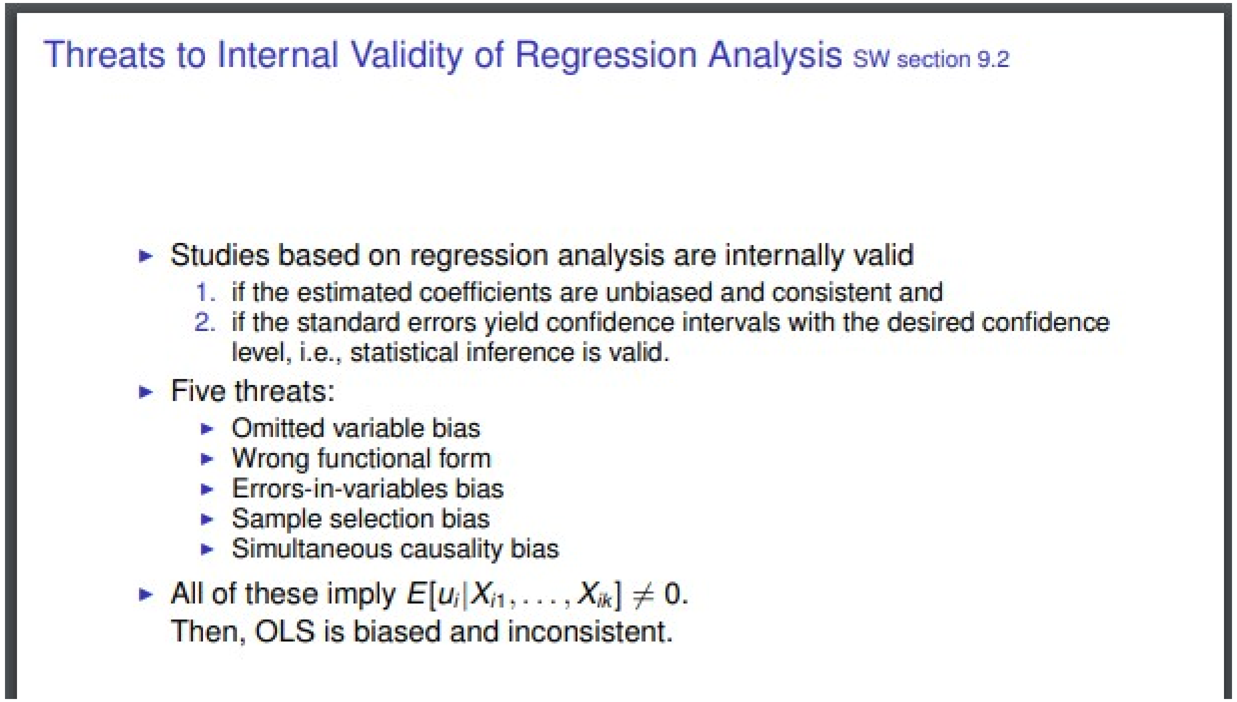
instead of $Y_i = \beta_0 + \beta_1 X_i + u_i$ (correct one). Then, $v_i = \beta_1(X_i - \widetilde{X}_i) + u_i$.

- As long as $(X_i - \widetilde{X}_i)$ is correlated with $\widetilde{X}_i$, we have $E[v_i|\widetilde{X}_i] \neq 0$.



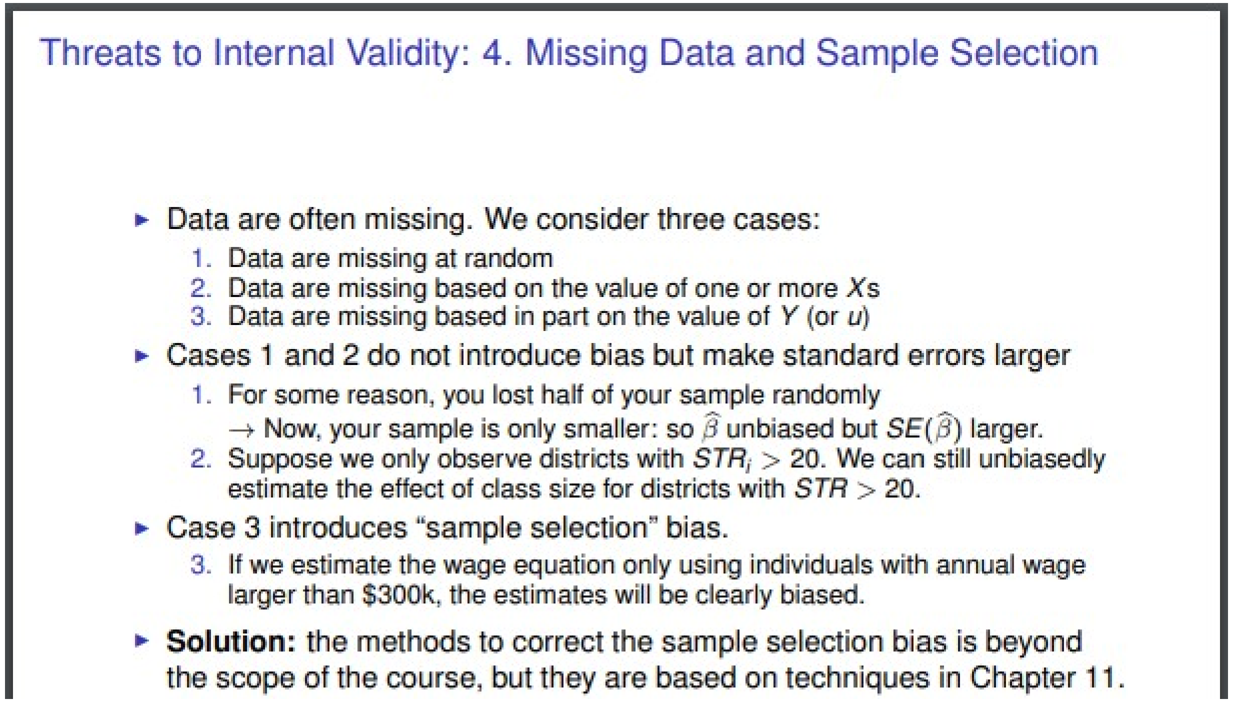
- Errors-in-variables: Age is included in the regression as a "proxy" for experience. Workers with more experience are expected to earn more because their productivity increases with experience. But Age is an imperfect measure of experience. (One worker might start his career at age 22, while another might start at age 25. Or, one worker might take a year off to start a family, while another might not). There is also potential measurement error in AHE as these data are collected by retrospective survey in which workers in March 2013 are asked about their average earnings in 2012.

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b



Threats to Internal Validity of Regression Analysis SW section 9.2

- Studies based on regression analysis are internally valid
  1. if the estimated coefficients are unbiased and consistent and
  2. if the standard errors yield confidence intervals with the desired confidence level, i.e., statistical inference is valid.
- Five threats:
  - Omitted variable bias
  - Wrong functional form
  - Errors-in-variables bias
  - Sample selection bias
  - Simultaneous causality bias
- All of these imply $E[u_i|X_{i1}, \dots, X_{ik}] \neq 0$. Then, OLS is biased and inconsistent.

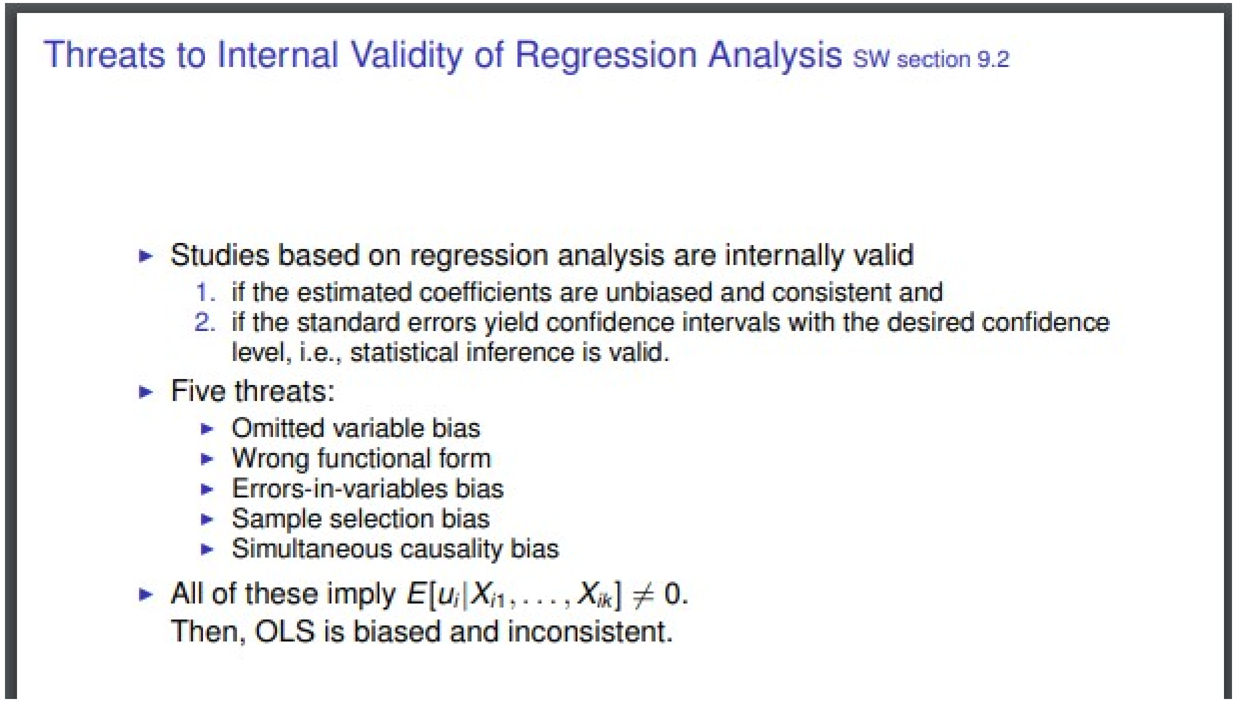


Threats to Internal Validity: 4. Missing Data and Sample Selection

- Data are often missing. We consider three cases:
  1. Data are missing at random
  2. Data are missing based on the value of one or more Xs
  3. Data are missing based in part on the value of Y (or u)
- Cases 1 and 2 do not introduce bias but make standard errors larger
  1. For some reason, you lost half of your sample randomly $\rightarrow$ Now, your sample is only smaller: so $\widehat{\beta}$ unbiased but $SE(\widehat{\beta})$ larger.
  2. Suppose we only observe districts with $STR_i > 20$. We can still unbiasedly estimate the effect of class size for districts with $STR > 20$.
- Case 3 introduces "sample selection" bias.
  3. If we estimate the wage equation only using individuals with annual wage larger than \$300k, the estimates will be clearly biased.
- **Solution:** the methods to correct the sample selection bias is beyond the scope of the course, but they are based on techniques in Chapter 11.

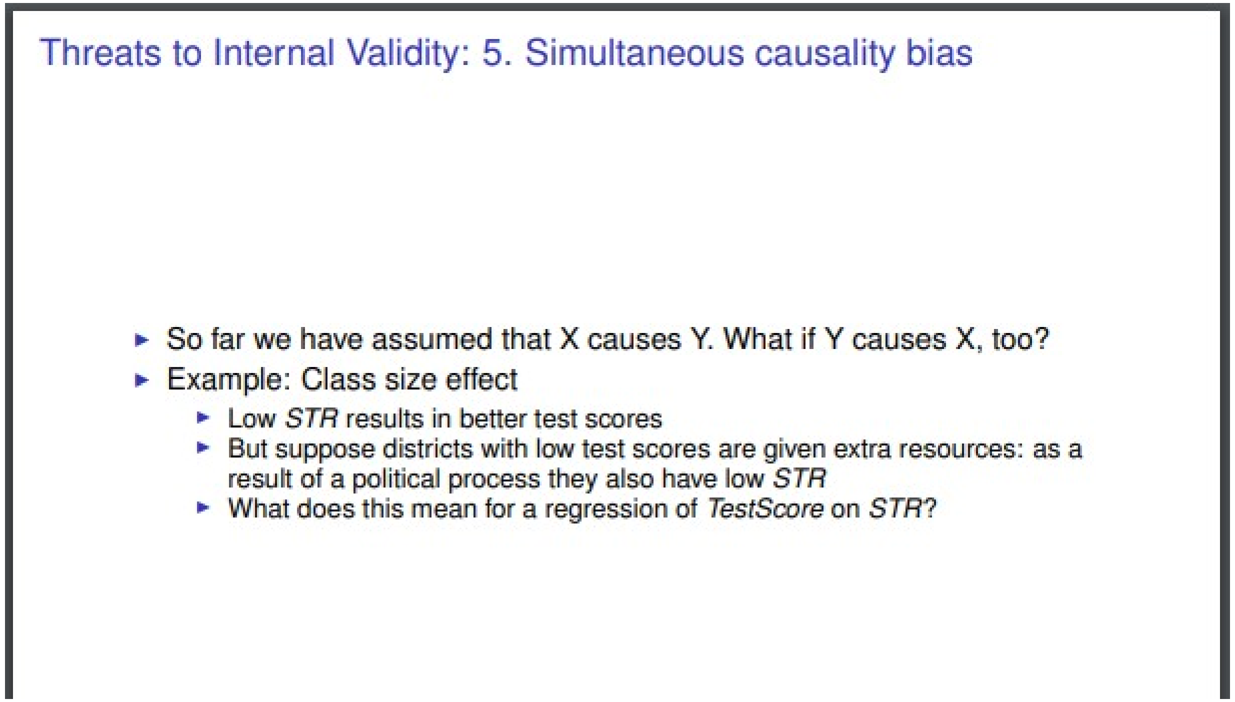
- Sample selection: The data are full-time workers only, so there is potential for sample-selection bias.

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b



Threats to Internal Validity of Regression Analysis SW section 9.2

- Studies based on regression analysis are internally valid
  1. if the estimated coefficients are unbiased and consistent and
  2. if the standard errors yield confidence intervals with the desired confidence level, i.e., statistical inference is valid.
- Five threats:
  - Omitted variable bias
  - Wrong functional form
  - Errors-in-variables bias
  - Sample selection bias
  - Simultaneous causality bias
- All of these imply $E[u_i|X_{i1},\ldots,X_{ik}] \neq 0$.
  Then, OLS is biased and inconsistent.

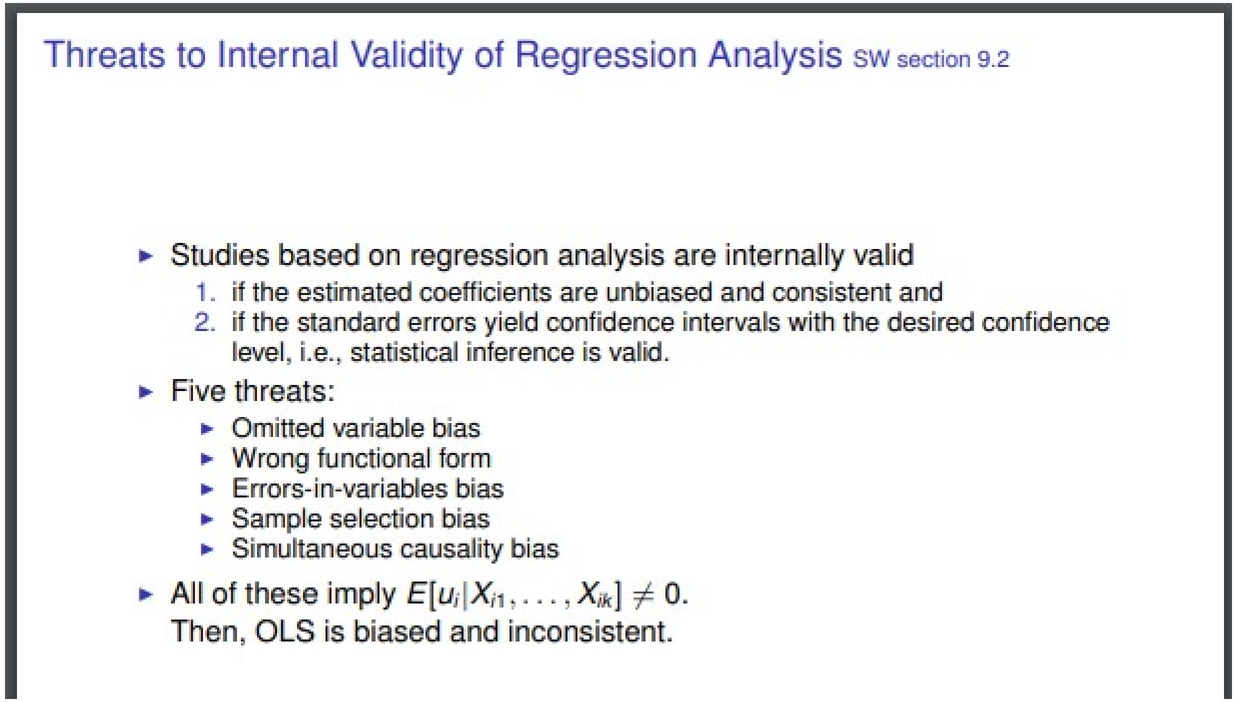


Threats to Internal Validity: 5. Simultaneous causality bias

- So far we have assumed that X causes Y. What if Y causes X, too?
- Example: Class size effect
  - Low *STR* results in better test scores
  - But suppose districts with low test scores are given extra resources: as a result of a political process they also have low *STR*
  - What does this mean for a regression of *TestScore* on *STR*?

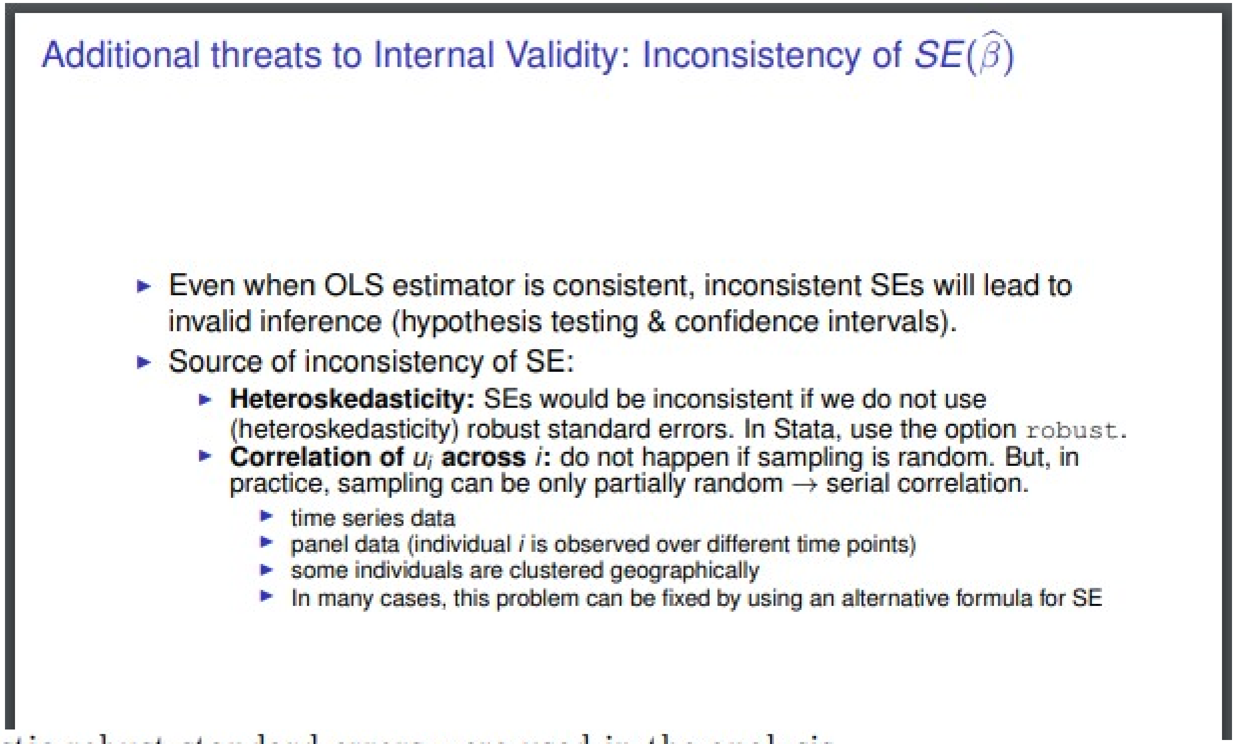
- Simultaneous causality: This is unlikely to be a problem. It is unlikely that AHE affects Age or gender.

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b



Threats to Internal Validity of Regression Analysis SW section 9.2

- Studies based on regression analysis are internally valid
  1. if the estimated coefficients are unbiased and consistent and
  2. if the standard errors yield confidence intervals with the desired confidence level, i.e., statistical inference is valid.
- Five threats:
  - Omitted variable bias
  - Wrong functional form
  - Errors-in-variables bias
  - Sample selection bias
  - Simultaneous causality bias
- All of these imply $E[u_i|X_{i1}, \ldots, X_{ik}] \neq 0$.
  Then, OLS is biased and inconsistent.



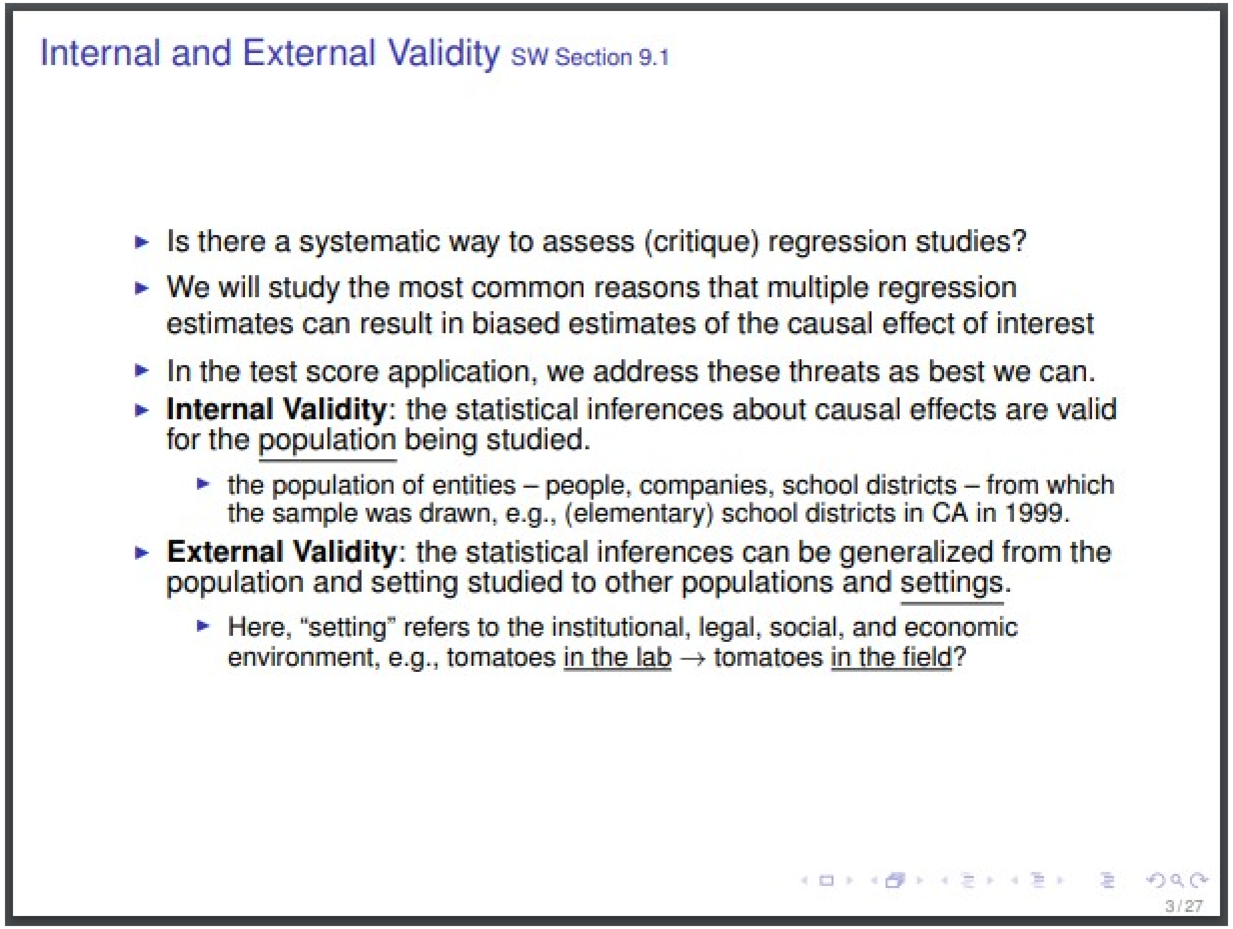
Additional threats to Internal Validity: Inconsistency of $SE(\widehat{\beta})$

- Even when OLS estimator is consistent, inconsistent SEs will lead to invalid inference (hypothesis testing & confidence intervals).
- Source of inconsistency of SE:
  - **Heteroskedasticity:** SEs would be inconsistent if we do not use (heteroskedasticity) robust standard errors. In Stata, use the option `robust`.
  - **Correlation of u_i across i:** do not happen if sampling is random. But, in practice, sampling can be only partially random → serial correlation.
    - time series data
    - panel data (individual i is observed over different time points)
    - some individuals are clustered geographically
    - In many cases, this problem can be fixed by using an alternative formula for SE

- 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, at least approximately, using i.i.d. sampling, so that correlation across the errors is unlikely to be a problem.

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b

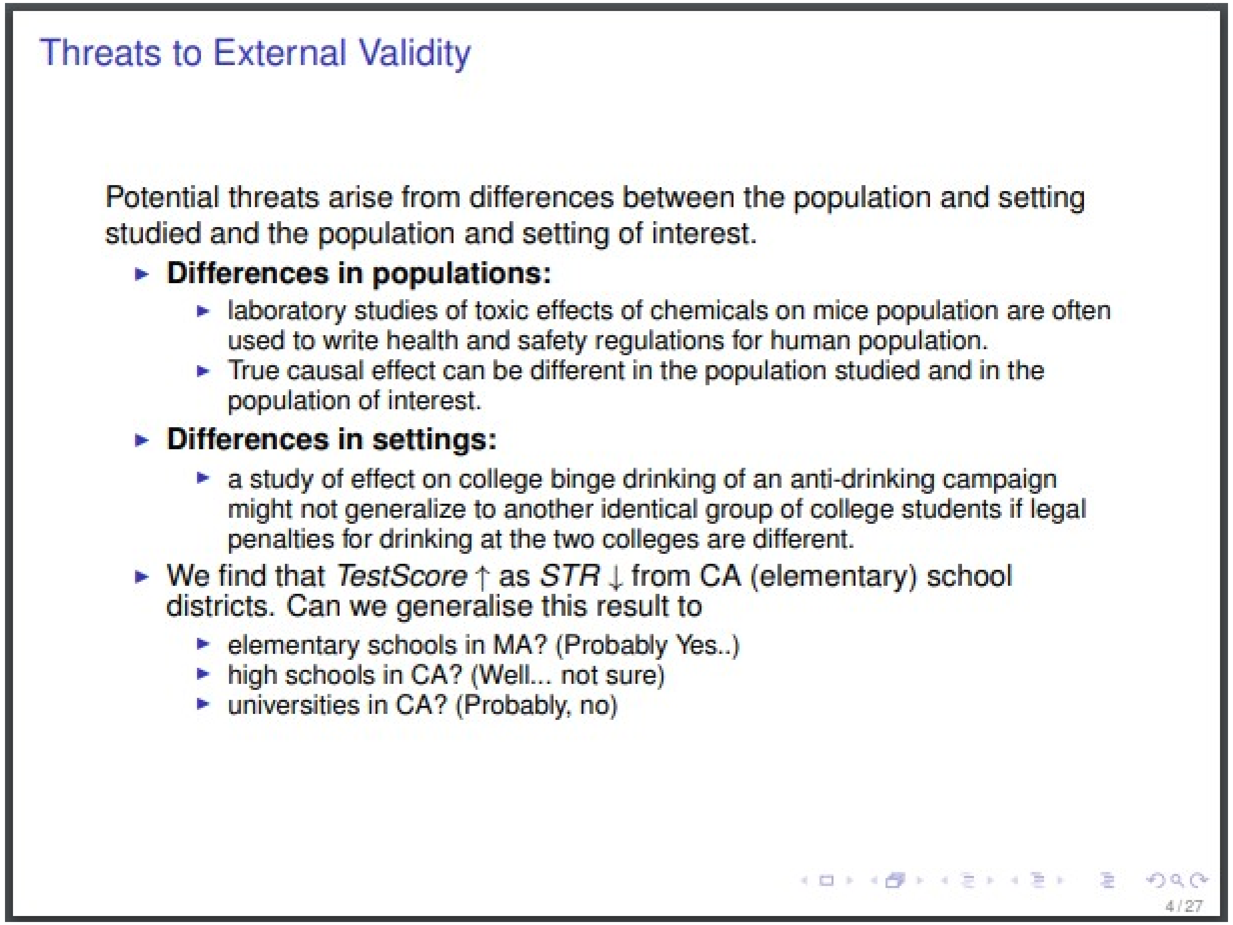
(b) The data set `cps92_12.csv` described in `cps92_12_description.pdf` includes data from 2012 and 1992. Use these data to investigate the (temporal) external validity of the conclusions that you reached in E8.2(l). [Note: In 2012 the value of the Consumer Price Index (CPI) was 229.6. In 1992, the value of the CPI was 140.3. Adjust the 1992 data for the price inflation that occurred between 1992 and 2012.]



Internal and External Validity SW Section 9.1

- Is there a systematic way to assess (critique) regression studies?
- We will study the most common reasons that multiple regression estimates can result in biased estimates of the causal effect of interest
- In the test score application, we address these threats as best we can.
- **Internal Validity**: the statistical inferences about causal effects are valid for the population being studied.
  - the population of entities – people, companies, school districts – from which the sample was drawn, e.g., (elementary) school districts in CA in 1999.
- **External Validity**: the statistical inferences can be generalized from the population and setting studied to other populations and settings.
  - Here, "setting" refers to the institutional, legal, social, and economic environment, e.g., tomatoes in the lab → tomatoes in the field?

3 / 27



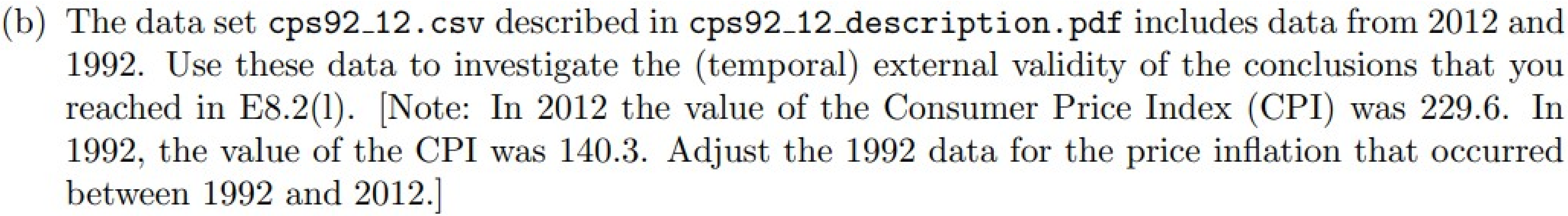
Threats to External Validity

Potential threats arise from differences between the population and setting studied and the population and setting of interest.

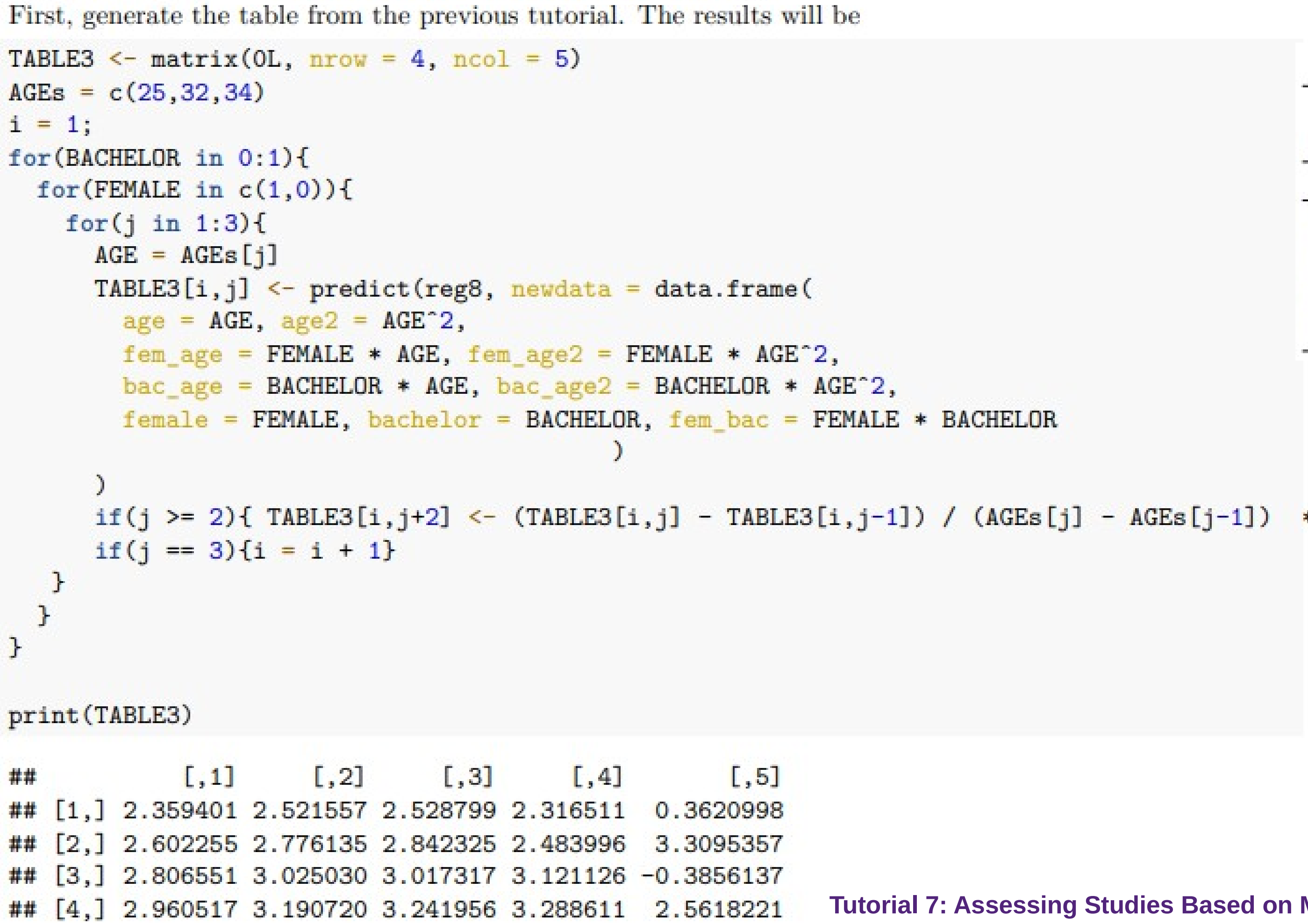
- **Differences in populations:**
  - laboratory studies of toxic effects of chemicals on mice population are often used to write health and safety regulations for human population.
  - True causal effect can be different in the population studied and in the population of interest.
- **Differences in settings:**
  - a study of effect on college binge drinking of an anti-drinking campaign might not generalize to another identical group of college students if legal penalties for drinking at the two colleges are different.
- We find that *TestScore* ↑ as *STR* ↓ from CA (elementary) school districts. Can we generalise this result to
  - elementary schools in MA? (Probably Yes..)
  - high schools in CA? (Well... not sure)
  - universities in CA? (Probably, no)

4 / 27

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b



(b) The data set cps92_12.csv described in cps92_12_description.pdf includes data from 2012 and 1992. Use these data to investigate the (temporal) external validity of the conclusions that you reached in E8.2(l). [Note: In 2012 the value of the Consumer Price Index (CPI) was 229.6. In 1992, the value of the CPI was 140.3. Adjust the 1992 data for the price inflation that occurred between 1992 and 2012.]



First, generate the table from the previous tutorial. The results will be

```
TABLE3 <- matrix(0L, nrow = 4, ncol = 5)
AGEs = c(25,32,34)
i = 1;
for(BACHELOR in 0:1){
  for(FEMALE in c(1,0)){
    for(j in 1:3){
      AGE = AGEs[j]
      TABLE3[i,j] <- predict(reg8, newdata = data.frame(
        age = AGE, age2 = AGE^2,
        fem_age = FEMALE * AGE, fem_age2 = FEMALE * AGE^2,
        bac_age = BACHELOR * AGE, bac_age2 = BACHELOR * AGE^2,
        female = FEMALE, bachelor = BACHELOR, fem_bac = FEMALE * BACHELOR
                                            )
      )
      if(j >= 2){ TABLE3[i,j+2] <- (TABLE3[i,j] - TABLE3[i,j-1]) / (AGEs[j] - AGEs[j-1])  * 100}
      if(j == 3){i = i + 1}
    }
  }
}

print(TABLE3)
```

```
##          [,1]     [,2]     [,3]     [,4]       [,5]
## [1,] 2.359401 2.521557 2.528799 2.316511  0.3620998
## [2,] 2.602255 2.776135 2.842325 2.483996  3.3095357
## [3,] 2.806551 3.025030 3.017317 3.121126 -0.3856137
## [4,] 2.960517 3.190720 3.241956 3.288611  2.5618221
```

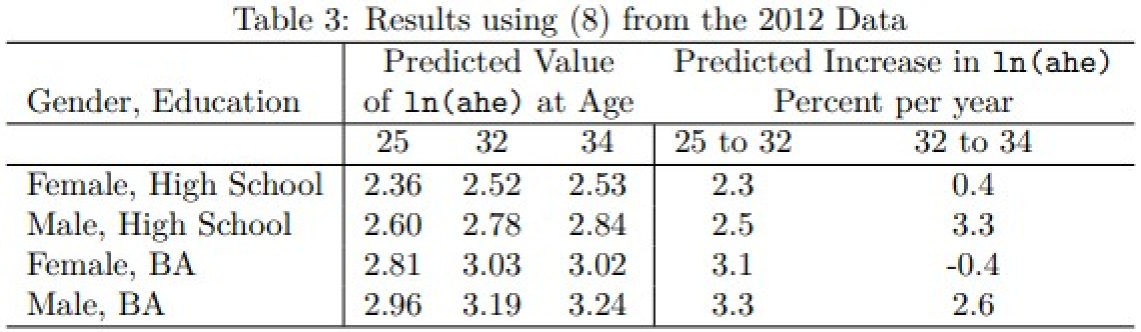
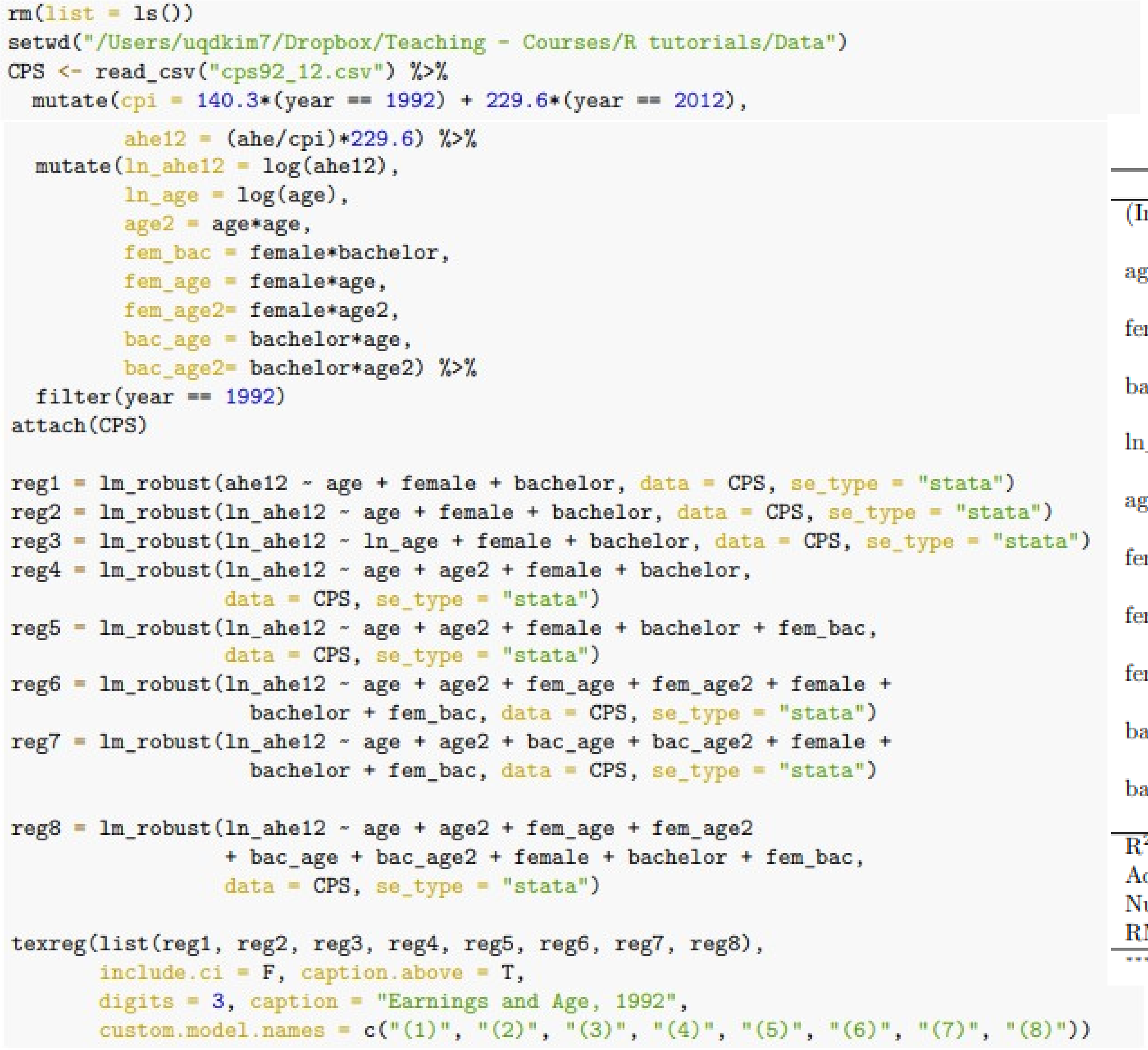


Table 3: Results using (8) from the 2012 Data

Gender, Education	Predicted Value of ln(ahe) at Age			Predicted Increase in ln(ahe) Percent per year	
	25	32	34	25 to 32	32 to 34
Female, High School	2.36	2.52	2.53	2.3	0.4
Male, High School	2.60	2.78	2.84	2.5	3.3
Female, BA	2.81	3.03	3.02	3.1	-0.4
Male, BA	2.96	3.19	3.24	3.3	2.6

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b



```
rm(list = ls())
setwd("/Users/uqdkim7/Dropbox/Teaching - Courses/R tutorials/Data")
CPS <- read_csv("cps92_12.csv") %>%
  mutate(cpi = 140.3*(year == 1992) + 229.6*(year == 2012),
         ahe12 = (ahe/cpi)*229.6) %>%
  mutate(ln_ahe12 = log(ahe12),
         ln_age = log(age),
         age2 = age*age,
         fem_bac = female*bachelor,
         fem_age = female*age,
         fem_age2= female*age2,
         bac_age = bachelor*age,
         bac_age2= bachelor*age2) %>%
  filter(year == 1992)
attach(CPS)

reg1 = lm_robust(ahe12 ~ age + female + bachelor, data = CPS, se_type = "stata")
reg2 = lm_robust(ln_ahe12 ~ age + female + bachelor, data = CPS, se_type = "stata")
reg3 = lm_robust(ln_ahe12 ~ ln_age + female + bachelor, data = CPS, se_type = "stata")
reg4 = lm_robust(ln_ahe12 ~ age + age2 + female + bachelor,
                 data = CPS, se_type = "stata")
reg5 = lm_robust(ln_ahe12 ~ age + age2 + female + bachelor + fem_bac,
                 data = CPS, se_type = "stata")
reg6 = lm_robust(ln_ahe12 ~ age + age2 + fem_age + fem_age2 + female +
                   bachelor + fem_bac, data = CPS, se_type = "stata")
reg7 = lm_robust(ln_ahe12 ~ age + age2 + bac_age + bac_age2 + female +
                   bachelor + fem_bac, data = CPS, se_type = "stata")

reg8 = lm_robust(ln_ahe12 ~ age + age2 + fem_age + fem_age2
                 + bac_age + bac_age2 + female + bachelor + fem_bac,
                 data = CPS, se_type = "stata")

texreg(list(reg1, reg2, reg3, reg4, reg5, reg6, reg7, reg8),
       include.ci = F, caption.above = T,
       digits = 3, caption = "Earnings and Age, 1992",
       custom.model.names = c("(1)", "(2)", "(3)", "(4)", "(5)", "(6)", "(7)", "(8)"))
```

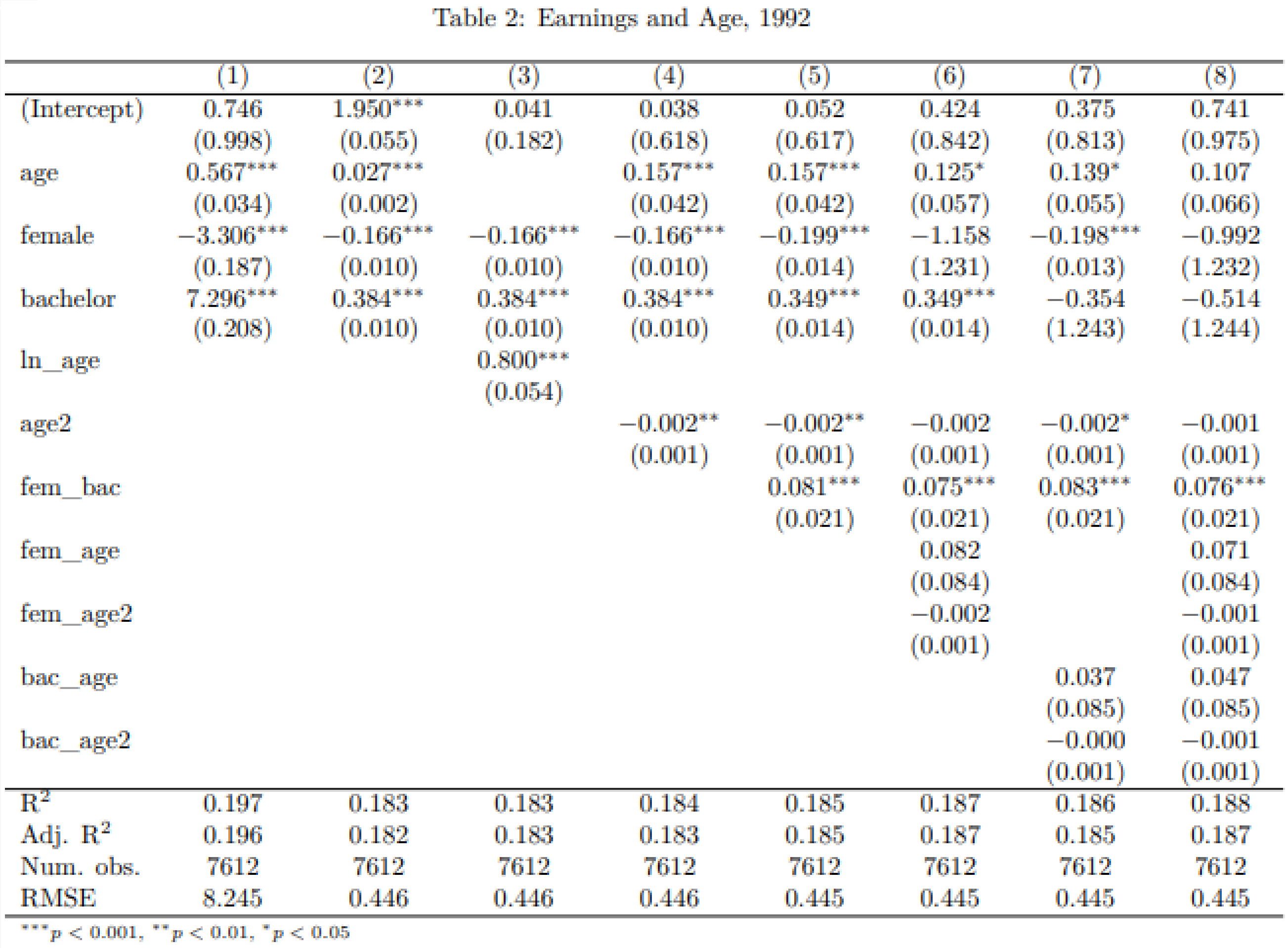
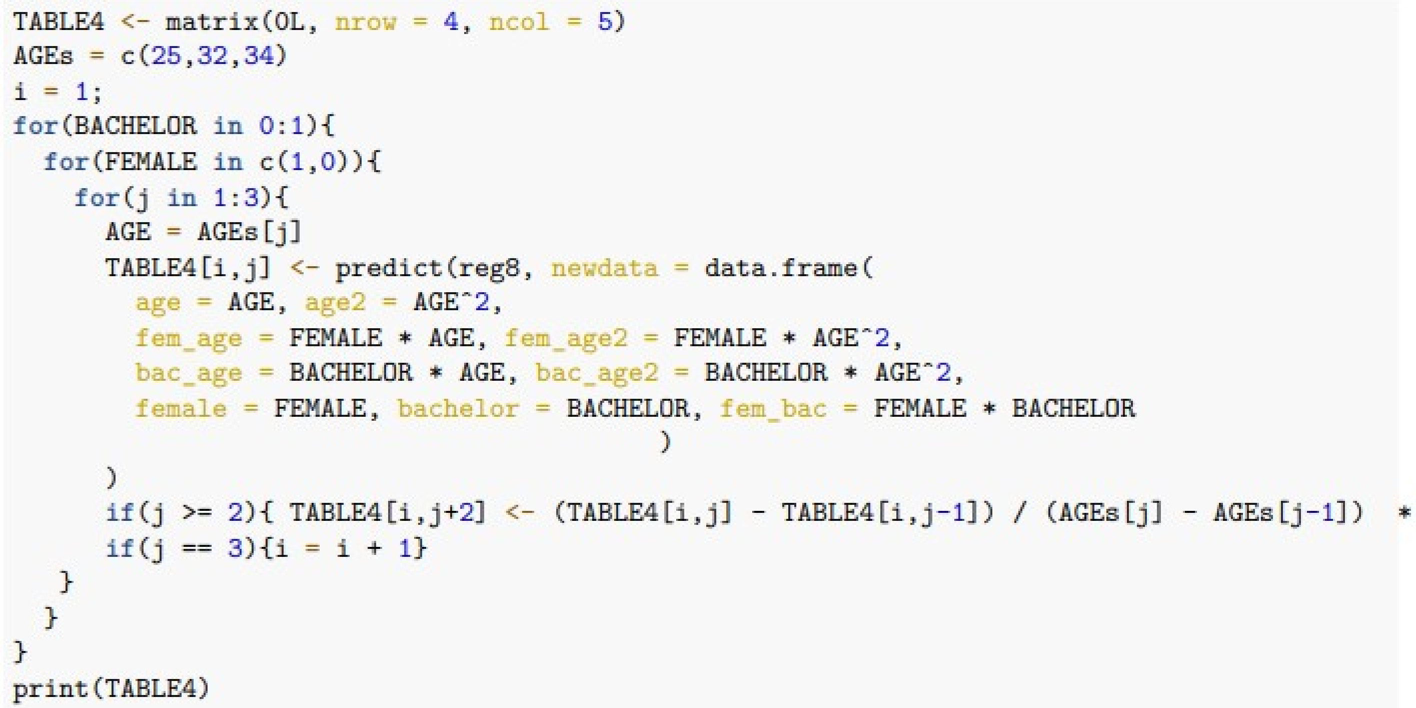


Table 2: Earnings and Age, 1992

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(Intercept)	0.746	1.950***	0.041	0.038	0.052	0.424	0.375	0.741
	(0.998)	(0.055)	(0.182)	(0.618)	(0.617)	(0.842)	(0.813)	(0.975)
age	0.567***	0.027***		0.157***	0.157***	0.125*	0.139*	0.107
	(0.034)	(0.002)		(0.042)	(0.042)	(0.057)	(0.055)	(0.066)
female	−3.306***	−0.166***	−0.166***	−0.166***	−0.199***	−1.158	−0.198***	−0.992
	(0.187)	(0.010)	(0.010)	(0.010)	(0.014)	(1.231)	(0.013)	(1.232)
bachelor	7.296***	0.384***	0.384***	0.384***	0.349***	0.349***	−0.354	−0.514
	(0.208)	(0.010)	(0.010)	(0.010)	(0.014)	(0.014)	(1.243)	(1.244)
ln_age			0.800***					
			(0.054)					
age2				−0.002**	−0.002**	−0.002	−0.002*	−0.001
				(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
fem_bac					0.081***	0.075***	0.083***	0.076***
					(0.021)	(0.021)	(0.021)	(0.021)
fem_age						0.082		0.071
						(0.084)		(0.084)
fem_age2						−0.002		−0.001
						(0.001)		(0.001)
bac_age							0.037	0.047
							(0.085)	(0.085)
bac_age2							−0.000	−0.001
							(0.001)	(0.001)
R^2	0.197	0.183	0.183	0.184	0.185	0.187	0.186	0.188
Adj. R^2	0.196	0.182	0.183	0.183	0.185	0.187	0.185	0.187
Num. obs.	7612	7612	7612	7612	7612	7612	7612	7612
RMSE	8.245	0.446	0.446	0.446	0.445	0.445	0.445	0.445

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b



```
TABLE4 <- matrix(0L, nrow = 4, ncol = 5)
AGEs = c(25,32,34)
i = 1;
for(BACHELOR in 0:1){
  for(FEMALE in c(1,0)){
    for(j in 1:3){
      AGE = AGEs[j]
      TABLE4[i,j] <- predict(reg8, newdata = data.frame(
        age = AGE, age2 = AGE^2,
        fem_age = FEMALE * AGE, fem_age2 = FEMALE * AGE^2,
        bac_age = BACHELOR * AGE, bac_age2 = BACHELOR * AGE^2,
        female = FEMALE, bachelor = BACHELOR, fem_bac = FEMALE * BACHELOR
                                              )
      )
      if(j >= 2){ TABLE4[i,j+2] <- (TABLE4[i,j] - TABLE4[i,j-1]) / (AGEs[j] - AGEs[j-1])  *
      if(j == 3){i = i + 1}
   }
  }
}
print(TABLE4)
```

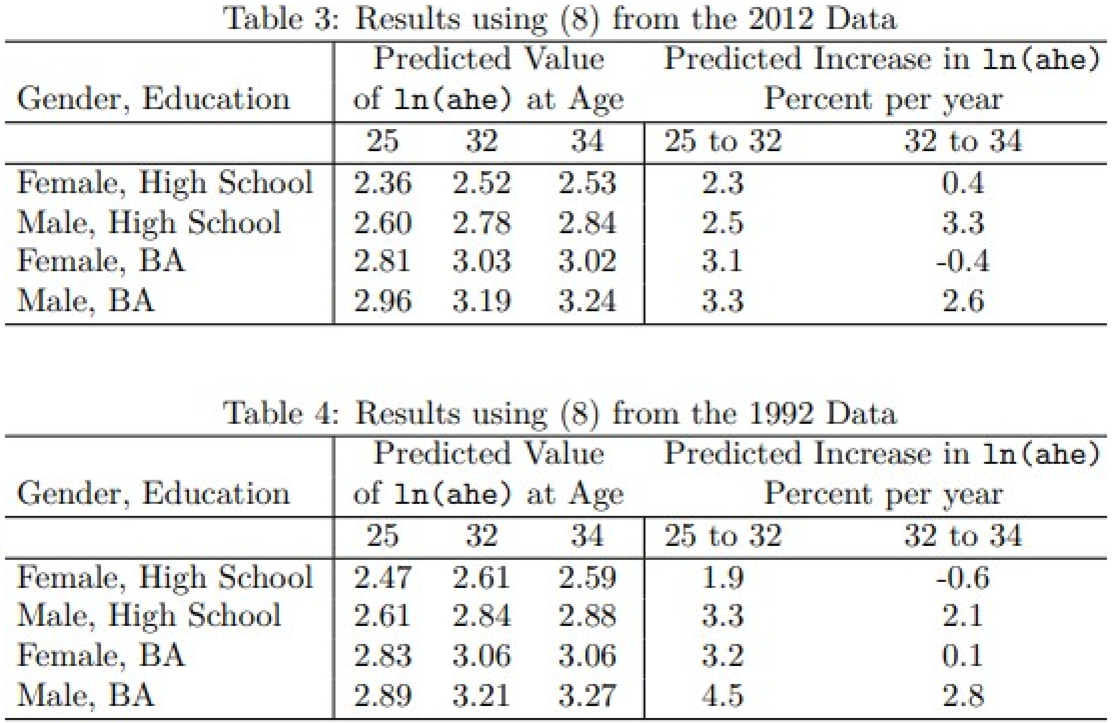
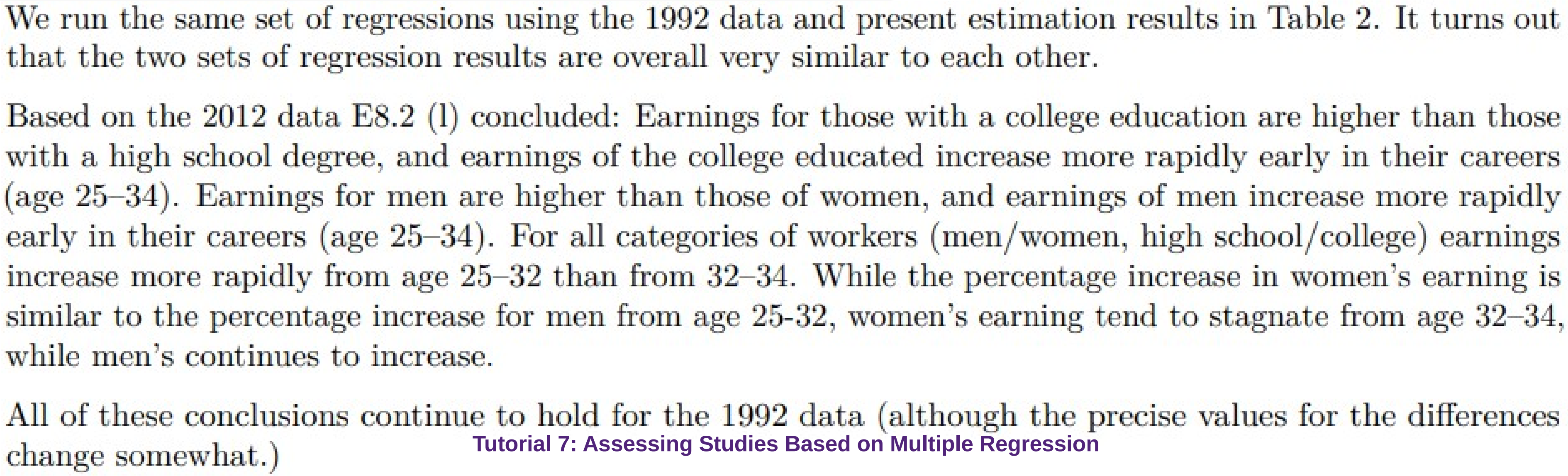


Table 3: Results using (8) from the 2012 Data

Gender, Education	Predicted Value of ln(ahe) at Age			Predicted Increase in ln(ahe) Percent per year	
	25	32	34	25 to 32	32 to 34
Female, High School	2.36	2.52	2.53	2.3	0.4
Male, High School	2.60	2.78	2.84	2.5	3.3
Female, BA	2.81	3.03	3.02	3.1	-0.4
Male, BA	2.96	3.19	3.24	3.3	2.6

Table 4: Results using (8) from the 1992 Data

Gender, Education	Predicted Value of ln(ahe) at Age			Predicted Increase in ln(ahe) Percent per year	
	25	32	34	25 to 32	32 to 34
Female, High School	2.47	2.61	2.59	1.9	-0.6
Male, High School	2.61	2.84	2.88	3.3	2.1
Female, BA	2.83	3.06	3.06	3.2	0.1
Male, BA	2.89	3.21	3.27	4.5	2.8



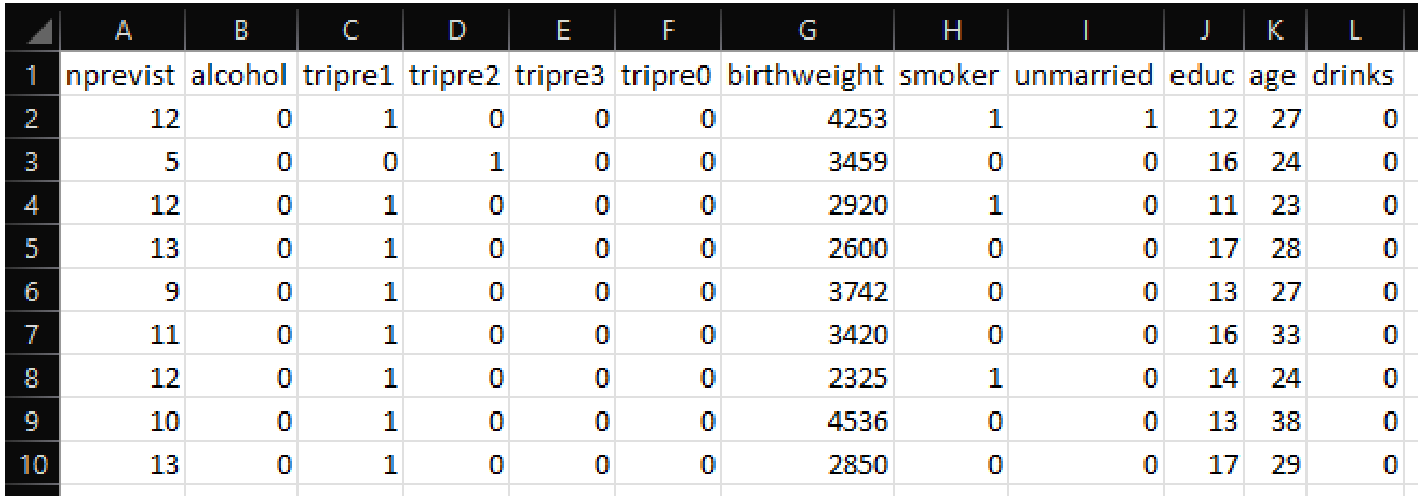
We run the same set of regressions using the 1992 data and present estimation results in Table 2. It turns out that the two sets of regression results are overall very similar to each other.

Based on the 2012 data E8.2 (l) concluded: Earnings for those with a college education are higher than those with a high school degree, and earnings of the college educated increase more rapidly early in their careers (age 25–34). Earnings for men are higher than those of women, and earnings of men increase more rapidly early in their careers (age 25–34). For all categories of workers (men/women, high school/college) earnings increase more rapidly from age 25–32 than from 32–34. While the percentage increase in women's earning is similar to the percentage increase for men from age 25-32, women's earning tend to stagnate from age 32–34, while men's continues to increase.

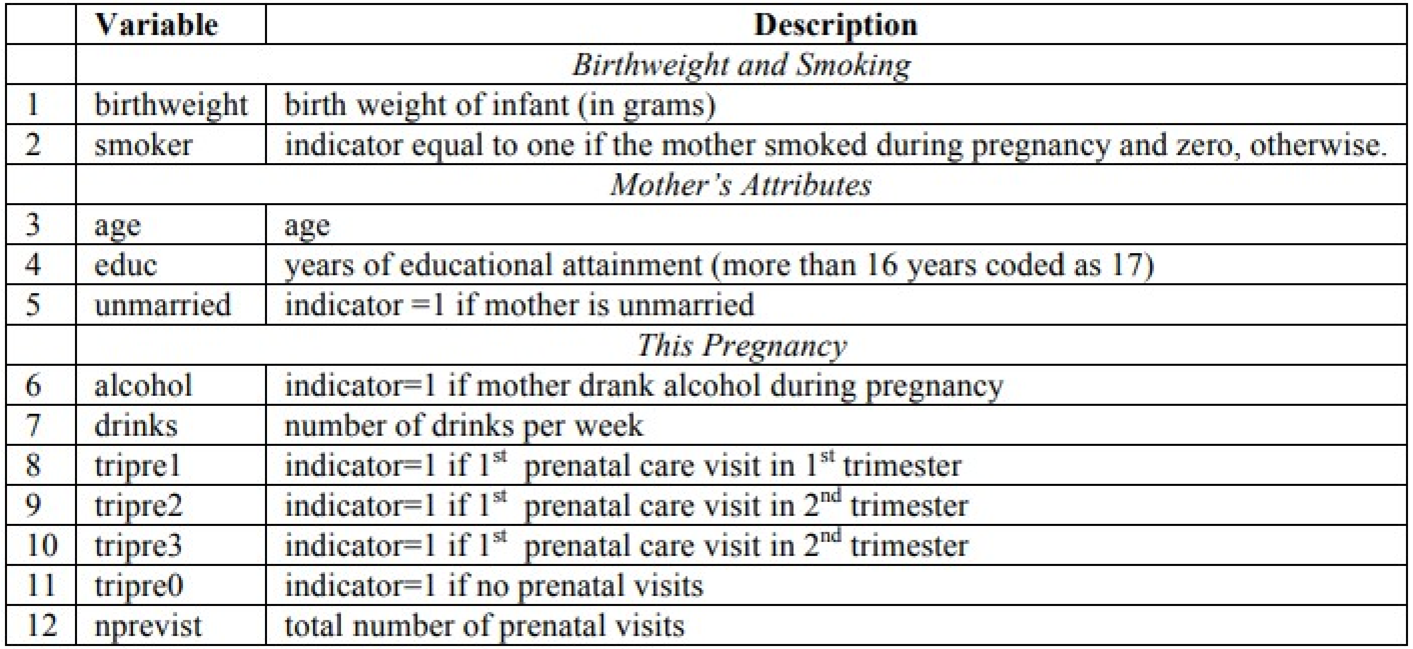
All of these conclusions continue to hold for the 1992 data (although the precise values for the differences change somewhat.)

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b

E9.2 Use the data set birthweight_smoking.csv introduced in [E5.1] to answer the following questions.

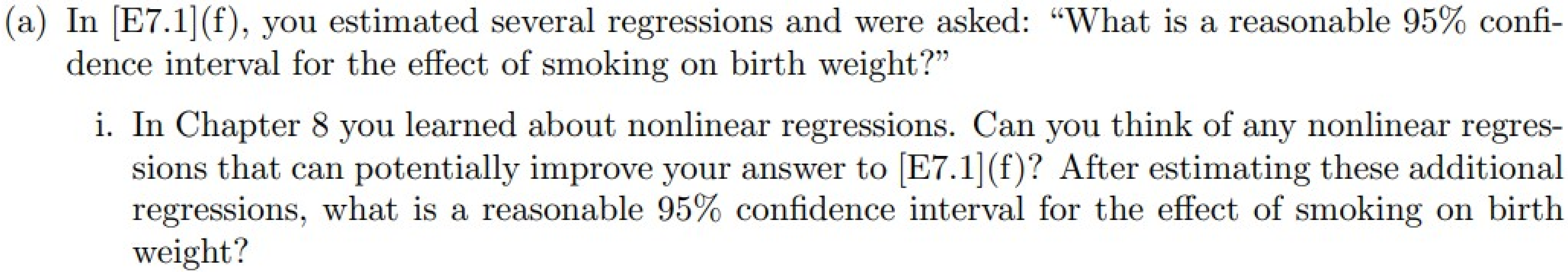


	A	B	C	D	E	F	G	H	I	J	K	L
1	nprevist	alcohol	tripre1	tripre2	tripre3	tripre0	birthweight	smoker	unmarried	educ	age	drinks
2	12	0	1	0	0	0	4253	1	1	12	27	0
3	5	0	0	1	0	0	3459	0	0	16	24	0
4	12	0	1	0	0	0	2920	1	0	11	23	0
5	13	0	1	0	0	0	2600	0	0	17	28	0
6	9	0	1	0	0	0	3742	0	0	13	27	0
7	11	0	1	0	0	0	3420	0	0	16	33	0
8	12	0	1	0	0	0	2325	1	0	14	24	0
9	10	0	1	0	0	0	4536	0	0	13	38	0
10	13	0	1	0	0	0	2850	0	0	17	29	0



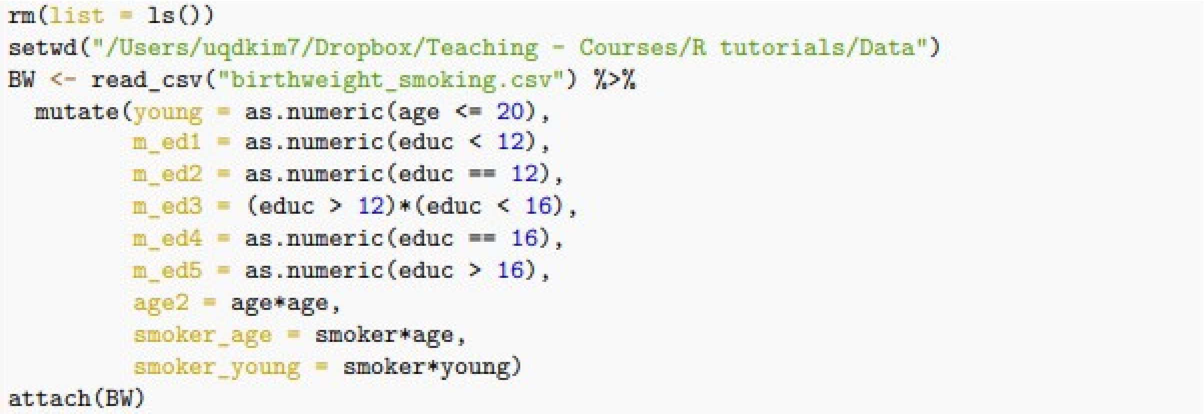
	Variable	Description
		Birthweight and Smoking
1	birthweight	birth weight of infant (in grams)
2	smoker	indicator equal to one if the mother smoked during pregnancy and zero, otherwise.
		Mother's Attributes
3	age	age
4	educ	years of educational attainment (more than 16 years coded as 17)
5	unmarried	indicator =1 if mother is unmarried
		This Pregnancy
6	alcohol	indicator=1 if mother drank alcohol during pregnancy
7	drinks	number of drinks per week
8	tripre1	indicator=1 if 1st prenatal care visit in 1st trimester
9	tripre2	indicator=1 if 1st prenatal care visit in 2nd trimester
10	tripre3	indicator=1 if 1st prenatal care visit in 2nd trimester
11	tripre0	indicator=1 if no prenatal visits
12	nprevist	total number of prenatal visits

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b

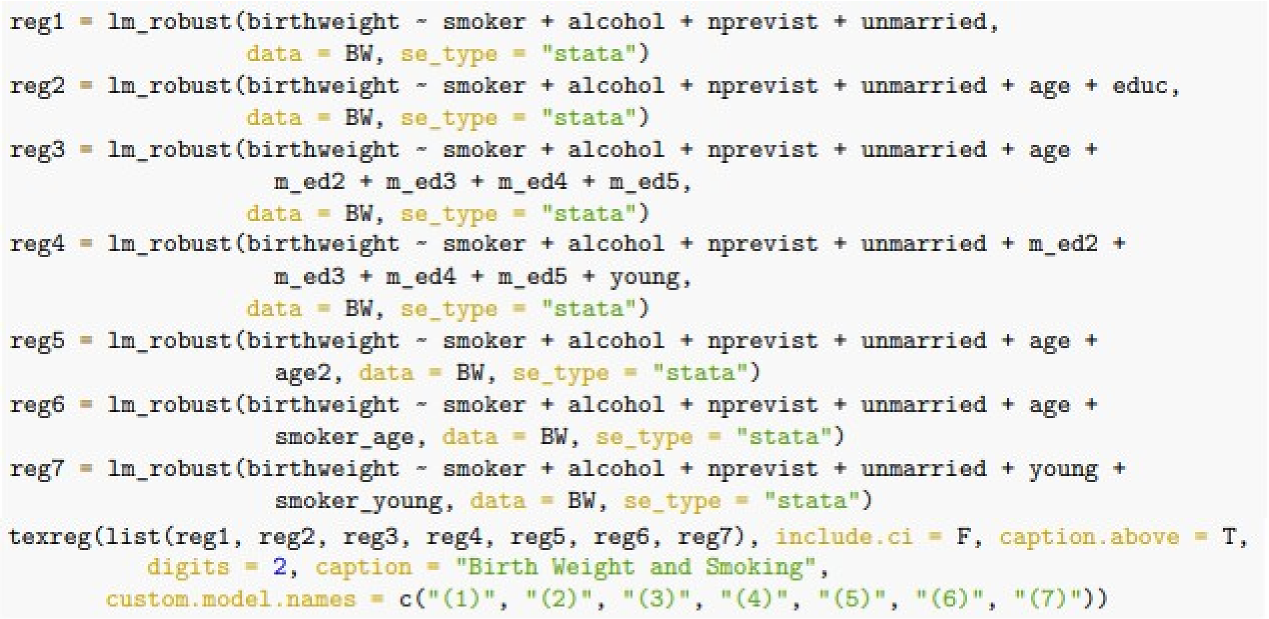


(a) In [E7.1](f), you estimated several regressions and were asked: “What is a reasonable 95% confidence interval for the effect of smoking on birth weight?”

i. In Chapter 8 you learned about nonlinear regressions. Can you think of any nonlinear regressions that can potentially improve your answer to [E7.1](f)? After estimating these additional regressions, what is a reasonable 95% confidence interval for the effect of smoking on birth weight?



```
rm(list = ls())
setwd("/Users/uqdkim7/Dropbox/Teaching - Courses/R tutorials/Data")
BW <- read_csv("birthweight_smoking.csv") %>%
  mutate(young = as.numeric(age <= 20),
         m_ed1 = as.numeric(educ < 12),
         m_ed2 = as.numeric(educ == 12),
         m_ed3 = (educ > 12)*(educ < 16),
         m_ed4 = as.numeric(educ == 16),
         m_ed5 = as.numeric(educ > 16),
         age2 = age*age,
         smoker_age = smoker*age,
         smoker_young = smoker*young)
attach(BW)
```



```
reg1 = lm_robust(birthweight ~ smoker + alcohol + nprevist + unmarried,
                 data = BW, se_type = "stata")
reg2 = lm_robust(birthweight ~ smoker + alcohol + nprevist + unmarried + age + educ,
                 data = BW, se_type = "stata")
reg3 = lm_robust(birthweight ~ smoker + alcohol + nprevist + unmarried + age +
                   m_ed2 + m_ed3 + m_ed4 + m_ed5,
                 data = BW, se_type = "stata")
reg4 = lm_robust(birthweight ~ smoker + alcohol + nprevist + unmarried + m_ed2 +
                   m_ed3 + m_ed4 + m_ed5 + young,
                 data = BW, se_type = "stata")
reg5 = lm_robust(birthweight ~ smoker + alcohol + nprevist + unmarried + age +
                   age2, data = BW, se_type = "stata")
reg6 = lm_robust(birthweight ~ smoker + alcohol + nprevist + unmarried + age +
                   smoker_age, data = BW, se_type = "stata")
reg7 = lm_robust(birthweight ~ smoker + alcohol + nprevist + unmarried + young +
                   smoker_young, data = BW, se_type = "stata")
texreg(list(reg1, reg2, reg3, reg4, reg5, reg6, reg7), include.ci = F, caption.above = T,
          digits = 2, caption = "Birth Weight and Smoking",
       custom.model.names = c("(1)", "(2)", "(3)", "(4)", "(5)", "(6)", "(7)"))
```

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b

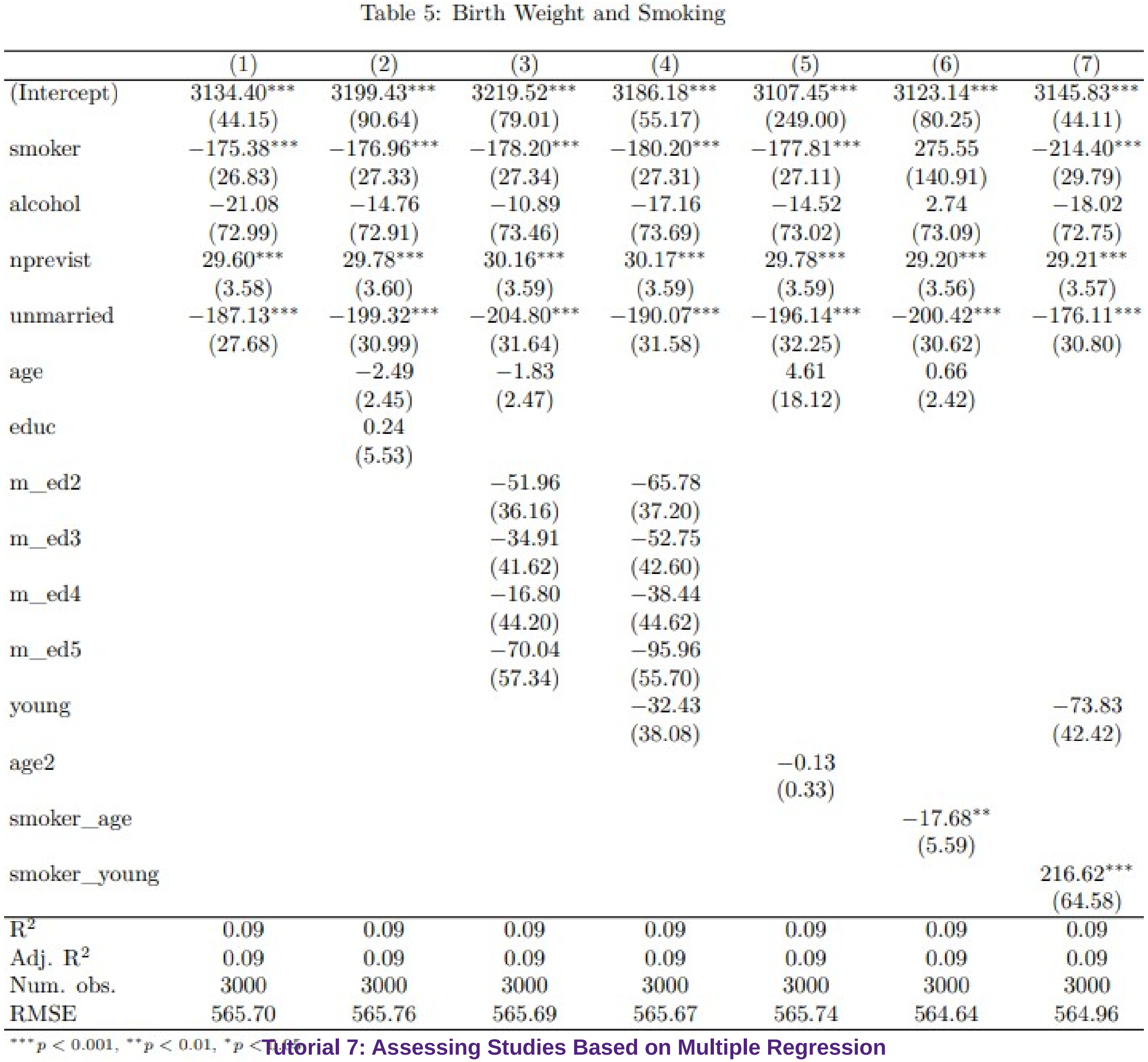
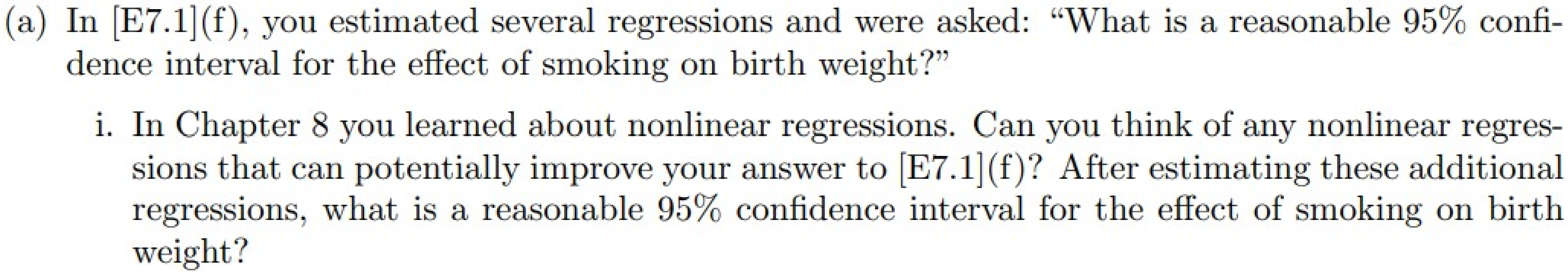


Table 5: Birth Weight and Smoking

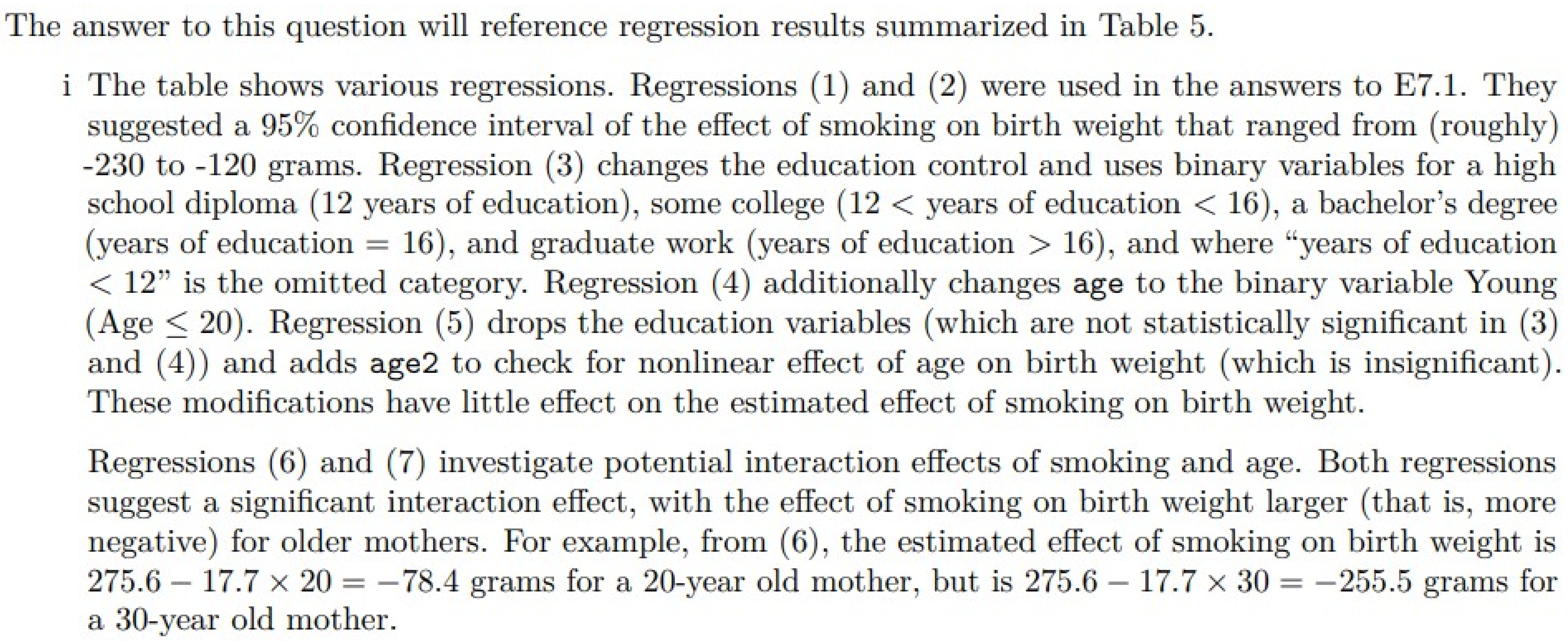
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(Intercept)	3134.40***	3199.43***	3219.52***	3186.18***	3107.45***	3123.14***	3145.83***
	(44.15)	(90.64)	(79.01)	(55.17)	(249.00)	(80.25)	(44.11)
smoker	−175.38***	−176.96***	−178.20***	−180.20***	−177.81***	275.55	−214.40***
	(26.83)	(27.33)	(27.34)	(27.31)	(27.11)	(140.91)	(29.79)
alcohol	−21.08	−14.76	−10.89	−17.16	−14.52	2.74	−18.02
	(72.99)	(72.91)	(73.46)	(73.69)	(73.02)	(73.09)	(72.75)
nprevist	29.60***	29.78***	30.16***	30.17***	29.78***	29.20***	29.21***
	(3.58)	(3.60)	(3.59)	(3.59)	(3.59)	(3.56)	(3.57)
unmarried	−187.13***	−199.32***	−204.80***	−190.07***	−196.14***	−200.42***	−176.11***
	(27.68)	(30.99)	(31.64)	(31.58)	(32.25)	(30.62)	(30.80)
age		−2.49	−1.83		4.61	0.66	
		(2.45)	(2.47)		(18.12)	(2.42)	
educ		0.24					
		(5.53)					
m_ed2			−51.96	−65.78			
			(36.16)	(37.20)			
m_ed3			−34.91	−52.75			
			(41.62)	(42.60)			
m_ed4			−16.80	−38.44			
			(44.20)	(44.62)			
m_ed5			−70.04	−95.96			
			(57.34)	(55.70)			
young				−32.43			−73.83
				(38.08)			(42.42)
age2					−0.13		
					(0.33)		
smoker_age						−17.68**	
						(5.59)	
smoker_young							216.62***
							(64.58)
R^2	0.09	0.09	0.09	0.09	0.09	0.09	0.09
Adj. R^2	0.09	0.09	0.09	0.09	0.09	0.09	0.09
Num. obs.	3000	3000	3000	3000	3000	3000	3000
RMSE	565.70	565.76	565.69	565.67	565.74	564.64	564.96

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$



(a) In [E7.1](f), you estimated several regressions and were asked: "What is a reasonable 95% confidence interval for the effect of smoking on birth weight?"

i. In Chapter 8 you learned about nonlinear regressions. Can you think of any nonlinear regressions that can potentially improve your answer to [E7.1](f)? After estimating these additional regressions, what is a reasonable 95% confidence interval for the effect of smoking on birth weight?

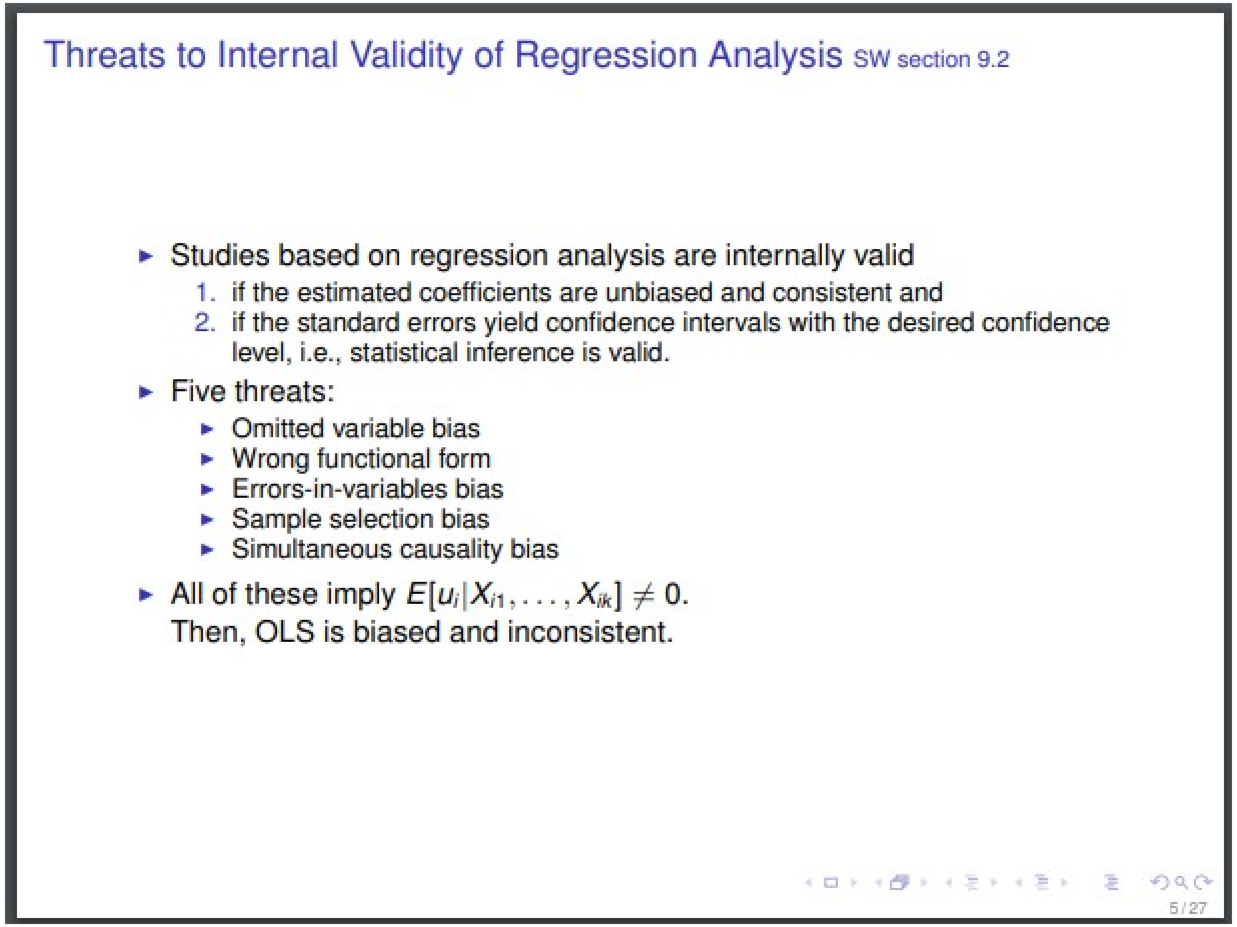


The answer to this question will reference regression results summarized in Table 5.

i The table shows various regressions. Regressions (1) and (2) were used in the answers to E7.1. They suggested a 95% confidence interval of the effect of smoking on birth weight 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 `age2` to check for nonlinear effect of age on birth weight (which is insignificant). These modifications have little effect on the estimated effect of smoking on birth weight.

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 birth weight larger (that is, more negative) for older mothers. For example, from (6), the estimated effect of smoking on birth weight is $275.6 - 17.7 \times 20 = -78.4$ grams for a 20-year old mother, but is $275.6 - 17.7 \times 30 = -255.5$ grams for a 30-year old mother.

ii. Discuss the internal validity of the regressions you used to construct the confidence interval. Include a discussion of possible omitted variable bias, misspecification of the functional form of the regression, errors in variables, sample selection, simultaneous causality, and inconsistency of the OLS standard errors.

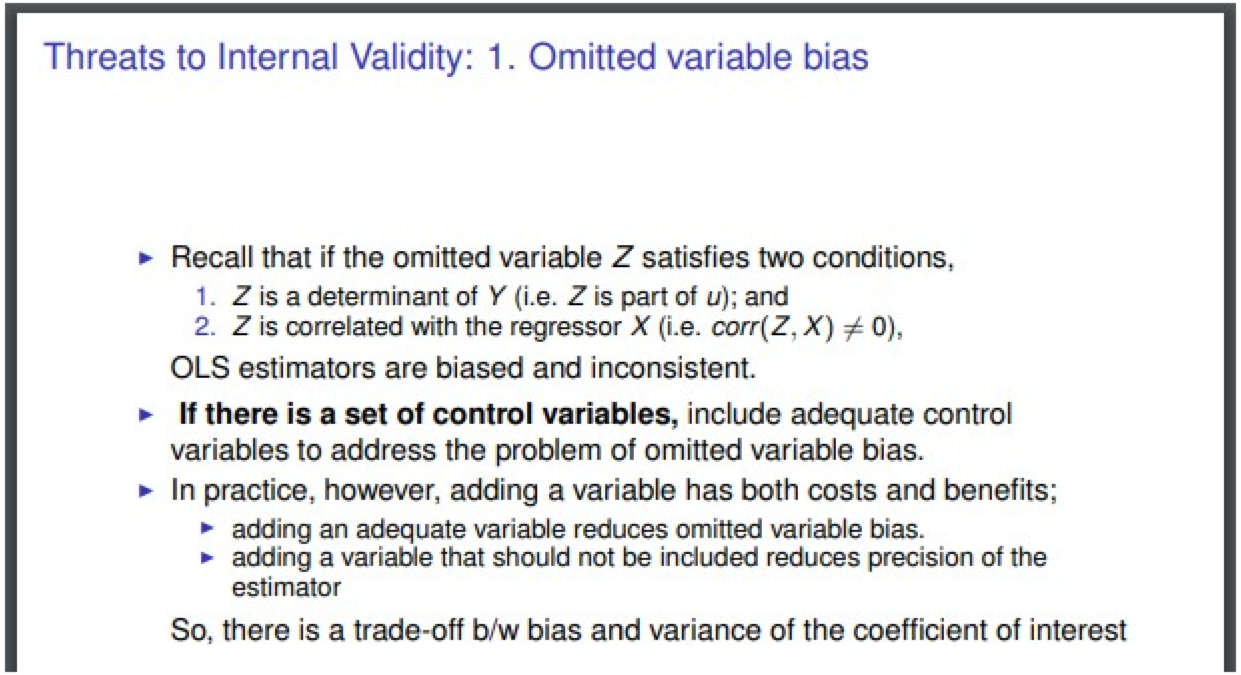


Threats to Internal Validity of Regression Analysis SW section 9.2

- Studies based on regression analysis are internally valid
  1. if the estimated coefficients are unbiased and consistent and
  2. if the standard errors yield confidence intervals with the desired confidence level, i.e., statistical inference is valid.
- Five threats:
  - Omitted variable bias
  - Wrong functional form
  - Errors-in-variables bias
  - Sample selection bias
  - Simultaneous causality bias
- All of these imply $E[u_i|X_{i1}, \dots, X_{ik}] \neq 0$.
  Then, OLS is biased and inconsistent.

5 / 27

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b



Threats to Internal Validity: 1. Omitted variable bias

- Recall that if the omitted variable Z satisfies two conditions,
  1. Z is a determinant of Y (i.e. Z is part of u); and
  2. Z is correlated with the regressor X (i.e. $corr(Z, X) \neq 0$),

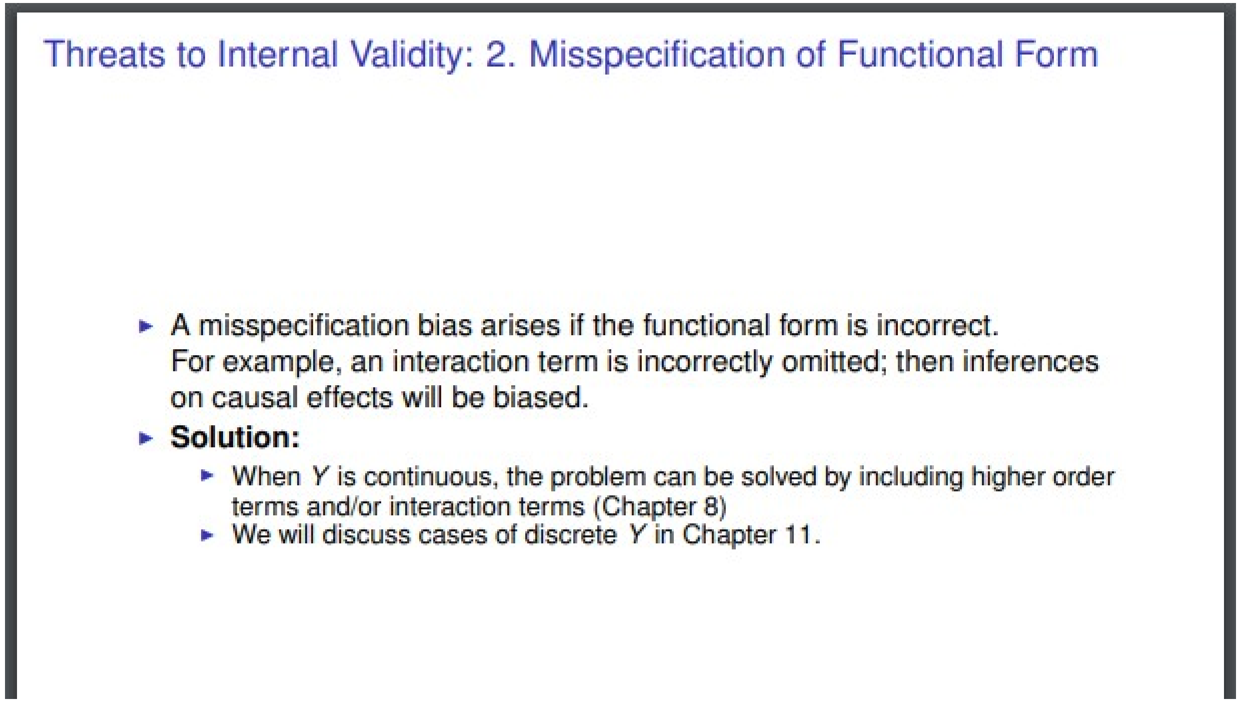
  OLS estimators are biased and inconsistent.
- **If there is a set of control variables,** include adequate control variables to address the problem of omitted variable bias.
- In practice, however, adding a variable has both costs and benefits;
  - adding an adequate variable reduces omitted variable bias.
  - adding a variable that should not be included reduces precision of the estimator

  So, there is a trade-off b/w bias and variance of the coefficient of interest

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 birth weight 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 birth weight 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 birth weight 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.

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b

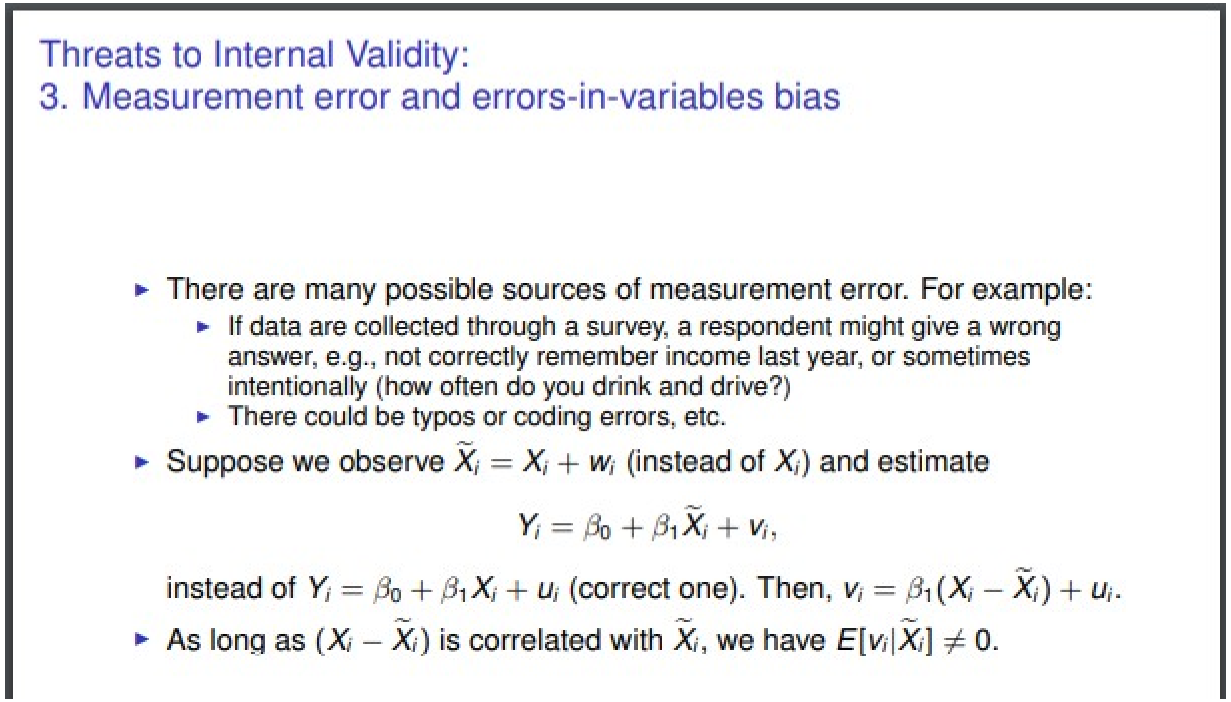


Threats to Internal Validity: 2. Misspecification of Functional Form

- A misspecification bias arises if the functional form is incorrect. For example, an interaction term is incorrectly omitted; then inferences on causal effects will be biased.
- **Solution:**
  - When Y is continuous, the problem can be solved by including higher order terms and/or interaction terms (Chapter 8)
  - We will discuss cases of discrete Y in Chapter 11.

- 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.

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b



Threats to Internal Validity:
3. Measurement error and errors-in-variables bias

- There are many possible sources of measurement error. For example:
  - If data are collected through a survey, a respondent might give a wrong answer, e.g., not correctly remember income last year, or sometimes intentionally (how often do you drink and drive?)
  - There could be typos or coding errors, etc.
- Suppose we observe $\widetilde{X}_i = X_i + w_i$ (instead of X_i) and estimate

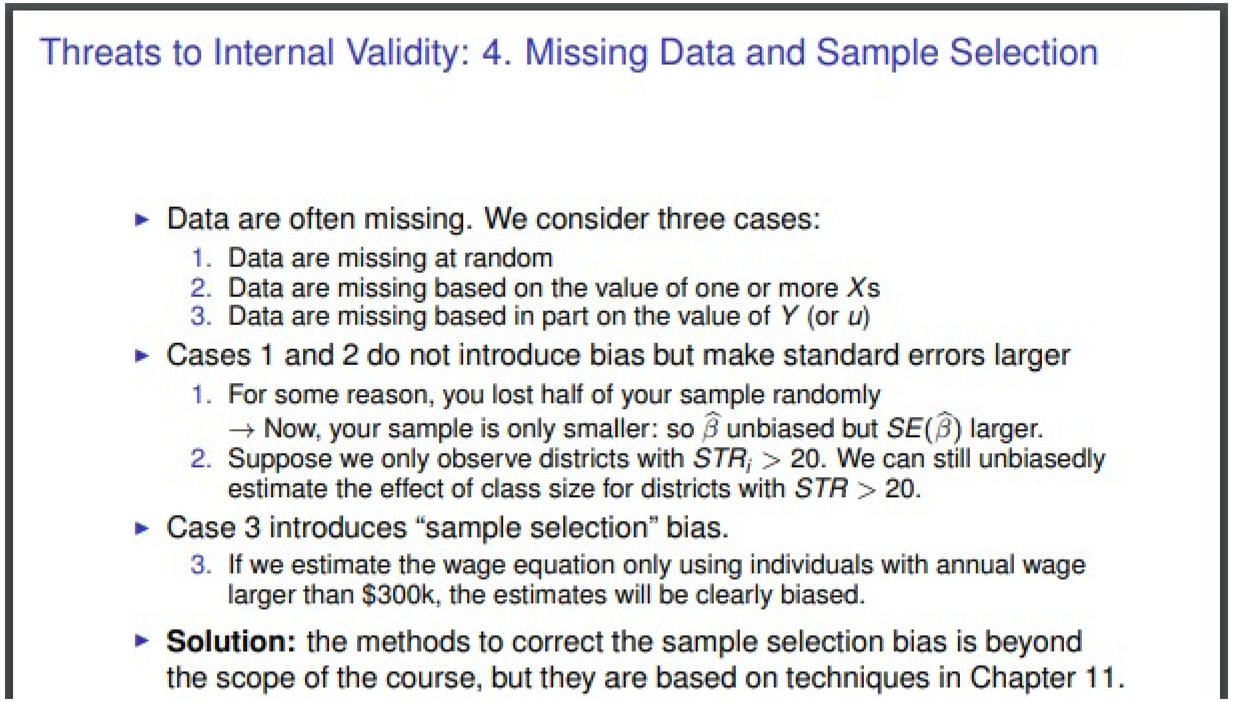
$$Y_i = \beta_0 + \beta_1 \widetilde{X}_i + v_i,$$

instead of $Y_i = \beta_0 + \beta_1 X_i + u_i$ (correct one). Then, $v_i = \beta_1(X_i - \widetilde{X}_i) + u_i$.

- As long as $(X_i - \widetilde{X}_i)$ is correlated with $\widetilde{X}_i$, we have $E[v_i|\widetilde{X}_i] \neq 0$.

- 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.

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b

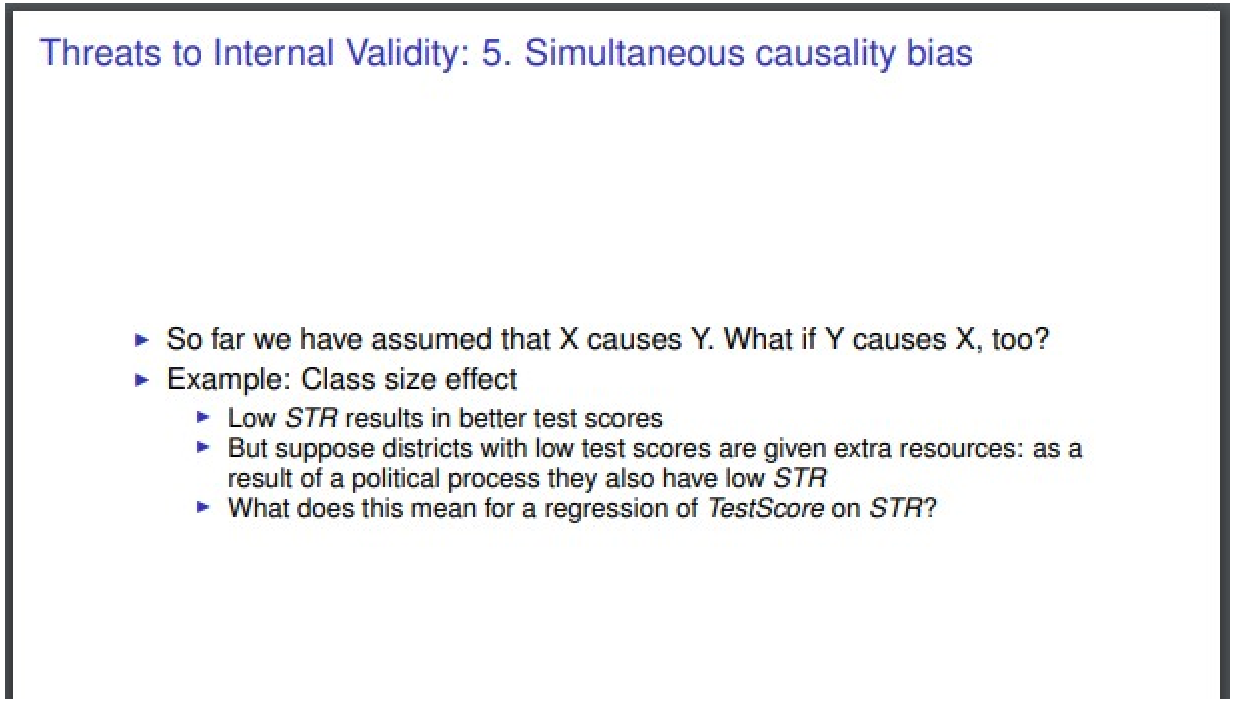


Threats to Internal Validity: 4. Missing Data and Sample Selection

- Data are often missing. We consider three cases:
  1. Data are missing at random
  2. Data are missing based on the value of one or more Xs
  3. Data are missing based in part on the value of Y (or u)
- Cases 1 and 2 do not introduce bias but make standard errors larger
  1. For some reason, you lost half of your sample randomly
     $\rightarrow$ Now, your sample is only smaller: so $\widehat{\beta}$ unbiased but $SE(\widehat{\beta})$ larger.
  2. Suppose we only observe districts with $STR_i > 20$. We can still unbiasedly estimate the effect of class size for districts with $STR > 20$.
- Case 3 introduces "sample selection" bias.
  3. If we estimate the wage equation only using individuals with annual wage larger than \$300k, the estimates will be clearly biased.
- **Solution:** the methods to correct the sample selection bias is beyond the scope of the course, but they are based on techniques in Chapter 11.

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

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b

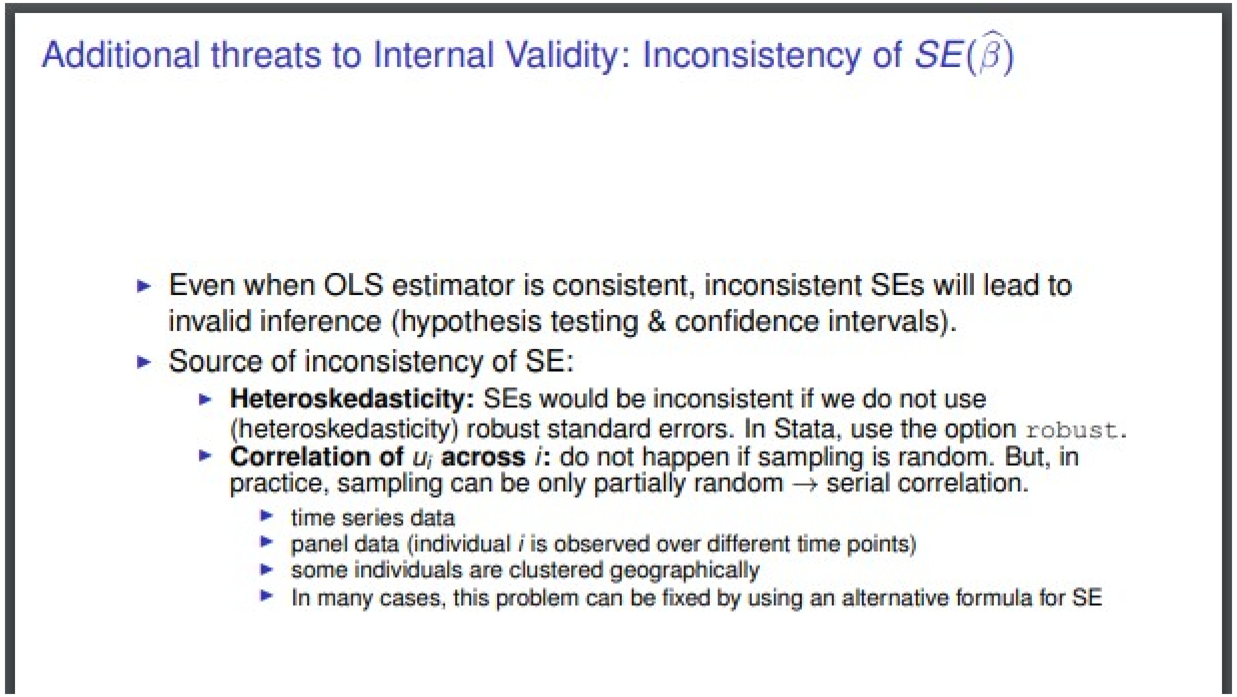


Threats to Internal Validity: 5. Simultaneous causality bias

- So far we have assumed that X causes Y. What if Y causes X, too?
- Example: Class size effect
  - Low *STR* results in better test scores
  - But suppose districts with low test scores are given extra resources: as a result of a political process they also have low *STR*
  - What does this mean for a regression of *TestScore* on *STR*?

- Simultaneous causality: This is a problem to the extent that women who are more likely to have low-birth weight 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.

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b

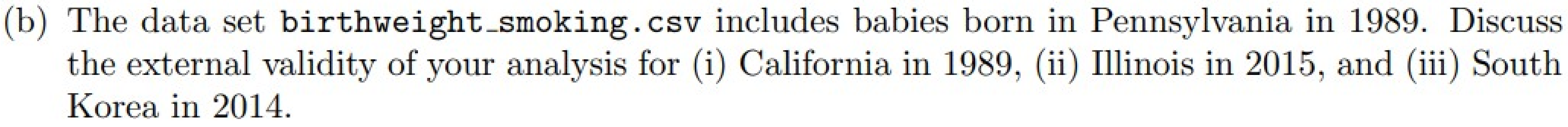


Additional threats to Internal Validity: Inconsistency of $SE(\widehat{\beta})$

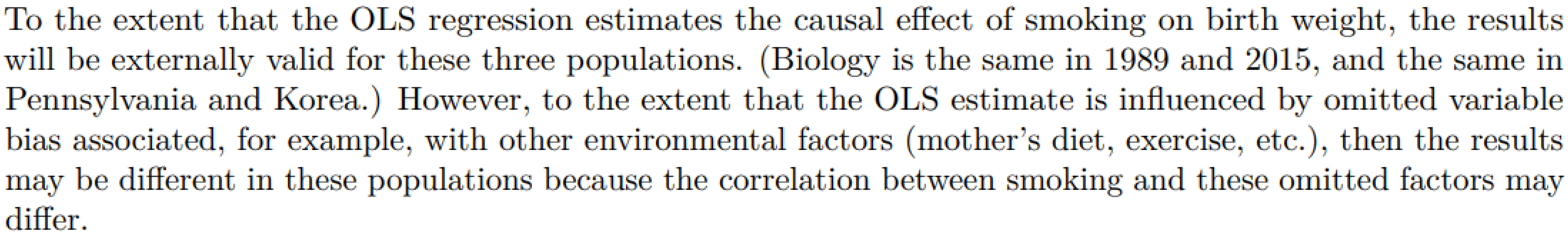
- Even when OLS estimator is consistent, inconsistent SEs will lead to invalid inference (hypothesis testing & confidence intervals).
- Source of inconsistency of SE:
  - **Heteroskedasticity:** SEs would be inconsistent if we do not use (heteroskedasticity) robust standard errors. In Stata, use the option `robust`.
  - **Correlation of u_i across i:** do not happen if sampling is random. But, in practice, sampling can be only partially random → serial correlation.
    - time series data
    - panel data (individual i is observed over different time points)
    - some individuals are clustered geographically
    - In many cases, this problem can be fixed by using an alternative formula for SE

- 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.

Adm · Tut 07 · E9.1 · a · b · E9.2 · a · b



(b) The data set birthweight_smoking.csv includes babies born in Pennsylvania in 1989. Discuss the external validity of your analysis for (i) California in 1989, (ii) Illinois in 2015, and (iii) South Korea in 2014.



To the extent that the OLS regression estimates the causal effect of smoking on birth weight, 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.

The University of Queensland
AUSTRALIA

CREATE CHANGE

Thank you

Francisco Tavares Garcia
Academic Tutor | School of Economics
tavaresgarcia.github.io

Reference

Stock, J. H., & Watson, M. W. (2019). Introduction to Econometrics, Global Edition, 4th edition. Pearson Education Limited.