Introductory Econometrics
Multicollinearity and Heteroskedasticity
Suggested Solution

by Hieu Nguyen

Fall 2024

1.

We estimate a linear regression model for the years 1972 to 1991:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \epsilon_t,$$

where ϵ_t are normally and independently distributed, but we suspect that the variance of the error term is heteroskedastic and depends on x_{t1}. We estimate the following regression where e_t are residuals from regression (1):

$$e_t^2 = \delta_0 + \delta_1 x_{t1} + u_t.$$

We find that R^2 for regression (2) is 0.201. Use these results to test for the presence of heteroskedasticity.

Extract from statistical table of χ^2 distribution (area under right-hand tail):

d.f.	0.05	0.025	0.01
1	3.841	5.324	6.635
2	5.991	7.378	9.210
3	7.815	9.348	11.345
4	9.488	11.143	13.277

2.

Use data `htv_selected.gdt` to estimate the returns to education in the 'wage equation.'

(a) Estimate the baseline model of the impact of education and experience on wages:

$$\ln(\text{wage}_i) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \epsilon_i.$$

Interpret the estimated coefficient $\hat{\beta}_1$.

(b) Re-estimate the model using robust standard errors, comment on the differences.

(c) Test for heteroskedasticity in the model in part (a). Is it necessary to use robust standard errors in this case?

(d) Perform RESET (specification test) and discuss the results.

(e) Generate variable exper^2. Why we include this variable in the model and what is the expected sign of its coefficient?

(f) Estimate the model with quadratic specification (polynomial functional form) of experience:

$$\ln(\text{wage}_i) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + u_i.$$

Comment on how and why the estimated coefficient $\hat{\beta}_2$ changed with respect to part (a). Did the estimated coefficient $\hat{\beta}_1$ change as well? Why or why not? Compare R^2 and R^2_{adj} with the previous specification. Perform RESET again.

(g) Find $\frac{\partial \ln(\text{wage})}{\partial \text{exper}}$, which describes the marginal effect of a 1 year increase in work experience on wage. Compare the result with the marginal effect from the estimated model without exper^2.

(h) Do you believe that the coefficient β_1 is correctly estimated? Is there any issue that could create a bias in this equation? If yes, how would you solve for this problem? What is the expected sign of this bias?

(i) In the dataset, there are two proxies for inherent abilities and skills of the observed individuals, `abil1` and `abil2`. Estimate the model with just one of those. Is there an impact on the coefficient $\hat{\beta}_1$? Does this signalize there likely was a problem with bias in the model from part (f)?

Estimate the model with both proxies and discuss the differences and potential multicollinearity. Which Classical Assumption might be violated in this case? How do we check for this assumption?

(j) Include in the model from part (f) the education of the mother and of the father of the observed individuals:

$$\ln(\text{wage}_i) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + \beta_4 \text{motheduc}_i + \beta_5 \text{fatheduc}_i + v_i.$$

i. What is the idea beyond including these variables in the model?

ii. Is there an impact on the estimated coefficient $\hat{\beta}_1$? Does this signalize there likely was a problem with bias in the model from part (f)? Comment on the sign of this bias.

iii. Are both motheduc and fatheduc individually significant? Are they jointly significant? Check potential multicollinearity.

iv. What happens if you exclude one these variables from the regression? Which one would you keep?

(k) Compare the final models from parts (i) and (j). Which is a better model (based on the dataset in hand)? Try RESET again to potentially support your answer.