Natchaya Laokham
6104641425

1. Using the data in SLEEP75 (see also Problem 3 in Chapter 3), we obtain the estimated equation

$$\begin{aligned}\widehat{sleep} &= 3{,}840.83 - .163\,totwrk - 11.71\,educ - 8.70\,age \\ &\quad (235.11)\ (.018)\qquad (5.86)\qquad (11.21) \\ &\quad + .128\,age^2 + 87.75\,male \\ &\quad (.134)\qquad (34.33) \\ n &= 706,\ R^2 = .123,\ \overline{R}^2 = .117.\end{aligned}$$

The variable *sleep* is total minutes per week spent sleeping at night, *totwrk* is total weekly minutes spent working, *educ* and *age* are measured in years, and *male* is a gender dummy.

i. All other factors being equal, is there evidence that men sleep more than women? How strong is the evidence?

The coefficient on male is 87.75, so a man is estimated to sleep almost one and one-half hours more per week than a comparable woman. Further, $t_{male} = \frac{87.75}{34.33} \approx 2.56$, which is close to the 1% critical value against a two-sided alternative (about 2.58). Thus, the evidence for a gender differential is fairly strong.

ii. Is there a statistically significant tradeoff between working and sleeping? What is the estimated tradeoff?

The t statistic on totwrk is $-\frac{.163}{.018} \approx -9.06$, which is very statistically significant. The coefficient implies that one more hour of work (60 mins) is associated with $.163(60) \approx 9.8$ minutes less sleep.

iii. What other regression do you need to run to test the null hypothesis that, holding other factors fixed, age has no effect on sleeping?

$$F \equiv \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n-k-1)}$$

8. Suppose you collect data from a survey on wages, education, experience, and gender. In addition, you ask for information about marijuana usage. The original question is: "On how many separate occasions last month did you smoke marijuana?"

   i. Write an equation that would allow you to estimate the effects of marijuana usage on wage, while controlling for other factors. You should be able to make statements such as, "Smoking marijuana five more times per month is estimated to change wage by *x*%."

- To be able to interpret the variables in that way, we need to build a log-linear model. The regression equation would look like that:

$$\log(wage) = \beta_0 + \beta_1 marijuna\_wage + \beta_2 education + \beta_3 experience + \delta_1 gender + u$$

   ii. Write a model that would allow you to test whether drug usage has different effects on wages for men and women. How would you test that there are no differences in the effects of drug usage for men and women?

- We need to add an interaction variable between the gender and the marijuana variables. The new regression equation would look like that:

$$\log(wage) = \beta_0 + \beta_1 marijuna\_usage + \beta_2 education + \beta_3 experience + \delta_1 gender + \delta_2 gender.marijuna\_usage + u$$

- To test whether there are differences in the effects of drug usage for men and women, we could test the following hypothesis with a t-test:

$$H_0: \delta_2 = 0 \qquad H_1: \delta_2 \neq 0$$

- To perform the t-test, we would first need to calculate the t-statistic with the following formula:

$$t = \frac{\overline{gender \times marijuna} - 0}{/\sqrt{n}}$$

- We would then look for the critical value based on $(1-\alpha/2)$ percentile in the t distribution with $n-1$ degree of freedom. If the absolute value of the t-statistic is greater than the critical value, we would them reject H_0.

iii. Suppose you think it is better to measure marijuana usage by putting people into one of four categories: nonuser, light user (1 to 5 times per month), moderate user (6 to 10 times per month), and heavy user (more than 10 times per month). Now, write a model that allows you to estimate the effects of marijuana usage on wage.

- Incorporating this change into the model in q.i, we would have:

$$\log(wage) = \beta_0 + \beta_2 education + \beta_3 experience + \delta_1 gender + \delta_2 light\_user + \delta_3 moderate\_user + \delta_4 heavy\_user + u$$

- It is now easy to estimate each of the coefficients by running the regression normally.

iv. Using the model in part (iii), explain in detail how to test the null hypothesis that marijuana usage has no effect on wage. Be very specific and include a careful listing of degrees of freedom.

- We would need to test the following (i.e. we want to test whether delta 2, delta 3 and delta 4 are together jointly significant), using a F-test:

$$H_0: \delta_2 = 0 \text{ and } \delta_3 = 0 \text{ and } \delta_4 = 0$$

$$H_1: H_0 \text{ is false}$$

- Let's call the model in (iii) the "unrestricted model". The "restricted model" would then be:

$$\log(wage) = \beta_0 + \beta_2 education + \beta_3 experience + \delta_1 gender + u$$

- We then calculate the F-statistic, using the following formula:

$$\frac{SSR_{restricted} - \frac{SSR_{unrestricted}}{q_f}}{SSR_{unrestricted}/(n-k-1)}$$

where q_f = member of restrictions = 3 (because we test three parameter),
k = number of variables in the unrestricted model = 6

- We would then reject H_0 if the F-statistic is higher than the critical value (based on the Fisher distribution at $d_1 = q_f$, $d_2 = n-k-1$).

v. What are some potential problems with drawing causal inference using the survey data that you collected?

The survey data might have multiple problems that would make it non presentative of the population. One of the biggest issues is self-selection and social desirability bias. In the case of this study, we could expect for example individuals to voluntarily (or unconsciously) report lower values than their actual marijuna consumption, by fear of looking like an addict/junkie (social desirability). Other issues might be linked to the way the data has been collected. For example, if the survey has been conducted in a particular area or at a particular time of the day, the respondants might not be a truly random sample of the population; this will be the case for example if the survey is conducted by phone during the day, at times when the active population is at work (which would result in in a overrepresentation of unemployed people, housewives, retired people, etc.) There are of course many other response bias that could make the data inaccurate, such as the acquiescence bias.

11. The following equations were estimated using the data in ECONMATH, with standard errors reported under coefficients. The average class score, measured as a percentage, is about 72.2; exactly 50% of the students are male; and the average of *colgpa* (grade point average at the start of the term) is about 2.81.

$$\widehat{score} = 32.31 + 14.32\, colgpa$$
$$(2.00) \quad (0.70)$$
$$n = 856, R^2 = .329, \overline{R}^2 = .328.$$

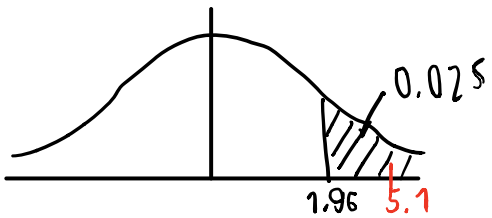
$$\widehat{score} = 29.66 + 3.83\, male + 14.57\, colgpa$$
$$(2.04) \quad (0.74) \quad (0.69)$$
$$n = 856, R^2 = .349, \overline{R}^2 = .348.$$

$$\widehat{score} = 30.36 + 2.47\, male + 14.33\, colgpa + 0.479\, male \cdot colgpa$$
$$(2.86) \quad (3.96) \quad (0.98) \quad (1.383)$$
$$n = 856, R^2 = .349, \overline{R}^2 = .347.$$

$$\widehat{score} = 30.36 + 3.82\, male + 14.33\, colgpa + 0.479\, male \cdot (colgpa - 2.81)$$
$$(2.86) \quad (0.74) \quad (0.98) \quad (1.383)$$
$$n = 856, R^2 = .349, \overline{R}^2 = .347.$$

i. Interpret the coefficient on *male* in the second equation and construct a 95% confidence interval for β_{male}. Does the confidence interval exclude zero? when male increase by 1 people, the score will increase by 3.83 percent.

$$t_{male} = \frac{\hat{\beta}_{male} - \beta_{male}}{s.e.\hat{\beta}_{male}}$$
$$= \frac{3.83 - 0}{0.74}$$
$$= 5.18$$



95% confidence interval = 5% level of significance

one side = 0.025

$= 0.5 - 0.025$

$= 0.475 + 1.96$

ii. In the second equation, how come the estimate on *male* is so imprecise? Should we now conclude that there are no gender differences in *score* after controlling for *colgpa*? [*Hint*: You might want to compute an *F* statistic for the null hypothesis that there is no gender difference in the model with the interaction.]

iii. Compared with the third equation, how come the coefficient on *male* in the last equation is so much closer to that in the second equation and just as precisely estimated?