Stat 414 - Day 6
Adjusted Associations

Last Time

- To add a categorical variable to our linear model, can create *k*-1 binary indicator (0,1) variables or *k*-1 (-1, 0, 1)/"sum to zero" variables.
- Interpretations:
- With indicator coding, the intercept is the predicted mean for the reference category and the slope coefficients represent the difference in means to the reference category.
- With effect coding, the intercept is the "least squares" estimate of the overall mean and the slope coefficients represent the "effects" (differences between group means and overall means). You can solve for the missing coefficient by making them sum to zero.
- To test the statistical significance of the variable, test $H_0: \beta_1 = \ldots = \beta_{k-1} = 0$ in a partial *F*-test or likelihood ratio test (df = *k* -1). Still need normality and equal variance of the responses in each group.
- Measures of 'effect size': R^2 vs. ω^2 vs. ICC - different ways to measure the proportion of total variation in the response due to the categorical variable (bewteen vs. within groups) $(\frac{\sigma_g^2}{\sigma_g^2+\sigma^2})$

Example 1: Squid revisited

We started with simple explorations of the data.

```
   Missing   n   Min    Q1 Median     Q3    Max   Mean    SD Skewness
1        0  45 0.152 9.382 10.661 11.769 14.981 10.200 2.866   -1.335
2        0  34 0.006 2.914  3.497  5.221 13.633  4.809 3.081    1.213
3        0  75 1.975 3.977  5.156  7.472 16.240  6.106 3.075    1.202
4        0  46 0.113 2.505  4.096  6.299  9.400  4.591 2.581    0.297
5        0  38 0.013 2.316  3.385  4.826 10.847  3.686 2.348    1.001
6        0  38 0.023 0.189  0.300  5.850  9.282  2.623 3.299    0.814
7        0  37 0.015 0.166  0.337  0.862 11.269  1.353 2.576    2.584
8        0  52 0.012 0.363  0.605  0.955  7.270  1.107 1.508    2.878
9        0 134 0.008 1.071  4.000  9.783 37.811  6.225 6.442    1.478
10       0 134 0.012 1.228  3.093  8.239 24.746  6.090 6.784    1.343
11       0  88 0.011 2.993  4.505  7.630 22.468  5.604 4.182    1.644
12       0  47 0.008 3.406  4.572  8.061 20.340  5.826 4.211    1.214
```

We saw that variation in Testisweight measurements varied by month and by DML. But I admit I was a little confused by this output.

```
library(nlme)
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
model3REML$modelStruct$varStruct
Variance function structure of class varIdent representing
    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
```

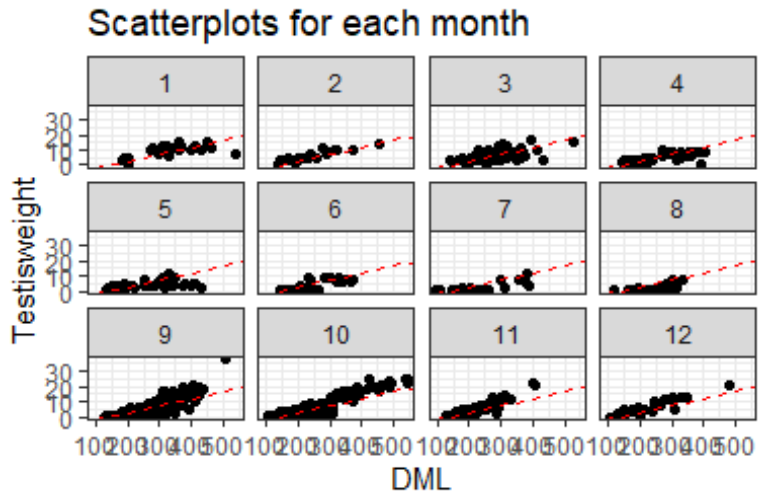
I thought months 9 and 10 had the larger variances and month 8 had one of the smallest. So why is Month 8 getting the largest multiplier?

(a) What are some possible explanations?

ⓘ (code)

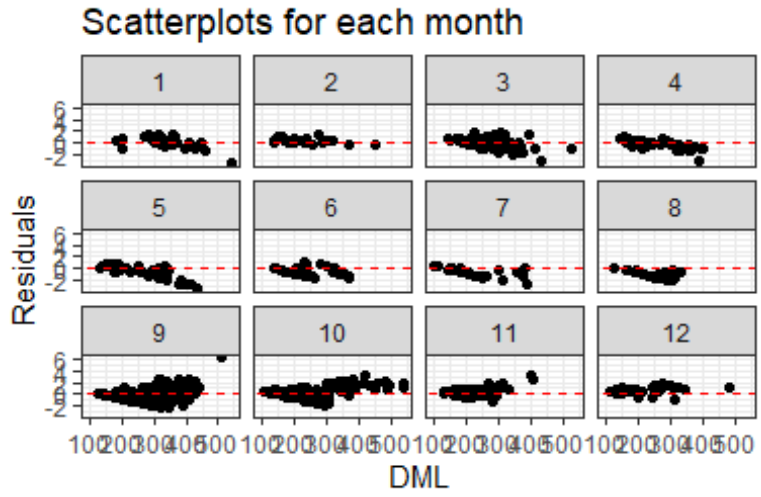
```
library(tidyverse)
model1REML <- gls(Testisweight ~ DML, data = Squid, method = "REML")

ggplot(Squid, aes(x = DML, y = Testisweight)) +
  geom_point() +
  geom_abline(intercept = -6.53, slope = .04660, color = "red", linetype = "dashe
d") + # Add the overall regression line
  facet_wrap(~ MONTH) +
  labs(title = "Scatterplots for each month",
       x = "DML",
       y = "Testisweight") +
  theme_bw()
```



ⓘ (code)

```
ggplot(Squid, aes(x = DML, y = residuals(model1REML, type = "normalized"))) +
  geom_point() +
  geom_abline(intercept = 0, slope = 0, color = "red", linetype = "dashed") + # A
dd the overall regression line
  facet_wrap(~ MONTH) +
  labs(title = "Scatterplots for each month",
       x = "DML",
       y = "Residuals") +
  theme_bw()
```



(b) How do we improve the model then?

```
vfbymonth <- varPower(form = ~ DML | MONTH)  #Don't use with quantitative predictor
s that can equal zero
fMonth = as.factor(Squid$MONTH)
```

```
model5REML = gls(Testisweight ~ DML + fMonth, data=Squid, weights = varIdent(form=
~ 1 | MONTH), method="REML")
#model5REML <-  gls(Testisweight ~ DML + fMonth, data = Squid, weights = vfbymonth)
#default is REML
model5REML$modelStruct$varStruct
Variance function structure of class varIdent representing
    2     9    12    11     8    10     5     7     6     4     1     3
1.000 3.251 1.412 1.818 1.168 2.585 2.680 1.673 1.622 1.797 2.248 2.220
summary(model5REML)
Generalized least squares fit by REML
  Model: Testisweight ~ DML + fMonth
  Data: Squid
   AIC  BIC logLik
  3744 3859  -1847

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 3.251 1.412 1.818 1.168 2.585 2.680 1.673 1.622 1.797 2.248 2.220

Coefficients:
             Value Std.Error t-value p-value
(Intercept) -4.041    0.5940   -6.80  0.0000
DML          0.043    0.0012   34.92  0.0000
fMonth2     -0.301    0.5066   -0.59  0.5527
fMonth3     -1.951    0.5472   -3.56  0.0004
fMonth4     -2.358    0.5584   -4.22  0.0000
fMonth5     -3.479    0.7127   -4.88  0.0000
fMonth6     -3.354    0.5623   -5.96  0.0000
fMonth7     -4.021    0.5755   -6.99  0.0000
fMonth8     -5.550    0.4903  -11.32  0.0000
fMonth9     -1.588    0.5676   -2.80  0.0053
fMonth10    -0.676    0.5282   -1.28  0.2011
fMonth11     0.081    0.5165    0.16  0.8756
fMonth12     0.619    0.5267    1.18  0.2401

Standardized residuals:
     Min       Q1      Med       Q3      Max
-4.19224 -0.63380 -0.07374  0.57826  5.13864

Residual standard error: 1.289
Degrees of freedom: 768 total; 755 residual
```

(c) How do we interpret the slope parameter estimates? How do the month parameter estimates match up with our descriptive analysis?

the slope estimate for month 8 tells us how much lower we predict the mean testisweight is compared to month 1 after adjusting for DML, now the slope estimate and the variance estimates are more consistent with the summary statistics

With *multiple regression*, we **always** have to interpret the slope coefficients conditional on the other variables in the model (e.g., the "effect" of DML after adjusting for MONTH). But what does that mean?

Example 2 - Applet demonstration (see handout)

Example 3 - Salary data cont.

Let's look at the salary data another way

```
saldata <-read.table("https://www.rossmanchance.com/stat414/data/saldata.txt", head
er=T)
```

(a) Calculate the correlation coefficient between *salary* and *semesters*. What does this number tell you?

```
cor(saldata$salary, saldata$semesters)
[1] 0.575
```

The strength of the linear association between the salaries and number of semesters across different people.

Key Idea

Rather than looking at the correlation between two variables, I want to measure how correlated measurements from different people in the same major are. Let's rearrange the data a bit:

```
#library(tidyverse)
salpairs <- saldata |>
  group_by(major) |>
  reframe(
    as.data.frame(t(combn(salary, 2))) |>
      rename(obs1 = V1, obs2 = V2)
  )

head(salpairs, 16)
# A tibble: 16 × 3
   major     obs1  obs2
   <chr>    <int> <int>
 1 business    40    44
 2 business    40    37
 3 business    40    39
 4 business    40    42
 5 business    40    38
 6 business    40    36
 7 business    40    32
 8 business    44    37
```

```
 9 business    44    39
10 business    44    42
11 business    44    38
12 business    44    36
13 business    44    32
14 business    37    39
15 business    37    42
16 business    37    38
```

(b) What has this code done?!

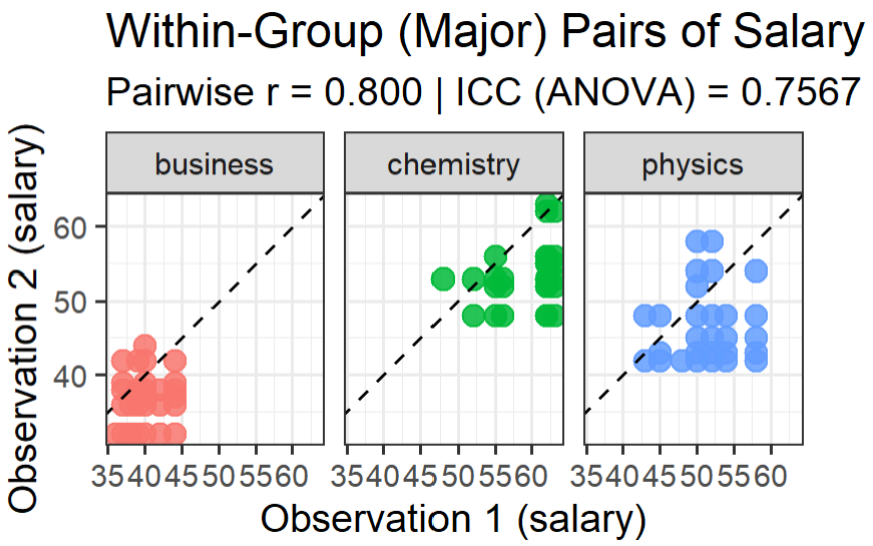
Created all possible pairs of observations of individuals within the same major

(c) Calculate the correlation between obs1 and obs2 in this new data frame.

```
cor(salpairs$obs1, salpairs$obs2)
[1] 0.7997
```

(d) How does the correlation in (c) compare to what you found for the ICC for these data?

Similar




```
#install.packages("ICC")
ICC::ICCbare(x = major, y = salary, data = saldata)
[1] 0.7567
```

Key Idea

The intraclass correlation coefficient (ICC) can also be interpreted as a measure of how correlated two responses are from individuals in the same "class." It measures the degree of "sameness" of individuals in the same group vs. across groups. The most traditional application is as a measure of "reliability" of repeat observations.

Notes:

- The reason the values don't match better is due to the small number of groups, so the ICC tends to underestimate the true correlation. There is also a distinction between the population ICC and the sample results which use the same observations many times.
- There are a number of different intraclass correlation coefficients out there