STAT3015/4030/7030 Generalised Linear Modelling

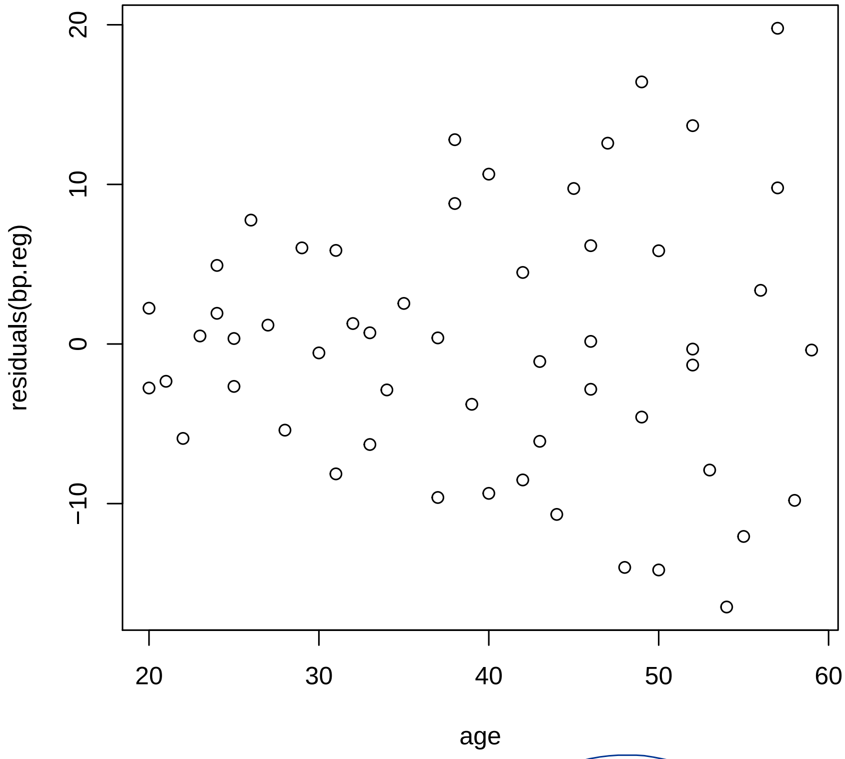
Tutorial 5

1. A study was conducted to examine the relationship between age and blood pressure. The data for 54 healthy adult women are contained in the data file `BP.txt`, the first column containing ages and the second containing diastolic blood pressures.

   (a) Fit a simple linear regression to this data and examine the residuals versus predictor plot. What do you notice?

   Solution: The following commands do the job.

```
> bp <- read.table("BP.txt", header=TRUE)
> names(bp); attach(bp)

[1] "age"    "diasbp"

> bp.reg <- lm(diasbp~age)
> coef(bp.reg)

(Intercept)         age
 56.1569294   0.5800308

> summary(bp.reg)$sigma^2

[1] 66.35317

> plot(age, residuals(bp.reg))
```



A plot of the residuals clearly shows heteroscedasticity in the data.

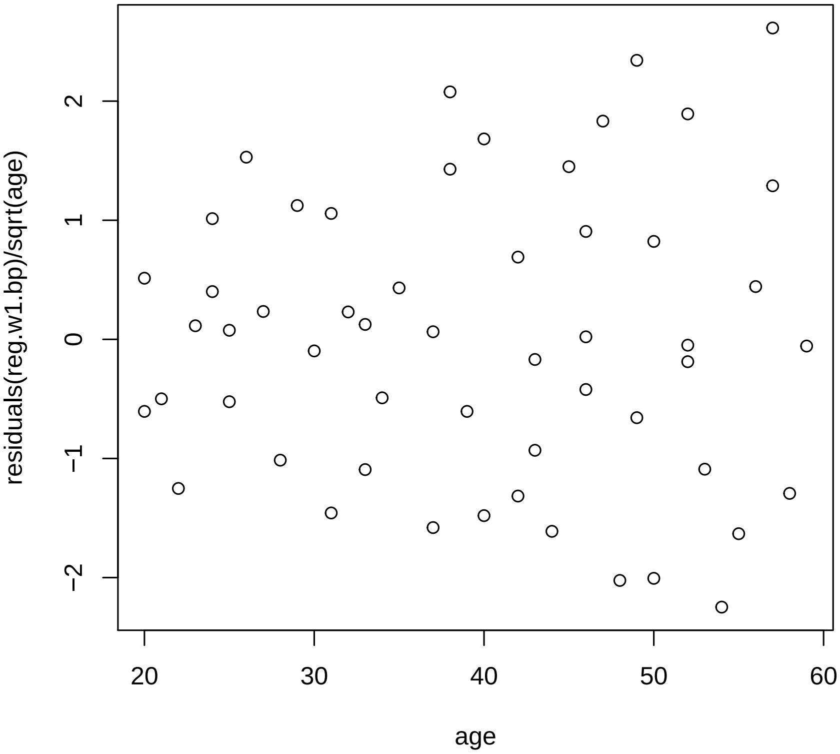
(b) To account for heteroscedasticity, it is sometimes suggested to fit the model:

$$\begin{aligned} Y_i =& \beta_0 + \beta_1 x_i + \epsilon_i \\ \epsilon_i \sim& \mathcal{N}(0, \sigma_i^2) \end{aligned}$$

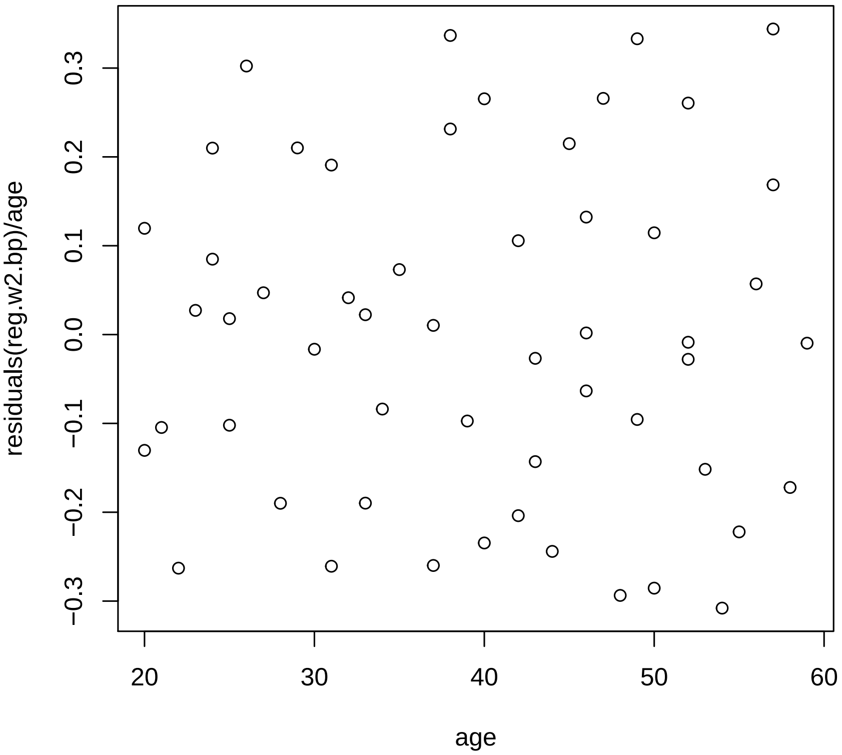
where $\sigma_i^2 = \sigma^2 / w_i^2$ and the w_i^2's being weights used to indicate how the variance of the data changes from data point to data point. For this data, we might try using $w_i = 1/\sqrt{x_i}$ or $w_i = 1/x_i$. Why might these be appropriate choices? Fit weighted regressions using both of these suggestions. How do the parameter estimates (including the estimate of σ^2) change? Plot the weighted residuals, $w_i e_i$, versus predictor values for each weighting scheme. Which set of weights seems more appropriate to this dataset?

Solution: The first of the suggested weights implies that $\sigma_i^2 = x_i \sigma^2$ while the second suggestion leads to $\sigma_i^2 = x_i^2 \sigma^2$. In each case, this implies that the variability of the response is increasing as the value of the predictor increases, which is what the residual plot in part (a) suggested. To fit the weighted regression:

```
> wgt1 <- 1/age
> wgt2 <- 1/age^2
> reg.w1.bp <- lm(diasbp~age, weights=wgt1)
> coef(reg.w1.bp)

(Intercept)         age
 56.0499634   0.5827337

> summary(reg.w1.bp)$sigma^2

[1] 1.491525

> anova(reg.w1.bp)

Analysis of Variance Table

Response: diasbp
          Df Sum Sq Mean Sq F value   Pr(>F)
age        1 66.951  66.951  44.888 1.49e-08 ***
Residuals 52 77.559   1.492
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> plot(age, residuals(reg.w1.bp)/sqrt(age))
```



```
> reg.w2.bp <- lm(diasbp~age, weights=wgt2)
> coef(reg.w2.bp)

(Intercept)         age
  55.831037    0.588828

> summary(reg.w2.bp)$sigma^2

[1] 0.03578945

> anova(reg.w2.bp)

Analysis of Variance Table

Response: diasbp
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1 1.8644 1.86441  52.094 2.226e-09 ***
Residuals 52 1.8611 0.03579
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> plot(age, residuals(reg.w2.bp)/age)
```



Clearly, the parameter estimates for the β's do not change dramatically, however, the estimate of σ^2 does change drastically. This should not be surprising as σ^2 has a somewhat new interpretation in the weighted models. In addition, the weighted residual plots show that the second set of weights appear to be achieving homoscedasticity, while the first set are not quite enough. [NOTE: The sums of squares in an ANOVA table for a weighted regression are actually a weighted sums of squares, i.e., the MSE is calculated as $\frac{1}{n-p}\sum_{i=1}^{n} w_i^2(Y_i - \widehat{Y}_i)^2$, so that it estimates σ^2. Calculating the value $\frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ for a weighted regression does not make sense any longer, since the variances of the individual data points are changing, and thus should not be averaged together.]

(c) Why must we use the weighted residuals rather than the ordinary residuals from the weighted regressions in our plots to assess which weights are the most appropriate? [HINT: Recall that the variance of the ordinary residuals, e_i, is approximately the same as the variance of the error random variable ϵ_i; namely σ_i^2.]

Solution: Suppose that the true variance structure of the error random variables is $\mathbb{V}\epsilon_i = \sigma_i^2 = \sigma^2/w_i^2$. Then, from the hint, we note that regardless of the model fit, $\mathbb{V}e_i \approx \sigma^2/w_i^2$, which are clearly not equal. Thus, a plot of the ordinary residuals from any weighted regression model would be expected to show a fanning shape, even if we get the weights correct! Conversely, given some weights w_i^*, we have $\mathbb{V}(w_i^* e_i) = (w_i^*)^2\,\mathbb{V}e_i \approx (w_i^*)^2\,(\sigma^2/w_i^2)$ which are all equal to σ^2, provided we have chosen the proper weights $w_i^* = w_i$. Thus, if we choose the proper weights, then the weighted residual plot should have no fanning shape. Of course, if we choose improper weights, then even the weighted residual plot will exhibit a fanning shape.

2. The potency of an anaesthetic agent is measured in terms of the minimum concentration at which at least 50% of patients exhibit no response to stimulation. Thirty patients are administered a particular anaesthetic at various predetermined concentrations for 15 minutes before a stimulus was applied. The response variable was simply an indicator as to whether the patient responded or not. A GLM model with binomial error structure and link function

$$p(x_i) = g^{-1}(\beta_0 + \beta_1 x_i)$$

was fit to predict the probability of response (p) given the level of anaesthetic (x_i) where g was either the probit, logistic, or complementary log-log function. The focus of the experiment which gathered this data was to find the concentration value x at which the probability of responding to the stimulus was 50%, i.e., to estimate the value of x which satisfies the equation $p(x_i) = 0.5$. The table below gives the coefficient estimates for the three different link functions:

	$\widehat{\beta}_0$	$\widehat{\beta}_1$
Probit	3.8579	−3.3245
Logit	6.4685	−5.5676
Comp log-log	3.7316	−3.6370

Estimate the 50%-response concentration level for each of the three different functions. Does our choice of link function influence our estimate in this problem?

Solution: Solving the equation $p(x) = g^{-1}(\beta_0 + \beta_1 x) = 0.5$ shows that

$$\beta_0 + \beta_1 x = g(0.5) \Rightarrow x = \frac{g(0.5) - \beta_0}{\beta_1}.$$

Now, for the probit link, we see that $g(0.5) = \Phi^{-1}(0.5) = 0$, since the median of the standard normal distribution is the origin. Therefore, for this link, our estimate of the 50%-response concentration is:

$$x = \frac{0 - \widehat{\beta}_0}{\widehat{\beta}_1} = \frac{0 - 3.8579}{-3.3245} = 1.1604.$$

Similarly, for the logistic link, we have

$$g(0.5) = \log\left(\frac{0.5}{1 - 0.5}\right) = \log(1) = 0.$$

So, the estimate of the 50%-response concentration for this model is:

$$x = \frac{0 - \widehat{\beta}_0}{\widehat{\beta}_1} = \frac{0 - 6.4685}{-5.5676} = 1.1618.$$

Thus, although the parameter estimates are quite different for these two models (which is to be expected since they have different meanings and interpretations in each model), the estimates of the 50%-response concentration are almost identical. Finally, for the complementary log-log link, we have:

$$g(0.5) = \log(-\log(1 - 0.5)) = -0.3665.$$

leading to a 50%-response concentration estimate of:

$$x = \frac{-0.3665 - 3.7316}{-3.6370} = 1.1268.$$

Notice that this estimate does differ slightly from those of the previous two models.