**37252 Regression and Linear Models**

**Lab 7: Weighted Least Squares Regression**

This lab is marked out of 24.

Please save your file in PDF format with name

**37252\_Lab7\_Surname\_FirstName**

**Due: 12 noon Wednesday 1 May 2024**

In this week’s lab we use regression in a different way, this time to compare the accuracy of estimates of flock count made by two observers. The data are available in **37252\_Lab7\_data.csv** which can be downloaded from Canvas.

The variables we now consider are summarised in the table below.

|  |  |  |
| --- | --- | --- |
| **Name** | **Role** | **Description** |
| $$photo$$ | response | true flock count |
| $$obs1$$ | predictor | estimated flock count observer 1 |
| $$obs2$$ | predictor | estimated flock count observer 2 |

**Scenario.**

Aerial survey methods have been used to estimate the number of snow geese in their summer range areas west of Hudson’s Bay in Canada. Small aircraft fly over the range, and when a flock of geese is spotted an experienced observer estimates the number of geese in the flock. The method can obviously be used for other types of birds, and in Australia for the numbers of kangaroos or buffaloes in a herd. But is it a reliable method? One study investigated this by using two independent observers in the aeroplane, backing up their observations with a photograph. The photo was later used to obtain an accurate count of the number of birds.

1. Obtain scatter plots of $photo$ versus $obs1$ and of $photo$ versus $obs2$ (a matrix scatter plot will do). If you carried out a simple linear regression of $photo$ versus $obs1$ or $photo$ versus $obs2$, what problems would you expect **[1 mark]**? Why is it appropriate to take $photo$ as the response variable **[1 mark]**? Assuming high quality observers, why is simple linear regression more appropriate than a multiple linear regression using $obs1$ and $obs2$ as independent variables **[1 mark]**?

> photodat <- read.csv("~/2024\_37252/Labs/Lab7/37252\_Lab7\_data.csv")

Fit a simple linear regression model of $photo$ versus $obs1$ and (separately) of $photo$ versus $obs2$.

1. What problems do the residual scatter plots show about the fit of the simple linear regression models **[1 mark]**? What action could you take **[1 mark]**?

Since the variability in the residuals seems to be increasing proportionally with the independent variable, we can try WLS.

Let $\hat{ε1}\_{i}$ represent the residuals from the simple OLS model with $obs1$ as predictor, and $\hat{ε2}\_{i}$ the residuals from the model with $obs2$ as predictor.

Suppose

$$var\left(\hat{ε1}\_{i}\right)=obs1\_{i}σ^{2}$$

and

$$var\left(\hat{ε2}\_{i}\right)=obs2\_{i}σ^{2}.$$

1. What would the weights be for the WLS regression of $photo$ versus $obs1$ **[1 mark]**? What would we multiply the data by if we wish to transform the model directly **[1 mark]**?

**Running a WLS regression**

Before we can run the WLS regression models we need to create the weight variables.

> wt1 <- 1/photodat$obs1

> mod2\_obs1 <- lm(photo ~ obs1, data = photodat, weights = wt1)

> summary(mod2\_obs1)

> resid1\_WLS <- mod2\_obs1$residuals\*sqrt(wt1)

> st.resid1\_WLS <- (resid1\_WLS - mean(resid1\_WLS))/sd(resid1\_WLS)

> plot(photodat$obs1, resid1\_WLS, xlab = "Obs1", ylab = "Standardised WLS residuals")

> wt2 <- 1/photodat$obs2

> mod2\_obs2 <- lm(photo ~ obs2, data = photodat, weights = wt2)

> summary(mod2\_obs2)

> resid2\_WLS <- mod2\_obs2$residuals\*sqrt(wt2)

> st.resid2\_WLS <- (resid2\_WLS - mean(resid2\_WLS))/sd(resid2\_WLS)

> plot(photodat$obs2, resid2\_WLS, xlab = "Obs2", ylab = "Standardised WLS residuals")

1. For both WLS models, analyse the Student-T version of the standardised, weighted residuals using scatter plots involving the independent variables. Has weighting improved the behaviour of the residuals **[2 marks]**?
2. For both WLS models, describe the results of the two-sided T-test with null hypothesis $β\_{0}=0$ **[2 marks]**. Explain, in the context of these models, why we are interested in such tests **[2 marks].**
3. For both WLS models, perform a two-sided T-test with null $β\_{1}=1$. Write down the null and alternative hypotheses **[2 marks]**, the test-statistic **[2 marks]** and the result of the test with reason **[2 marks]**. Explain, in the context of these models, why we are interested in such tests **[2 marks]**.
4. With reference to your answers in (e) and (f), which observer should be preferred **[1 mark]**? Why **[2 marks]**?