Probability, Statistics and Modelling II

Tutorial 5 – Regression diagnostics

We will continue with the 'Policing the pandemic' dataset from last week.

Constructs	*Variable*	*Item wording*	*Response options*
Duration it took for someone to finish the survey	duration	N/A	seconds
Confidence in the handling of the COVID-19 crisis	pm	Prime Minister	No confidence at all – A lot of confidence
	nhs	National Health Service	
	pol	Police	
	js	Justice System	
	gov	Government	
Coronavirus status	cov	Have you had Covid-19 (coronavirus)?	Yes, diagnosed and recovered ... Prefer not to say
Coronavirus attitudes	covknow	How would you rate your knowledge level on Covid-19?	Bad-Excellent
	covconc	How concerned are you about getting Covid-19?	Not concerned at all – Very concerned
	covidexp	How long do you expect it will be until the coronavirus outbreak is over and things are back to normal in the UK?	Less than 1 month - Never
Gender	gender	What is your gender?	Male/Female/Non-binary
	male	Binary variables created from gender	Male/Not
	female		Female/Not
	nonbin		Non-binary/Not
Age	age	Which of these age bands do you fall into?	16-24 – 65+
	age1	Binary variables created from age	16-24/Not
	age2		25-44/Not
	age3		45-64/Not
	age4		65+/Not
Area	area	Which city's metropolitan area do you live in?	Birmingham ... None of these
Ethnicity	ethnic	Please select the option which best describes your ethnic group:	Recoded to: Asian ... White
	asian	Binary variables created from ethnic	Asian/Not
	black		Black/Not
	mixed		Mixed/Not
	ethnico		Other ethnicity/Not
	white		White/Not
Key worker	keywork	Are you currently fulfilling any of the government's identified 'key worker' roles (listed below)?	Recoded to: Key worker/not

Table 1 Variables in the dataset

Please carry out the tasks and answer the questions below.

1.The behavioural science unit is concerned whether the final model for expectations regarding the pandemic from last week satisfies the required modelling assumptions[1]. You are asked to consider each of the following:

a) Homoskedasticity (i.e., that the residual variance is constant)
b) Linearity (i.e., the explanatory and outcome variables have a linear relationship)
c) Normality of the residuals (i.e., the residual variance is normally distributed)
d) (Multi)collinearity (i.e., there are no highly correlated variables which would have an undue influence on the results)
e) Influential outliers (i.e., the model is not thrown off by outliers)
f) Considering all these aspects, what kind of assurance can you provide regarding each of these? Were any of these assumptions violated? If they were, how would you fix them?

2.As the next step, carry out a similar analysis to what we did last week. This time, use confidence in the NHS's handling of the pandemic as the outcome variable, and the confidence in the prime minister, government, and the police as the explanatory variables. What do you find? Were any of the assumptions violated? If they were, how would you address them?

[1] Important side note: there are two further assumptions that need to be considered. The independence assumption will be revisited later in the course, while the correct model specification is a largely untestable assumption.