**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")

> pairs(~ photo + obs1 + obs2, data = photodat)



In both cases the variation is increasing. This increasing variance in the data will be reflected in the fitted model residuals (with respect to the predictors), contrary to the assumptions of OLS **[1 mark]**.

The observer’s *estimate* is used to gauge the *true* count, i.e. *estimate* → *true*. In a modelling sense this is *estimate* as predictor, *true* as response allowing testing of the significance of *estimate* in this role **[1 mark]**.

If we did model with two observers who were good at their jobs, we should expect high correlation between both predictors and potential multi-collinearity problems **[1 mark]**.

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

> mod1\_obs1 <- lm(photo ~ obs1, data = photodat)

> mod1obs1.st.resid<-rstandard(mod1\_obs1)

> plot(photodat$obs1, mod1obs1.st.resid, xlab = "Obs1", ylab = "Standardised residuals")

> mod1\_obs2 <- lm(photo ~ obs2, data = photodat)

> mod1obs2.st.resid<-rstandard(mod1\_obs2)

> plot(photodat$obs2, mod1obs2.st.resid, xlab = "Obs2", ylab = "Standardised residuals")

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]**?



In both models there appears to be increasing variance in the residuals as the independent variable increases **[1 mark]**.

We could solve this by fitting a weighted regression model (or by transforming the data appropriately) **[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]**?

To use the R WLS feature, use the weights $w\_{i}=\frac{1}{obs1\_{i}}$ for $i\in \{1,2\}$ **[1 mark]**.

To weight the data and use OLS, multiple the $i$-th observation by $\sqrt{w\_{i}}$ (including the “1” term for the intercept parameter) **[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]**?



For both models, weighting has removed the increasing variance **[2 marks]**.

1. 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].**



> summary(mod2\_obs1)

Call:

lm(formula = photo ~ obs1, data = photodat, weights = wt1)

Weighted Residuals:

 Min 1Q Median 3Q Max

-10.6968 -1.9668 -0.1675 2.7977 7.2061

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.93337 4.47915 1.994 0.0525 .

obs1 1.14451 0.09255 12.366 9.4e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.831 on 43 degrees of freedom

Multiple R-squared: 0.7805, Adjusted R-squared: 0.7754

F-statistic: 152.9 on 1 and 43 DF, p-value: 9.402e-16

> summary(mod2\_obs2)

Call:

lm(formula = photo ~ obs2, data = photodat, weights = wt2)

Weighted Residuals:

 Min 1Q Median 3Q Max

-5.0823 -1.7633 -0.1905 1.6731 8.4908

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.8330 3.5455 2.773 0.00817 \*\*

obs2 0.8403 0.0571 14.718 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.851 on 43 degrees of freedom

Multiple R-squared: 0.8344, Adjusted R-squared: 0.8305

F-statistic: 216.6 on 1 and 43 DF, p-value: < 2.2e-16

**Hypothesis tests**

For obs2 we can reject the null hypothesis $β\_{0}=0$ but the result is borderline insignificant for obs1. **[2 marks]**.

**Why this test?**

If the observers had perfect accuracy the population models would be

$$photo=0+1×obs+ϵ.$$

Therefore we hope to retain the null $β\_{0}=0$ so that when $obs=0$ we have $photo=0+ϵ$. If this is not the case then the observers are making some linear bias in their estimates **[2 marks]**.

1. 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]**.

**Model with** $obs1$

**Hypotheses**

H0: $β\_{1}=1$

HA: $β\_{1}\ne 1$

**[1 mark]**

**Test statistic and p-value**

Test stat $t^{\*} = \frac{1.145-1}{0.093}≈1.56$.

P-value: $p≈0.126.$

**[1 mark]**

**Test result**

Since $p>α$ we retain the null hypothesis **[1 mark]**.

**Model with** $obs2$

**Hypotheses**

H0: $β\_{1}=1$

HA: $β\_{1}\ne 1$

**[1 mark]**

**Test statistic and p-value**

Test stat $t^{\*} = \frac{0.840-1}{0.057}≈-2.80$.

P-value: $p≈0.007$

**[1 mark]**

**Test result**

Since $p<α$ swe reject the null hypothesis **[1 mark]**.

**Why this test?**

If the observers had perfect accuracy the population models would be

$$photo=0+1×obs+ϵ.$$

Therefore we hope to retain the null $β\_{1}=1$ so that when $obs=x$ we have $photo=x+ϵ$. If this is not the case then the observers are making some multiplicative bias in their estimates **[2 marks]**.

1. With reference to your answers in (e) and (f), which observer should be preferred **[1 mark]**? Why **[2 marks]**?

Observer 1 **[1 mark]**.

The hypothesis $β\_{0}=0$ was rejected for obs1 but borderline insignificant for obs2. **[1 mark]**.

The hypothesis $β\_{1}=1$ was also rejected for the model with $obs2$ as predictor, and retained for the model with $obs1$ as predictor. So Observer 1 should be preferred **[1 mark]**.