Stat 414 — Day 5 revised
Categorical Predictors/ANOVA

Last Time

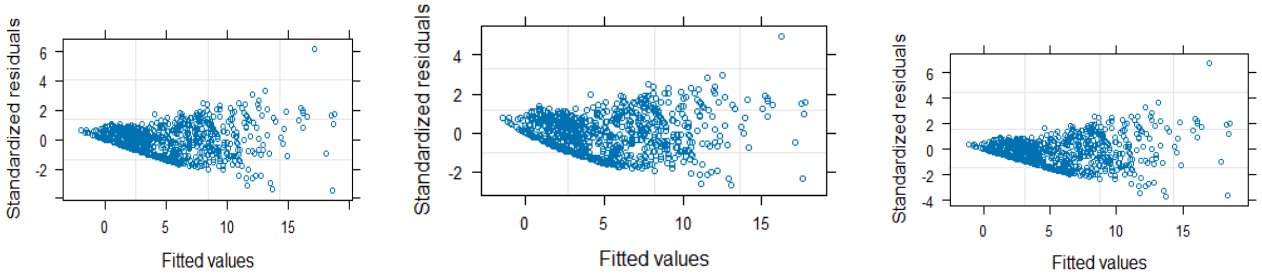
- Detecting unequal variance: Often focus on whether there is strong evidence of a linear relationship between $|e_i|$ or e^2 and the fitted value or vs. individual predictors (e.g., Scale-Location Plot). Other tests for heterogeneity include Barlett's test, Levene's test for comparing groups.
- *Transforming* the response variable can often better "scaled" where the variability in the response is more consistent.
- *Weighted least squares* can account for unequal variance in the response, e.g., allowing Predicting Intervals for have different widths depending on the weight values. Remember to compare validity using standardized residuals.
- *Generalized least squares* allows the modeler to specify different variance-covariance matrices for the residuals (e.g., σ_i^2)
- *Sandwich estimation* uses the observed residuals to specify a more complicated variance-covariance matrix to estimate the standard errors of the coefficients. "Allows you to take into account the heteroscedasticity without having to know about or model the functional form of the heteroscedasticity or use 'arbitrary' transformations."

Example 0: Squid revisited

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

We looked at several ways of modelling the unequal variance, e.g., increasing with DML, varying by month. Neither of these seemed to completely describe the variance pattern we were seeing.

```
library(nlme)
model1REML <- gls(Testisweight ~ DML, data = Squid, method = "REML")
model2REML = gls(Testisweight ~ DML, data = Squid, weights = varFixed(~DML), method
= "REML")
model3REML <- gls(Testisweight ~ DML, data = Squid, varIdent(form = ~1 | MONTH), me
thod = "REML")
par(mfrow=c(1,3))
plot(model1REML); plot(model2REML); plot(model3REML)
```

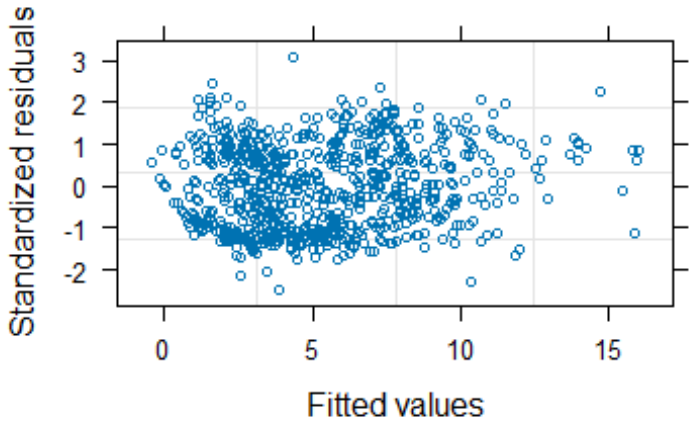


So let's trying something crazy: letting the variances increase with DML, perhaps differently (with a different power) for each month:

$$Var(\epsilon_i) = \sigma^2\left(DML^{\delta_j}\right)^2$$

```
library(nlme)
vfbymonth <- varPower(form = ~ DML | MONTH)  #Don't use with quantitative predictor
s that can equal zero
model4REML <-  gls(Testisweight ~ DML, data = Squid, weights = vfbymonth) #default
is REML
summary(model4REML)
Generalized least squares fit by REML
  Model: Testisweight ~ DML
  Data: Squid
   AIC  BIC logLik
  3694 3764  -1832

Variance function:
 Structure: Power of variance covariate, different strata
 Formula: ~DML | MONTH
 Parameter estimates:
    2     9    12    11     8    10     5     7     6     4     1     3
1.624 1.675 1.667 1.643 1.731 1.652 1.642 1.710 1.699 1.591 1.617 1.602

Coefficients:
             Value Std.Error t-value p-value
(Intercept) -3.986   0.24185  -16.48       0
DML          0.037   0.00124   29.69       0

 Correlation:
    (Intr)
DML -0.956

Standardized residuals:
     Min       Q1      Med       Q3      Max
-2.46793 -0.97112 -0.09893  0.77226  3.07010

Residual standard error: 0.0003004
Degrees of freedom: 768 total; 766 residual
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
vcmatrix4 = nlraa::var_cov(model4REML); vcmatrix4[1:5, 1:5]
       [,1]  [,2]   [,3]   [,4] [,5]
[1,] 0.7708 0.000 0.0000 0.0000 0.00
[2,] 0.0000 1.537 0.0000 0.0000 0.00
[3,] 0.0000 0.000 0.5431 0.0000 0.00
```

```
[4,] 0.0000 0.000 0.0000 0.7993 0.00
[5,] 0.0000 0.000 0.0000 0.0000 1.46
plot(model4REML)
```




```
anova(model1REML, model2REML, model3REML, model4REML)
           Model df  AIC  BIC logLik   Test L.Ratio p-value
model1REML     1  3 4055 4069  -2024
model2REML     2  3 3886 3900  -1940
model3REML     3 14 4012 4077  -1992 2 vs 3   104.6  <.0001
model4REML     4 15 3694 3764  -1832 3 vs 4   320.4  <.0001
```

(a) How many parameters are being estimated in this model? what do the parameter estimates at the bottom represent? Can you verify Squid 1's estimated variance?

```
  Specimen YEAR MONTH DML Testisweight
1     1017 1991     2 136        0.006
```

15: interept, slope, sigma, and 12 powers. The parameter estimates are the powers on the residual standard error for each month. Squid 1 had DML 136 in Month 2 and the estimated residual standard error is .003004 so the estimated standard deviation for Squid 1 is .0003004*136^1.6244 = .8778 which we then square to convert to variance .87785^2 = .7706.

(b) Has the residual plot improved? Are the additional parameter estimates statistically significant? Is the last likelihood ratio test appropriate?

The standardized residuals vs. fitted values is now beautiful. The likelihood ratio test gives a very small p-value comparing model 4 to model 3. Not technically nested but does have a different df value and is informative.

For fun:

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

===============================================================
```

```
                              Dependent variable:
                    ------------------------------------------
                                    Testisweight
                       (1)        (2)        (3)        (4)
-------------------------------------------------------------------
DML                  0.047***   0.043***   0.044***   0.037***
                     (0.001)    (0.001)    (0.001)    (0.001)

Constant            -6.534***  -5.624***  -5.421***  -3.986***
                     (0.393)    (0.338)    (0.344)    (0.242)

-------------------------------------------------------------------
Observations           768        768        768        768
Log Likelihood      -2,025.000 -1,940.000 -1,992.000 -1,832.000
Akaike Inf. Crit.   4,055.000  3,886.000  4,012.000  3,694.000
Bayesian Inf. Crit. 4,069.000  3,900.000  4,077.000  3,764.000
===================================================================
Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

Example 1: Pace of Life

Recall our pace of life data

```
PaceData = read.table("https://www.rossmanchance.com/stat414/data/Pace.txt", header
=TRUE)
head(PaceData)
             City Heart Walk Talk Bank Watch    Region
1       Boston,MA    24   28   24   31    30 Northeast
2      Buffalo,NY    29   23   23   30    33 Northeast
3      NewYork,NY    31   24   18   29    32 Northeast
4 SaltLakeCity,UT    26   28   23   28    23      West
5     Columbus,OH    26   22   30   27    23   Midwest
6    Worcester,MA    20   25   24   26    27 Northeast
```

Suppose I want to see whether the heart disease appears to differ significantly by region of the country (on average).

```
load(url("https://www.rossmanchance.com/iscam4/ISCAM.RData"))
iscamsummary(PaceData$Heart)
  Missing  n Min Q1 Median Q3 Max  Mean    SD Skewness
1       0 36  11 16     19 24  31 19.81 5.214    0.156
iscamsummary(PaceData$Heart, PaceData$Region)
          Missing n Min Q1 Median Q3 Max  Mean    SD Skewness
Midwest         0 9  13 20     21 24  26 21.44 4.126   -0.756
Northeast       0 9  14 18     20 26  31 22.00 5.788    0.279
South           0 9  14 16     19 23  27 19.67 4.528    0.350
West            0 9  11 11     16 18  26 16.11 4.910    0.675
```

(a) How would you suggest answering this question?

An 'Analysis of Variance' can be used to find a p-value to compare the means across the four regions.

(b) Carry out your analysis and summarize your conclusion.

```
summary(modela <- aov(PaceData$Heart ~ PaceData$Region))
                Df Sum Sq Mean Sq F value Pr(>F)
PaceData$Region  3    191    63.5    2.67  0.064 .
Residuals       32    761    23.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let μ_i represent the population mean heart disease rate across all cities in the same region.
H_0: $\mu_{midwest} = \mu_{northeast} = \mu_{south} = \mu_{west}$ vs. H_a: at least one μ differs from the others.
With a test statistic of F = 2.67 with degrees of freedom 3 and 27, we find a p-value of 0.0641, giving us weak evidence against the null hypothesis and in favor of the alternative hypothesis.

(c) Could we fit a basic regression model for this relationship? If not, why not? If so, how?

We have to convert the categorical variable into binary, numeric variables so we can fit lines between pairs of points. We will want to compare 3 pairs, if all 3 pairs are equal, then the means are the same.

How does R fit the model?

```
summary(model1 <- lm(Heart ~ Region, data = PaceData))

Call:
lm(formula = Heart ~ Region, data = PaceData)

Residuals:
   Min     1Q Median     3Q    Max
-8.444 -3.750 -0.556  3.000  9.889

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       21.444      1.626   13.19  1.7e-14 ***
RegionNortheast    0.556      2.299    0.24    0.811
RegionSouth       -1.778      2.299   -0.77    0.445
RegionWest        -5.333      2.299   -2.32    0.027 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.88 on 32 degrees of freedom
Multiple R-squared:   0.2,  Adjusted R-squared:  0.125
F-statistic: 2.67 on 3 and 32 DF,  p-value: 0.0641
logLik(model1)
'log Lik.' -106 (df=5)
anova(model1)
Analysis of Variance Table

Response: Heart
          Df Sum Sq Mean Sq F value Pr(>F)
Region     3    191    63.5    2.67  0.064 .
```

```
Residuals 32    761    23.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(d) How many parameters are estimated by the model? What are they? How do the ANOVA tables compare?

We have estimated an intercept, a coefficient for NE, a coefficient for S, and a coefficient for W. The F statistic and p-value for the overall F-test are the same. The null hypothesis is H_0: $\beta_{NE} = \beta_S = \beta_W = 0$. We recognize 21.44 as the mean for the Midwest, so R's default is to use indicator coding: the intercept estimates the mean for 'reference group' and the other coefficients estimate differences between that region's mean and the reference region's mean. For exmaple, 0.5556, tells us that the NE mean is 0.5556 larger than the MW mean = 0.56 + 21.44 = 22.0.

Convince R to use "effect coding" instead.

```
summary(model1b <- lm(Heart ~ Region, data = PaceData,
                      contrasts=list(Region = contr.sum))) # vs. contr.treatment

Call:
lm(formula = Heart ~ Region, data = PaceData, contrasts = list(Region = contr.sum))

Residuals:
   Min     1Q Median     3Q    Max
-8.444 -3.750 -0.556  3.000  9.889

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.806      0.813   24.37   <2e-16 ***
Region1        1.639      1.408    1.16     0.25
Region2        2.194      1.408    1.56     0.13
Region3       -0.139      1.408   -0.10     0.92
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.88 on 32 degrees of freedom
Multiple R-squared:   0.2,  Adjusted R-squared:  0.125
F-statistic: 2.67 on 3 and 32 DF,  p-value: 0.0641
```

(e) What is the difference between "indicator coding" and "effect coding"?

Now the intercept represents the overall average 19.86 and the coefficients represent the difference between the region and the overall average. For example, Region 1's group mean is 19.81 + 1.64 = 21.45, so that's the Midwest.

(f) Why doesn't R give us all four regions??

Because the four coefficients must sum to zero so if we know the values of 3 of them, then we know the value of the 4th.

(g) Which type of coding is better?

They are equivalent! Rather than one being better than another/they ultimately give you the same information, the exact same fit etc. Sometimes a research question might be more quickly answered with one than the other but you can always go back and forth between them.

(h) What do the *t*-tests tell you in each case?

With indicator coding, the t-test for a slope coefficient tells you whether the associated group mean is significantly different from the reference group mean. With effect coding, the t-test for a slope coefficient tells you whether the associated group mean is significantly different from the 'overall mean'.

(i) Do the model assumptions appear to be met? (What are the model assumptions?)

Still check for equal variance of the responses in the groups, normality of the responses in each group and independent observations (just don't really have a 'Linear' check.

(j) An 'advantage' to the regression model is automatic reporting of R^2. What is the R^2 value for each model and how is it interpreted?

R^2 = .2002, so about 20% of the variability in the heart disease rates is explained by which region the city is in.

More ANOVA table details

(a) Complete the handout providing the sum of squares formulas for the ANOVA table.

(b) So what does an ANOVA F-test compare?

The F-statistic is the ratio between the variability among the groups and the within group variability - how many times larger is the variability among the groups than the within group variability.

Definition

- When only comparing group means, the *coefficient of determination* is also referred to by some disciplines as η^2, *eta-squared*.
- Despite the popularity and wide-use of this statistic, it is a biased estimator of the population value. Another measure is $\frac{SSgroups - df(groups) * MSError}{SSTotal + MSError}$, (omega-squared)

These are considered two different measures of "effect size."

(c) Calculate the Omega-squared value.

```
summary(modela)
                Df Sum Sq Mean Sq F value Pr(>F)
PaceData$Region  3    191    63.5    2.67  0.064 .
Residuals       32    761    23.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(190.5 - 3*(23.78))/(190.5 + 761.1 + 23.78) = 0.122

Verify with R

```
#install.packages("effectsize")
library(effectsize)
omega_squared(model1b)
# Effect Size for ANOVA

Parameter | Omega2 |       95% CI
---------------------------------
Region    |   0.12 | [0.00, 1.00]

- One-sided CIs: upper bound fixed at [1.00].
```

(d) What is meant by effect size? Why is it important information?

Effect sizes are measures of 'how big is the effect' or 'how large (collectively) are the differences.' Rather than only using the p-value, they help us judge the 'practical significance' of the results.

Verify the ICC in R

```
#install.packages("ICC")
ICC::ICCbare(x = Region, y = Heart, data = PaceData)
[1] 0.1565
```

Computer Problem 5 - due Wednesday 7am

(You are encouraged to work with a partner and turn in one submission with both names)

A study by Foa et al. (1991) in the *Journal of Counseling and Clinical Psychology* looked at a study of 45 crime victims dealing with post traumatic stress disorder who were randomly assigned to one of four groups: 1) Stress Inoculation Therapy (SIT) in which subjects were taught a variety of coping skills; 2) Prolonged Exposure (PE) in which subjects reviewed the event in their mind repeatedly for seven sessions; 3) Supportive Counseling (SC) which was a standard therapy control group; and 4) a Waiting List (WL) control (no treatment). In this example, you will look at post-treatment data on PTSD Severity, the total number of symptoms endorsed by the subject.

```
foadata <-read.table("http://www.rossmanchance.com/stat414/data/foa.txt",header=T)
head(foadata)
  ID Treatment Score
1  1         1     3
2  2         1    13
3  3         1    13
4  4         1     8
5  5         1    11
6  6         1     9
load(url("https://www.rossmanchance.com/iscam4/ISCAM.RData"))
#if the above doesn't work, try load(url("https://www.rossmanchance.com/iscam3/ISCAM.RData"))
```

(a) Find the overall mean PTSD score for this sample.

(b) Use the iscamsummary function to see the basic descriptive statistics for each treatment group. Include your output.
(c) Find a linear regression model using only the treatment variable, what's the problem? How did you spot it?

Convert the Group variable to a factor, and make the control treatment the reference group:

```
GroupF = factor(foadata$Group, levels = c(4, 1, 2, 3))
```

(d) Convert the Group variable to a factor, and now fit the linear regression model with the treatment variable, using effect coding. (i) Interpret the intercept. Is it the same as your answer to (a)? (ii) Interpret one of the slope coefficients in context.
(e) Rerun the model using indicator coding and tell me which treatments appear significantly different from the control treatment. (Include relevant output, be clear how you are deciding.)