Stat 414 - Day 3
Unequal Variances

Last Time:

- When satisfy the basic regression model assumptions (LINE), then can carry out tests of significance and confidence intervals
  - Testing the significance of the slope is equivalent to to assessing the significance of the linear relationship between the explanatory variable and response variable and whether the full model is significantly better than the residual model (if $\beta\hat$ is significant, then have evidence β differs from zero)
- Like a drop in SSError test, we can compare "nested" models by looking at the change in the log likelihood.
  - $2(L_1 - L_0)$ follows a chi-square distribution with df = difference in number of parameters in the two models

---

Example 1: Modeling Heterogeneity/Weighted Least Squares

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).

(a) How does Testis weight appear to change with DML?

Fit a linear model predicting testis weight from dorsal mantel length.

(b) Give a rough desert island approximation of the prediction interval for $TestisWeight$ when DML = 200.

(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'?

(d) So what's the problem?

(e) What variables are used in the Scale-Location plot?

Key Idea: The Breusch-Pagan Test is a likelihood ratio 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)

(f) Why might a transformation <u>not</u> be helpful here?

Another approach is to *model* the heterogeneity. In other words, can we explain the variation in the variation! One exploration is the relationship between the variation in the residuals and the fitted values.

```
#regressing the absolute value of the residuals vs. x
ersd <- lm(abs(model1$residuals) ~ model1$fitted.values)
```

(g) What do these fitted values estimate?

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

So we could take $w_i = 1/\widehat{ersd}^2$ or we can take $w_i = 1/DML_i$ (the error variance is proportional to DML) which will give more "weight" in the least squares estimation to squid with smaller DML values.

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

Manually fitting a weighted least squares model:

```
wmod <- lm(Testisweight ~ DML , data = Squid, weights = 1/ersd$fitted.values^2)
model2 = lm(Testisweight ~DML, data = Squid , weights = 1/DML)
```

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 estimates of SD(residual).

```
plot(rstandard(model2)~Squid$DML)
```

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

(j) How did the slope coefficient change? Is this what you predicted? Did R^2 and $\hat{\sigma}$ change?

Computer Problem 3: Due Wednesday, 8am

Of course we don't know the true σ_j values, we have only estimated them from the sample data. Instead, we should use *generalized least squares* (Aiken, 1934) which is an iterative approach for simultaneously estimating the regression coefficients and estimating the variance terms….

Model 1:

```
library(nlme)
model1REML <- gls(Testisweight ~ DML, data = Squid, method = "REML")
summary(model1REML)
logLik(model1REML)
```

(k) How many parameters are being estimated by this model?

Model 2: Use gls to fit the weighted regression

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

(l) 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?

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)
boxplot(Squid$Testisweight~Squid$MONTH)
```

(m) Which months have the most variability? Which have the least?

Model 3: Fit a weighted regression allowing the variances to differ by month $Var(\epsilon_i) = \sigma_j^2$ for observation i and month j.

```
model3REML = gls(Testisweight ~ DML, data=Squid, weights = varIdent(form= ~ 1 |
MONTH), method="REML")
summary(model3REML)
logLik(model3REML)
```

(n) 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!

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

Model 4: Now we can go really crazy, we can let the variances increase by DML, in perhaps a different way (different power) for each month. $Var(\epsilon_i) = \sigma^2\left(DML^{\delta_j}\right)^2$

```
vfbymonth <- varPower(form =~ DML | MONTH ) #allows different powers of DML, per
month
model4REML = gls(Testisweight ~ DML, data=Squid, weights = vfbymonth) #default is
REML
summary(model4REML)
plot(model4REML)
anova(model1REML, model2REML, model3REML, model4REML)
```

```
##See also
#install.packages("stargazer")
library(stargazer)
stargazer(model1REML, model2REML, model3REML, model4REML, type = "text")
```

(p) How many parameters are being estimated in this model? What do the parameter estimates at the bottom represent? (Hint: Look back at the $Var(\epsilon)$ expression.) Is it worth it? (Has the residual plot improved? Are the additional parameter estimates statistically significant?)

Notes

- The last likelihood ratio test isn't quite appropriate (why?) but not a horrible idea
- The varPower used in model 4 should not be used with a quantitative predictor that can take the value of zero.
- Finding the right variance structure for a study like this would be largely trial-and-error, using tools like AIC to compare the models. Better yet, use subject-matter knowledge/past information to help inform your choice of model.
- Notice if we fail to reject a model when compared to the "basic" model then say we can assume homogeneity in the residuals; this is another (some say better) "test for heterogeneity" (e.g., vs. Breusch-Pagen, Bartlett's test, Levene's test). In particular, these likelihood ratio tests work when you have violations of normality.
- When explore other "forms" (e.g., powers) of the variance covariates, watch for zero and negative values
- The estimates of the coefficients will usually be nearly the same as the unweighted estimates, but the weights will impact the widths of prediction intervals (and their validity).