ECON306 – Final exam

2013 · 6 · 24

Name: Bruno Salcedo (Answer Key) PSU ID: bxs5142

There are 50 questions worth 2 points each. You have 1h 50min to answer all of them. Don't forget to write your name and PSU ID (e.g. bxs5142).

Part I. For problems 1–6, consider the random variables x, y with the following joint distribution

	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$
$x = -1$	0.00	0.15	0.01	0.07	0.02
$x = 0$	0.10	0.07	0.03	0.00	0.30
$x = 1$	0.01	0.03	0.01	0.15	0.05

1. Find the marginal distribution of x.

$$\begin{aligned}
\Pr(x = -1) &= \Pr(x = -1 \,\&\, y = 1) + \Pr(x = -1 \,\&\, y = 2) + \Pr(x = -1 \,\&\, y = 3)\dots \\
&\quad \dots \quad + \Pr(x = -1 \,\&\, y = 4) + \Pr(x = -1 \,\&\, y = 5) \\
&= 0 + 0.15 + 0.01 + 0.07 + 0.02 = 0.25 \\
\Pr(x = 0) &= 0.1 + 0.07 + 0.03 + 0 + 0.3 = 0.50 \\
\Pr(x = 1) &= 0.01 + 0.03 + 0.01 + 0.15 + 0.05 = 0.25
\end{aligned}$$

2. Compute the expected value of x.

$$\begin{aligned}
\mathbb{E}[x] &= 1 \cdot \Pr(x = 1) + 0 \cdot \Pr(x = 0) + (-1) \cdot \Pr(x = -1) \\
&= \Pr(x = 1) - \Pr(x = -1) = 0.25 - 0.25 = 0
\end{aligned}$$

3. Compute the variance of x. [Hint: what is the expected value of x^2?]

$$\begin{aligned}
\mathbb{E}\left[x^2\right] &= 1^2 \cdot \Pr(x = 1) + 0^2 \cdot \Pr(x = 0) + (-1)^2 \cdot \Pr(x = -1) \\
&= \Pr(x = 1) + \Pr(x = -1) = 0.25 + 0.25 = 0.5 \\
\mathbb{V}[x] &= \mathbb{E}\left[x^2\right] - \left(\mathbb{E}[x]\right)^2 = 0.5 - (0)^2 = 0.5
\end{aligned}$$

4. Find the distribution of x conditional on $y = 3$.

$$\Pr(y=3) = \Pr(x=-1 \,\&\, y=3) + \Pr(x=0 \,\&\, y=3) + \Pr(x=1 \,\&\, y=3) = 0.01 + 0.03 + 0.01 = 0.05$$

$$\Pr(x=-1|y=3) = \frac{\Pr(x=-1 \,\&\, y=3)}{\Pr(y=3)} = \frac{0.01}{0.05} = \frac{1}{5}$$
$$\Pr(x=0|y=3) = \frac{\Pr(x=0 \,\&\, y=3)}{\Pr(y=3)} = \frac{0.03}{0.05} = \frac{3}{5}$$
$$\Pr(x=1|y=3) = \frac{\Pr(x=1 \,\&\, y=3)}{\Pr(y=3)} = \frac{0.01}{0.05} = \frac{1}{5}$$

5. Compute the expected value of x conditional on $y = 3$.

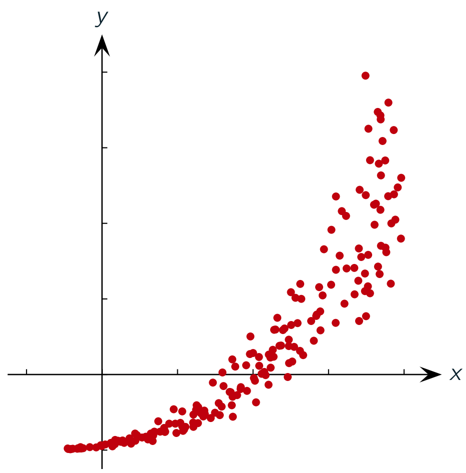
$$\begin{aligned}\mathbb{E}[x|y=3] &= 1 \cdot \Pr(x=1|Y=3) + 0 \cdot \Pr(x=0|Y=3) + (-1) \cdot \Pr(x=-1|Y=3) \\ &= \Pr(x=1|Y=3) - \Pr(x=-1|Y=3) = 0.2 - 0.2 = 0\end{aligned}$$

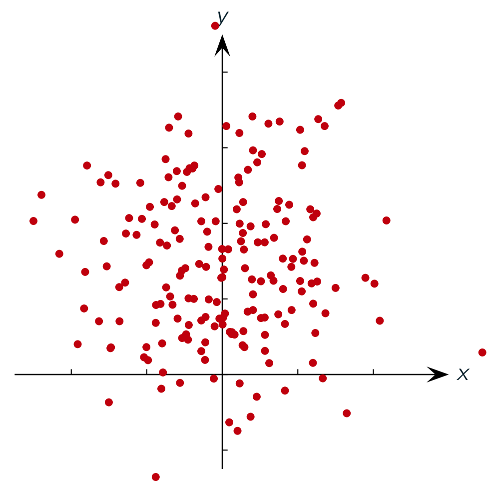
6. Are x and y independent? Briefly justify your answer (1–2 sentences).

No, e.g., $\Pr(x=0) = 0.5 \neq 0.6 = Pr(x=0|y=3)$.

Part II. (Questions 7-30) For each of the following scatterplots involving random samples for variables x and y, answer the following questions:

- Do x and y appear to be independent?
- Would the OLS intercept estimate $\hat{\beta}_0$ be positive, negative, or close to 0?
- The relationship between x and y appears to be positive and significant, negative and significant, or insignificant?
- Would the R^2 coefficient of the OLS estimated model be big (> 0.9), small (< 0.1), or intermediate?
- Does the data generating process appear to be linear? If not, which transformation could you use to obtain a better fit?
- Is there any classical assumption other than correct specification which appears to be violated? (Write at most one.)





7. Independence: NOT independent

8. β_0: negative

9. β_1: positive

10. R^2: high

11. Transformation: log–log

12. Assumptions: linearity

13. Independence: independent

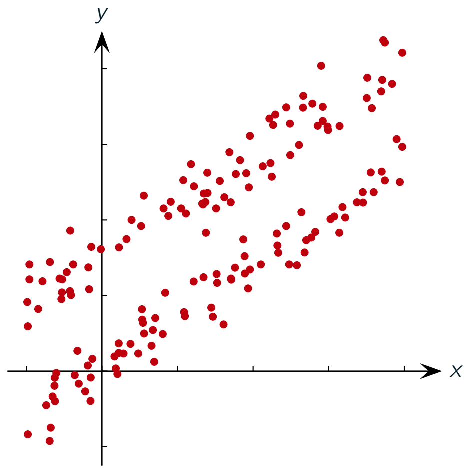
14. β_0: positive

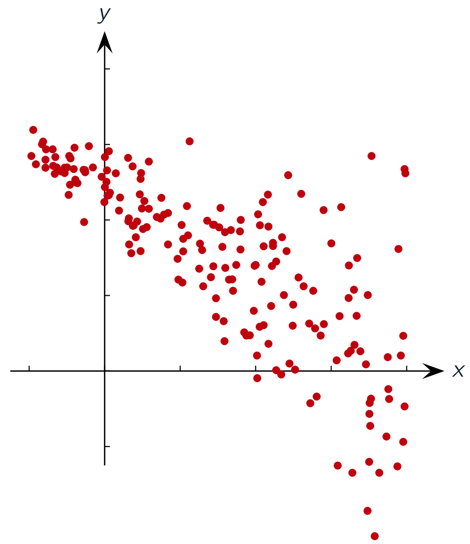
15. β_1: close to 0

16. R^2: low

17. Transformation: none

18. Assumptions: none





19. Independence: NOT independent

20. β_0: positive

21. β_1: positive

22. R^2: medium

23. Transformation: none

24. Assumptions: normality

25. Independence: NOT independent

26. β_0: positive

27. β_1: negative

28. R^2: medium

29. Transformation: mone

30. Assumptions: homoskedasticity

Part III. For problems 31–38, consider the following model for the weight of a person:

$$\text{WEIGHT}_i = \beta_0 + \beta_1\,\text{HEIGHT}_i \cdot \text{WAIST}_i + \beta_2\,\text{BACK}_i \cdot \text{HEIGHT}_i + \beta_3\,\text{NECK}_i \cdot \text{HEIGHT}_i + \varepsilon_i$$

where:

$$\begin{aligned}
\text{WEIGHT}_i &= \text{weight measured in kilograms}\\
\text{HEIGHT}_i &= \text{height measured in centimeters}\\
\text{NECK}_i &= \text{neck girth measured in centimeters (indicates build)}\\
\text{BACK}_i &= \text{width, measured from shoulder to shoulder in centimeters (indicates bone structure)}\\
\text{WAIST}_i &= \text{waist diameter measured in centimeters (indicates build)}
\end{aligned}$$

Suppose that you run a linear regression and obtain the following results (some values are missing)

weigh t	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
height*waist	-0.1632209			0.971		
height*neck	69.47699	2.290621				
height*back	33.42678				27.87425	38.97932
_cons	1.905555	7.177806	0.27	0.791	-12.17987	15.99098

31. Write down the estimated model (equation).

$$\widehat{\text{WEIGHT}}_i = \beta_0 - 0.16\,\text{HEIGHT}_i \cdot \text{WAIST}_i + 69.48\,\text{BACK}_i \cdot \text{HEIGHT}_i + 33.43\,\text{NECK}_i \cdot \text{HEIGHT}_i$$

32. What is the average difference in weight between people with the same values for NECK, WAIST and BACK, whose HEIGHT differs by one centimetre?

$$\begin{aligned}
\Delta\widehat{\text{WEIGHT}}_i &= \hat{\beta}1\,\text{WAIST}_i + \hat{\beta}2\,\text{BACK}_i + \hat{\beta}_3\text{NECK}_i\\
&= -0.16\,\text{WAIST}_i + 69.48\,\text{BACK}_i + 33.42\,\text{NECK}_i
\end{aligned}$$

33. Which variables are significant with 95% confidence?

There are three regressors: $\text{HEIGHT}_i \times \text{WAIST}_i$, $\text{HEIGHT}_i \times \text{BACK}_i$, and $\text{HEIGHT}_i \times \text{NECK}_i$.

$\text{HEIGHT}_i \times \text{WAIST}_i$ is not significant at the 95% confidence level since the corresponding p-value is 0.971 and $1 - 0.971 = 0.029 < 0.95$.

For $\text{HEIGHT}_i \times \text{WAIST}_i$, the corresponding t-statistic is $t_3 = \hat{\beta}_3/\text{SE}(\hat{\beta}_3) = 69.48/2.29 \approx 30.33 \gg 1.96$. Hence it is significant at the 95% confidence level.

For $\text{HEIGHT}_i \times \text{WAIST}_i$, we see that 0 does not belong to the 95% confidence interval, and it is significant at the 95% confidence level.

34. How can you explain that the coefficient of $\text{WAIST} \cdot \text{HEIGHT}$ is negative?

First, the coefficient is negative but not significant. This is at odds with my intuition because I would expect bigger people to weight more. Since both NECK_i and WAIST_i indicate build. I would expect them t be highly correlated with each other (specially after controlling for BACK_i). Therefore there strange estimate may be a result of multicolinearity.

35. What do you think would happen to R^2 if we excluded WAIST · HEIGHT from the regression?

Removing a regressor always reduces the R^2 coefficient.

36. What do you think would happen to *the adjusted* R^2? Briefly justify your answer (1–2 sentences).

Based on the previous answers, since there is multicolinearity, WAIST · HEIGHT is not helping to explain additional variation on WEIGHT. Hence, when it is removed I would expect $\bar{R}^2$ to go up.

37. What would happen to $\text{SE}(\hat{\beta}_1)$ if we increased the sample size?

It would decrease (on average) because:

$$\mathbb{E}\left[\,\text{SE}(\hat{\beta}_1)\,\right] = \mathbb{V}\left[\,\hat{\beta}_1\,\right] = \frac{1}{n} \cdot \frac{\mathbb{V}\left[\,\varepsilon_1\,\right]}{\mathbb{V}\left[\,x_1\,\right]}$$

(This formula is only correct if there is only one regressor, or if the different regressors are uncorrelated, but the general formula has a very similar structure and the intuition still applies).

Part IV. For problems 38–44, consider the following estimated model

$$\widehat{\text{WAGE}}_i = \hat{\beta}_0 + \hat{\beta}_1 \,\text{EXP}_i + \hat{\beta}_2 \,\text{DEG}_i + \hat{\beta}_3 \,\text{EXP}_i \cdot \text{DEG}_i$$

where:

$$\begin{aligned}
\text{WAGE}_i &= \text{annual income in thousands of dollars} \\
\text{EXP}_i &= \text{years of professional experience} \\
\text{DEG}_i &= 1 \text{ if the individual has a college degree and 0 otherwise} \\
\text{SKILL}_i &= \text{(unobserved) innate ability}
\end{aligned}$$

38. What sign (positive or negative) would you expect each *slope* coefficient to have? Briefly justify your answer (1–2 sentences per coefficient).

$\hat{\beta}_1 > 0$ – Positive, because the labor market rewards productivity which increases with experience. (There are other interesting reasons, I just picked one).

$\hat{\beta}_2 > 0$ – Positive, since (a) schooling signals high productivity, and (b) education increases productivity.

$\hat{\beta}_3 < 0$ – Because the effect of college is materialized mostly at entry level jobs, progression thereafter is mostly driven by performance. Therefore, the longer the experience, the smaller the wage differential between people with different degrees.

Alternative valid answer
$\hat{\beta}_3 > 0$ – Because people with college degrees are likely to go into managerial career tracks and will observe many significant wage increases throughout their career. In contrast, people without college degrees are more likely to end up in dead-end jobs with no progression possibilities and be stuck with a similar wage throughout their career. Therefore, the longer the experience, the larger the wage differential between people with different degrees.

39. For people with 10 years of experience, what is the difference in average income between those who went to college and those who didn't?

$$\Delta\widehat{\text{WAGE}}_i = \hat{\beta}_2 + \hat{\beta}_3 \cdot 10$$

40. How would the income of people with no college education and 10 years of experience change on average if they had gone to college? (They would have 4 less years of experience.)

Original wage predicted by the model:

$$\widehat{\text{WAGE}}_i = \hat{\beta}_0 + \hat{\beta}_1 \, 10$$

New wage predicted by the model:

$$\widehat{\text{WAGE}}_i' = \hat{\beta}_0 + \hat{\beta}_1 \, 6 + \hat{\beta}_2 \; + \hat{\beta}_3 \, 6$$

Difference:

$$\Delta\widehat{\text{WAGE}}_i = \widehat{\text{WAGE}}_i' - \widehat{\text{WAGE}}_i = -4\,\hat{\beta}_1 + \hat{\beta}_2 \; + 6\,\hat{\beta}_3$$

41. There may be endogeneity because we omitted the variable SKILL. Which variables would you expect to be endogenous? Which estimates (from $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$) would you expect to be biased?

I would expect DEG to be correlated with SKILL, since skilful people are more likely to go to college. Therefore $\hat{\beta}_2$ and $\hat{\beta}_3$ would be biased.

One could also argue that SKILL may be correlated with EXP since skilful people are more likely to keep their jobs longer, but this is debatable. For the rest of this answer key I will assume EXP is exogenous.

42. For each of the endogenous variables, would you expect a positive or a negative bias?

Since I expect a positive correlation between DEG and SKILL, and a positive effect of SKILL on WAGE, the bias of $\hat{\beta}_2$ and $\hat{\beta}_3$ would be positive.

43. Provide an example of an instrumental variable that could be used to correct this bias.

From the examples discussed in class, it could be distance to a college. People who love closer to a college are more likely to get a college degree (lower opportunity cost), and distance to college is probably otherwise unrelated to SKILL.

44. What are the three requirements that a valid instrument should satisfy?

Available (we can find data), relevant (explains sufficient variation of the endogenous regressor) and exogenous (uncorrelated with the error term).

Part V. For problems 45–50, consider the following estimated model for Geico (auto insurance company):

$$\widehat{\text{SWITCH}}_i = 0.055 + 0.001\,\text{AD}_i + 0.152\,\text{ACC}_i - 0.011\,\text{INC}_i + \varepsilon_i$$

where:

$$\begin{aligned} \text{SWITCH}_i &= \text{ 1 if a person switched from his current insurance provider to Geico and 0 otherwise} \\ \text{AD}_i &= \text{ 1 if the person received personalized advertisement from Geico and 0 otherwise} \\ \text{ACC}_i &= \text{ 1 if the person made an insurance claim on the previous year} \\ \text{INC}_i &= \text{ annual household income in thousands of dollars} \end{aligned}$$

45. What is the effect of personalized advertisement on the probability of switching to Geico?

$$\hat{\beta}_1 = 0.001$$

46. What is the probability of switching for a person who received personalized advertisement, made no insurance claim last year, and whose household annual income is 60,000 USD?

$$\widehat{\text{SWITCH}}_i = 0.055 + 0.001 - 0.011 \cdot 60 = -0.604$$

47. How can you interpret this number? Are there any difficulties involved?

The difficulty is that it is negative, which happens because we are using a linear probability model. We should interpret it as saying that, according to the estimated model, the probability of switching is close to 0.

48. How could you transform the model to avoid these difficulties?

Taking a distribution function F and estimating the model:

$$\Pr(\text{SWITCH}_i = 1) = F\Big(\beta_0 + \beta_1\,\text{AD}_i + \beta_2\,\text{ACC}_i + \beta_3\,\text{INC}_i\Big)$$

For example, if F was a standard normal distribution, this would be the probit model.

49. Do you see any potential source of endogeneity?

Geico chooses the audience of its advertisement carefully. People who are more likely to switch for exogenous reasons, are also more likely to be exposed to advertisement. Hence, there may be a positive correlation between AD and ε_i.

50. How could you solve it?

You would need an instrumental variable that helps to explain the probability of being exposed to a Geico add, but is not correlated with ε_i.