Stat 414 — Day 4
Violating Unequal Variance

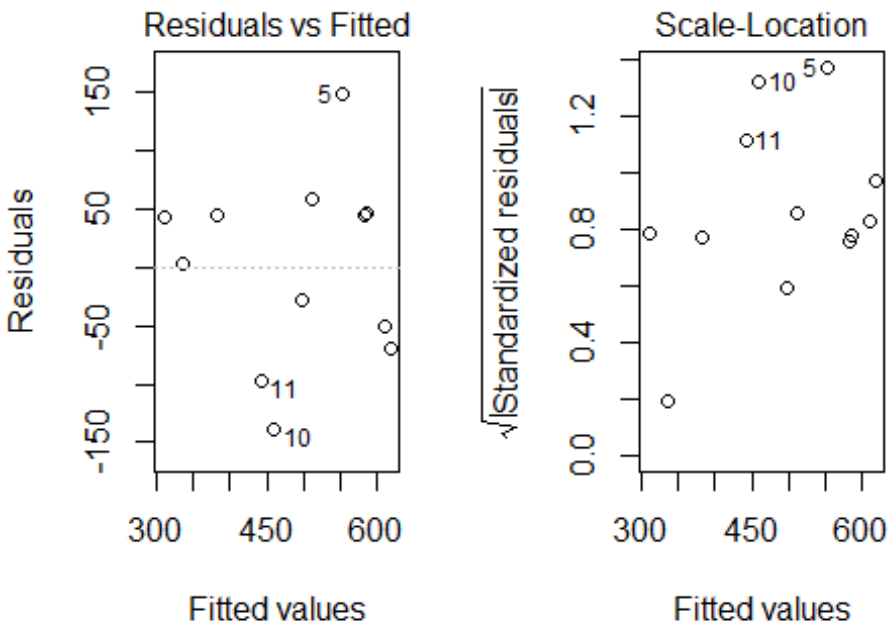
Last Time

- When the basic regression model assumptions are met, the distribution of slopes is approximately normal with mean equal to the population slope and standard deviation $\sigma/\left(SD(X)\times\sqrt{n-1}\right)$ (for simple regression, this formu1la is more complicated with multiple predictors and will depend on the level of correlation in those predictors, e.g., $Var\left(\hat{\beta}\right)=\sigma^2(X'X)^{-1}$). This means we can also use t-tests, F-tests, confidence intervals for slopes, and confidence/prediction intervals for future observations.
- F-test for one variable is equivalent to the t-test, and more generally is a "drop in SSError" comparison between a "full" model and a (nested) "reduced" model.
- When using ML or REML, we instead use likelihood ratio tests and intervals for testing our parameter estimate. REML is better to use when you want to focus on estimates of the "random part" of the model (e.g., σ = population variance about the regression line).
- $2(L_1-L_0)$ follows a chi-square distribution with df = difference in number of parameters in the two models

Example 1: Review Scale-Location Plot

Reconsider the residual plots from the airfare data (n = 12).

```
airfare = read.table("https://www.rossmanchance.com/stat414/data/airfare.txt", head
er=TRUE)

model1 <- lm(price ~ distance, data = airfare)
par(mfrow=c(1,2))
plot(model1, which = c(1, 3), add.smooth=FALSE)
```



While we look for "fanning" or "megaphones" (etc) in the residual vs. fits graph, we look more for an increasing or decreasing trend in the Location-Scale plot - are the 'magnitudes' of the residuals increase or decreasing with the fitted values.

For the airfare data, we saw (CP 2) a little bit of evidence that the variability in the prices increases with the distances traveled, which makes some sense in context. The main consequence is that we might want to specify wider prediction intervals for large distances than for smaller distances. But the unequal variance can also throw off our estimation of σ which then impacts our slope p-values and confidence intervals as well.

We can follow-up our visual inspection of the residual plots with a more formal test (as long as the linearity, independence, and normality assumptions are met!).

Definition

The *Breusch-Pagan Test* is a likelihood ratio type test to assess the linear relationship between e_i^2 and x_i. If the p-value is small, reject the null hypothesis of homoscedasticity. (df = number of slopes in the model)

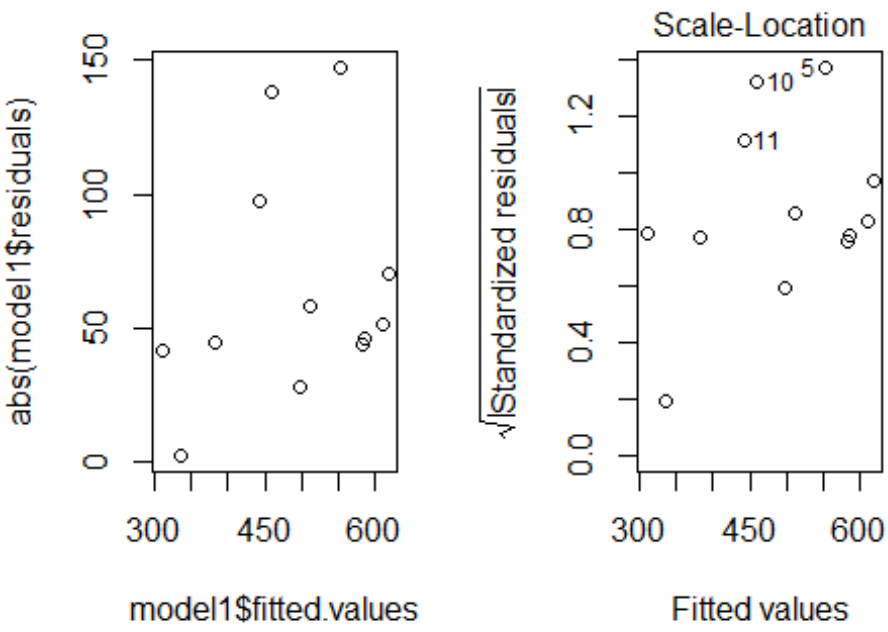
```
#install.packages("lmtest")
lmtest::bptest(model1)

    studentized Breusch-Pagan test

data:  model1
BP = 0.21, df = 1, p-value = 0.6
```

(a) What does the Breusch-Pagan Test reveal about the airfare data?

The variability is not increasing significantly with the distances

This test is essentially looking at the significance of the relationship between the magnitude of the residuals and the fitted values…. so why not just regress the residuals on the fitted values to see whether that linear association is significant?

```
rvar <- lm(I(model1$residuals^2) ~ model1$fitted.values); summary(rvar)

Call:
lm(formula = I(model1$residuals^2) ~ model1$fitted.values)

Residuals:
   Min     1Q Median     3Q    Max
 -5103  -4431  -2707   -534  15293

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           1315.25   10735.50    0.12     0.90
model1$fitted.values     9.15      21.42    0.43     0.68

Residual standard error: 7530 on 10 degrees of freedom
Multiple R-squared:  0.0179,    Adjusted R-squared:  -0.0803
F-statistic: 0.182 on 1 and 10 DF,  p-value: 0.678
rsd <- lm(abs(model1$residuals) ~ model1$fitted.values); summary(rsd)

Call:
lm(formula = abs(model1$residuals) ~ model1$fitted.values)

Residuals:
   Min     1Q Median     3Q    Max
-46.45 -27.89  -8.27   5.87  77.23

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           16.2157    62.6144    0.26     0.80
model1$fitted.values   0.0974     0.1249    0.78     0.45

Residual standard error: 43.9 on 10 degrees of freedom
Multiple R-squared:  0.0574,    Adjusted R-squared:  -0.0369
F-statistic: 0.608 on 1 and 10 DF,  p-value: 0.453
#The graph for the last model is essentially the Scale-Location plot!
par(mfrow=c(1,2))
plot(abs(model1$residuals)~model1$fitted.values)
plot(model1, which = c(3), add.smooth=FALSE)
```




```
#The Scale-Location plot uses the square root of the absolute residuals to mitigate
the impact of extreme values to make a trend more easy to detect. Using the standar
dized residuals also uses a common scale/accounts for leverage values.
```

In fact, we could then use the fitted values from these regressions as our estimates of the variances (or standard deviations) of the observations.

Instead, we will focus on four different options:

1. Transformations

2. Weighted Least Squares

3. Specifying variance covariates

4. Robust standard errors

In other words, rexpress the data in a scale that has equal variances (1) or we can *model* the variation in the variances (2 and 3) or at least fix the issue with the slope standard errors (4).

Transformations

For transformations, keep in mind that transforming predictor variables won't help the unequal variances, but transforming the response may help or hinder the normality and linearity assumptions. Choosing a transformation is often a trial-and-error process, re-examining the plots and BP-test to see whether the conditions are better met. But there is an ordering to the impact of different transformations (e.g., a log transformation is a more severe transformation

than a square-root transformation). Also keep in mind that computerized optimization should not replace subject-matter-expert intuition or considerations of interpretibility and parsimony.

Weighted Least Squares

The main idea of weighted least squares is to not treat all the observations equally when estimating the regression coefficients, but to "downweight" observations that are less reliable/have more variability. To find the new parameter estimates, we slightly alter the ordinary least squares process. Just remember to evaluate the assumptions using the "weighted residuals."

Definition

Weighted least squares assumes $V(Y_i|x_i) = \sigma^2/w_i$ and if we know the w_i then we minimize $\Sigma w_i(Y_i - \beta_0 - \beta_1 X_i)^2$.

Our parameter estimates may not change much, but our estimate of the (one) σ^2 may change quite a bit. The weights will also impact the prediction intervals.

Specifying Variance Covariates

We can use more interesting models of the variance, e.g., $V(Y_i|x_i) = \sigma^2 \times distance_i$. We will see how to do this using "generalized least squares" where we can indicate a structure to the variance-covariance matrix when specifying the model. Weighted least squares is a special case.

Robust Standard Errors

Let's expand our formula a bit, $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}X'\Sigma X(X'X)^{-1}$. This simplifies to the above equation when we use $\sigma^2 I$ in the middle, but now we can specify other values along the diagonal of Σ, the variance-covariance matrix of the residuals. In particular, we can estimate the variances of errors from the residuals when we don't have natural covariates in mind. The HC0 method (White, 1980) uses the squared residuals.

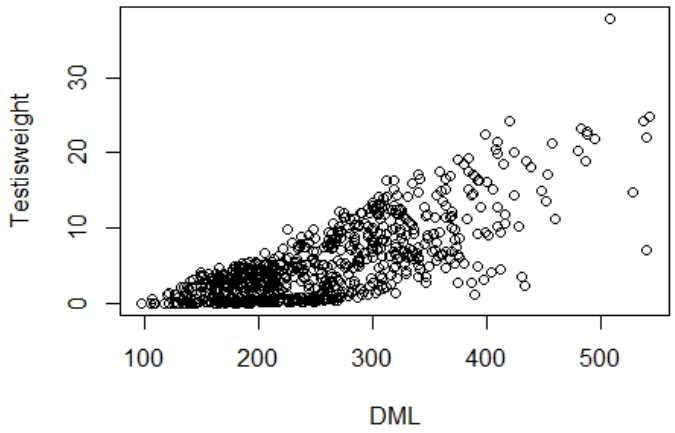
- HC1: scales the residuals by the df (Huber-White)
- HC2: scales the residuals by the leverage values $(1 - h_{ii})$
- HC3: scales the residuals by $(1 - h_{ii})^2$

The main idea is you have taken into account the heteroscedasticity in the coefficient standard errors without having to know about or model the functional form of the heteroscedasticity or use "arbitrary" transformations.

Example 2: Squid

Smith et al. (2005) examined reproductive and somatic tissues in the squid *Loligo forbesi*. The data in Squid.txt include the dorsal mantel length (in mm) and testis weight from 768 male squid, over different months. ''The idea behind the original analysis was to investigate the role of endogenous and exogenous factors affecting sexual maturation, more specifically to determine the extent to which maturation is size-related and seasonal'' (Zuur et al., 2009).

```
Squid<-read.table("http://www.rossmanchance.com/stat414/data/Squid.txt",header=T)
plot(Testisweight~DML, data = Squid)
```



(a) How does Testis weight appear to change with DML? For which DML values do we have less variable measurements of Testis weight?

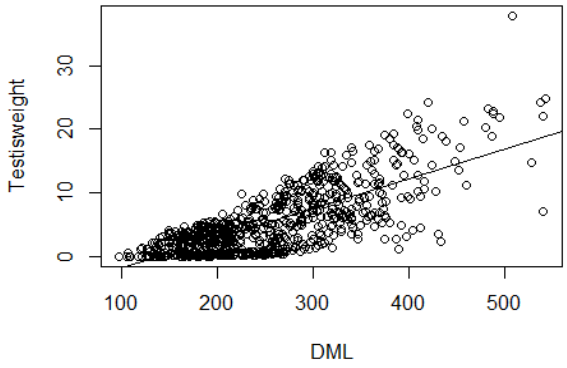
positive association though smaller DML values have less variation in Testis weight measurements

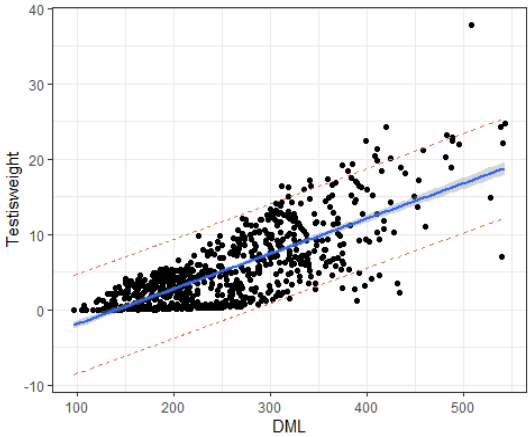
Fit an OLS model to predict testis weight from dorsal mantel length.

```
Call:
lm(formula = Testisweight ~ DML, data = Squid)

Residuals:
   Min     1Q Median     3Q    Max
-11.55  -2.28   0.16   2.08  20.64

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.53423    0.39259   -16.6   <2e-16 ***
DML          0.04666    0.00147    31.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.35 on 766 degrees of freedom
Multiple R-squared:  0.566, Adjusted R-squared:  0.566
F-statistic: 1e+03 on 1 and 766 DF,  p-value: <2e-16
```

PI short-cut: $\hat{y} \pm 2 \times s = -6.54 + .04666(200)(= 2.79) \pm 2(3.352) = (-3.91, 0.49)$

Visualizations:





(c) How do we interpret the confidence interval at DML = 200? How do we interpret the prediction interval at DML = 200? Why are the bands 'curved'?

CI: 95% confident the population mean testisweight for all squid with dml = 200 is in the interval

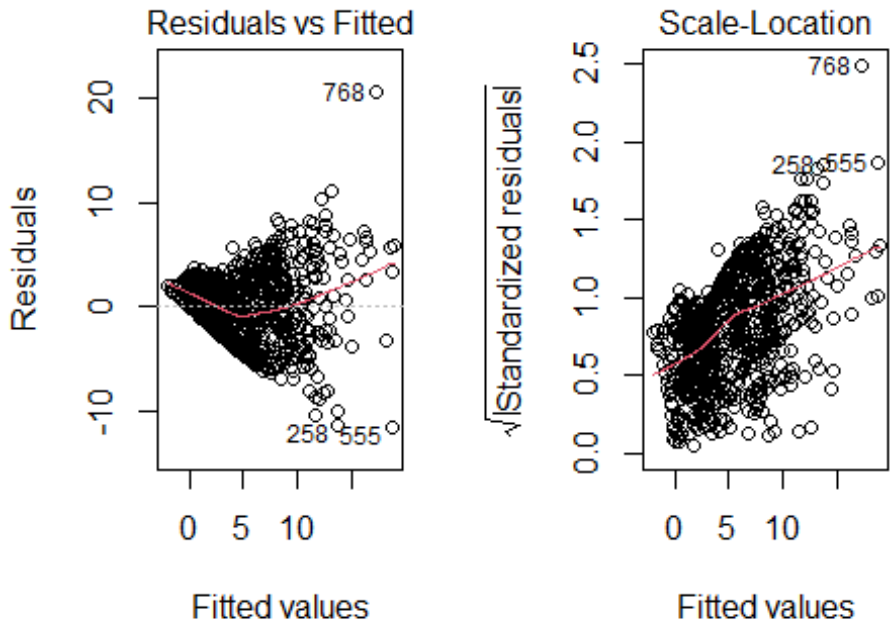
PI: 95% confident the testisweight of a squid with dml = 200 is in the interval

The bands are curved because of the 'bowtie' shape we saw to different sample regression lines there will be more 'uncertainly' at x-values further away from x-bar

(d) So what's the problem?

ⓘ Code

```
par(mfrow=c(1,2))
plot(model1, which = c(1, 3))
```




```
lmtest::bptest(model1)

    studentized Breusch-Pagan test

data:  model1
BP = 155, df = 1, p-value <2e-16
```

These interval procedures are not valid because we have violated equal variance!

(e) Why might a transformation not **be helpful here?**

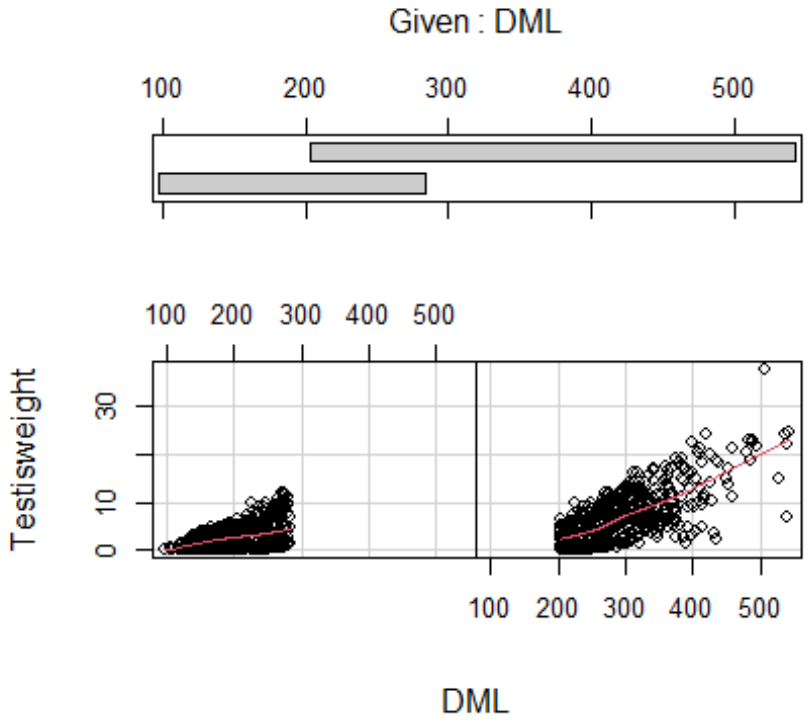
Because other (distributional) assumptions are mostly ok (linearity, normality) and transformations on the response variable (to fix equal variance) could make them no longer ok.

Because the variability appears to increase with the fitted values, which are positively related to our predictor variable, we could consider weighted regression using $w_i = 1/DML_i$ which will give more "weight" in the least squares estimation to squid with smaller DML values.

(f) *Prediction*: What should be the impact of using these weights on the least squares estimate of the "effect" of DML?

The estimated regression line will pay more attention to the data with smaller DML which had a shallower association (smaller slope - if you think there is some nonlinearity here) and so we predict a smaller slope with the WLS model

```
#Hint
coplot(Testisweight ~ DML | DML, data = Squid, num = 2, panel = function(x,y,...){
panel.smooth(x,y)      })
```



Manually fitting a weighted least squares model:

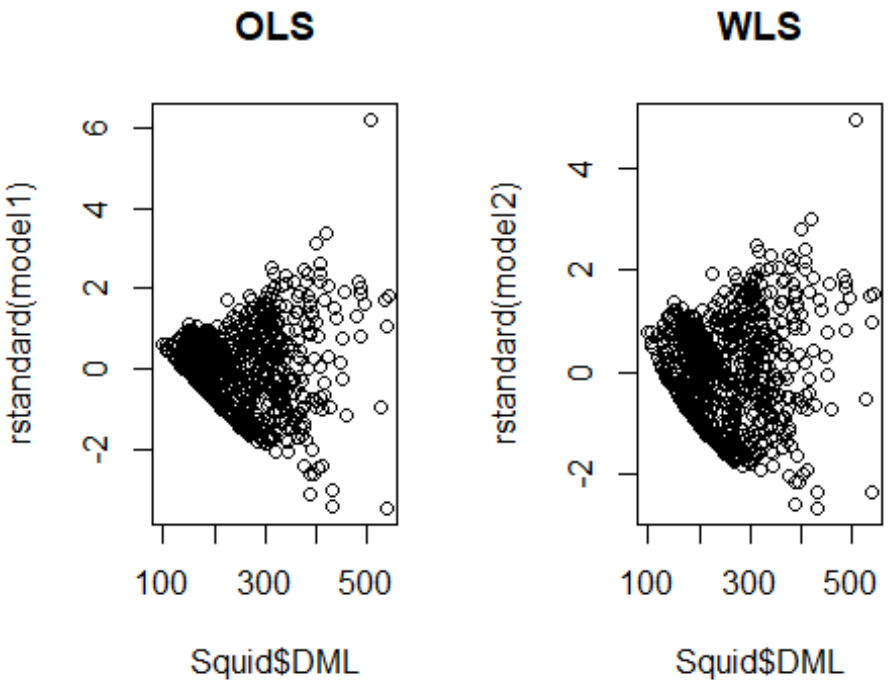
```
model2 = lm(Testisweight ~DML, data = Squid , weights = 1/DML)
summary(model2)

Call:
lm(formula = Testisweight ~ DML, data = Squid, weights = 1/DML)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-0.5164 -0.1457  0.0025  0.1381  0.9565

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.62394    0.33829   -16.6   <2e-16 ***
DML          0.04307    0.00141    30.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.194 on 766 degrees of freedom
Multiple R-squared:  0.55,  Adjusted R-squared:  0.55
F-statistic:  938 on 1 and 766 DF,  p-value: <2e-16
```

(g) How did the slope coefficient change? Is this what you predicted? Did R^2 and $\hat{\sigma}$ change? What do these numbers now represent? What about $SE(\hat{\beta})$?

The slope is smaller as predicted but only a little (.0466 to .043). Maybe don't compare R^2 values as we have kinda changed the response here but $\hat{\sigma}$ is a lot smaller (3.35 to .194) because now it represents the unexplained variation when DML is 1. The standard error for the slopoe coefficient has barely changed.

To see whether this has sufficiently addressed the heterogeneity we saw in the residuals, we want to look again at the residual plots. However, with weighted least squares, we need to look at the *standardized* residuals rather than the non-standardized residuals. You can think of these like z-scores, though there are different versions that divide by slightly different SD(residual).

```
par(mfrow=c(1,2))
plot(rstandard(model1)~Squid$DML, main = "OLS")
plot(rstandard(model2)~Squid$DML, main = "WLS")
```



(h) Have things improved? Be clear how you are deciding.

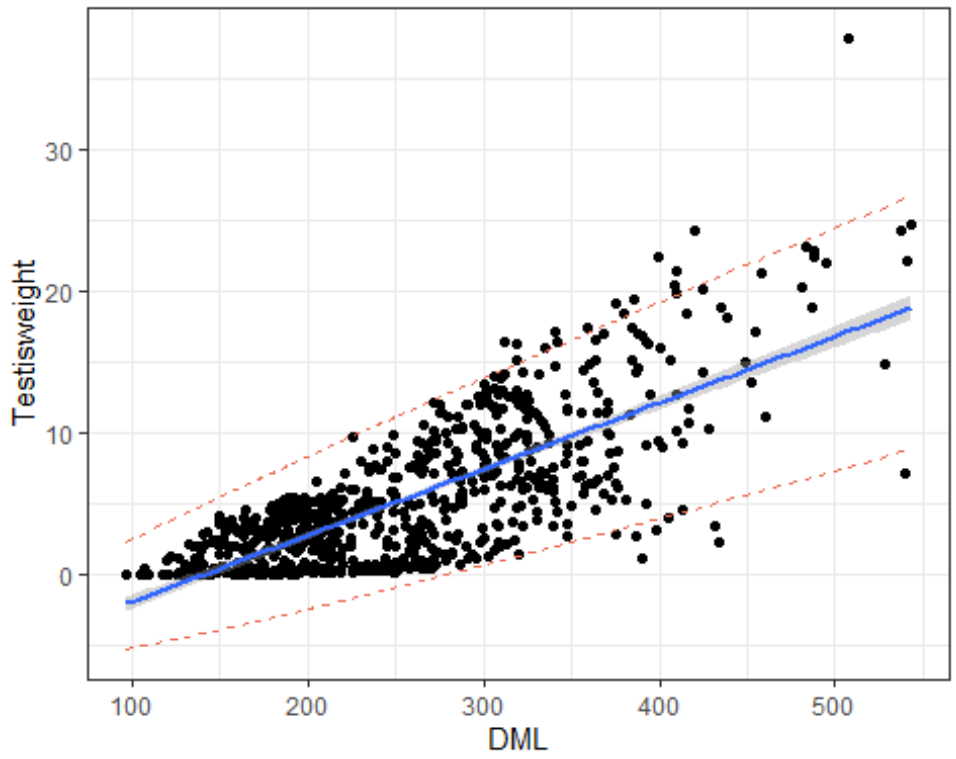
Still a fair bit of fanning but less?

What about the prediction bands?

ⓘ Code

```
predictions <- predict(model2, newdata = data.frame(DML = Squid$DML), interval =
"prediction", weights = 1/Squid$DML)
```

```
alldata <- cbind(Squid, predictions)

ggplot(alldata, aes(x = DML, y = Testisweight)) + #define x and y axis variables
  geom_point() + #add scatterplot points
  stat_smooth(method = lm) + #confidence bands
  geom_line(aes(y = lwr), col = "coral2", linetype = "dashed") + #lwr pred interv
al
  geom_line(aes(y = upr), col = "coral2", linetype = "dashed") + #upr pred interv
al
  theme_bw()
```



Clearly there seems to be more going on (see CP 4). Now try the "sandwich estimators" for producing "robust" ("heterogeneity corrected") standard errors. The second line produces the "variance-covariance matrix for the estimated coefficients" - the values on the diagonal of vcovHC are the "corrected" $Var(\hat{\beta})$ values. Then these are used in the regression table.

```
library(sandwich)
sqrt(diag(vcovHC(model1, "HC1"))) # HC1 gives us the White-Huber standard errors
(Intercept)         DML
   0.476599    0.002135
library(lmtest)
coeftest(model1, vcov = vcovHC(model1, type = "HC1"))   #updates the significance t
ests

t test of coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.53423    0.47660   -13.7   <2e-16 ***
DML          0.04666    0.00213    21.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(i) How do the standard errors of the regression coefficients change? Does the statistical significance of any of the variables change? (If not, then can claim analysis was not being affected by the heterogeneity.)

In this case, things don't change too much, but they could! DML SE is now .0021 but still statistically significant.

Notes

- "The easiest solution is a data transformation, but we try to avoid this for as long as possible. In our view, heterogeneity is interesting ecological information that you should not throw away, just because it is statistically inconvenient. With a 'little' bit of extra mathematical effort, heterogeneity can be incorporated into the models and can provide extra biological information." (Zuur et al.)
- When heteroscedasticity is discovered, we should not simply ask "What can I do to make the problem go away?" without also asking "What does heteroscedasticity tell me about the process I am studying?" (Hayes & Cai, 2007).
- Keep in mind that non-constant variance could be due to a misspecified model (e.g., missing key predictors, interactions, or non-linear effects).
- Which model is better? Then one that has the right variance structure! But we often don't know that in advance… "The best we can do is apply theory and judgment in a thoughtful way."

Computer Problem 4:

Due ideally by Friday but it's longer and worth more so if you need more time that's fine.

Let's use *generalize least squares* (Aiken, 1934) to try several different variance structures. Because we are focusing on the variance component rather than the "fixed" part of the model (the coefficients), we will use REML.

Model 1: $V(Y_i) = \sigma^2$

```
library(nlme)
model1REML <- gls(Testisweight ~ DML, data = Squid, method = "REML")
summary(model1REML)
Generalized least squares fit by REML
  Model: Testisweight ~ DML
  Data: Squid
   AIC  BIC logLik
  4055 4069  -2025

Coefficients:
             Value Std.Error t-value p-value
(Intercept) -6.534    0.3926  -16.64       0
DML          0.047    0.0015   31.64       0

 Correlation:
    (Intr)
DML -0.951

Standardized residuals:
     Min       Q1      Med       Q3      Max
-3.44695 -0.67972  0.04775  0.61890  6.15753

Residual standard error: 3.352
Degrees of freedom: 768 total; 766 residual
logLik(model1REML)
'log Lik.' -2025 (df=3)
```

(a) How many parameters are being estimated by the model? Identify them.

Use the nlraa package to examine the *estimated* variance-covariance matrix for this model for the first five observations (specimen 1017, 1034, 1070, 1070, 1019)

```
#install.packages("nlraa")
library(nlraa)
head(Squid, 5)
  Specimen YEAR MONTH DML Testisweight
1     1017 1991     2 136        0.006
2     1034 1990     9 144        0.008
3     1070 1990    12 108        0.008
4     1070 1990    11 130        0.011
5     1019 1990     8 121        0.012
```

```
vcmatrix1 = nlraa::var_cov(model1REML); vcmatrix1[1:5, 1:5]
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 11.24  0.00  0.00  0.00  0.00
[2,]  0.00 11.24  0.00  0.00  0.00
[3,]  0.00  0.00 11.24  0.00  0.00
[4,]  0.00  0.00  0.00 11.24  0.00
[5,]  0.00  0.00  0.00  0.00 11.24
```

(b) Where are these *variances* coming from (see above!)

Now fit the weighted least squares model using REML. (Should match the OLS model.)

Model 2: weighted least squares $V(Y_i) = \sigma^2 DML_i$

```
#Notice have varFixed works differently from weights!)
model2REML = gls(Testisweight ~ DML, data=Squid, weights = varFixed(~DML), method="
REML")
```

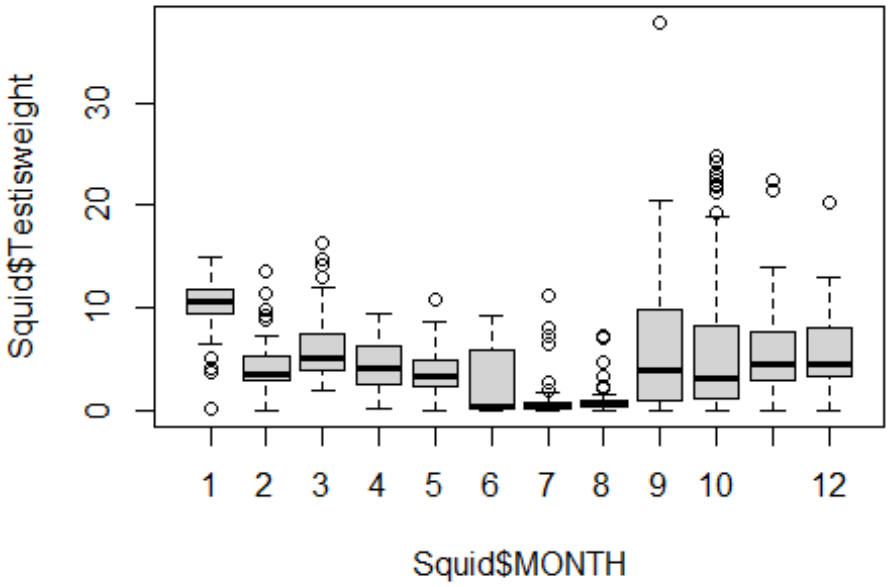
(c) How many parameters are estimated by this model? How do the likelihoods of model 1 and model 2 compare? What does this tell you? Can you carry out a likelihood ratio test?

Show the estimated variance-covariance matrix for the first 5 squid for this model.

(d) How is the very first value (σ_1^2) related to the residual standard error for model 2? Which observation has the largest variance in the first 5 rows of vcmatrix2? Why?

We also note pretty different variances in y across the different months.

```
load(url("http://www.rossmanchance.com/iscam3/ISCAM.RData"))
iscamsummary(Squid$Testisweight, Squid$MONTH)
   Missing   n   Min    Q1 Median     Q3    Max   Mean    SD   skew
1        0  45 0.152 9.382 10.661 11.769 14.981 10.200 2.866 -1.382
2        0  34 0.006 2.914  3.497  5.221 13.633  4.809 3.081  1.269
3        0  75 1.975 3.977  5.156  7.472 16.240  6.106 3.075  1.227
4        0  46 0.113 2.505  4.096  6.299  9.400  4.591 2.581  0.307
5        0  38 0.013 2.316  3.385  4.826 10.847  3.686 2.348  1.042
6        0  38 0.023 0.189  0.300  5.850  9.282  2.623 3.299  0.848
7        0  37 0.015 0.166  0.337  0.862 11.269  1.353 2.576  2.694
8        0  52 0.012 0.363  0.605  0.955  7.270  1.107 1.508  2.964
9        0 134 0.008 1.071  4.000  9.783 37.811  6.225 6.442  1.495
10       0 134 0.012 1.228  3.093  8.239 24.746  6.090 6.784  1.358
11       0  88 0.011 2.993  4.505  7.630 22.468  5.604 4.182  1.673
12       0  47 0.008 3.406  4.572  8.061 20.340  5.826 4.211  1.254
boxplot(Squid$Testisweight~Squid$MONTH)
```



(e) Which two months have the most variability? Which two have the least?

Model 3: $Var(\epsilon_i) = \sigma_j^2$

Fit a weighted regression allowing the variances to differ by month for observation i and month j.

```
model3REML = gls(Testisweight ~ DML, data=Squid, weights = varIdent(form= ~ 1 | MON
TH), method="REML")
summary(model3REML)
Generalized least squares fit by REML
  Model: Testisweight ~ DML
  Data: Squid
   AIC  BIC logLik
  4012 4077  -1992

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | MONTH
 Parameter estimates:
    2     9    12    11     8    10     5     7     6     4     1     3
1.000 2.681 1.616 1.680 3.004 2.121 2.705 2.310 1.949 1.703 1.986 1.932

Coefficients:
             Value Std.Error t-value p-value
(Intercept) -5.421    0.3437  -15.77       0
DML          0.044    0.0013   33.12       0
```

```
 Correlation:
    (Intr)
DML -0.949

Standardized residuals:
    Min      Q1     Med      Q3     Max
-3.9828 -0.7930 -0.1288  0.5329  4.9889

Residual standard error: 1.555
Degrees of freedom: 768 total; 766 residual
logLik(model3REML)
'log Lik.' -1992 (df=14)
```

(f) How many parameters are being estimated in this model? What are they? Include and explain the "variance structure" output. (What does this model estimate for the standard deviation in month 2? What about month 9? Compare back to the summary data if you aren't sure!

(g) How does the likelihood of model 3 compare? What does this tell you? Can you carry out a likelihood ratio test? DF?

Show the estimated variance-covariance matrix for the first 5 squid for this model.

(h) For vcmatrix3, how is the very first value (σ_1^2) related to the residual standard error for model 3? How is the second value (σ_2^2) related to the residual standard error for model 3? Which observation has the largest variance in the first 5 rows of vcmatrix3? Why?