Quiz 4

2023-07-11

Question 1

You are making a regression model for a physical process that you know follows the form $y_i = A + Bx_i + \epsilon_i$. You wish to use regression to find values for A and B. However, you know the that the errors follow the relationship $\epsilon_i = \rho\epsilon_i - 1 + \omega_i$, where the ω_i are i.i.d. drawn from a normal distribution ($N(0, \theta^2)$) and $\rho \neq 0$

(a) What violation of the regression assumptions will this create? **Name the problem and draw a sketch** of a graph that might be used to diagnose the problem. Be sure to label your axes!

(b) Could a standardized residuals vs fitted values be used to diagnose this problem? Explain why or why not.

(c) Which of the following do you expect to be systematically affected i.e. biased if you fit the model without correcting this problem? Check all that apply:

i. Estimate of A
ii. Estimates of B
iii. Standard error of A
iv. Standard error of B
v. Predictions of Y
vi. Confidence Intervals for predictions

(d) To test if the violation in (a) is significant, we did runs test: we observed that there are 12 positive and 8 negative residuals and there are 8 runs in the residuals (parametrize with z-statistics). Perform you hypothesis test. Please specify your H_0 and H_A and perform the analysis. You may find the following statistic useful:

```
qnorm(0.05, mean = 0, sd = 1, lower.tail = TRUE) = -1.64
```

```
qnorm(0.025, mean = 0, sd = 1, lower.tail = TRUE)=-1.96
```

Question 2

You wish to fit a model with predictor variables Paper, Machine, and Labor. Consider the following result:

```
Variable: VIF:
```

```
Paper 15.5 Machine 3.5 Labor 6.0
```

(a) Is there any evidence of multicollinearity? Why or why not?

(b) What could you do if there is multicollinearity? Describe at least two strategies.

Question 3

You are trying to decide which variables to include in a model, and you create the following three fits. You may assume $n = 1,000$ for each of the models, and use $\alpha = 0.05$ for significance. Please note, however, that for several of the following questions I'm looking for reasoning which goes beyond simple tests of significance.

Table 1

Model 1:	$\hat{Y}_i = \underset{(0.11)}{0.324} + \underset{(0.47)}{1.84}\,X_{1i} + \underset{(0.71)}{5.43}\,X_{2i} - \underset{(0.75)}{0.85}\,X_{3i}$
Model 2:	$\hat{Y}_i = \underset{(0.10)}{0.354} + \underset{(0.46)}{1.85}\,X_{1i} + \underset{(0.70)}{5.41}\,X_{2i}$
Model 3:	$\hat{Y}_i = \underset{(0.05)}{0.222} + \underset{(0.46)}{0.76}\,X_{1i} - \underset{(0.74)}{0.87}\,X_{3i}$

[Note: standard errors appear in parentheses]

(a) Is the coefficient for X3 statistically significant in Model 1? Use a t-test. ($t_{1000,0.025} = 1.96$)

(b) What is the relationship between an F-test comparing Models 1 and 2 and the t-test you used to answer part (a)?

(c) Look at Models 1 and 3. What about these two models should make you concerned about a possible violation of regression assumptions? How would you check this if you had this data in R?

(d) Assuming you've determined that there is not a violation of regression assumptions, would you say that X2 a relevant variable in Model 1? Describe the strategy you want to use to make this decision.