Analysis of Covariance with Applications in R

Haiyang (Jason) Jiang, Sichen (Coco) Liu

Department of Statistics and Actuarial Science
Simon Fraser University

November 23, 2018

Outline

Motivation

Introduction

- Recall that Analysis of Variance (ANOVA) is used to test the difference in means between groups regarding treatment(s)
- However, ANOVA deals with categorical variables only
- What if we have a continuous variable and we want to include it into our model?

What is ANCOVA?

Introduction

- A potential **covariate** is any continuous variable that is significantly correlated with the dependent variable(s)
- **Analysis of Covariance (ANCOVA)** can be used for regression problems with a mixture of quantitative and qualitative predictors
- The continuous variables are used to adjust the dependent variable(s) before comparing the difference in group means [7]

Purpose of ANCOVA

Introduction

Covariates are used for [5]:

- **Elimination of Confounding Variables**: Once a possible confounding variable has been identified, ANCOVA is ideally suited to remove the systematic bias by entering the variable into the analysis as a covariate.
- **Reducing within-group error variance**: If we can use covariates to explain some of the unexplained variance in our data, then reducing the within-group error variance will allow us to assess the effect of the categorical variable more accurately.

Model Assumptions

Methodology

The assumptions for a one-way ANCOVA model (one-way ANOVA + one covariate) are:

1. Assumptions for ANOVA and linear regression: **normality, equal variance between treatments** [3]
2. **Homogeneity of regression slopes.** [7]

Model Estimations

Methodology

In general, an ANCOVA model can be written as:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= (\boldsymbol{Z}, \boldsymbol{X}) \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} + \boldsymbol{\epsilon} \end{aligned}$$

where

- $\boldsymbol{Z}\boldsymbol{\alpha}$ is the ANOVA part. $\boldsymbol{Z}$ is the model matrix which contains dummy variables with appropriate constraint; $\boldsymbol{\alpha}$ contains μ and $k-1$ treatment levels $\alpha_i, i = 1, ..., k-1$
- $\boldsymbol{X}\boldsymbol{\beta}$ is the regression part. $\boldsymbol{X}$ contains the observed values for the covariate, and $\boldsymbol{\beta}$ contains coefficients of the covariates (no β_0 here as we have added the baseline μ in the ANOVA part)
- $\boldsymbol{\epsilon}$ is the random error part. As usual, $\boldsymbol{\epsilon} \overset{iid}{\sim} N(0, \sigma^2 I)$

One-way (balanced) model with one covariate

Model Estimations

Suppose we have 3 treatment levels for α, and one covariate

$$\boldsymbol{y} = \boldsymbol{Z\alpha} + \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \epsilon_{ij}, i = 1, 2, 3, j = 1, 2, ..., n$$

Where

$$\boldsymbol{Z}_{nxk} = \begin{pmatrix} 1 & -1 & -1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \end{pmatrix} \quad \boldsymbol{\alpha}_{kx1} = \begin{pmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{pmatrix} \quad \boldsymbol{X}_{nx1} = \begin{pmatrix} x_{11} \\ \vdots \\ x_{21} \\ \vdots \\ x_{3n} \end{pmatrix}$$

- Set up the model matrix $\boldsymbol{Z}_{nx3}$ with the **sum to zero constraint**

General Procedure in ANCOVA

Applications in R

- The following steps for ANCOVA are recommended [2] [3]:
  - Import data
  - Explore the data through graphing the relationship between y and the covariate, grouped by treatment levels
  - Check if the assumption of normality and homoscedasticity are met for both the factor and the covariate by building models (aov() and lm()) separately, also check if there is any outlier or high leverage point
  - Check if the assumption of homogeneity of regression slopes is met by testing the significance of the interaction between the factor and the covariate
  - If everything goes well, we build the model and perform an ANCOVA test
  - Perform the Tukey's test

Import data

Applications in R

- **Fruitfly data**[4]: A study on whether the sexual activity of male fruit flies will affect their lifespan by Partridge and Farquhar (1981)
- Key points:
  - $n = 124$
  - Response variable: **longevity** (days)
  - Factor: **activity** (5 levels)
  - Covariate: **thorax** lengths of male fruit flies
- We first import data and load necessary packages:

```
library(car)
library(ggplot2)
library(multcomp)
library(faraway)

data("fruitfly",package='faraway')
```

Explore the data by graphing

Applications in R

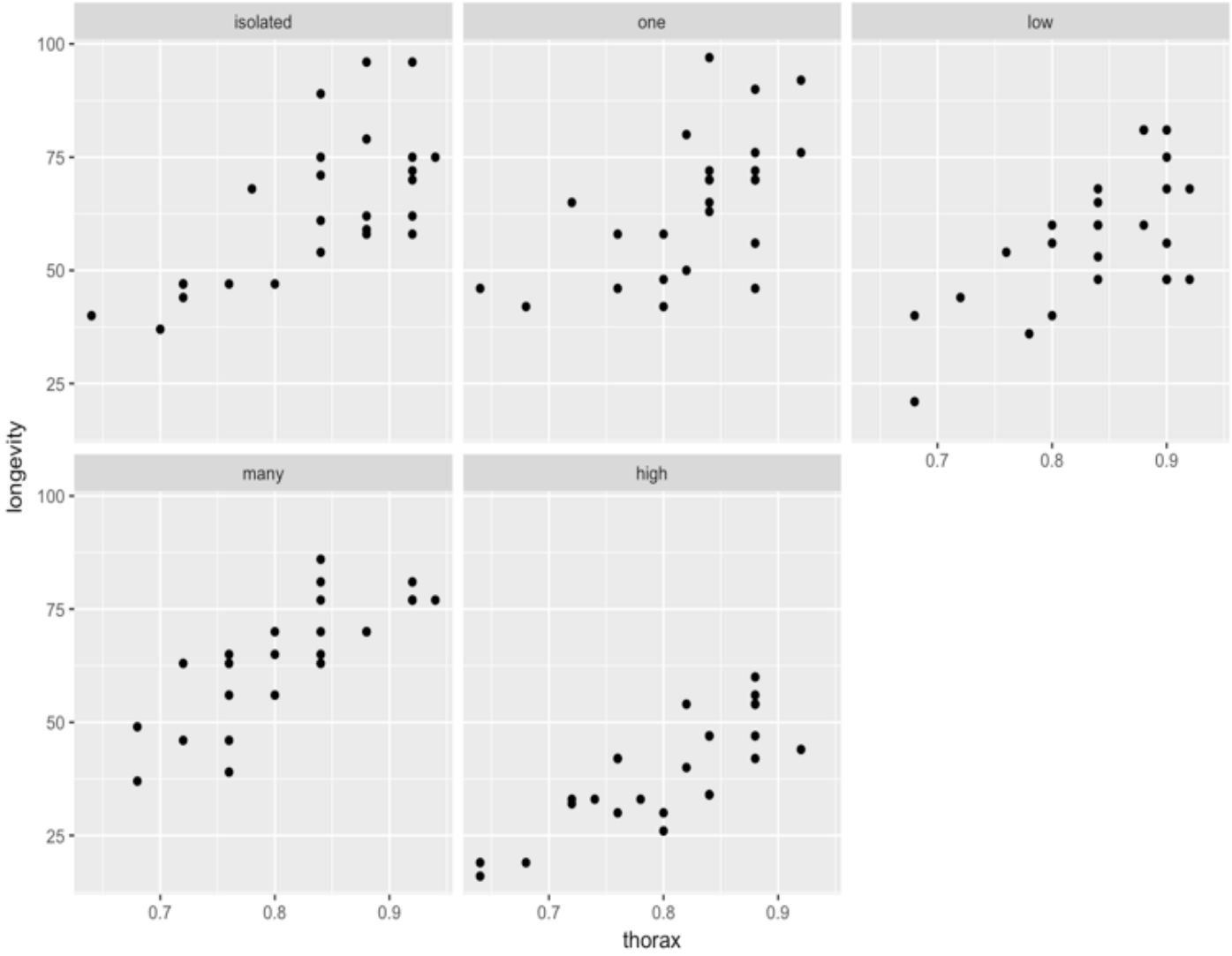
- We can first check the structure of our data

```
'data.frame': 124 obs. of  3 variables:
 $ thorax   : num  0.68 0.68 0.72 0.72 0.76 0.76 0.76 0.76 0.76 0.8 ...
 $ longevity: int  37 49 46 63 39 46 56 63 65 56 ...
 $ activity : Factor w/ 5 levels "isolated","one",..: 4 4 4 4 4 4 4 4 4 4 ...
```

- Since we are interested in whether the means will be different among all groups with the existence of the covariate, we would like to make some plots on the next slide. Since we have 5 groups, we cannot really tell the differences explicitly. The good news is that we know that there is a linear relationship between the covariate and the response variable

Explore the data by graphing

Applications in R



Check if the assumptions are met for both the factor and the covariate

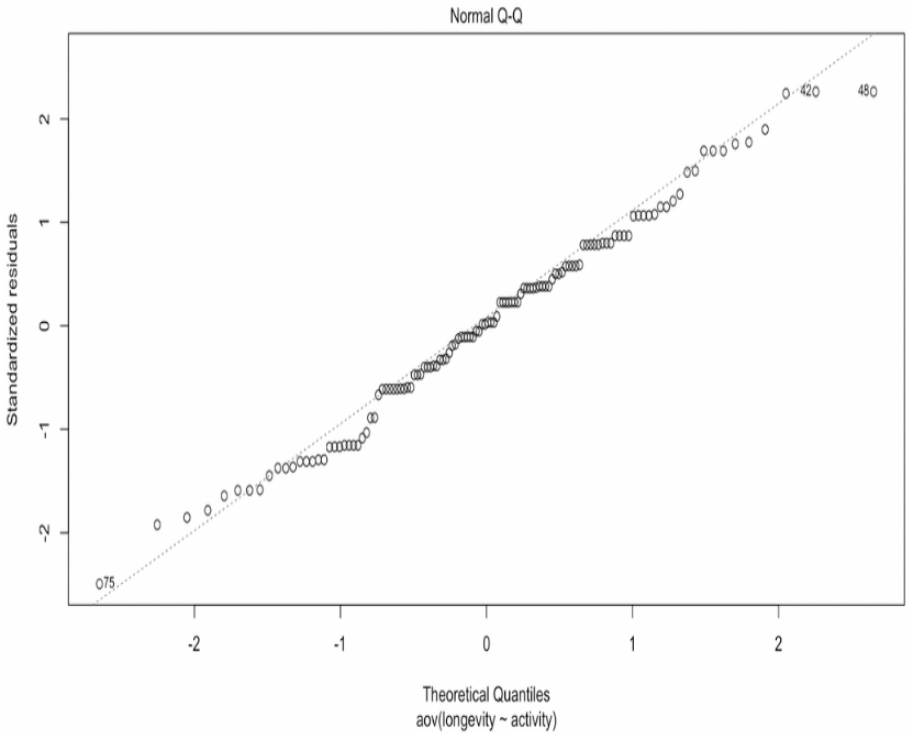
Applications in R

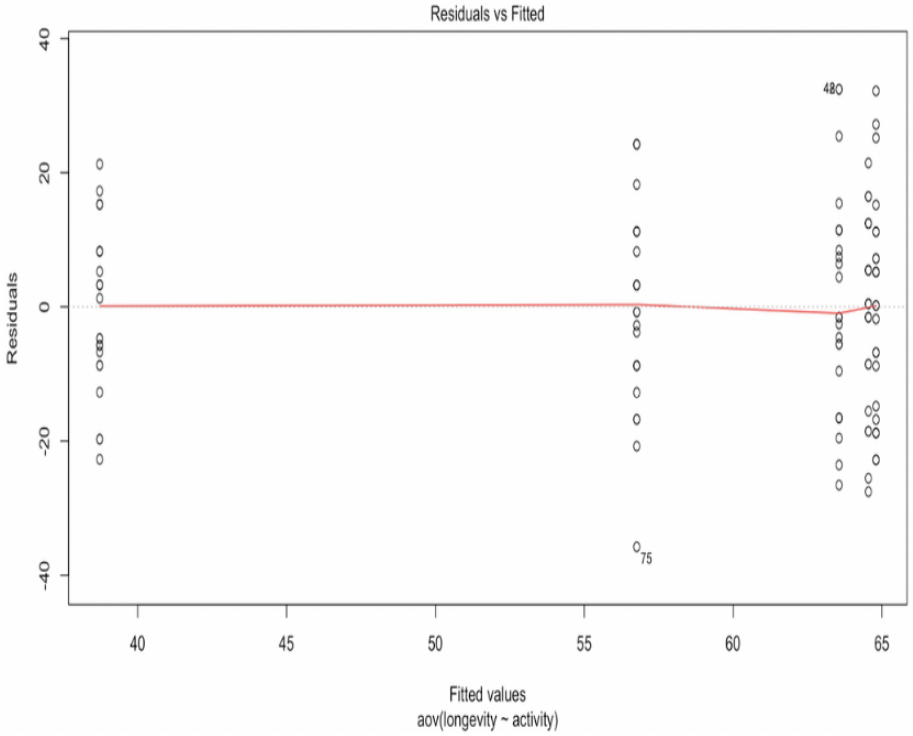
- As mentioned, we first need to see if the normality and equal variance assumptions are met for both the factor and the covariate
- The way of doing this is to build an ANOVA model by the aov() function for **activity** and build a linear model by the lm() function for **thorax**
- Based on the model diagnostics plots, there is no obvious outlier or high leverage point for both models
- The following plots are the Q-Q plot and residual plot for the ANOVA model

Check if the assumptions are met for both the factor and the covariate

Applications in R

Q-Q plot and residual plot for the ANOVA model:

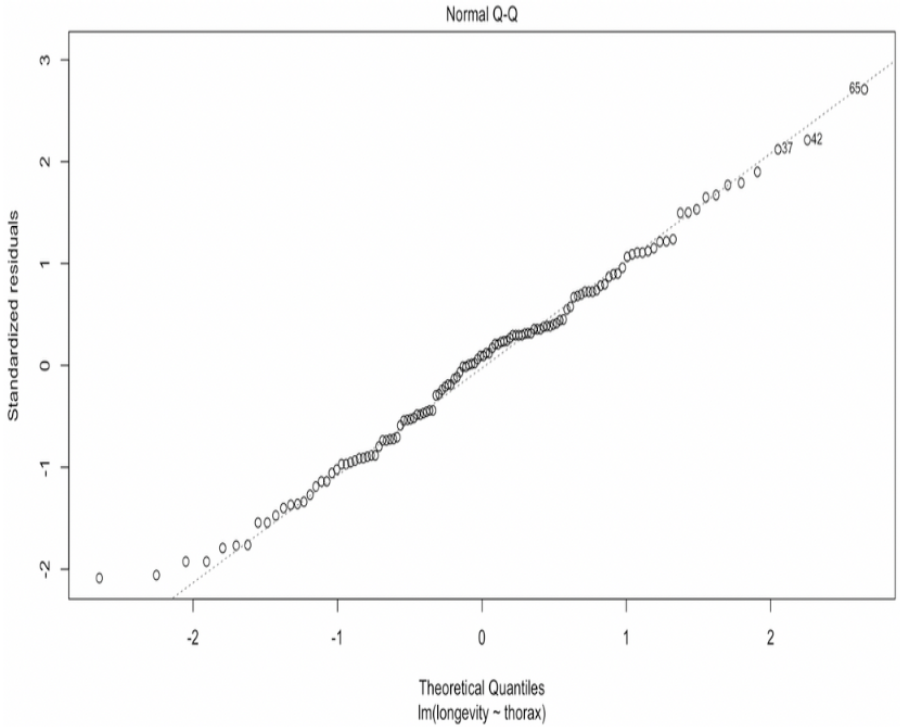


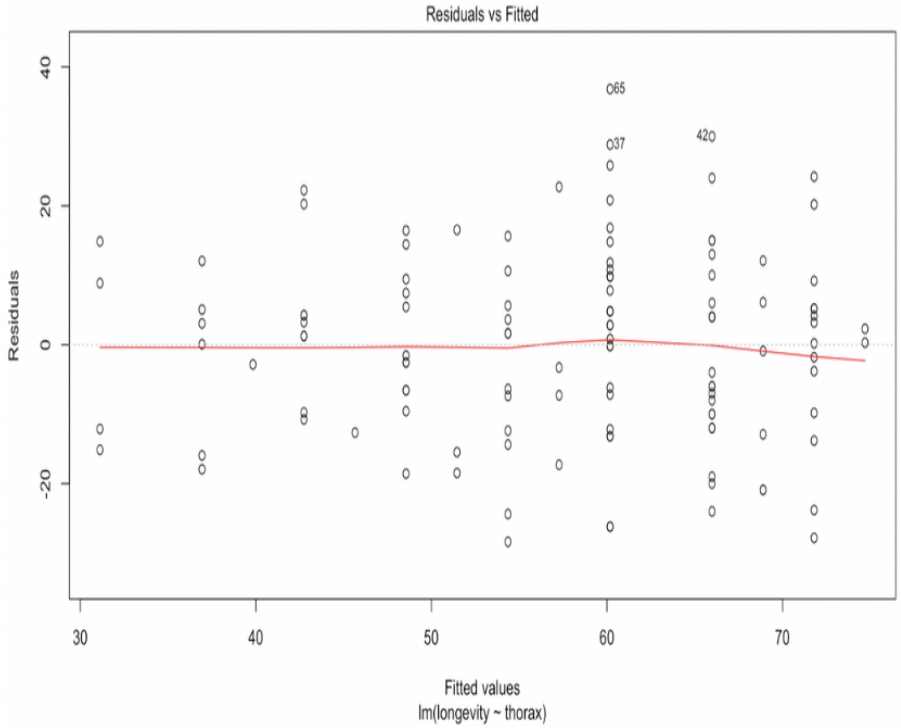


Check if the assumptions are met for both the factor and the covariate

Applications in R

Q-Q plot and residual plot for the linear regression model:





If the assumptions are not met, see the first presentation for transformations :)

Check if the assumption of homogeneity of regression slopes is met

Applications in R

- We need to make sure that including the covariate(s) into the model will not affect the relationship between factor(s) and the response variable.
- The reason behind it can be illustrated by the graph [3] on the next slide

Check if the assumption of homogeneity of regression slopes is met

Applications in R

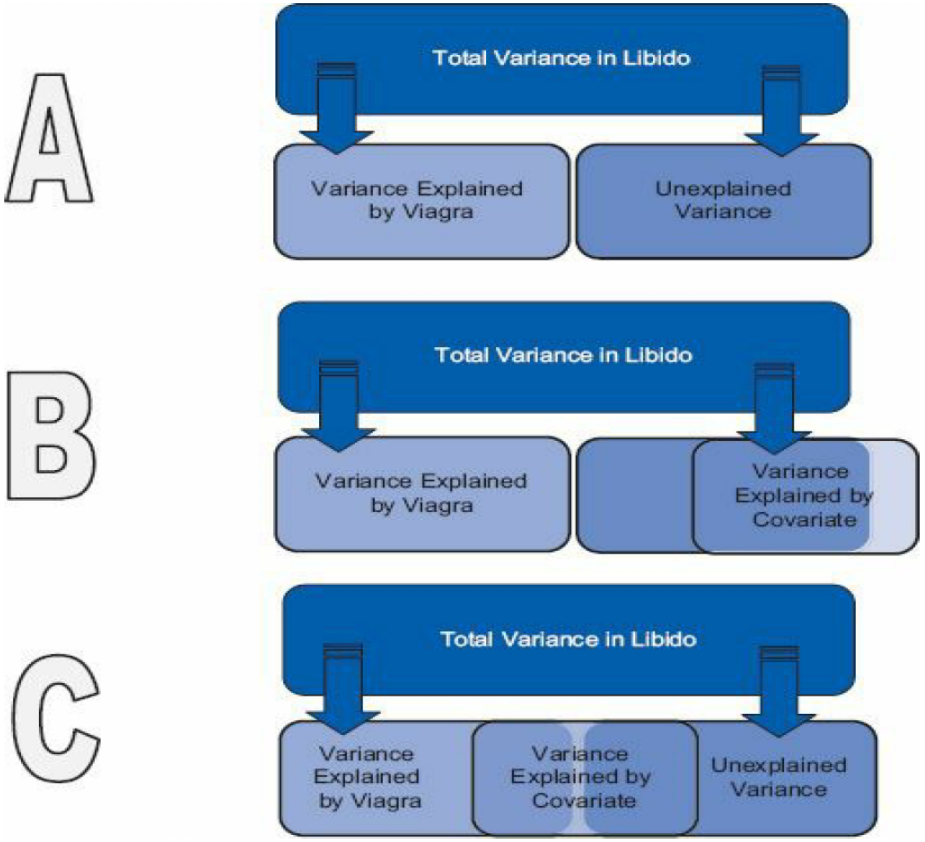
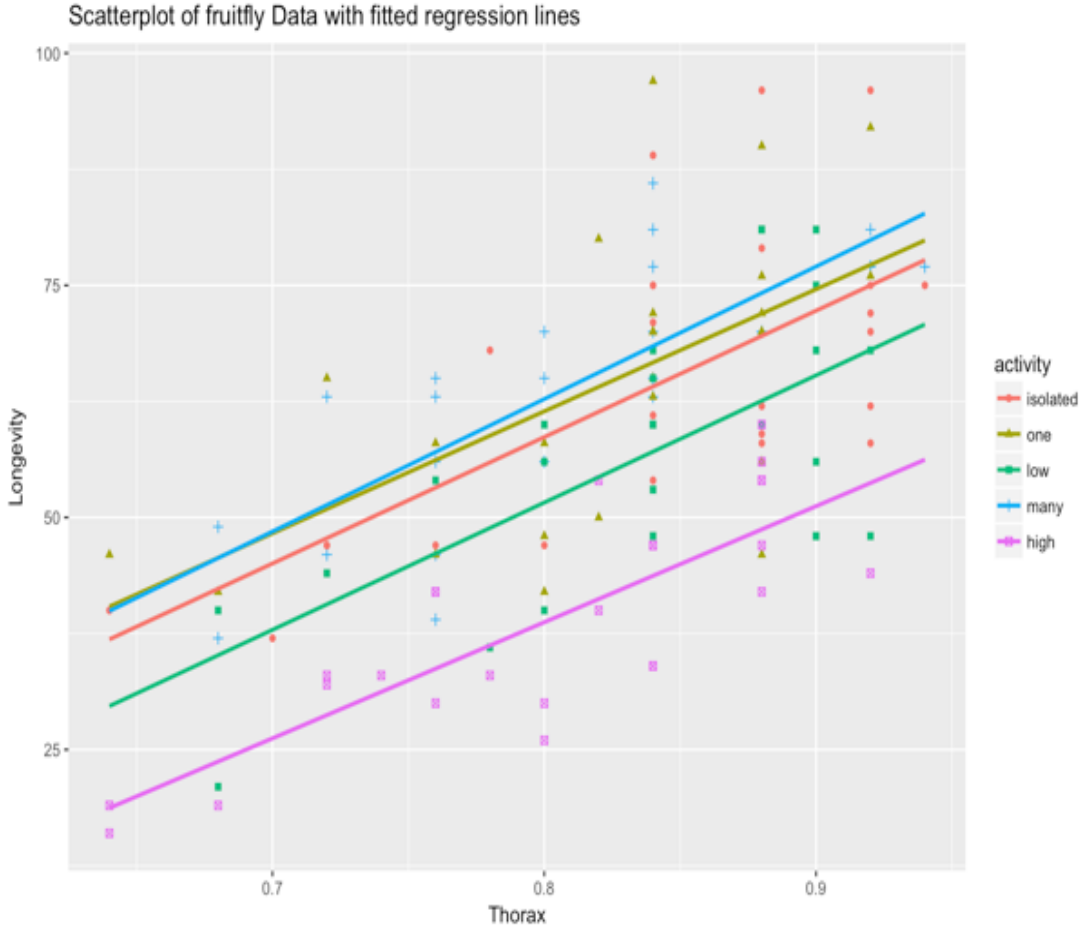


Figure: Illustration of ANCOVA [3]

Check if the assumption of homogeneity of regression slopes is met

Applications in R

- Graphically, we can plot the covariate vs the response variable for all five levels of activity, and we expect to see five parallel lines:



Check if the assumption of homogeneity of regression slopes is met

Applications in R

- Alternatively, we can perform a formal statistical test by checking whether or not the interaction between **activity** and **thorax** is significant

```
> full_mod <- aov(longevity ~ activity * thorax, data=fruitfly)
> summary(full_mod)
             Df Sum Sq Mean Sq F value  Pr(>F)
activity      4  12269    3067  26.728 1.2e-15 ***
thorax        1  12368   12368 107.774 < 2e-16 ***
activity:thorax  4     24       6   0.053   0.995
Residuals   114  13083     115
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- Based on the test result (or the graphics), we can conclude that there is no significant interaction between **thorax** and **activity**

Build the model and perform the ANCOVA test

Applications in R

- Intuitively, we can fit the model with both the factor and the covariate in R through the aov() function. But...

```
> activityfirst <- aov(longevity ~ activity + thorax, data=fruitfly)
> thoraxfirst <- aov(longevity ~ thorax + activity, data=fruitfly)
> summary(activityfirst)
            Df Sum Sq Mean Sq F value   Pr(>F)
activity     4  12269    3067   27.61 3.48e-16 ***
thorax       1  12368   12368  111.35  < 2e-16 ***
Residuals  118  13107     111
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> summary(thoraxfirst)
            Df Sum Sq Mean Sq F value   Pr(>F)
thorax       1  15003   15003  135.07  < 2e-16 ***
activity     4   9635    2409   21.68 1.97e-13 ***
Residuals  118  13107     111
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

What just happened?

Applications in R

- Observation: Although both **thorax** and **activity** are significant in these two models, all Sum of Squares and p-values in these two models are different
- To explain this phenomenon, we need to talk about three types of Sum of Squares [6][1]

Three types of Sum of Squares

Applications in R

Type I Sum of Squares (Sequential)

- Default setting for lm() and aov() functions in R
- The SS for each factor is the incremental improvement in the error SS as each factor effect is added to the regression model

Type II Sum of Squares (Partially Sequential)

- It is the reduction in residual error due to adding the term to the model after all other terms **except those that contain it**

Type III Sum of Squares (Marginal)

- It gives the sum of squares that would be obtained for each variable if it were entered **last** into the model

Notes

Applications in R

Both Type II and Type III Sum of Squares are:

- Invariant to the order in which effects are entered into the model
- Available in the Anova() function in the 'car' package

Since we are not allowed to have the interaction effect between the covariate and the factor, both Type II or Type III SS are the same

Perform an ANCOVA test

Applications in R

- In our example, I will use the Anova() function in the 'car' package with the Type II Sum of Squares

```
> red_mod <- aov(longevity ~., data=fruitfly)
> Anova(red_mod,type='2')
Anova Table (Type II tests)

Response: longevity
           Sum Sq  Df F value    Pr(>F)
thorax    12368.4   1 111.348 < 2.2e-16 ***
activity   9634.6   4  21.684 1.974e-13 ***
Residuals 13107.3 118
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- Interpretation: we can clearly see that both the factor and covariate are significant. That means the longevity of the fruitfly are different for different levels of sexual activity, and the thorax length of each male fruitfly is an important predictor to the longevity

Perform the Tukey's test

Applications in R

- We would like to find out which two groups' **adjusted means** are different
- **Adjusted** because the covariate comes into play
- The usual TukeyHSD() function will not work

```
> TukeyHSD(red_mod)
Error in rep.int(n, length(means)) : unimplemented type 'NULL' in 'rep3'
In addition: Warning message: In replications(paste(" ", xx), data = mf) :
non-factors ignored: thorax
```

- Use the glht() function in the **multcomp** package instead (glht stands for General Linear Hypothesis Testing)

Perform the Tukey's test

Applications in R

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = longevity ~ ., data = fruitfly)

Linear Hypotheses:
                     Estimate Std. Error t value Pr(>|t|)
one - isolated == 0     2.637      2.984   0.884  0.90237
low - isolated == 0    -7.015      2.981  -2.353  0.13578
many - isolated == 0    4.139      3.027   1.367  0.64961
high - isolated == 0  -20.004      3.016  -6.632  < 0.001 ***
low - one == 0         -9.652      2.985  -3.234  0.01349 *
many - one == 0         1.502      3.016   0.498  0.98743
high - one == 0       -22.641      2.999  -7.550  < 0.001 ***
many - low == 0        11.154      3.029   3.683  0.00312 **
high - low == 0       -12.989      3.019  -4.302  < 0.001 ***
high - many == 0      -24.142      3.016  -8.005  < 0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Conclusion

Applications in R

- We demonstrated the necessity of ANCOVA and how we can run an ANCOVA test
- There could be two or more covariates that we would like to include into our model. Variable selection schemes such as Information Criterion or stepwise regression, could be used to help select the best model
- For your interest, **Discovering Statistics using R** also talked about the remedy to the situation where the equal slope assumption is violated, which is called the robust version of ANCOVA

Reference

- Matt Cooper. *Anova fffdfffdfffd Type I/II/III SS explained*. May 2012. URL: https://mcfromnz.wordpress.com/2011/03/02/anova-type-iiiiii-ss-explained/.
- Julian J Faraway. *Linear models with R*. Chapman and Hall/CRC, 2016.
- Andy Field, Jeremy Miles, and Zoë Field. *Discovering statistics using R*. Sage publications, 2012.
- Linda Partridge and Marion Farquhar. "Sexual activity reduces lifespan of male fruitflies". In: *Nature* 294.5841 (1981), p. 580.
- Keenan A Pituch and James P Stevens. *Applied multivariate statistics for the social sciences: Analyses with SAS and IBMfffdfffdfffds SPSS*. Routledge, 2015.
- Nancy Reid. *STA442/2101F: Applied Statistics I*. Oct. 2009. URL: http://www.utstat.utoronto.ca/reid/442F09.html.