UNIVERSITY of VIRGINIA HEALTH SYSTEM

REGRESSION IN R

DATA SERVICES @ HSL

MARIEKE K JONES, PHD

MARIEKE@VIRGINIA.EDU

UNIVERSITY of VIRGINIA HEALTH SYSTEM

OUTLINE

- Linear models and Multiple regression
- Measures of model fit
- Logistic (and Multinomial regression)
- Mixed effects models

- Variable selection

UNIVERSITY of VIRGINIA HEALTH SYSTEM

LINEAR REGRESSION

Simple linear regression

- With a Categorical Variable (ANOVA)
    - Covered last class
- With a Continuous Variable (review)

Multiple regression

- Can seamlessly combine categorical and continuous predictors

LINEAR REGRESSION

SINGLE PREDICTOR X

$$Y = \beta_0 + \beta_1 X + \epsilon$$

MULTIPLE PREDICTORS X_1 AND X_2

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Y = RESPONSE VARIABLE

X = PREDICTOR

β_0 = Y-INTERCEPT

$\beta_{1,2}$ = SLOPE. What effect does one unit change in X do to Y?

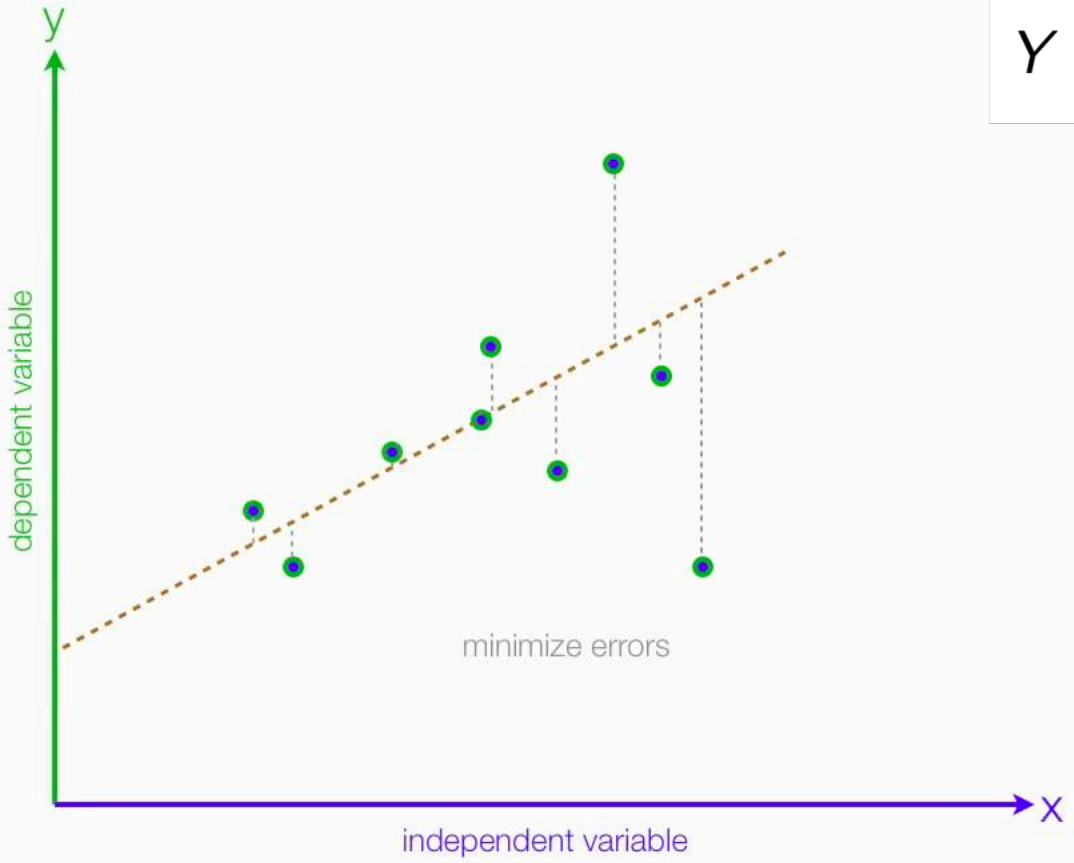
ϵ = RESIDUAL ERROR

Given X, slope, and y-intercept, model cannot perfectly predict Y. These are assumed to be normally distributed with a mean of 0 and a standard deviation σ

UNIVERSITY of VIRGINIA HEALTH SYSTEM

LINEAR REGRESSION

$$Y = \beta_0 + \beta_1 X + \epsilon$$



UNIVERSITY of VIRGINIA HEALTH SYSTEM

ASSUMPTIONS OF LINEAR REGRESSION

- Random Sampling
- Residuals are independent
- Residuals are normally distributed
- X and Y have linear relationship
- Residuals show constant variance across levels of X

UNIVERSITY of VIRGINIA HEALTH SYSTEM

MEASURES OF MODEL FIT

- R^2 / Adjusted R^2
- RMSE (Root Mean Squared Error)
- AIC / BIC

- Special case: compare 2 nested models
  - Likelihood Ratio test

MEASURES OF MODEL FIT

- R^2
  - Percentage of variation in Y explained by model
  - Increases with number of predictors
- Adjusted R^2
  - Proportion of total variance explained by model
  - Increases with number of predictors if improvement in model fit is worthwhile
- RMSE (root-mean-square error)
  - Square root of the mean of the squared residuals
  - Indicates absolute fit of the model to the data
    - How close are observed data to model's predicted values
  - Lower values indicate better fit
  - Best measure of model fit for predictive models

MEASURES OF MODEL FIT

- AIC / BIC
  - Akaike Information Criterion / Bayesian Information Criterion
  - Estimate relative quality of models for a given set of data
  - When a model is used to represent a process that generated data, some information will be lost by using the model rather than the process itself
    - AIC / BIC estimate relative information lost by a given model
  - Include a tradeoff between model fit and model complexity
  - Lower is better

The optimal model has a high R^2, low RMSE, and low AIC / BIC

UNIVERSITY of VIRGINIA HEALTH SYSTEM

MULTIPLE REGRESSION

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Rarely can biological phenomena be explained by one variable
- Simple extension of linear model with 1 predictor
- Can add continuous and / or categorical predictors
- ** Interpretation of coefficient is “average effect on Y of a one unit increase in Xj, holding all other predictors constant”

MULTIPLE REGRESSION TIPS

- Use biological knowledge to inform which predictors are included
- Do not add predictors that correlate too highly with other predictors (~+/- .8)
- The addition of new predictors will change the model fit and coefficients of other predictors

LOGISTIC AND MULTINOMIAL REGRESSION

- Outcome variable is binary → logistic regression
  - e.g. model whether someone is insured based on several other variables
  - Model **log odds** of being insured

- Outcome variable is categorical (3+) → multinomial regression

UNIVERSITY of VIRGINIA HEALTH SYSTEM

LOGISTIC REGRESSION

p is the probability of being insured

$\frac{p}{1-p}$ is the odds of being insured

$$p = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

The generalized linear model is therefore:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Where β_0 is the intercept, β_1 is the increase in the log odds of outcome for every unit increase in x_1.

UNIVERSITY of VIRGINIA HEALTH SYSTEM

LOGISTIC REGRESSION

- Logistic regression is a type of *generalized linear model* (GLM)
- using glm() function in R
- glm() works like lm() except we specify which distribution to use with the family argument
- Logistic regression uses family=binomial

```
mod <- glm(y ~ x, data=yourdata, family='binomial')
summary(mod)
```

UNIVERSITY of VIRGINIA HEALTH SYSTEM

MULTINOMIAL REGRESSION

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+...+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_{1\ell}X_1+...+\beta_{p\ell}X_p}}$$

- Extension of logistic regression where each class (k) is modeled by its own linear function

- multinom() function in nnet package
  - Tutorial at https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/

MIXED EFFECTS REGRESSION

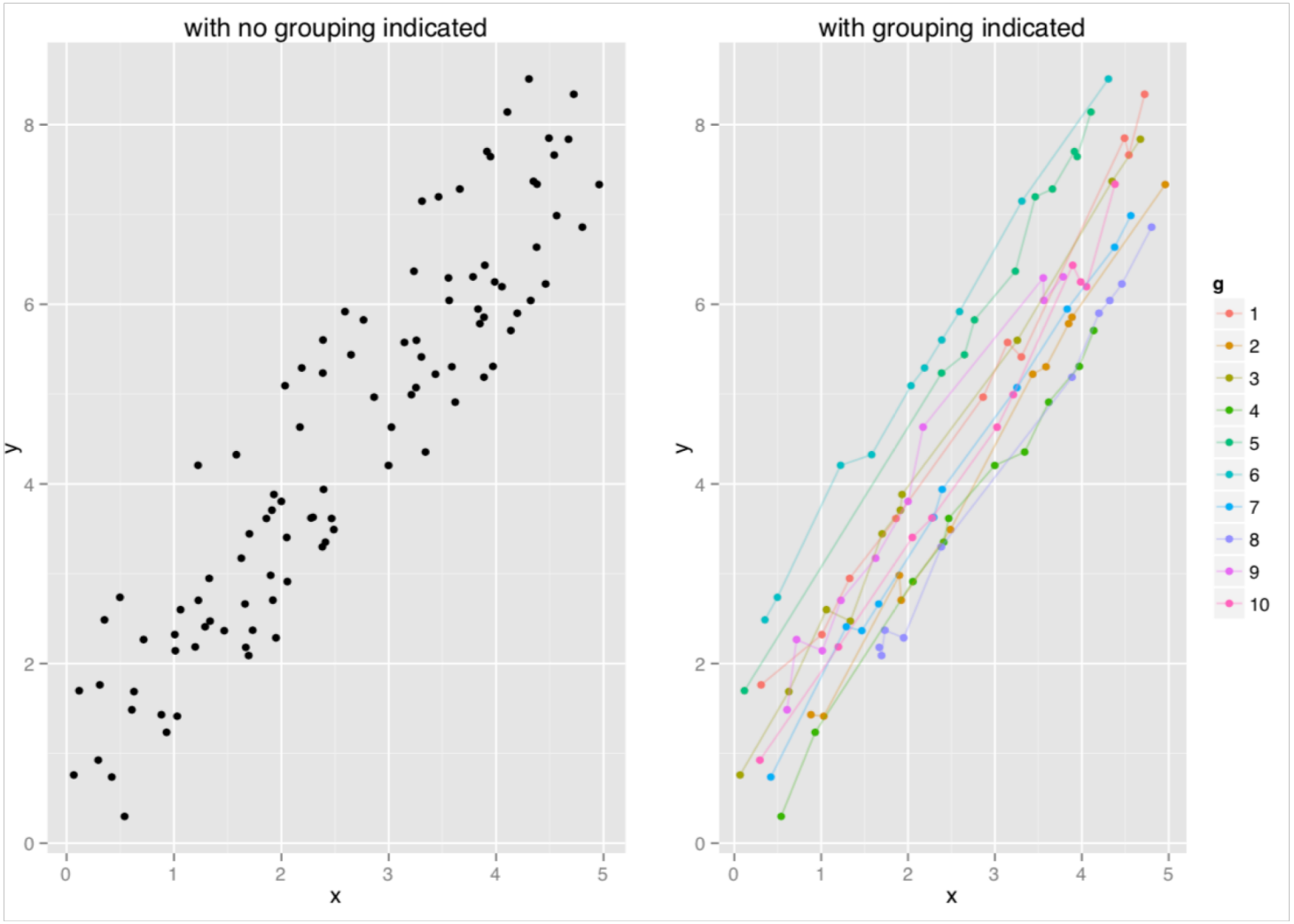
- Linear model with both fixed effects and random effects
- LME (**L**inear **M**ixed **E**ffects) models are suitable for clustered, longitudinal, or repeated measures data where data are grouped by *random levels* and response variable is continuous
- *Random levels* means that there are observations that belong to single subject or group that can be thought of as randomly selected from the population
- **Fixed effects have estimates of their effect on Y (slope / coefficient)**
- **Random effects have estimates of variation**

UNIVERSITY of VIRGINIA HEALTH SYSTEM

MIXED EFFECTS REGRESSION

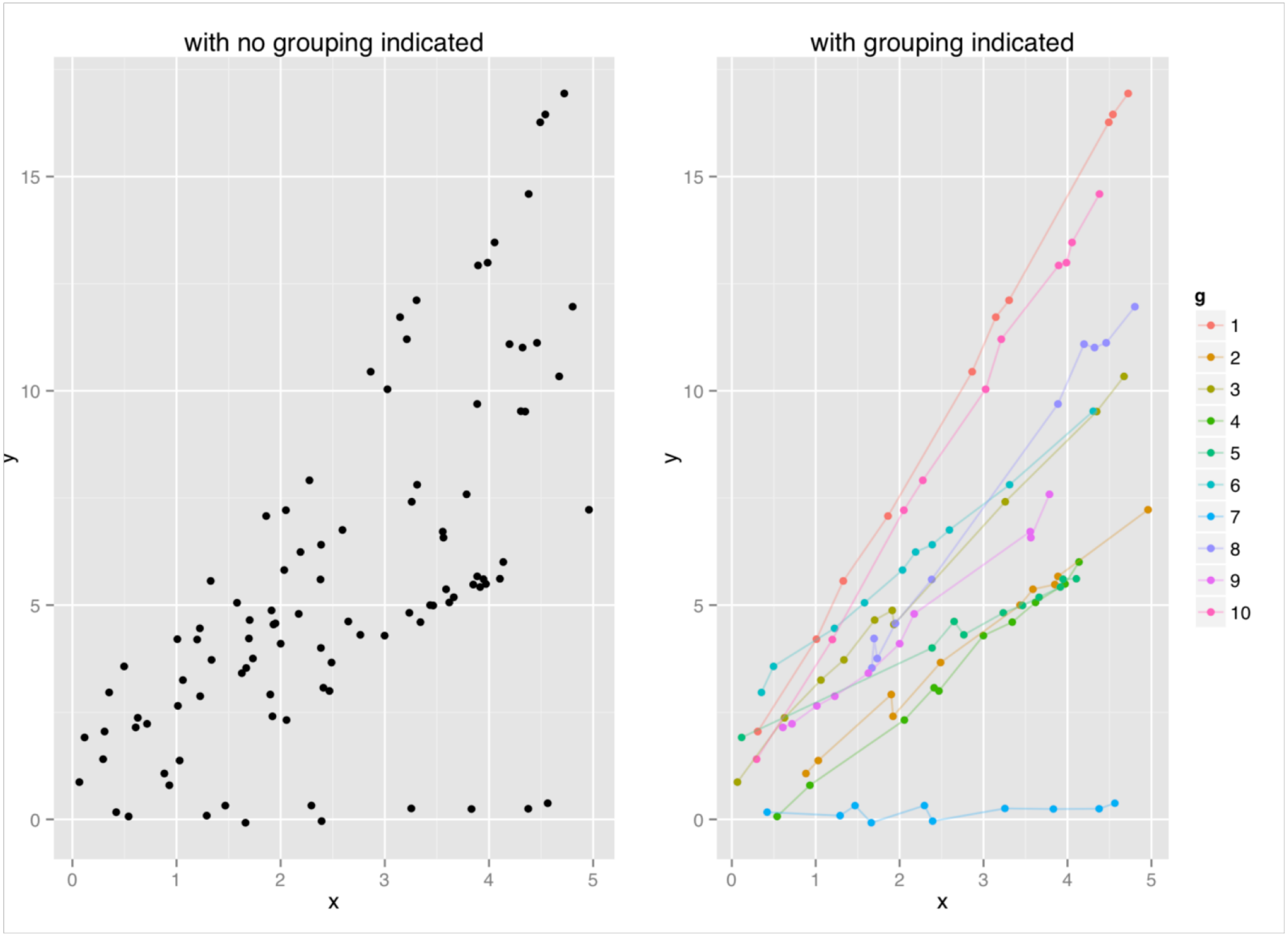
- Example: Treatment applied to subjects and outcome measured every 2 weeks for 2 months
- Treatment is a **fixed effect.** Other fixed effects could be age, sex, cell type, time. If we were to repeat this experiment, we would use these variables again
- These data are grouped by subject. The subjects we choose are a random sample from the population, so there are **random effects** associated with subjects. If we repeated the experiment, we would not use these subjects again
- In R: lmer() function in lme4 package and lmerTest package
- May need glht() function in multcomp package for multiple comparisons

NEED FOR MIXED EFFECTS



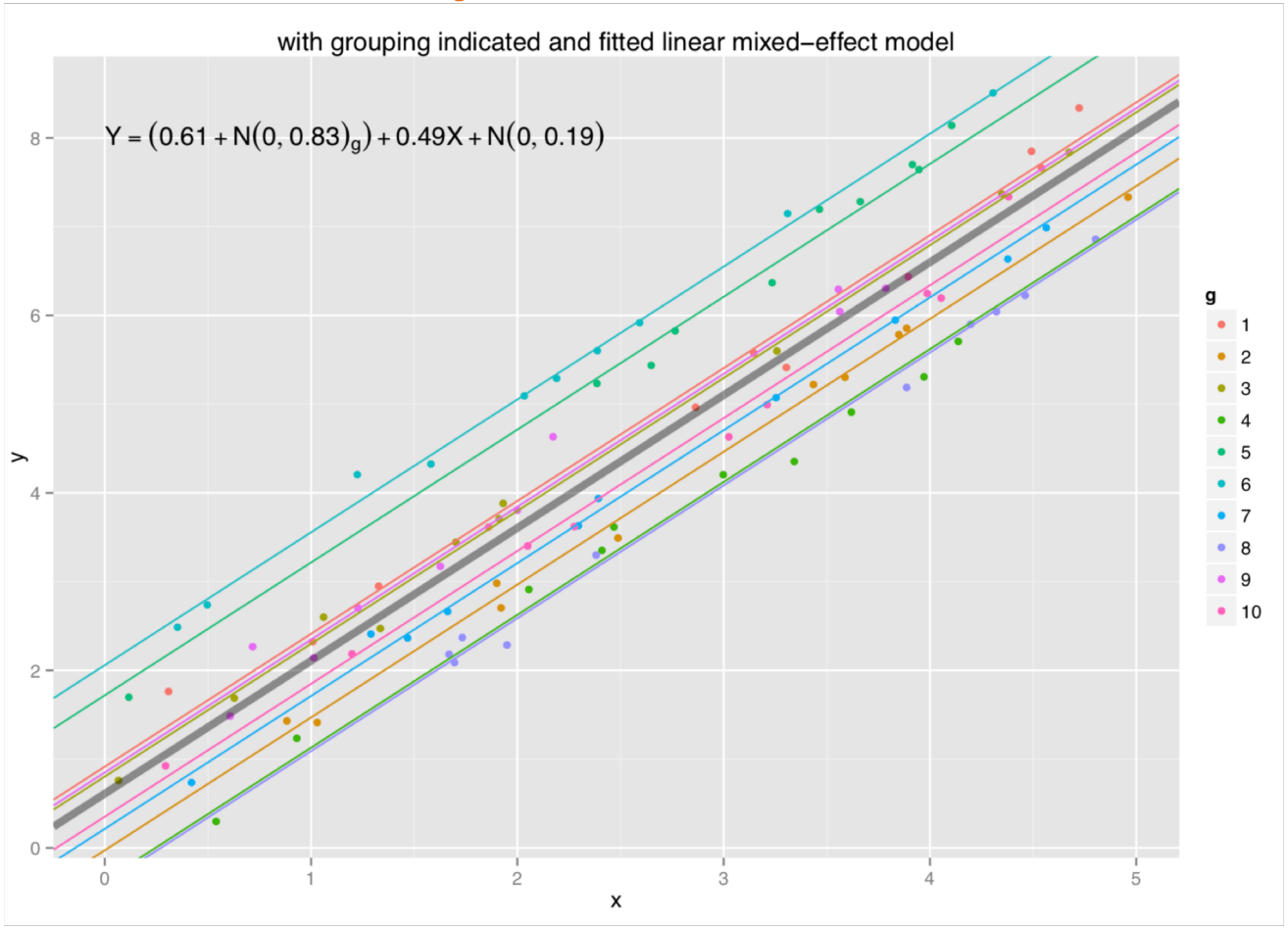
UNIVERSITY of VIRGINIA HEALTH SYSTEM

NEED FOR MIXED EFFECTS



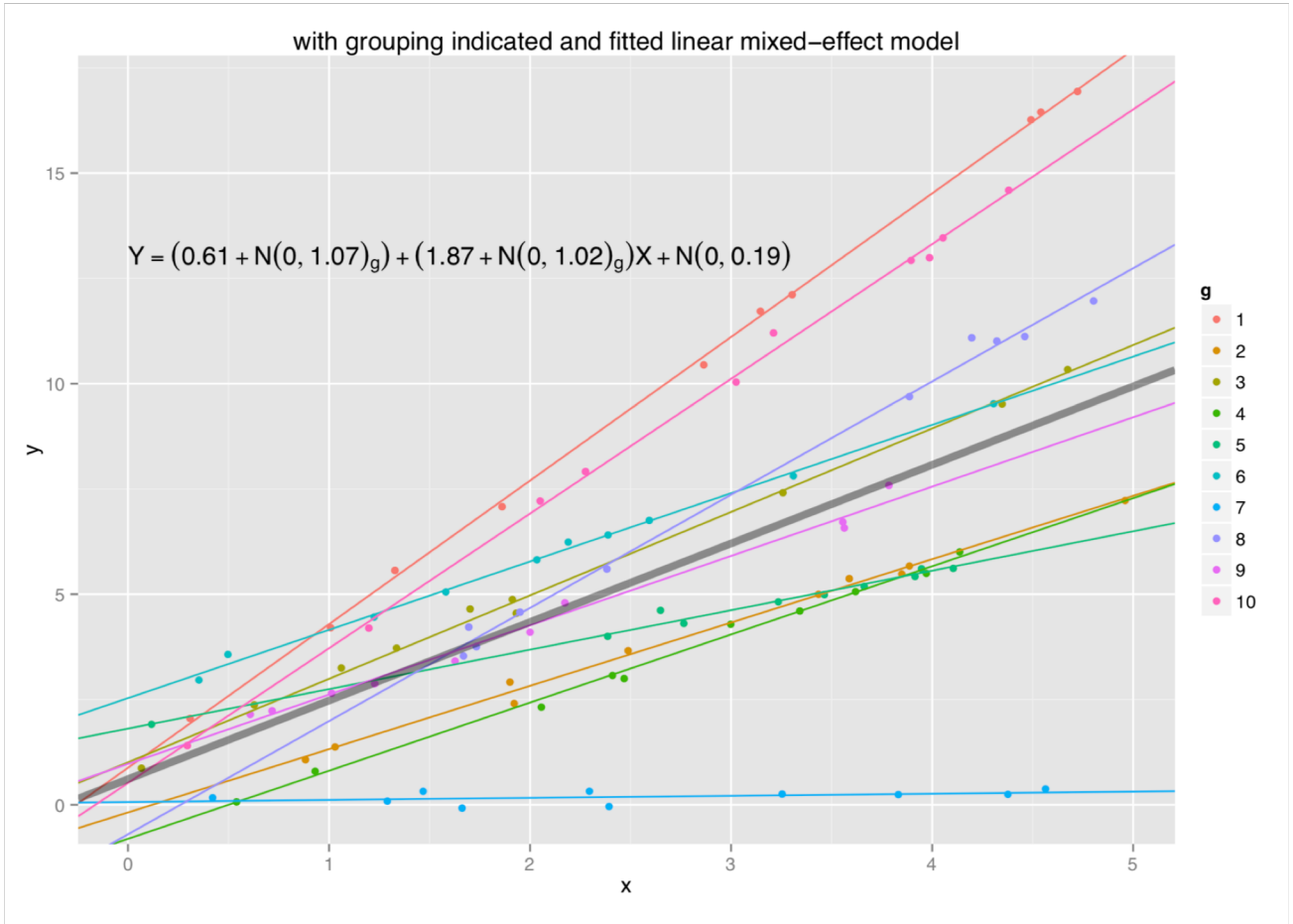
UNIVERSITY of VIRGINIA HEALTH SYSTEM

SAME SLOPE, DIFFERENT INTERCEPT




```
lmer(y ~ x + (1 | g), data = dat)
```

DIFFERENT SLOPE, DIFFERENT INTERCEPT




```
lmer(y ~ x + (x | g), data = dat)
```