Homework #5

answers

Problem 1 Alice is interested in how salaries for similar work differ across the public and private sectors. She obtains employment data on 30,000 workers randomly sampled from the entire country.

She uses the following regression model, where $SALARY$ is annual salary in dollars, and $PUBLIC$ equals 1 if an individual is employed in the public sector, and 0 otherwise:

$$SALARY = \beta_0 + \beta_1 * PUBLIC + \epsilon \tag{1}$$

She obtains the following results:

R^2 .091

coefficient	estimate	p-value
Intercept	45,000	.00012
β_1	12,000	.03549

a. Interpret Alice's estimate of β_1.

β_1 is the amount by which public sector salaries differ from private sector salaries.

b. Alice suggests that her results prove that public sector wages are excessive, and that governments could lower the salaries they offer and attract the same workers. Do you agree? Explain.

The problem with this argument is that there is a possibility that worker characteristics differ between the public and private sectors, and that this is responsible for some of the salary gap. So, to study whether or not public sector employees make more or less than their private sector counterparts, a good strategy would be to do multiple regression analysis, controlling for worker characteristics such as education.

Bob thinks that public sector employees may have different individual characteristics than private sector employees. Using Alice's data, Bob estimates the following regression:

$$SALARY = \beta_0 + \beta_1 * PUBLIC + \beta_2 * EDUCATION + \beta_3 * AGE + \beta_4 * AGE^2 + \epsilon \tag{2}$$

where $EDUCATION$ is years of education and AGE is age of the employee. He obtains the following results:

R^2 .63

coefficient	estimate	p-value
β_0	25,000	.00005
β_1	4,000	.00312
β_2	1,200	.01579
β_3	300	.02697
β_4	−3	.00049

c. Bob's estimate of β_1 is lower than Alice's. Explain in simple, intuitive terms why this might be the case.

Following part (b), Bob's regression controls for education and age, whereas Alice's does not. The correct interpretation of β_1 in Bob's regression is therefore the amount by which salary is higher for a public sector

worker, controlling for education level. In other words, for two workers with the same education, one in the private sector and one in the public sector, we expect the public sector worker to earn about $4,000 more. As Alice's regression does not control for education, the interpretation of her β_1 is the amount by which an average public sector worker outearns an average private sector employee. The fact that Bob's estimate of β_1 is smaller than Alice's suggests that the average public sector worker has more education than the average private sector employee.

d. Based on a comparison of Bob's and Alice's results, do you think that public sector employees are more educated, or less educated than private sector employees, on average?

See part c.

e. Do Bob's results predict that a 40-year old public-sector employee or a 60-year old public sector employee will make more, all else equal?

Bob predicts that a 40-year old worker will earn the same salary as a 60-year old worker, on average.

f. At approximately what age do Bob's results predict that SALARY will peak?

Bob's results predict that salary is related to age via the function $SALARY = A + 300 * AGE - 3AGE^2$, where $A$is a constant. This function has a peak at $AGE = 50$ (to see this, take the derivative with respect to age, and set to zero, or experiment numerically in Excel.

Problem 2 (use dataset 18-17) The manager of an amusement park would like to predict daily attendance in order to develop more accurate plans about how much food to order and how many ride operators to hire. After some consideration, he decided that the following three factors are critical:

- Yesterday's attendance
- Weekday or weekend
- Predicted weather

He then took a random sample of 40 days. For each day, he recorded the attendance, the previous day's attendance, day of the week, and weather forecast. The first independent variable is interval, but the other two are nominal. Accordingly, he created the following sets of indicator variables:

$$I_1 = \begin{cases} 1 & \text{(if weekend)} \\ 0 & \text{(if not)} \end{cases}$$

$$I_2 = \begin{cases} 1 & \text{(if mostly sunny is predicted)} \\ 0 & \text{(if not)} \end{cases}$$

$$I_3 = \begin{cases} 1 & \text{(if rain is predicted)} \\ 0 & \text{(if not)} \end{cases} \tag{3}$$

a. Conduct a regression analysis in which you regress attendance on yesterday's attendance, weekend/weekday status, and weather. Which coefficients are significant?

Including both weather indicators generate p-values of .000036, .0023, .071, and .123 for yesterday's attendance, weekend, sunny, and rain, respectively. Including only a dummy for the first weather variable

(as whether or not it is sunny is highly correlated with whether or not it is likely to rain, so it is reasonable to exclude one of the weather variables) generates p-values of .000013, .0073, and .093, which seem similar to those when both weather variables were included. So there is at best weak evidence that either of the weather variables are significant. Weekend status has somewhat stronger evidence for being significant (p-value below .05 in one regression, close in the other). Previous day's attendance is highly significant either way.

b. Can we conclude that weather is a factor in determining attendance?

See answer to (a)

c. Do these results provide sufficient evidence that weekend attendance is, on average, larger than weekday attendance?

See answer to (a)

Problem 3 The General Social Survey asks respondents to report the number of hours per average dat of television viewing (TVHOURS). Conduct a regression analysis using the following independent variables:

- Education (EDUC)
- Age (AGE)
- Hours of work (HRS)
- Number of children (CHILDS)
- Number of family members earning money (EARNRS)
- Occupation prestige score (PRESTG80)

a. For each coefficient, say whether or not it is significant.

Removing all observations with at least one missing value for the variables listed yields 780 usable observations. Estimates are as follows:

	Coefficients	tandard Err	t Stat	P-value	Lo
Intercept	4.718604062	0.483934	9.750502	2.90428E-21	3
AGE	0.01361368	0.005813	2.342089	0.019429095	0
EDUC	-0.091261678	0.02932	-3.11259	0.001922835	-
HRS	-0.011523822	0.0049	-2.35166	0.018940058	-
PRESTG80	-0.020874718	0.006013	-3.47174	0.000545813	-
CHILDS	-0.042704996	0.050945	-0.83826	0.402143823	-
EARNRS	-0.079560334	0.081598	-0.97503	0.329851441	-

b. What is the coefficient of determination, and what does it tell you?

The R^2, or coefficient of determination, is .066. This means that only 6.6% of the variation in TV hours watched is explained by the included regressors. Hence, it seems that while several of the included variables are significant explainers of TV viewing habits, there are a lot of other factors in how much TV people watch.

c. A friend suggests that men watch more TV than women. Test this suggestion by including sex (SEX) as an independent variable, and interpreting its coefficient.

The SEX variable has an estimated coefficient of .0194, with a p-value of .8924. Hence, it seems that men and women watch similar amounts of TV, controlling for other factors.

d. Another friend suggests that men and women may be differentially affected by the number of children. Test this claim by including a term interacting sex and number of children, and interpret its results.

The interaction term has an estimated coefficient of .048, with a p-value of .617, meaning there is no clear evidence that men's and women's TV habits are differentially affected by number of children.

Problem 4 The following table is taken from Klick and Tabarrok, "Using terror alert levels to estimate the effect of police on crime". (1) and (2) denote two separate regressions run by the authors.

	(1)	(2)
High alert	-7.316^{*}	-6.046^{*}
	(2.877)	(2.537)
Log(midday ridership)		17.341^{**}
		(5.309)
R^2	.14	.17

NOTE: The dependent variable is the daily total number of crimes in Washington D.C. during the period March 12,2002–July 30,2003. Both regressions contain day-of-the-week fixed effects. The number of observations is 506. Standard errors are in parentheses.
* Significantly different from zero at the 5 percent level.
** Significantly different from zero at the 1 percent level.

a. In specification (1), interpret the coefficient estimate in plain English. Is it significant? What does it tell us?

On high alert days, there are an estimated 7.316 fewer crimes in Washington D.C. This result is significant with $\alpha = .05$.

b. Why did the authors choose midday ridership (on the DC subway system) as a control variable in specification (2)?

Perhaps fewer tourists visit Washington D.C. on high alert days, wanting to avoid either burdensome security and delays or actual risk to their person. Midday metro ridership is a reasonable (though imperfect) way of measuring how many tourists are in the District on a given day.

The following appears as column 1 of Table 4 in the same paper:

	Coefficient (standard error)
High Alert * District 1	−2.621**
	(.044)
High Alert * Other Districts	−.571
	(.455)
Log(midday ridership)	2.477*
	(.364)
Constant	−11.058**
	(4.211)

NOTE: $R^2 = .28$. The dependent variable is the daily total number of crimes by district. The regression contains day-of-the-week fixed effects. Standard errors are in parentheses. The number of observations is 3,542.

* Significantly different from zero at the 5 percent level.

** Significantly different from zero at the 1 percent level.

c. Interpret the coefficient to High Alert * District 1 in plain English. Is it significant? What does it tell us about the effect of police in District 1?

In District 1 alone, there are 2.621 fewer crimes on high alert days. This coefficient estimate is significant with $\alpha = .01$. The fact that this coefficient estimate is significant, but the interaction term "High Alert * Other Districts" is not suggests that much of the effect of extra police on crime is concentrated in District 1.

d. Explain why the authors chose to single out District 1. What did they conclude about the magnitude of the elasticity of crime with respect to police in District 1?

Most of the National Mall, including the Capital, the White House, the Washington Monument, and the Smithsonian Museums are located in District 1, thus making it the most likely target of a terrorist attack. The authors argue that crime decreases by about 15% in District 1 on high alert days. Given that police shift from 8- to 12-hour shifts on high alert days (a 50% increase in police manpower), the implied elasticity in District 1 is .3. If additional police are deployed disproportionately to District 1, so that the District 1 increase in police is greater than 50%, the elasticity will be lower.