STA502: Math & Stat for MBA

Lecture Note 7 & 8

Question 1. A sample of students from a large university is used to obtain the following regression result in an attempt to explain the university grade point average ($UGPA_i$).

$$\widehat{UGPA}_i = \underset{(0.33)}{1.39} + \underset{(0.094)}{0.412}HGPA_i + \underset{(0.011)}{0.15}\,SAT_i - \underset{(0.026)}{0.083}SK_i$$

where $N = 141$, $R^2 = 0.234$, $HGPA_i$ is the high school GPA, SAT_i is an SAT score, and SK_i is the average number of lectures missed per week. The standard errors are between brackets.

1. Interpret this equation. Do the parameters have the expected signs?

2. compute the adjsted R^2 and test the significance of the regression.

3. Which slope coefficients are significantly different from zero at the 5% level of significance? What difference does it make whether we test using a one or two sided alternative?

4. Provide the p-value associated with testing the significance of skipping classes on the college GPA using a two sided alternative. What does the p-value tell us? (Hint: You may want to make use of the standard normal table in light of the "large" sample size.)

5. Find the 95% confidence interval for β_{HGPA}, where β_{HGPA} is the true parameter associated with high school GPA in this model. Can you reject the hypothesis that $\beta_{HGPA} = 1$ against a two-sided alternative at the 5% level?

Solution.

1. Need to interpret these marginal effects ceteris paribus, yes

2.
$$\bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - (1 - R^2) \times \frac{n-1}{n-k} = 1 - (1 - 0.234) \times \frac{140}{137} = 0.217$$

   The significance of the regression, assuming all classical linear regression assumption inclusive of normality of the errors can be tested using the F-test, which under the null is distributed as $F_{3,137}$ (3 restrictions, and $n-k$ degrees of freedom). The null tested is that all slopes are equal to zero against the two sided alternative that at least one is non-zero, the 5% critical value equals about 2.86, the test statistic

$$F = \frac{(RRSS - URSS)/3}{URSS/137} = \frac{(TSS - RSS)/3}{RSS/137} = \frac{R^2/3}{(1 - R^2)/137} = 13.85$$

   indicating a strong rejection of the null, a highly significant regression.

3. The significance of individual coefficients (here the slopes), assuming all classical linear regression assumptions inclusive of normality of the errors can be tested using the t-test, which under the null is distributed as t_{137}. The null tested $H_0 : \beta_i = 0$ against the two-sided alternative $H_A : \beta_i \neq 0$, the 5% critical value equals 1.96, the test statistic $t_{\hat{\beta}_i} = \frac{\hat{\beta}_i}{S.E.(\hat{\beta}_i)}$.

   Using a one-sided test has improved power implications, the critical value changes, since e.g., we don't expect skipping classes to improve college GPA we are only concerned about too high negative values indicative of any non-negligible effect on college GPA (ceteris paribus)

4. The p-value (two sided) gives the lowest level of signifance at which we want to reject the null given the null is true, i.e.,

$$\begin{aligned}\Pr(\mid \frac{\hat{\beta}_i}{S.E.(\hat{\beta}_i)} \mid > 3.19) &= \Pr(\mid t \mid > 3.19) = 2 \times \Pr(t_{137} > 3.19) \\ &\cong 2 \times \Pr(N(0,1) > 3.19) = 0.0012\end{aligned}$$

5. Answer

$$\Pr(\mid \frac{\hat{\beta}_{HGPA} = \beta_{HGPA}}{S.E.(\hat{\beta}_{HGPA})} \mid > 1.96) = 5\%$$

The range matches to $[0.22776 \;\; 0.59624]$. Since 1 does not lie in this 95% confidence region, we reject the hypothesis at the 5% level of significance.

Question 2. (A question for mathematical derivation) A researcher from SIAI using cross-sectional data hypothesizes that two variables Y and X are jointly determined by a simultaneous equations model consisting of the following two relationships:

$$Y = \beta_1 + \beta_2 X + \beta_3 Z + u \qquad (1)$$
$$X = \alpha_1 + \alpha_2 Y + v \qquad (2)$$

where Z may be assumed to be an exogenous variables and u and v are identically and independently distributed disturbance terms with zero means. The observations for Z are drawn from a fixed population with finite mean and variance.

1. Derive the reduced form equation for Y

2. Demonstrate that the OLS estimator of α_2 is, in general, inconsistent. How is your conclusion affected in the special case $\beta_2 = 0$? How is your conclusion affected in the special case $\alpha_2\beta_2 = 1$?

3. Demonstrate that the instrumental variables (IV) estimator of α_2, using Z as an instrument for Y, is consistent. Why do you need an IV estimator?

4. Instead of using IV, the researcher decides to use 2-Stage-Least-Square (2SLS) in the expectation of obtaining a more efficient estimator of α_2. He fits the reduced form equation for Y:

$$\hat{Y} = h_1 + h_2 Z \qquad (3)$$

saves the fitted values, and uses them as an instrument for Y in equation (2). Demonstrate that the 2SLS estimator is consistent

5. Determine whether the researcher is correct in believing that the 2SLS estimator is more efficient than the IV estimator

6. How do you prove that IV (or 2SLS) estimation is superior to OLS?

7. Now that another researcher from a coding institution claims that adding a quadratic term, instead of IV or 2SLS, is far more superior estimation strategy, because he believes non-linear & non-parametric estimation by computers are better than human logic, as was witnessed by Alpha-Go. Provide your rebuttal.

Solution.

1.

$$Y = \frac{1}{1 - \alpha_2\beta_2}(\beta_1 + \alpha_1\beta_2 + \beta_3 Z + u + \beta_2 v)$$

2.

$$
\begin{aligned}
a_2^{OLS} &= \frac{\sum(Y_i-\bar{Y})(X_i-\bar{X})}{\sum(Y_i-\bar{Y})^2} \\
&= \frac{\sum(Y_i-\bar{Y})([\alpha_1+\alpha_2 Y_i+v_i]-\big[\alpha_1+\alpha_2\bar{Y}+\bar{v}\big])}{\sum(Y_i-\bar{Y})^2} \\
&= \alpha_2 + \frac{\sum(Y_i-\bar{Y})(v_i-\bar{v})}{\sum(Y_i-\bar{Y})^2}
\end{aligned}
$$

It is not possible to obtain a closed-form expression for the expectation of the error term because v is a component of Y. Instead, we investigate the limiting value, first dividing the numerator and denominator by n so that they have limits:

$$
\begin{aligned}
p\lim a_2^{OLS} &= \alpha_2 + p\lim \frac{\sum(Y_i-\bar{Y})(v_i-\bar{v})}{\sum(Y_i-\bar{Y})^2} \\
&= \alpha_2 + \frac{p\lim \frac{1}{n}\sum(Y_i-\bar{Y})(v_i-\bar{v})}{p\lim \frac{1}{n}\sum(Y_i-\bar{Y})^2} \\
&= \alpha_2 + \frac{Cov(Y,v)}{Var(Y)} \\
&= \alpha_2 + \frac{1}{Var(Y)} Cov\left[\frac{1}{1-\alpha_2\beta_2}(\beta_1+\alpha_1\beta_2+\beta_3 Z+u+\beta_2 v), v\right] \\
&= \alpha_2 + \frac{\beta_2}{1-\alpha_2\beta_2}\frac{Var(v)}{Var(Y)}
\end{aligned}
$$

Since $Cov(Z,v) = Cov(u,v) = 0$. Thus a_2^{OLS} is, in general, an inconsistent estimator of α_2. If $\beta_2 = 0$, X is not influenced by Y, there is no simultaneity, and OLS will be a consistent estimator. (What about bias?) If $\alpha_2\beta_2 = 1$, the lines are parallel in the $\{X,Y\}$ dimensions and they do not interesect. Thus, the reduced form relationship is undefined.

3.

$$
\begin{aligned}
a_2^{IV} &= \frac{\sum(Z_i-\bar{Z})(X_i-\bar{X})}{\sum(Z_i-\bar{Z})(Y_i-\bar{Y})} \\
&= \frac{\sum(Z_i-\bar{Z})([\alpha_1+\alpha_2 Y_i+v_i])-\big[\alpha_1+\alpha_2\bar{Y}+\bar{v}\big]}{\sum(Z_i-\bar{Z})(Y_i-\bar{Y})} \\
&= \alpha_2 + \frac{\sum(Z_i-\bar{Z})(v_i-\bar{v})}{\sum(Z_i-\bar{Z})(Y_i-\bar{Y})}
\end{aligned}
$$

Again, v influences Y and expectations cannot be taken. Instead, taking $p\lim$s,

$$
\begin{aligned}
p\lim a_2^{IV} &= \alpha_2 + p\lim \frac{\frac{1}{n}\sum(Z_i-\bar{Z})(v_i-\bar{v})}{\frac{1}{n}\sum(Z_i-\bar{Z})(Y_i-\bar{Y})} \\
&= \alpha_2 + \frac{p\lim \sum(Z_i-\bar{Z})(v_i-\bar{v})}{p\lim \sum(Z_i-\bar{Z})(Y_i-\bar{Y})} \\
&= \alpha_2 + \frac{Cov(Z,v)}{Cov(Z,Y)} = \alpha_2
\end{aligned}
$$

Since $Cov(Z,v) = 0$ and $Cov(Z,y) \neq 0$.

4.

$$
\begin{aligned}
a_2^{2SLS} &= \frac{\sum(\hat{Y}_i - \bar{\hat{Y}})(X_i - \bar{X})}{\sum(\hat{Y}_i - \bar{\hat{Y}})(Y_i - \bar{Y})} = \frac{\sum([h_1 + h_2 Z_i] - \left[h_1 + h_2 \bar{Z}\right])(X_i - \bar{X})}{\sum([h_1 + h_2 Z_i] - \left[h_1 + h_2 \bar{Z}\right])(Y_i - \bar{Y})} \\
&= \frac{\sum(h_2 \left[Z_i - \bar{Z}\right])(X_i - \bar{X})}{\sum(h_2 \left[Z_i - \bar{Z}\right])(Y_i - \bar{Y})} \\
&= \frac{\sum(Z_i - \bar{Z})(X_i - \bar{X})}{\sum(Z_i - \bar{Z})(Y_i - \bar{Y})} = a_2^{IV}
\end{aligned}
$$

Hence a_2^{2SLS} is equivalent to the IV estimator and, accordingly, consistent.

5. Incorrect. Because the estimators are the same.

6. Hausman test, or testing for β_3.

7. Inherent endogeneity generated by simultaneity is the source of critical violation of A3, which cannot be resolved by simple non-linear & non-parametric transformation of the regression function.

Question 3. A researcher from SIAI has data on the number of crimes per 1,000 inhabitants, C, number of police per 1,000 inhabitants, P, and average household income, Y, measured in thousands of pounds, for 30 cities for year 2100. She hypothesizes that C is likely to be negatively related to Y, and that P may be related to Y:

$$
\begin{aligned}
C &= \beta_1 + \beta_2 P + u \\
P &= \alpha_1 + \alpha_2 Y + v
\end{aligned}
$$

where u and v are both identically and independently distributed disturbance terms and unrelated to each other. Putting the two relationships together, the researcher thinks C is related to Y:

$$C = \alpha_1\beta_2 + \beta_1 + \alpha_2\beta_2 Y + u + \beta_2 v$$

Accordingly she fits the regressions (standard errors in parentheses):

$$
\begin{aligned}
\hat{C} &= \underset{(0.46)}{3.79} + \underset{(0.10)}{0.22}\,P \qquad R^2 = 0.16 \;\; (1) \\
\hat{P} &= \underset{(0.80)}{9.50} - \underset{(0.02)}{0.08}\,Y \qquad R^2 = 0.38 \;\; (2) \\
\hat{C} &= \underset{(0.49)}{0.59} - \underset{(0.01)}{0.03}\,Y \qquad R^2 = 0.26 \;\; (3) \\
\hat{C} &= \underset{(0.60)}{4.20} - \underset{(0.02)}{0.08}\,Y - \underset{(0.20)}{0.63}\,P \quad R^2 = 0.16 \;\; (4)
\end{aligned}
$$

1. The investigator makes the following statement: ”Crime rates per 1,000 inhabitants can be expected to be lower in cities with more police per 1,000 inhabitants. I am therefore entitled to perform a one-sided test on the coefficient of P in regression (1). The t statistic for the coefficient is 2.20. The critical value of t, using a one-sided test, is about 1.70 at the 5% level and 2.47 at the 1% level. I can therefore conclude at the 5% level (but not the 1% level) that the number of police per 1,000 inhabitants has a significant effect on crime rates.” Explain whether you agree or disagree with this statement. Why?

2. (Optional) Explain mathematically why the coefficients of Y and P in regression (4) are different from their coefficients in regressions (1) and (3).

3. In regression (3) the residual sum of squares was 3,700. In regression (4) it was 2,700. Perform an F test on the reduction in the residual sum of squares, stating clearly your null hypothesis and conclusion. How is this test related to a one-sided test on the coefficient of P in regression (4)?

4. Another researcher B offers the following explanation for the coefficient of Y begin less negative in regression (3) than in regression (4). ”In regression (4), the coefficient is an estimate of the marginal effect of Y, holding P constant. In regression (3), the coefficient estimates the overall effect of Y on C, taking account of the fact that in reality P decreases with Y.” Explain whether this interpretation is correct.

5. Researcher C suggests that the real reason for the negative correlation between P and Y is that more police are needed where there is more crime, and there is less crime in higher-income areas. If this is correct, how would this affect the researcher’s conclusions? What are the necessary preconditions needed to be satisfied for a variable that researcher C proposes?

Solution.

1. Given the highly significant coefficient of Y in (4), the regression is subject to OVB and hence any test would be invalid. In addition, if the researcher was committed to a one-sided test, the fact that the coefficient has the wrong sign means that she would not reject H_0.

2. Regression (4) has the specification

$$C = \delta_1 + \delta_2 Y + \delta_3 P + u$$

Assuming that this is the true model, the slope coefficient in an OLS regression of (1), which omits Y, will be given by

$$\begin{aligned} d_3 &= \frac{\sum(P_i - \bar{P})(C_i - \bar{C})}{\sum(P_i - \bar{P})^2} \\ &= \frac{\sum(P_i - \bar{P})\left([\delta_1 + \delta_2 Y_i + \delta_3 P_i + u_i] - \left[\delta_1 + \delta_2 \bar{Y} + \delta_3 \bar{P} + \bar{u}\right]\right)}{\sum(P_i - \bar{P})^2} \\ &= \delta_3 + \delta_2 \frac{\sum(P_i - \bar{P})(y_i - \bar{Y})}{\sum(P_i - \bar{P})^2} + \frac{\sum(P_i - \bar{P})(u_i - \bar{u})}{\sum(P_i - \bar{P})^2} \end{aligned}$$

Therefore,

$$\begin{aligned} E\left[d_3\right] &= \delta_3 + \delta_2 \frac{\sum(P_i - \bar{P})(Y_i - \bar{Y})}{\sum(P_i - \bar{P})^2} + E\left[\frac{\sum(P_i - \bar{P})(Y_i - \bar{Y})}{\sum(P_i - \bar{P})^2}\right] \\ &= \delta_3 + \delta_2 \frac{\sum(P_i - \bar{P})(Y_i - \bar{Y})}{\sum(P_i - \bar{P})^2} + \frac{\sum(P_i - \bar{P})E(u_i - \bar{u})}{\sum(P_i - \bar{P})^2} \\ &= \delta_3 + \delta_2 \frac{\sum(P_i - \bar{P})(Y_i - \bar{Y})}{\sum(P_i - \bar{P})^2} \end{aligned}$$

Regression (2) reveals that P and Y are negatively correlated in the sample. Hence the second factor in the bias term is positive. One would anticipate $\delta_2 < 0$, and hence that the bias is positive. The bias is so large that the coefficient of P in regression (1) is actually positive. The analysis for the case where P is omitted is similar. The coefficient of Y is upwards biased, and the coefficient in (3) is indeed less negative than in (4)

3. The null hypothesis is $H_0 : \delta_3 = 0$ and the alternative hypothesis is $H_1 : \delta_3 \neq 0$

$$F(1, 27) = \frac{(3700 - 2700)/1}{2700/27} = 10$$

At the 1% significance level, the critical value of $F(1, 27)$ is 7.68. Hence the null hypothesis is rejected. The test is equivalent to a two-sided t test with the same null and alternative hypotheses. A one-sided t test, which is justified here since you would not expect P to have a positive effect on C, is more powerful.

4. Correct. If one wishes to estimate the marginal effect of Y, controlling for P, then the coefficient of Y in regression (3) is subject to OVB, ut nevertheless it does provide an estimate of the total effect, direct and indirect, taking account of the fact that P is related to Y.

5. The model might be re-specified on the following lines:

$$C = \beta_1 + \beta_2 P + \beta_3 Y + u$$
$$P = \alpha_1 + \alpha_2 C + v$$

OLS would no longer be appropriate for either equation. The first equation would be under-identified and the second equation would have to be fitted using IV, with Y as an instrument for C.

Question 4. SIAI is interested in the effect of social media use on earnings. It has a survey of 12,487 individuals aged 25 to 35 from the planet Earth. The key variable “social media use” (SM) is a dummy variable indicating whether the individual has been using social media regularly over the past three years. Running regressions with the log of annual earnings on the left hand side he obtains the following results:

Regressor	(1)	(2)	(3)	(4)	(5)
Social media use (SM)	0.118 (0.021)	0.182 (0.025)	0.138 (0.030)	0.169 (0.021)	0.027 (0.022)
Female * SM	–	-0.120 (0.023)	0.015 (0.044)	–	–
Female	–	–	-0.135 (0.036)	-0.123 (0.019)	-0.246 (0.019)
Age	–	–	–	0.070 (0.053)	0.065 (0.053)
Age squared	–	–	–	-0.0002 (0.0009)	-0.0001 (0.0009)
Number of friends	–	–	–	–	-0.050 (0.009)

All regressions also contain a constant term. Robust standard errors are reported in parentheses.

1. Give a story why social media use may have a positive causal effect on earnings, and a different story why it may have a negative causal effect.
2. Interpret the coefficient in column (1). Explain whether this estimate is likely to have a causal interpretation.
3. SIAI2 reacts to these results: ”The results in column (2) say that women who use social media have lower earnings than those who don’t. That’s really strange.” Explain why SIAI2 is wrong on two counts.
4. How would you carry out a statistical test that social media use has no effect on the earnings of males using column (3)? If you can, derive the result of this test from the information in the table or explain why you can’t.
5. Explain whether you can interpret the results in column (5) causally.
6. Do women have more or fewer friends than men? Explain how you arrive at your answer.

Solution.

1. For example, social networks may be helpful in finding better jobs, resulting in higher earnings. Social media users may us SM at work rather than concentrating on their job and hence not get a raise.

2. SM users have 12% higher earnings than non-users. This estimate is unlikely to have a causal interpretation. There could be many confounders. E.g. more gregarious indivdiuals may use SM more and be more successful in the job market. There could also be reverse causality. Those earning very little can't affort a smartphone or computer. It is important that stories about confounders etc. are distinct from the causal stories in (a)

3. The effect of SM use of females is the sum of the coefficients on SM and Female*SM (0.182 - 0.120 = 0.062), which is positive. The regression has no female main effect. As col. (3) demonstrates, women earn less and use SM more, accounting for all of the negative estimate on Female*SM in col. (2)

4. t-test on the SM main effect, which is the effect of SM on males. 0.138/0.030 = 4.6, well above 2, so reject.

5. Comapred to col. (1), this controls for female, age and the number of friends, variables which may both be correlated with SM use and earnings. Female and age is probably not enough to control for. Friends are likely a good proxy for something like gregariousness, which may be a confounde. But SM use may also affect friendship networks which would make it a bad control (2 marks). So not clear (5) is a better spec than say (4).

6. Women have more friends. OVB on female comparing cols (4) and (5): friends has a positive effect on earnings, and OVB omitting friends on the coefficient of female is positive (-0.123 - (-0.246) = 0.123). Hence OVB = +*(relationship btw female and friends), so females must have more friends.

Question 5. The galaxy of SIAI consists of 34 planets. Each planet has their own university and students attend uni on their home planet. Students in SIAI either have to attend lectures in person or watch them online through hyper-space real-time communication. Researcher A has obtained data for the students from all the universities and would like to study the effect of watching lectures online on the students' exam scores. For each student, she has the exam score (0 - 100, 70+ is an A, 40 and below a fail), which she uses as her left hand side variable, the fraction of lectures watched online, and how many days the student visited the library each week. A obtains the following regression results:

Regressor	(1) - OLS	(2) - IV	(3) - IV	(4) - IV	(5) - IV
Fraction online	-4.91 (0.11)	-2.08 (0.22)	-0.56 (0.40)	-0.75 (0.90	6.09 (3.22)
Library visits	–	–	0.91 (0.08)	–	2.09 (0.43)
Sample	All	All	All	Unis with lottery	Unis with lottery

All regressions also contain a constant term and a dummy variable for each universities. Robust standard errors are reported in parentheses.

1. What is the interpretation of the coefficient in column (1). If this were a causal effect, would it be big or small? Explain whether this estimate is likely to have a causal interpretation.

2. Researcher B observes that some halls of residence are close to the lecture rooms while others are further away and suggests to use the distance of halls from lecture halls as an instrumental variable for the fraction of lectures watched online.

Results for this IV regression are displayed in column (2). Explain why instrumental variables may produce a better estimate of the causal effect of watching lectures online, and which assumptions need to be satisfied for this to be the case. Discuss the validity of the assumptions in this case. Can you ascertain whether any of these assumptions are true from the results in the table above?

3. Researcher C points out to B that students can pick which hall they want to live in and that some halls are located in Study Village, close to lecture halls and the library while others are located further away in Party Town, surrounded by pubs. How does this information affect your assessment of the IV strategy?

4. Researcher D realises that the data also include a variable for the number of times a student has checked into the library per week. He suggests to rerun the instrumental variables regression adding this variable as a control. Results for this regression are displayed in column (3). Assess D's strategy.

5. Researcher E notices that there are five universities which assign students to their halls of residence by a lottery. She suggests to run the IV model from columns (2) and (3) for the subsample of students from these universities only. Results are displayed in columns (4) and (5). Assess E's regressions.

6. Drawing on the results in the table above, what have you learned from this exercise about the causal effect of watching lectures online on students' exam results?

Solution.

1. It is the average difference in marks for a student watching all lectures online compared to a student attending all lectures live. A 5 mark difference for going to live lectures seems like a big effect, given that the content of live and online lectures is the same. Not likely causal. For example, students who go to live lectures may be more engaged in the course and study harder in other respects.

2. If there are unobserved confounders in col. (1) then IV has the potential to fix the OVB problem from those. Assumptions: First stage; seems plausible that students living closer go to more live lectures. Quasi random assignment; may be reasonable a priori that location of halls has no relationship to confounders, though maybe more engaged students want to live closer. Exclusion: the effect of distance only works through online, and not other channels like going to the library to study more often. More dubious. It is not possible to check any of these. First stage can be checked but info is not the table. (Actually, the first stage must be reasonable because the std. error in col. (2) doesn't go up too much – this suggests a viable first stage.)

3. This is bad news because it either affects random assignment (students choose in which halls to live leading to selection) or if students were randomly assigned to halls it would affect exclusion if students in Party Town are encouraged by their location to visit the pub more often.

4. This could potentially solve the problem in (c) to the degree that library visits captures the type of student who is more likely to go to live lectures, a confounder. But library visits are also an alternative channel for the instrument, which makes it a bad control

5. Random assignment to halls resolves issues about that IV assumption, and hence the need for a control like library. It doesn't resolve issues about the exclusion restriction (note that including library doesn't fix the problem if exclusion is violated through a library channel). Standard errors are large now, even in col (4) further limiting the usefulness of this IV exercise on the smaller sample.

6. If exclusion was believable, the result in col (4) would be the most plausible. The confidence interval is [-2.5, 1], consistent with moderate negative effects as well as small positive effects. As a result, we don't learn all that much and exclusion remains a worry.

Question 6. The lack of access to sufficient credit may be an important impediment for the growth of small businesses. If these businesses cannot borrow freely they may have to rely on the private resources of the owners to finance their business activities. Data scientists therefore have investigated whether the private wealth or collateral of owners of small businesses matters for business growth.

The following regressions are for 9,125 individuals who started being self-employed in 2100. The empirical strategy of the study is to compare various business outcomes (assets, sales, and number of employees) for homeowners and renters in 2105. The regressions also interact homeownership status with house price appreciation in the region of residence. Homeowners, who see their houses become more valuable, may often have access to mortgage financing which could be used to invest in their business. House price growth is coded so that growth of 5% would be 0.05.

	Dependent variable		
	$ln(Assets)$	$ln(Sales)$	$ln(Employment)$
	(1)	(2)	(3)
D_h	0.079 (0.027)	−0.131 (0.026)	−0.108 (0.015)
$D_h \times HP_g$	1.21 (0.18)	0.94 (0.17)	0.37 (0.11)
HP_g	0.58 (0.28)	0.29 (0.26)	0.21 (0.22)

where D_h is for Dummy for homeowner, and HP_g is for house price growth rate. Standard errors are displayed in parentheses. All regressions also contain a constant term.

(a) Explain why a simple regression of business outcomes on the wealth of the owner may not answer the question data scientists are interested in.

(b) Explain how the use of house price growth may circumvent the problem you described in part (a).

(c) Explain verbally what the coefficient of 0.079 on the dummy for homeowners in column (1) means.

(d) If house price growth is 10 percentage points higher, how much higher are the sales of renters in the sample on average? Explain whether this effect is statistically different from zero.

(e) What do you conclude from the results in the table about the effect of owner collateral on business outcomes?

(f) Suppose you also have data for assets, sales, and employment in these businesses in 2100. Suppose you were to run analogous regressions with these dependent variables to the regressions in the table above. Explain how the new regressions would help you interpret the results above.

Solution.

(a) The simple relationship btwn self-employment outcomes and wealth may be due to reverse causality or confounding. Business growth may not be due to wealth but rather the owners of the most successful businesses might have accumulated the most wealth. Alternatively, confounders like skills, effort, or ingenuity of the business owners might account for the relationship; they could have accumulated the wealth because they were successful in a previous job due to these traits.

(b) House price growth in the area where the business owner lives should not be caused by the success of the particular business (these businesses are small). To the degree that house price growth reflects aggregate demand or general growth in the area, the regression controls for house price growth and only considers the difference between homeowners and renters. Conditional on homeownership and houseprice growth, the interaction should be as good as randomly assigned.

(c) In an area where house price growth was zero, ln(Assets) of homeowners are 0.079 or about 8% higher than for renters in a similar area.

(d) We want

$$E\left[ln(Sales)|D_h = 0, \triangle HP_g = 0.1\right] = 0.29 \times 0.1 = 0.029$$

or about 3%, where 0.29 is the coefficient on House price growth in column (2). This effect is not statistically different from zero: the t-statistic on the House price growth coefficient is $0.29/0.26 = 1.1$, which is below conventional significance levels (note that the t-statistic for the 3% is the same as we are just multiplying the coefficient by the fixed number 0.1).

(e) House price growth seems to have a sizeable and statistically significant effect on assets, sales, and employment. A 10 percentage points higher house price growth is associated with 12% higher assets, 9% higher sales, and 4% higher employment for home owners than renters. Private resources or collateral seem to matter for these business owners.

(f) Future house price growth for the homeowners should not have any differential effect on the 2012 variables. Hence, these should be balanced and we would expect a zero effect on the homeowner x house price growth interaction.

Question 7. Are rent rates influenced by the student population in a college town? Let *rent* be the average monthly rent paid on rental units in a college town in the United States. Let *pop* denote the total city population, *avginc* the average city income, and *pctstu* the student population as a percentage of the total population. One model to test for a relationship is

$$log(rent) = \beta_0 + \beta_1 log(pop) + \beta_2 log(avginc) + \beta_3 pctstu + u.$$

The equation estimated using data from 64 college towns is

$$\widehat{log(rent)} = \underset{(.844)}{.043} + \underset{(.039)}{.066}\, log(pop) + \underset{(.081)}{.507}\, log(avginc) + \underset{(.0017)}{.0056} pctstu$$
$$n = 64, \quad R^2 = .458.$$

(i) Explain briefly how to interpret the estimate for β_1. If we replace *pop* in the above regression with $popthousand = pop/1000$ (measured in thousands), what will happen to the OLS estimates for β_0 and β_1?

(ii) State the null hypothesis that size of the student body relative to the population has no ceteris paribus effect on monthly rents. State the alternative that there is an effect. Then explain how to compute the p-value for this test. Draw a sketch to illustrate.

(iii) Suppose you want to test the joint hypothesis $H_0 = \beta_1 = \beta_2 = 0$. In addition to the above regression output, explain what kind of output do you need. Then explain how to implement the test for H_0.

Question 8. Three researchers from SIAI are investigating the effects of time spent studying on the examination marks earned by students on Machine Learning II. For a sample of 100 students, they have the examination mark, M, total hours spent studying, H, hours on primary study, P, and hours spent on revision, R. By definition, $H = P + R$. The sample means of H, P and R are 100 hours, 95 hours, and 5 hours, respectively. The sample correlation coefficients are 0.98 for H and P, 0.10 for H and R, and -0.11 for P and R. The standard deviations of the distributions of H, P and R are 10.1, 10.1, and 2.1, respectively.

Researcher A decides to regress M on P and R and fits the following regression (standard errors in parentheses; t statistics in square brackets):

$$\hat{M} = \underset{\substack{(2.8)\\ [16.30]}}{45.6} + \underset{\substack{(0.03)\\ [5.49]}}{0.15}\, P + \underset{\substack{(0.14)\\ [1.51]}}{0.21}\, R \qquad R^2 = 0.243 \quad (1)$$

Researcher B decides to regress M on H, P, and R. However, the regression application refuses to fit the regression with all three explanatory variables. Instead, it drops R and the regression output is

$$\hat{M} = \underset{\substack{(2.8)\\ [16.30]}}{45.6} + \underset{\substack{(0.14)\\ [1.51]}}{0.21}\,H - \underset{\substack{(0.14)\\ [-0.40]}}{0.05}\,P \qquad R^2 = 0.243 \quad (2)$$

1. Researcher A says that her specification has better explanatory power than that of Researcher B because the coefficient of her main variable, P, has a high t statistic. Explain whether this assertion is correct.

2. She says that the insignificant coefficient of R in (1) is to be expected because the students, on average, spent much less time on revision than on primary study. Explain whether this assertion is correct.

3. Researcher B says that, assuming that his specification is in fact correct, not being able to include R in the regression has given rise to omitted variable bias, and this is responsible for the negative coefficient of P. Explain whether this assertion is correct.

4. A commentator, drawing attention to the high correlation between H and P, says that the real reason for the implausible negative coefficient of P obtained by Researcher B is multicollinearity. Explain whether this assertion is correct. Is the negative coefficient of P implausible?

5. Researcher C says that it would be better to keep the specification simple and just regress M on H. He has done so, with the following reults:

   $$\hat{M} = \underset{\substack{(2.8)\\ [16.56]}}{45.8} + \underset{\substack{(0.03)\\ [5.58]}}{0.15}\,H \qquad R^2 = 0.241 \quad (3)$$

   He says that his results are actually more satisfactory than those of Researchers A and B. Explain whether this assertion is correct.

6. The commentator says that , given the low R^2, it is obvious that there are other important determinants of the marks and they may be associated with willingness to spend time studying, causing all the regression results to be distorted. Explain whether this assertion might be correct.

7. Another commentator says that, since the hours-of-study data are self-reported, it is likely that many of the students have made up the numbers. He says that this will cause t statistics to be lower than they would have been, if the reporting were accurate, and this may be one of the reasons that the R^2 is so low. Explain whether this assertion might be correct.

8. Now you are given exam marks for Machine Learning I for all students. How can you exploit the information? Is it exogenous to this regression?

9. You believe that the problem of lower R^2 is due to lack of sufficient data. The chair of SIAI gives you an option either to access the same set data for other courses, or to be a tenured professor for Machine Learning II for the next 30 years to collect the same set of data for the same course. Evaluate the options.