Correlated Data Models

BDSI 2019

Holly Hartman

6/26/2019

Outline/Learning objectives

- Why do we need to account for correlation?
- Simulation study
- What does correlation look like?
- Mixed effect models
- Real data example of mixed effect model

Review

What are the assumptions for a linear regression model?

Review

What are the assumptions for a linear regression model?

- Independence between observations
- Linearity between covariates and outcome
- Constant variance of errors
- Normally distributed errors

What is correlated data?

Correlated data arises when the assumption of independence is violated.

Example 1:

A researcher is conducting a study on a new treatment for hypertension. Clinics are randomly assigned to either prescribe the standard of care treatment or the new treatment. New patients seen at the clinics have their baseline blood pressure measured and then have their blood pressure measured again at 3 follow up appointments at set times. What are potential sources of correlation within the resulting data set for this study?

Example 2:

A researcher is interested in if there is a subtype of cancer more susceptible to a specific type of treatment. The cancer is genetically inheritable. Patients are recruited through a hospital and any family members with the cancer are also recruited. Samples of the tumors are genotyped. People in the study are randomly assigned to receive the new treatment or standard of care and their tumor response is measured. What are potential sources of correlation within the resulting data set for this study?

Section 1

What happens if you ignore correlation in data sets? A simulation example

Simulating a data set - load libraries

```
library(MASS)  #Required to generate multivariate normal data
library(tidyr)  #Required for formatting data **
library(lme4)  #Required for mixed models **
library(ggplot2)  #Required for the pretty plots **
library(PerformanceAnalytics)  #For cool correlation plot **
```

** indicates we will use this library for the real data example later too

Simulating a data set - set parameters

```r
set.seed(7789)  #Seed to ensure identical results of
# simulation.

# P.S. 7/7/89 is my birthday

# covariance matrix
(covMatrix <- matrix(c(1, 0.8, 0.5, 0.8,
    1, 0.8, 0.5, 0.8, 1), byrow = T, nrow = 3))
```

```
##      [,1] [,2] [,3]
## [1,]  1.0  0.8  0.5
## [2,]  0.8  1.0  0.8
## [3,]  0.5  0.8  1.0
```

```r
contMean <- c(0, 1, 2)  #means for the control group

trtMean <- contMean + 1  #treatment group means
```

Simulating a data set - Generate data

```
# Control group data
cont <- cbind(seq(101, 200), 0, mvrnorm(n = 100,
    mu = contMean, Sigma = covMatrix))

# treatment group data
trt <- cbind(seq(1, 100), 1, mvrnorm(n = 100,
    mu = trtMean, Sigma = covMatrix))

# Combine trt and control data
trialDataWide <- data.frame(rbind(cont, trt))
names(trialDataWide) <- c("ID", "trt", "0",
    "1", "2")  #name columns
```

Simulating a data set - View data

```
head(trialDataWide)
```

```
##    ID trt          0           1        2
## 1 101   0  0.5815913  1.33241014 2.151547
## 2 102   0 -1.1633814 -0.04971235 1.437077
## 3 103   0  2.2052564  3.21665116 3.791609
## 4 104   0 -1.2877354  0.21725424 1.027111
## 5 105   0 -0.2124468  0.79627989 2.556943
## 6 106   0 -1.7497573 -1.07734386 1.585896
```

Simulating a data set - Cleaning data

```
# Make data in long format from wide
# format
trialData <- gather(trialDataWide, time,
    measurement, "0":"2", factor_key = F)

# Change data type of 'time'
trialData$time <- as.numeric(trialData$time)

head(trialData)
```

```
##    ID trt time measurement
## 1 101   0    0   0.5815913
## 2 102   0    0  -1.1633814
## 3 103   0    0   2.2052564
## 4 104   0    0  -1.2877354
## 5 105   0    0  -0.2124468
## 6 106   0    0  -1.7497573
```

Results!

```
# Model not accounting for correlation
summary(lm(measurement ~ trt + time, data = trialData))$coeffi
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -0.0685503 0.08072985 -0.849132 3.961482e-01
## trt          0.9794843 0.08630384 11.349255 3.625688e-27
## time         0.9445598 0.05285009 17.872434 1.527042e-57
```

```
# Model accounting for correlation
summary(lmer(measurement ~ trt + time + (1 |
    ID), data = trialData))$coefficients
```

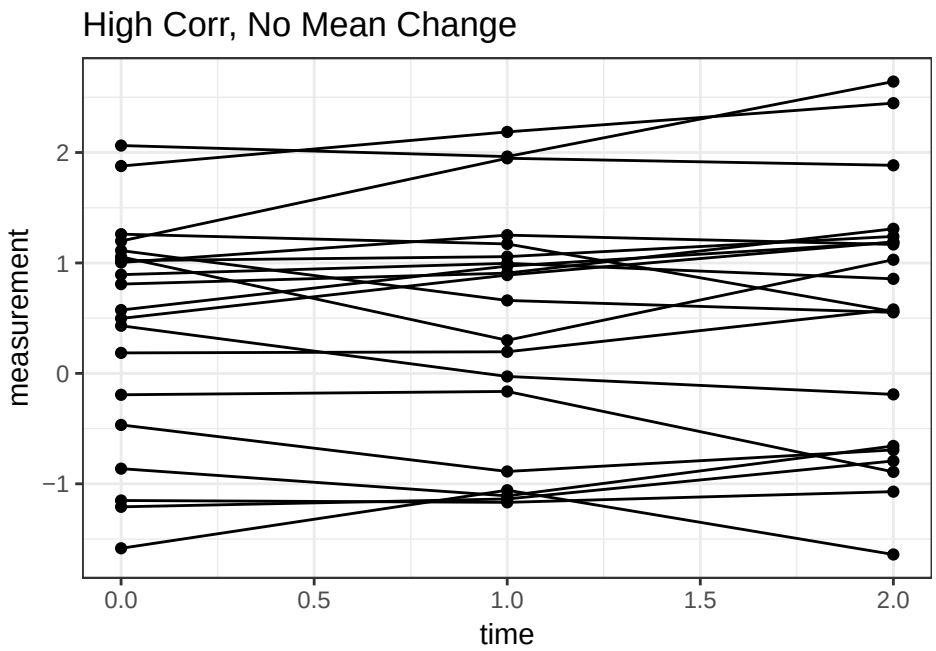
```
##                Estimate Std. Error    t value
## (Intercept) -0.0685503 0.09984972 -0.6865347
## trt          0.9794843 0.13582061  7.2116032
## time         0.9445598 0.02731936 34.5747400
```

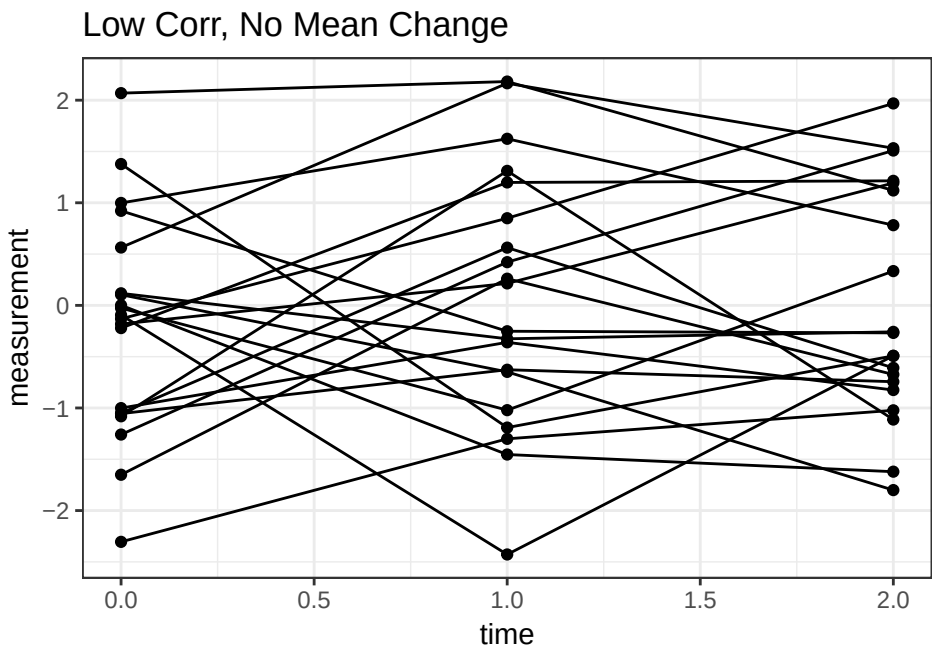
What happened?

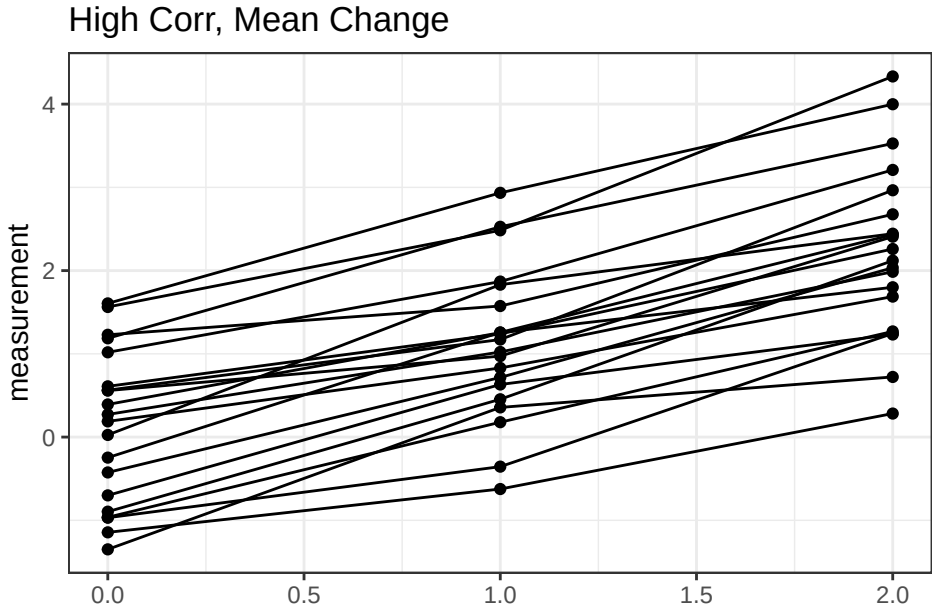
Between subject (trt) - If we assume all observations are independent, this artificially inflates the number of observations. This decreases the standard error estimates.

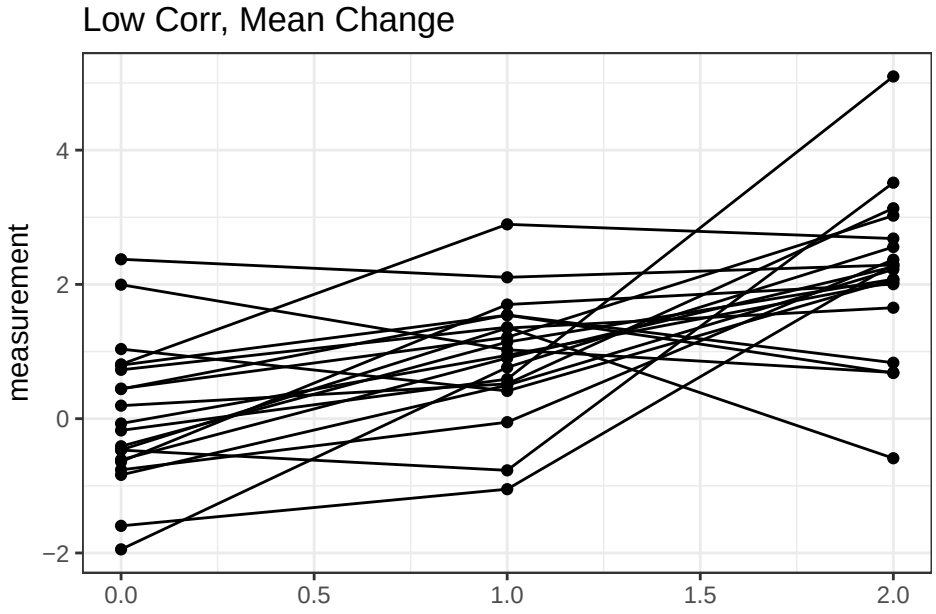
Within subject (time) - Subjects have a higher correlation with themselves than with other subjects and this causes lower variability in reality. If we assume all observations are independent, then this overestimates variability.

What does correlation look like?









Clustered Data

- Clustered data can arise from the study design
- Classic example: children in a classroom where the classrooms are randomized to treatment. Children within the same classroom are more related than children not in the same classroom because they have their teacher in common.
- Do the individuals share something that may make them more similar? (ex: same doctor, same household)

Longitudinal Data

- The same patients are at multiple time points and have the outcome of interest measured multiple times
- This is called "repeated measures."
- Can also occur when multiple measurements are taken on the same person that are likely correlated.

Mixed effects models

Traditional model:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{\epsilon}_i$$

Mixed effects model:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

- i is the index for the individual. Since we have repeated measures, $\boldsymbol{Y}_i$ has multiple elements since there are multiple outcomes observed for each individual.
- β are fixed effects, the same for every person.
- $\boldsymbol{b_i}$ are random effects, different for each person. We require $E(\boldsymbol{b_i}) = 0$ so that the the estimated $\boldsymbol{b_i}$ are interpreted as individual deviances from the population mean when $\boldsymbol{Z}$ is a subset of $\boldsymbol{X}$.

Random effects

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

$\boldsymbol{b}_i \sim MNV_q(\boldsymbol{0}, \boldsymbol{G})$ and $\boldsymbol{\epsilon}_i \sim MVN_{n_i}(\boldsymbol{0}, \boldsymbol{R}_i))$ with $\boldsymbol{b}_i$ being independent of $\boldsymbol{\epsilon}_i$.

Random effects - Intercepts

$$y_{ij} = \beta_0 + \beta_1 Trt_i + \beta_2 Time_{ij} + b_{0i} + \epsilon_{ij}$$

- β_1 and β_2 are fixed effects.
- b_{0i} is a random effect.
- $b_{0i} \sim N(0, \sigma_0^2)$
- Each patient has their own intercept $\beta_0 + b_{0i}$.

Random effects - Slopes

$$y_{ij} = \beta_0 + \beta_1 Trt_i + \beta_2 Time_{ij} + b_{0i} + b_{1i} Time_{ij} + \epsilon_{ij}$$

- Each patient has their own intercept $\beta_0 + b_{0i}$.
- Each patient has their own slope over time $\beta_1 + b_{1i}$.
- $\boldsymbol{b}_i \sim MVN(\boldsymbol{0}, \boldsymbol{G})$

Expected values in a random effects model

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

$\boldsymbol{b}_i \sim MNV_q(\boldsymbol{0}, \boldsymbol{G})$ and $\boldsymbol{\epsilon}_i \sim MVN_{n_i}(\boldsymbol{0}, \boldsymbol{R}_i))$

$$E(\boldsymbol{Y}_i|\boldsymbol{b}_i) = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i$$

$$E(\boldsymbol{Y}_i) = \boldsymbol{X}_i\beta$$

Covariance in a random effects model

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

$\boldsymbol{b}_i \sim MNV_q(\boldsymbol{0}, \boldsymbol{G})$ and $\boldsymbol{\epsilon}_i \sim MVN_{n_i}(\boldsymbol{0}, \boldsymbol{R}_i))$

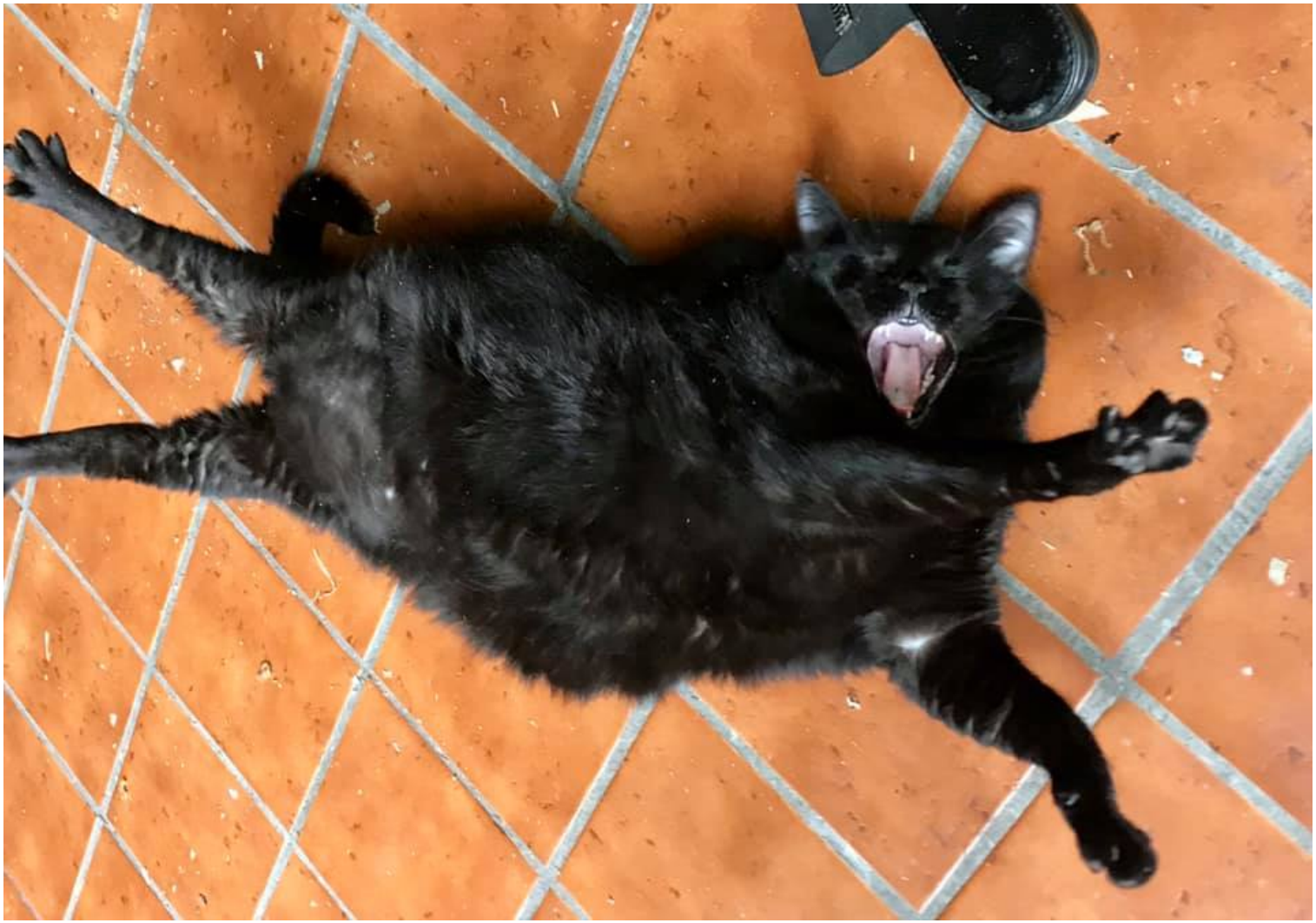
$$Cov(\boldsymbol{Y}_i|\boldsymbol{b}_i) = Cov(\boldsymbol{\epsilon}_i) = \boldsymbol{R}_i$$

$$\begin{aligned} Cov(\boldsymbol{Y}_i) &= Cov(\boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i) \\ &= Cov(\boldsymbol{Z}_i\boldsymbol{b}_i) + Cov(\boldsymbol{\epsilon}_i) \\ &= \boldsymbol{Z}_i Cov(\boldsymbol{b}_i)\boldsymbol{Z}_i' + Cov(\boldsymbol{\epsilon}_i) \\ &= \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}_i \end{aligned}$$

This means that $Cov(\boldsymbol{Y}_i)$ will, in general, have non-zero off diagonal elements accounting for the correlation among the repeated measures for the same individual.

Data example

Cats!



Data example

http://users.stat.ufl.edu/~winner/data/cats_anxiety1.txt

- Data is from a 2007 study on cat anxiety.
- Cats were given Zylkene or placebo and emotional state was measured 5 times.
- Demographic info is included (although we won't be using this)

Data information

Variables/Columns:

- ID # 7-8
- Weight (kg) 12-16
- Age (Months) 23-24
- Gender 32 /* 1=Neutered Female, 2=Neutered Male, 3=Female */
- Environment 40 /* 1=House, 2=Apartment */
- Origin 48 /* 1=House/Apt, 2=Humane Society,3=Market,4=Barn,5=Street,6=Breeder */
- Treatment 56 /* 1=Zylkene, 0=Placebo */
- Result 64 /* 1=Success, 0=Failure */
- Emotional score @ time 1 71-72
- Emotional score @ time 2 79-80
- Emotional score @ time 3 87-88
- Emotional score @ time 4 95-96
- Emotional score @ time 5 103-104

Read in data and clean

Link to data (since code gets cut off):
http://users.stat.ufl.edu/~winner/data/cats_anxiety1.dat

```
cats <- read.table("http://users.stat.ufl.edu/~winner/data/cat

names(cats) <- c("id", "wt", "age", "sex",
    "env", "org", "trt", "result", "emscore1",
    "emscore2", "emscore3", "emscore4", "emscore5")

summary(cats)
```

Data descriptives

```
##        id              wt             age       
##  Min.   : 1.00   Min.   :2.000   Min.   :  9.00  
##  1st Qu.: 9.25   1st Qu.:3.925   1st Qu.: 34.00  
##  Median :17.50   Median :4.500   Median : 49.00  
##  Mean   :17.50   Mean   :4.762   Mean   : 56.24  
##  3rd Qu.:25.75   3rd Qu.:6.000   3rd Qu.: 70.75  
##  Max.   :34.00   Max.   :7.800   Max.   :143.00  

##       sex             env             org       
##  Min.   :1.0   Min.   :1.000   Min.   :1.000  
##  1st Qu.:1.0   1st Qu.:1.000   1st Qu.:1.000  
##  Median :1.0   Median :2.000   Median :2.000  
##  Mean   :1.5   Mean   :1.529   Mean   :2.912  
##  3rd Qu.:2.0   3rd Qu.:2.000   3rd Qu.:5.000  
##  Max.   :3.0   Max.   :2.000   Max.   :6.000
```

Data descriptives

```
##        trt             result          emscore1
##  Min.   :0.0   Min.   :0.0000   Min.   : 2.0
##  1st Qu.:0.0   1st Qu.:0.0000   1st Qu.: 8.0
##  Median :0.5   Median :0.0000   Median : 9.5
##  Mean   :0.5   Mean   :0.4118   Mean   :10.0
##  3rd Qu.:1.0   3rd Qu.:1.0000   3rd Qu.:13.0
##  Max.   :1.0   Max.   :1.0000   Max.   :17.0

##     emscore2        emscore3        emscore4        emscore
##  Min.   : 2.00   Min.   : 2.00   Min.   : 2.00   Min.   : 2
##  1st Qu.: 9.00   1st Qu.: 9.25   1st Qu.:10.00   1st Qu.:10
##  Median :11.50   Median :13.00   Median :13.00   Median :14
##  Mean   :11.15   Mean   :12.24   Mean   :13.29   Mean   :13
##  3rd Qu.:14.00   3rd Qu.:14.75   3rd Qu.:16.00   3rd Qu.:17
##  Max.   :18.00   Max.   :19.00   Max.   :23.00   Max.   :23
```

Go from wide form to long form

```
head(cats)
```

```
##   id  wt age sex env org trt result emscore1 emscore2 emscore3 emscore4
## 1  1 4.0  15   1   1   1   1      0        8        9        8        9
## 2  2 4.0  67   1   1   2   0      0        9        9        9        9
## 3  3 6.0  55   1   1   1   1      1        9       11       12       13
## 4  4 3.5  78   2   1   1   0      0        9        9        9        9
## 5  5 6.2  50   1   2   3   1      0       10       10       10       10
## 6  6 2.0  10   2   2   4   0      0        6        6        6        6
##   emscore5
## 1        9
## 2        9
## 3       16
## 4        9
## 5       10
## 6        6
```

```
dim(cats)
```

```
## [1] 34 13
```

Go from wide form to long form

```
catsw <- gather(cats, time, measurement,
    "emscore1":"emscore5")
dim(catsw)
```

```
## [1] 170  10
```

```
head(catsw)
```

```
##   id  wt age sex env org trt result     time measurement
## 1  1 4.0  15   1   1   1   1      0 emscore1           8
## 2  2 4.0  67   1   1   2   0      0 emscore1           9
## 3  3 6.0  55   1   1   1   1      1 emscore1           9
## 4  4 3.5  78   2   1   1   0      0 emscore1           9
## 5  5 6.2  50   1   2   3   1      0 emscore1          10
## 6  6 2.0  10   2   2   4   0      0 emscore1           6
```

Go from wide form to long form

```
catsw$time <- as.numeric(gsub("emscore",
    "", catsw$time))

head(catsw)
```

```
##    id  wt age sex env org trt result time measurement
## 1   1 4.0  15   1   1   1   1      0    1             8
## 2   2 4.0  67   1   1   2   0      0    1             9
## 3   3 6.0  55   1   1   1   1      1    1             9
## 4   4 3.5  78   2   1   1   0      0    1             9
## 5   5 6.2  50   1   2   3   1      0    1            10
## 6   6 2.0  10   2   2   4   0      0    1             6
```

```
dim(catsw)
```

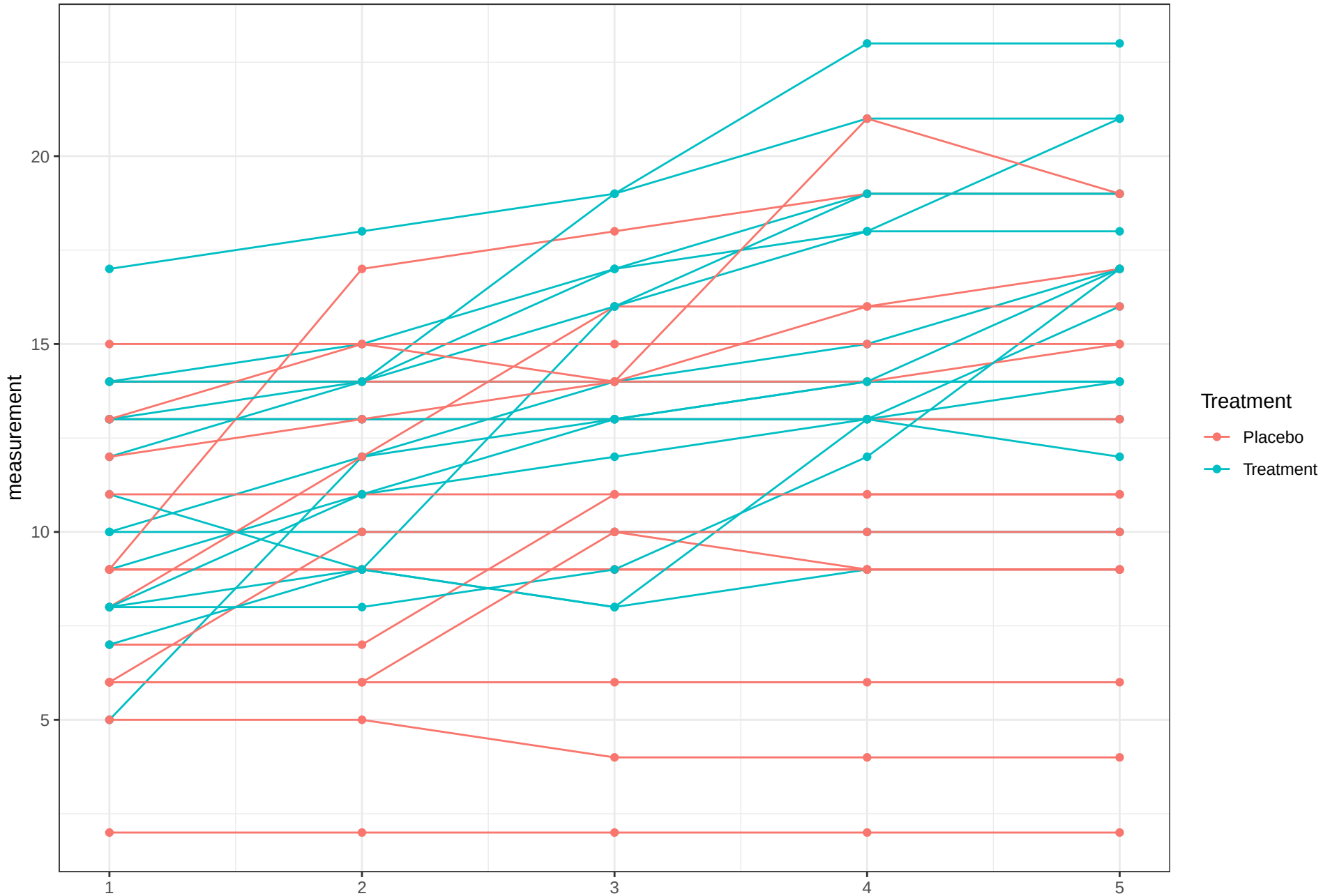
```
## [1] 170  10
```

Plots Code

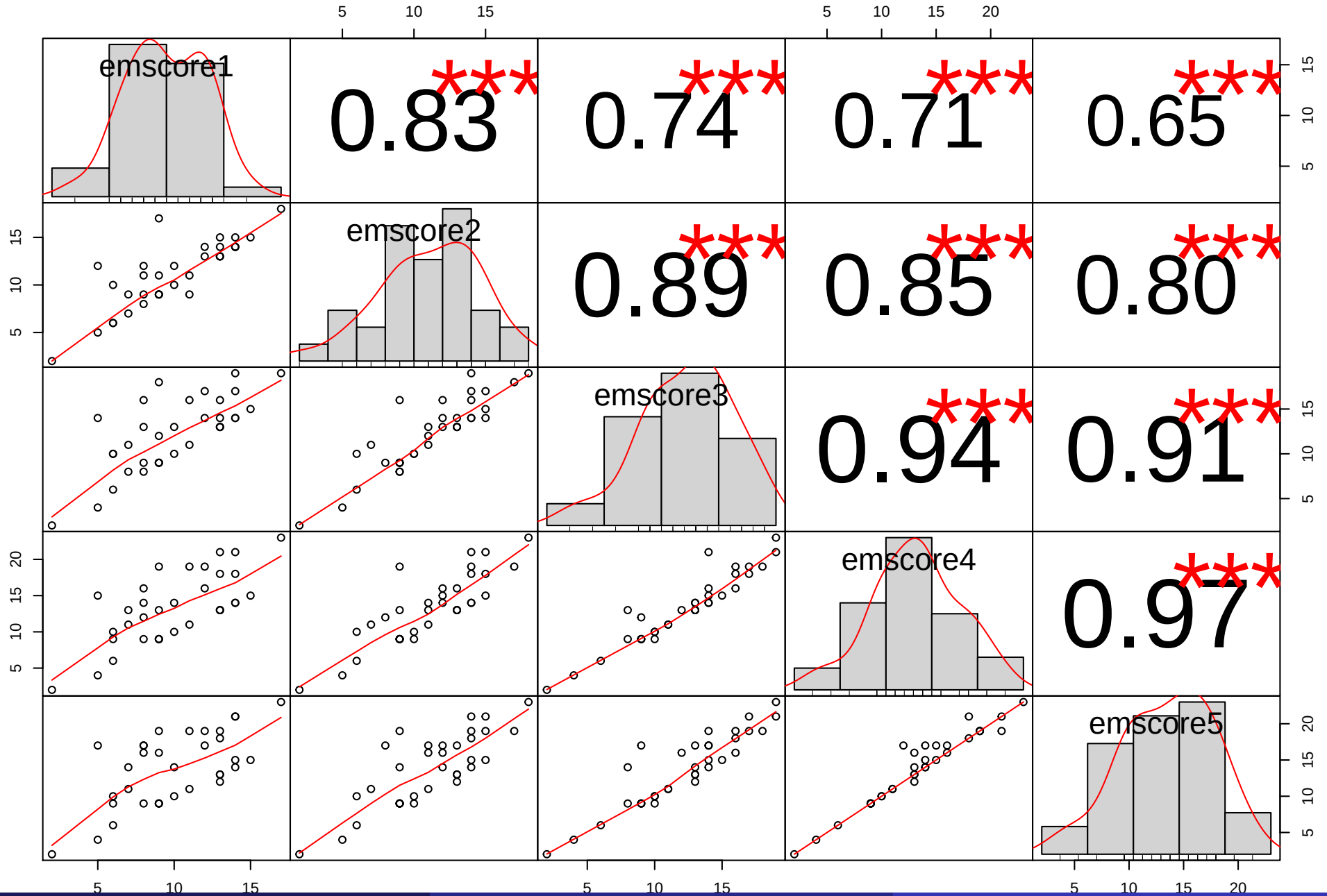
```
ggplot(data = catsw, aes(x = time, y = measurement,
    group = id, color = as.factor(trt))) +
    scale_colour_discrete(name = "Treatment",
        breaks = c("0", "1"), labels = c("Placebo",
            "Treatment")) + geom_line() +
    geom_point()


chart.Correlation(cats[, 9:13], histogram = TRUE,
    pch = 19)
```

Plot Anxiety Scores Over Time



Correlated data plot



Model the data

```
mod <- lmer(measurement ~ trt + time + (1 +
    time | id), data = catsw)
summary(mod)
```

Model the data

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: measurement ~ trt + time + (1 + time | id)
##    Data: catsw
##
## REML criterion at convergence: 717.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.03830 -0.28055 -0.01111  0.32886  2.55610
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 10.5550  3.2489
##           time         0.8126  0.9014   -0.25
##  Residual              1.3088  1.1440
## Number of obs: 170, groups:  id, 34
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   8.3588     0.8160  10.244
## trt           1.6529     1.1190   1.477
## time          0.9676     0.1666   5.809
##
## Correlation of Fixed Effects:
##      (Intr) trt
## trt  -0.686
## time -0.244  0.000
```

Interpreting the effect estimates

Since we have random effects, the interpretation of the fixed effect estimates are for population averages.

For a typical cat, Zylkene increases the emotional score by 1.65.

For a typical cat, the emotional score increases by 0.97 over each time period.

What is missing from this model?

If we were to do more on diagnostics and more modeling, what would you want to change/examine in our model?

What is missing from this model?

If we were to do more on diagnostics and more modeling, what would you want to change/examine in our model?

- Add interaction between time and trt
- Model the time effect differently (categorical, nonlinear, etc)
- Add demographics
- Examine interactions between demographics and trt or time
- Examine diagnostics (not covered here)
- Test if random effects are needed

Test if random effects are needed

```
fullMod <- lmer(measurement ~ trt + time +
    (1 + time | id), data = catsw, REML = F)
# redMod1<-lmer(measurement ~ trt + time
# + (1 | id ), data=catsw, REML=F)
# redMod2<-lmer(measurement ~ trt + time
# + (0+time | id ), data=catsw, REML=F)
redMod3 <- lm(measurement ~ trt + time, data = catsw)

anova(mod, redMod3)
```

Test if random effects are needed

```
## refitting model(s) with ML (instead of REML)

## Data: catsw
## Models:
## redMod3: measurement ~ trt + time
## mod: measurement ~ trt + time + (1 + time | id)
##         Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## redMod3  4 952.59 965.13 -472.29   944.59
## mod      7 732.25 754.20 -359.13   718.25 226.33      3  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

THANK YOU!

Thank you all for being a wonderful audience! Good luck with your projects and the rest of the BDSI program!