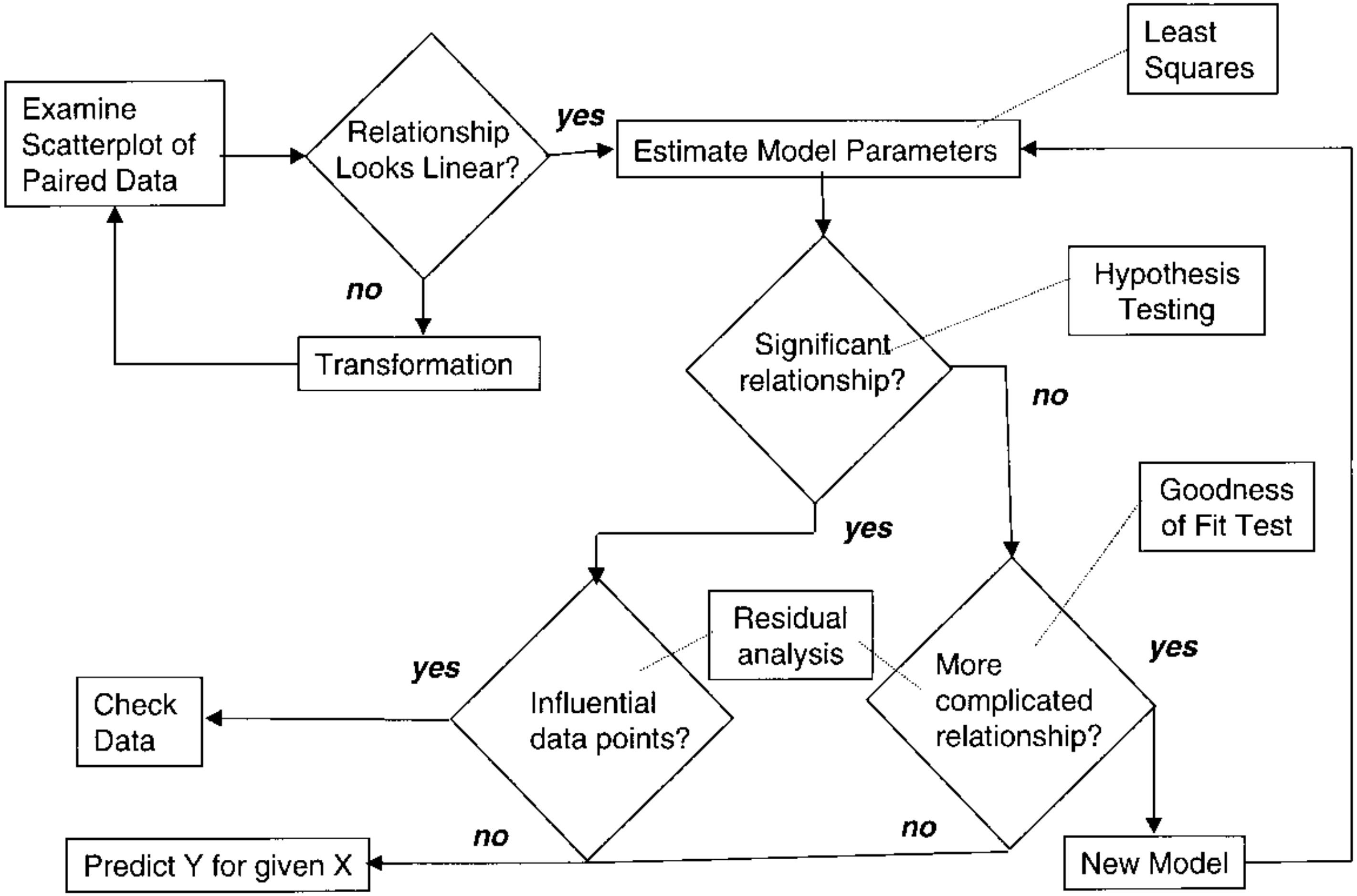
Data Analytics

Yong Zheng

Illinois Institute of Technology
Chicago, IL, 60616, USA

School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY

Multiple Linear Regression



11-16

Steps for Linear Regression Models

- Understand your data, figure out x and y variables
- According to the size of data, make a decision about which evaluation strategy you are going to use, hold-out or N-fold cross validation
- Examine the relationship between y and x. Apply transformations to x or y variables, if necessary
- Split the data if necessary
  - Hold-out evaluation: split data to train and test sets, build models based on train, and test it over test set
  - N-fold: use full data or a sample of the data to build models first, and evaluate models by N-fold

Steps for Linear Regression Models

- ❑ Build Models by different feature selections
- ❑ For each model, validate they are qualified or not
  - ❑ Vif function to examine multi-collinearity problem
  - ❑ F-test to examine at least one x variable is influential
  - ❑ Residual analysis to examine the assumptions of residual
- ❑ Evaluate your models
  - ❑ Hold-out: evaluate models based on the testing set
  - ❑ N-fold: using cv.glm()
  - ❑ Metrics: MAE, RMSE, MSE

Steps for Linear Regression Models

- ❑ Improve your models
  - ❑ Using nominal data?
  - ❑ Try higher-order terms, if necessary
  - ❑ Try interaction terms, especially dummy vs numerical variable
  - ❑ Identify and remove influential points?
- ❑ Final Steps
  - ❑ Write down the model
  - ❑ Try to explain it
  - ❑ Use the best model to make predictions

Clarification

- Hold-out evaluation
  - Split data into train and test
  - Build models based on train
  - Predict and evaluate it based on the test set
- You can use lm or glm to build models
  - fit = lm (y~x1+x2+x3+…., data=train)
  - You should use the column names in your data set to represent y, x1, x2, x3, …., In this case, you build a general model which can use predict() on test set
  - Otherwise, you may get some error message, 'newdata had 16 rows but variables found have 36 rows."

ANOVA

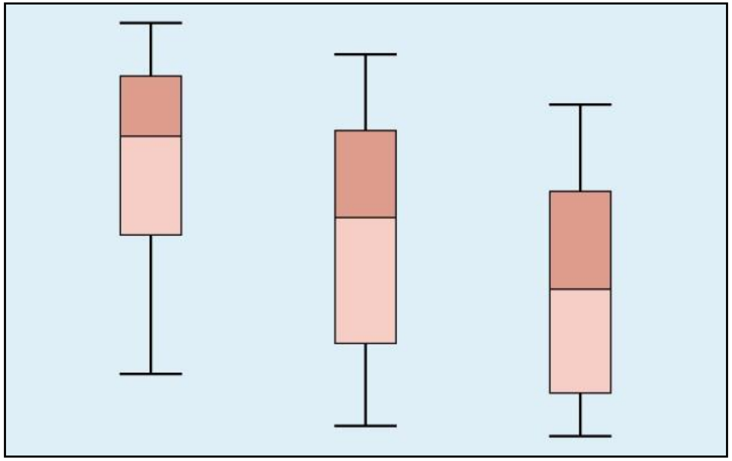
Comparing two groups

If we are going to compare two groups, such as group means, We can use two-sample t-test.

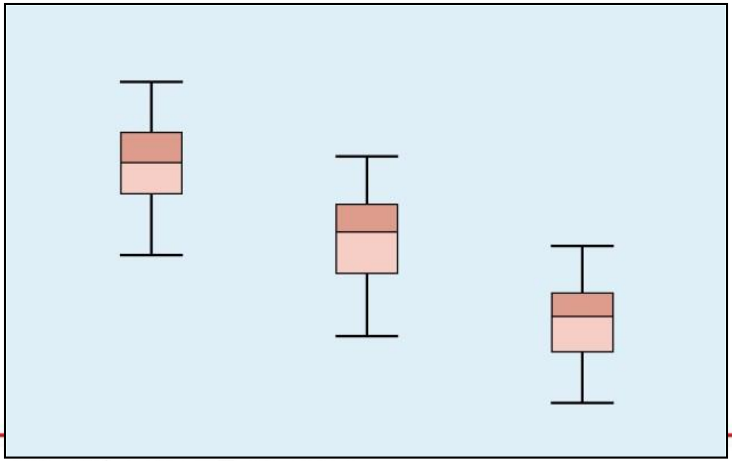
But, how about comparison among more than two groups?

Side-by-side box plots – how to interpret them

The sample medians are the same in both figures – but the variation is different. Which figure shows the highest variation?



The large variation within groups in the top figure suggests that the difference among the sample medians could be simply due to chance variability.



The data in the picture below are much more convincing that the populations differ.

ANOVA

Analysis of variance (ANOVA) is an approach to compare statistics (such as means) among more than two groups.

From the name, we can see it is not the analysis on the group means, but the variance. Why????????? ➔ because the variance matters!!!! See the box plots

The goal in ANOVA: compare group means among more than two groups by analyze the variances!! The two-sample t-test can be considered as a special case of ANOVA, when there are only two groups.

ANOVA, is also an application of linear regression models.

Comparing more than two groups

Problem:

- We have K **independent** simple random samples from each of K populations.
- Each population is **normal** with unknown average μ_t
- All the populations have the same standard deviation σ

Question: *Are the observed differences among the sample means statistically significant? (or just due to chance variability?)*

Answer: The ANOVA F-test tests the hypotheses***:***

Ho: $\mu_1 = \mu_2 = \ldots = \mu_K$ - *the averages are all equal*

Ha: not all the μ_t *are equal*

Steps for comparing K groups

1. Be sure that the observations arise from independent groups!

2. Draw side-by-side box plots for the groups, to visualize the differences among the groups and the within-group variation

3. Estimate the ANOVA regression model for t=1,…,K

   $$y_{it} = \mu_t + e_{it}$$

   where the errors e_{it} are normally distributed and with constant standard deviation σ. Use the regression F-test to check the hypothesis that the averages are equal.

4. Examine the residuals to verify that the model assumptions are satisfied.

Example: Stepping up your heart rate

Consider the following data, collected in a study to explore the relationship between a person’s heart rate and the frequency at which that person stepped up and down on steps of various heights. Each subject was assigned at random to a combination of step height and frequency of stepping. The subject performed the activity for three minutes, and his/her pulse was counted after 20 seconds.

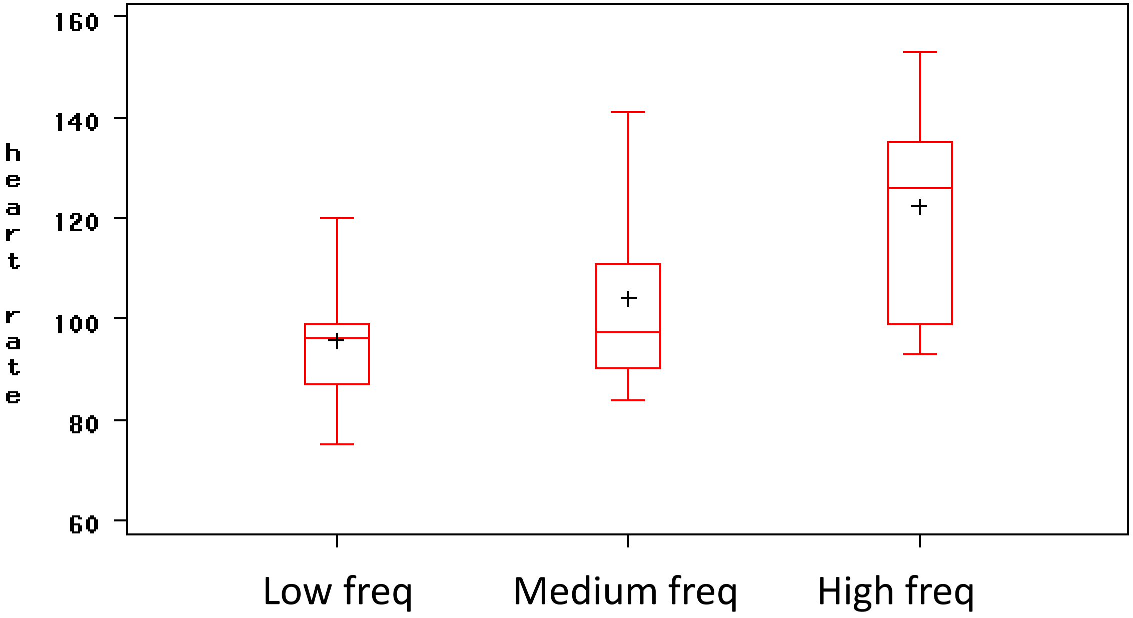
Response variable: heart rate measured in beats per minute after exercise.

Factors:

- *Step heights*: 5.75 inches (low - 1), and 11.5 inches (high - 2).
- *Rates of stepping*: 14 steps/min. (low - 1), 21 steps/min. (medium - 2), and 28 steps/min. (high - 3).

Question: how does stepping rate affect heart beat?

Step 1: Box plots



What does the figure indicate about:

1. The differences among the averages?
2. The within-group variation?

Summary statistics for the three groups

```
                         Stepping up the heart rate
                        The MEANS Procedure
                Analysis Variable : HR heart rate

                          Lower      Upper
      N    Mean   Std Dev    Median  Quartile  Quartile  Min    Max
      ------------------------------------------------------------
Low     10    95.70  12.61    96.00     87.00       99.00    75.00 120.00
Medium  10   104.10  18.44    97.50     90.00      111.00    84.00 141.00
High    10   122.40  20.82   126.00     99.00      135.00    93.00 153.00
          ------------------------------------------------------------
```

Define the ANOVA linear model $y_{it} = \mu_t + e_{it}$

μ_1=average heart rate for subjects in the **low** stepping rate group
μ_2=average heart rate for subjects in the **medium** stepping rate group
μ_3=average heart rate for subjects in the **high** stepping rate group

Fit regression model for y_{it}, by introducing two dummy variables:
X_1=1 for FREQ=2 (Medium) and X_1=0 otherwise
X_2=1 for FREQ=3 (High) and X_2=0 otherwise

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

GROUP 1 (low freq) : X_1=0, X_2=0 $\qquad \hat{\mu}_1 = \hat{\beta}_0$

GROUP 2 (medium freq): X_1=1, X_2=0 $\qquad \hat{\mu}_2 = \hat{\beta}_0 + \hat{\beta}_1$

GROUP 3 (high freq): X_1=0, X_2=1 $\qquad \hat{\mu}_3 = \hat{\beta}_0 + \hat{\beta}_2$

ANOVA model ≡ regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

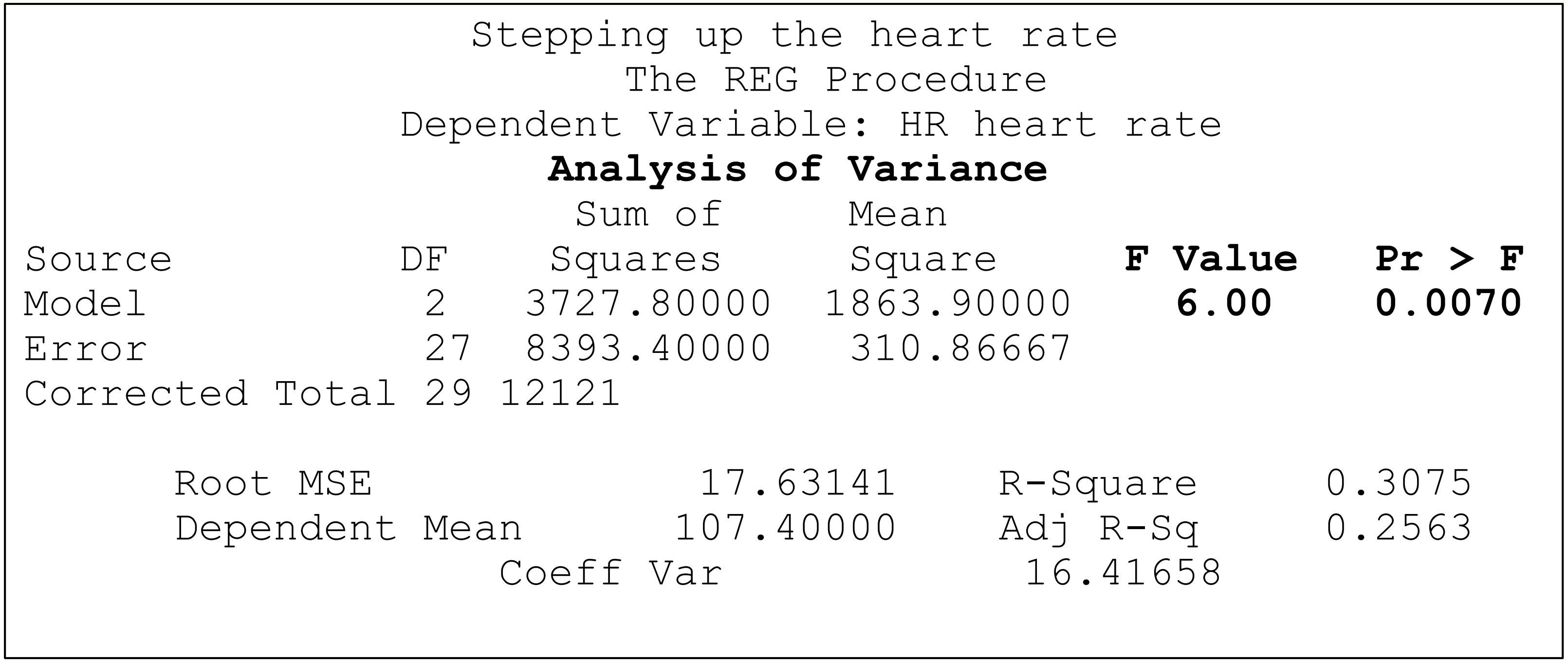
$$\hat{\beta}_1 = \hat{\mu}_2 - \hat{\mu}_1$$

Estimated difference between the average heart rates in the ***medium and low*** *frequency groups*

$$\hat{\beta}_2 = \hat{\mu}_3 - \hat{\mu}_1$$

Estimated difference between the average heart rates between the ***high and low*** *frequency groups*

Fitted model – ANOVA table



```
                     Stepping up the heart rate
                          The REG Procedure
                   Dependent Variable: HR heart rate
                        Analysis of Variance
                        Sum of        Mean
Source             DF     Squares       Square       F Value    Pr > F
Model               2   3727.80000  1863.90000      6.00     0.0070
Error              27   8393.40000   310.86667
Corrected Total 29 12121

      Root MSE                17.63141    R-Square     0.3075
      Dependent Mean         107.40000    Adj R-Sq     0.2563
                   Coeff Var                16.41658
```

The ANOVA table displays the value of the F-test statistic and its p-value for the test on the hypotheses:

Ho: $\mu_1 = \mu_2 = \mu_3$ (all the averages are equal if all betas = 0, i.e., $\beta_1 = \beta_2 = 0$)

Ha: not all the averages are equal.

Interpretation of the ANOVA table

The F-test statistic is F=6.00 with p-value=0.0070 (< 0.05).

What does this test indicate?

Does the heart rate change with a different frequency of stepping?

Further analysis using the parameter estimates

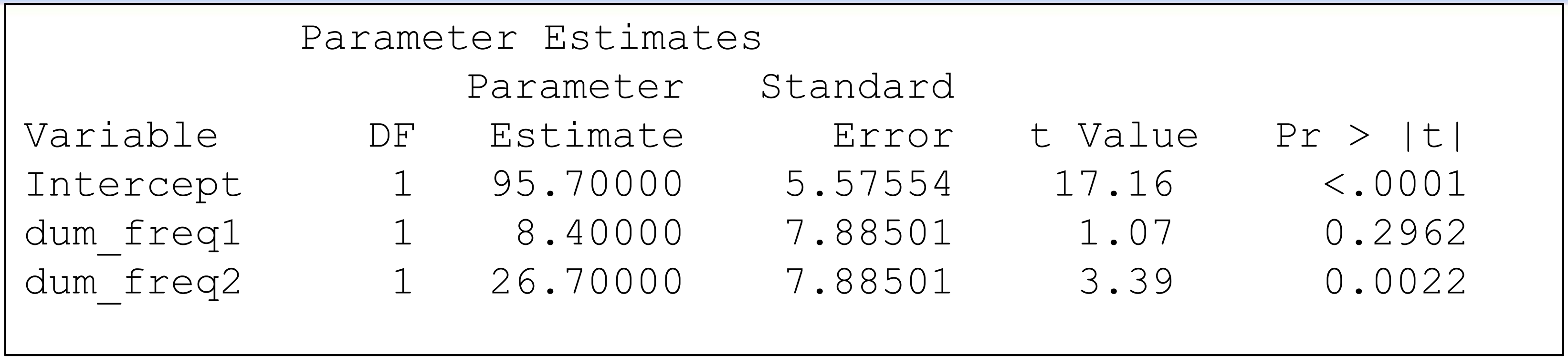
From the ANOVA table, we conclude that the population averages are not all equal. The next question is, are they all different?

The fitted regression model can be used to answer that question.

```
           Parameter Estimates
                          Parameter    Standard
  Variable          DF     Estimate       Error    t Value    Pr > |t|
  Intercept          1     95.70000     5.57554     17.16       <.0001
  dum_freq1          1      8.40000     7.88501      1.07       0.2962
  dum_freq2          1     26.70000     7.88501      3.39       0.0022
```

Examine the t-test to see if some variables are not statistically significant. If the dummy variable is not significant, then the corresponding coefficient can be assumed equal to zero. This implies that the corresponding difference between the averages is likely to be zero.

Further analysis using the parameter estimates

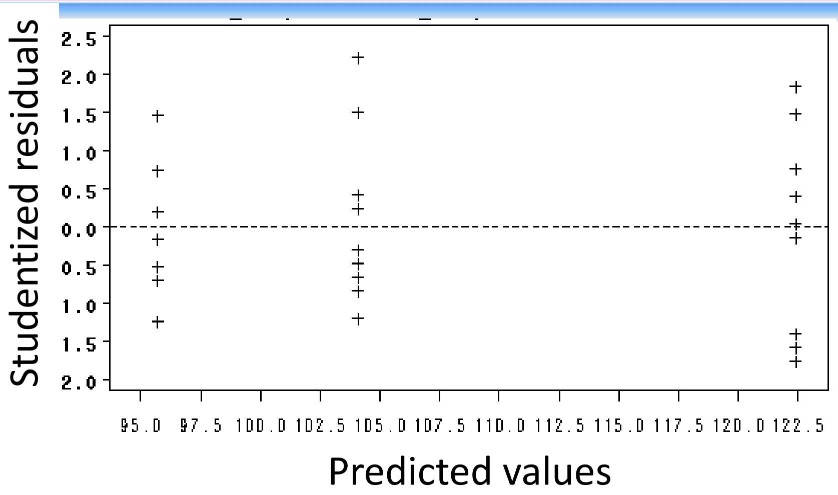


```
                     Parameter Estimates
                            Parameter    Standard
Variable          DF     Estimate       Error    t Value    Pr > |t|
Intercept          1     95.70000     5.57554     17.16       <.0001
dum_freq1          1      8.40000     7.88501      1.07       0.2962
dum_freq2          1     26.70000     7.88501      3.39       0.0022
```

1. T-test on coefficient of dum_freq1 is not significant → The heart rate averages for the low stepping rate and the medium stepping rate are not significantly different.

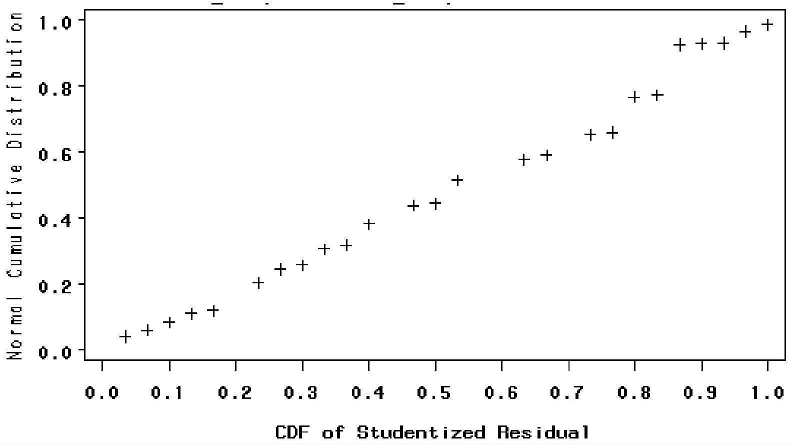
2. The t-test for β_2 indicates that the parameter is not equal to zero. Heart rate average varies significantly for low stepping rate and high stepping rate. *In particular, the average heart rate of a subject involved in the high frequency exercise is about 26.7 beats for minute higher than the average heart rate of the subject involved in the low frequency exercise*.

Check model assumptions: residual analysis



Is the within-groups variability constant?

Normal probability plot



Is the normality assumption satisfied?

Summary

- We examined the statistical analyses of data arising from a completely randomized experiment: K independent random samples for the K treatments or K levels of a factor variable.
- Perform an analysis of variance to check if the K population averages of the groups are equal.
- Check that the ANOVA assumptions are satisfied, if not use alternative methods
- Know how to observe the p-value of the individual parameter test, and how to interpret the coefficient in different situations.

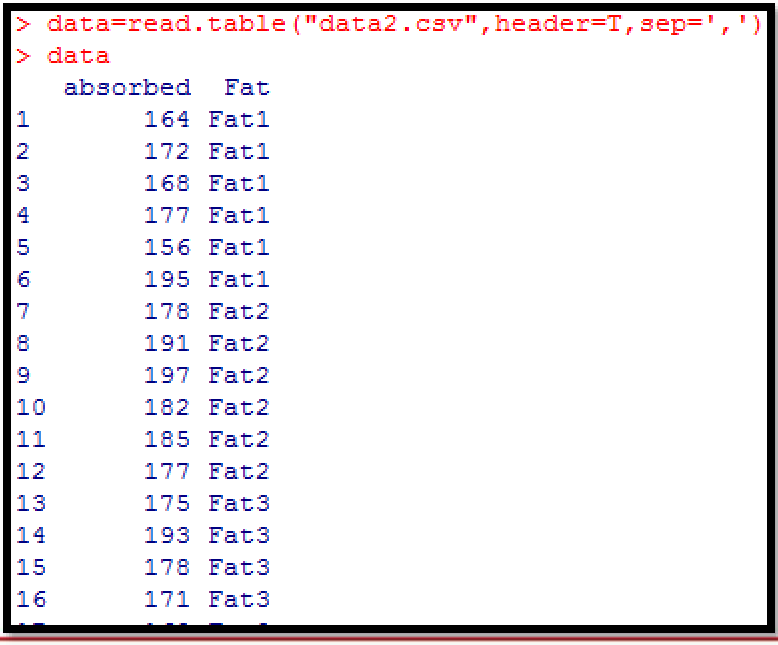
Example: Donuts

Research tries to explore the relation between the number of fat absorbed and different fat groups.

Response variable: The number of fat absorbed

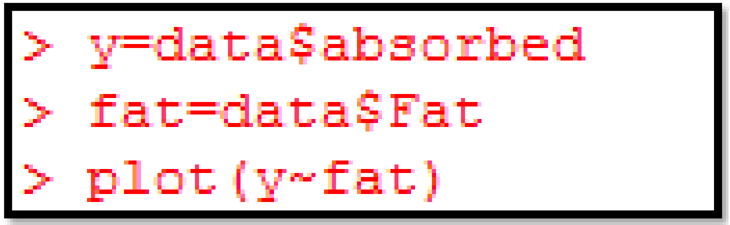
Fat groups: FAT1, FAT2, FAT3, FAT4

The data can be shown in either data1.csv or data2.csv

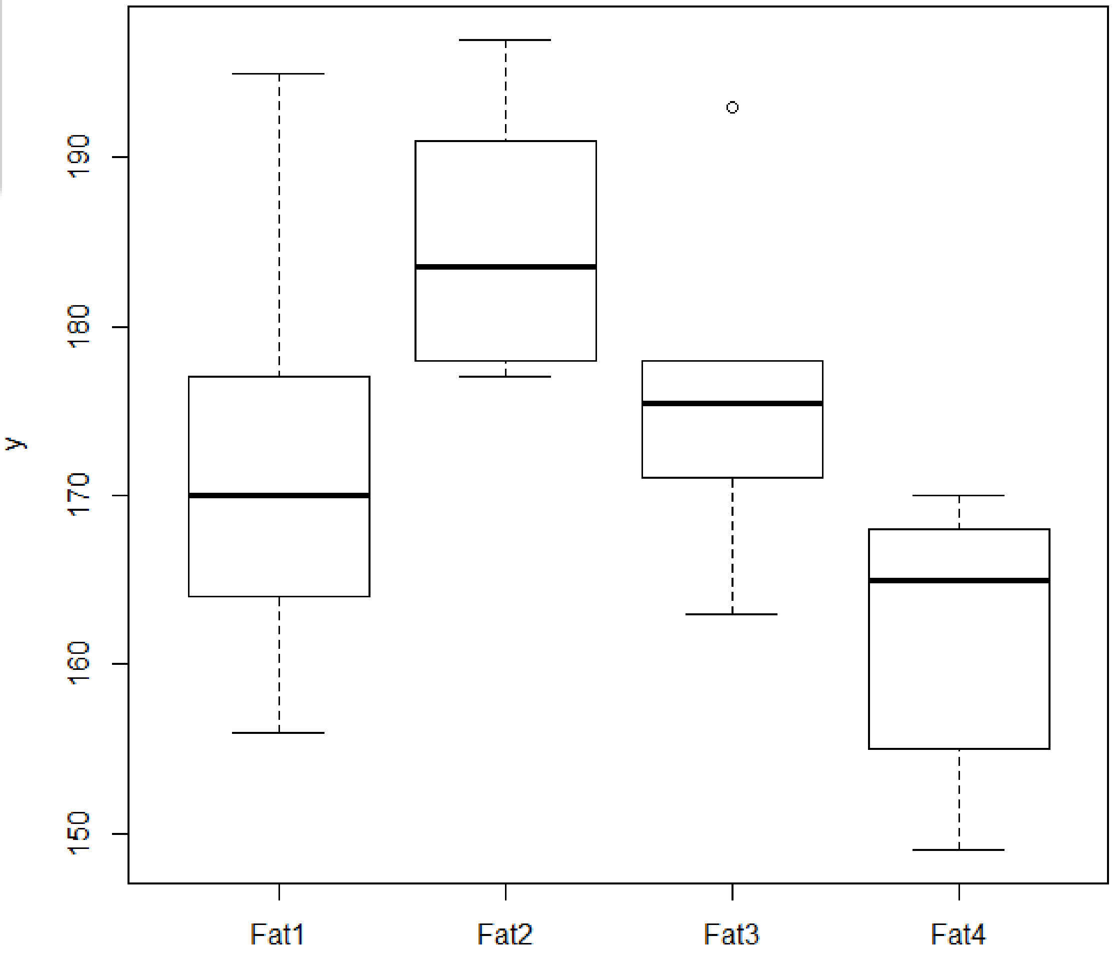


```
> data=read.table("data2.csv",header=T,sep=',')
> data
   absorbed  Fat
1       164 Fat1
2       172 Fat1
3       168 Fat1
4       177 Fat1
5       156 Fat1
6       195 Fat1
7       178 Fat2
8       191 Fat2
9       197 Fat2
10      182 Fat2
11      185 Fat2
12      177 Fat2
13      175 Fat3
14      193 Fat3
15      178 Fat3
16      171 Fat3
```

Step 1: Box plots

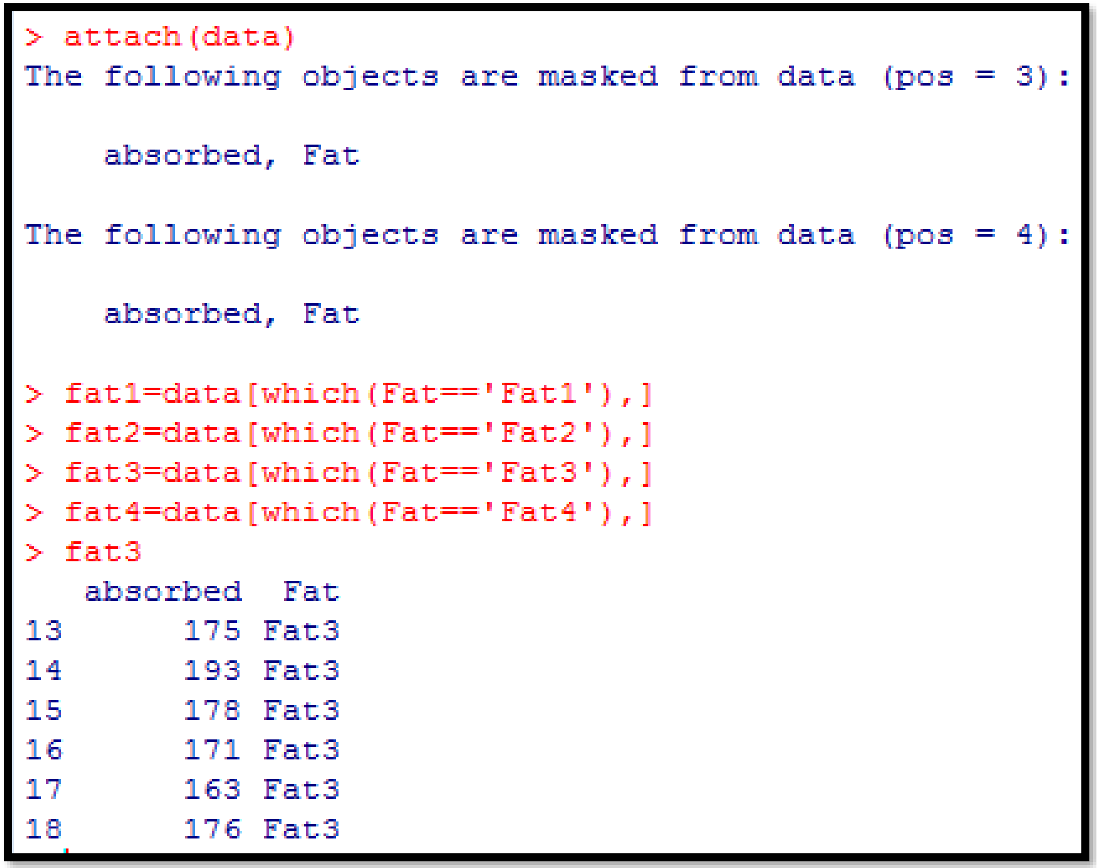


```
> y=data$absorbed
> fat=data$Fat
> plot(y~fat)
```

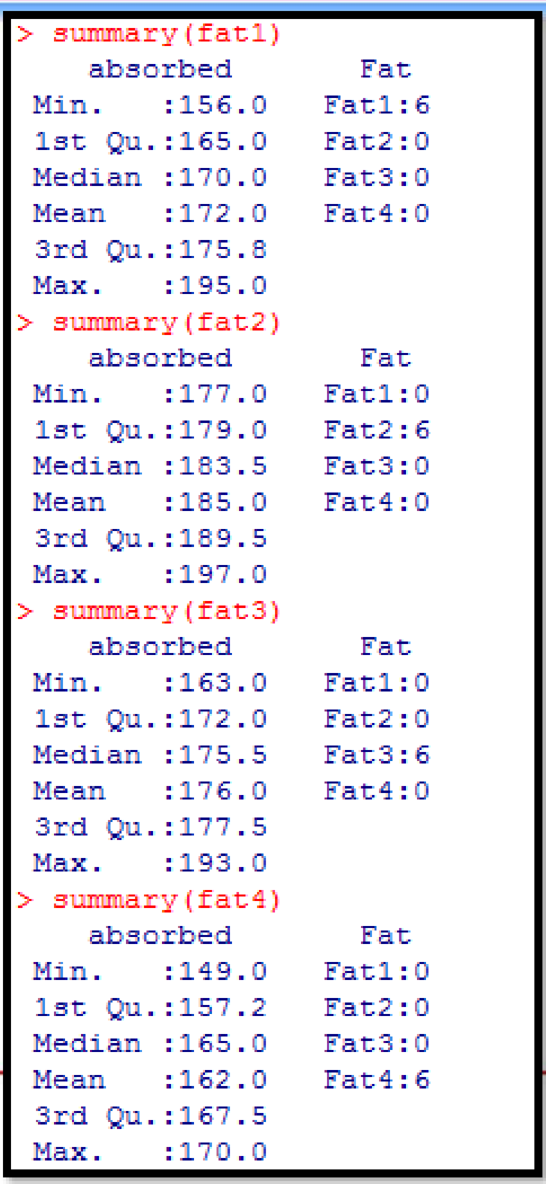


Can you observe the group mean
Are they equal means?
How about in-group variation?

Step 1: Try Group Statistics

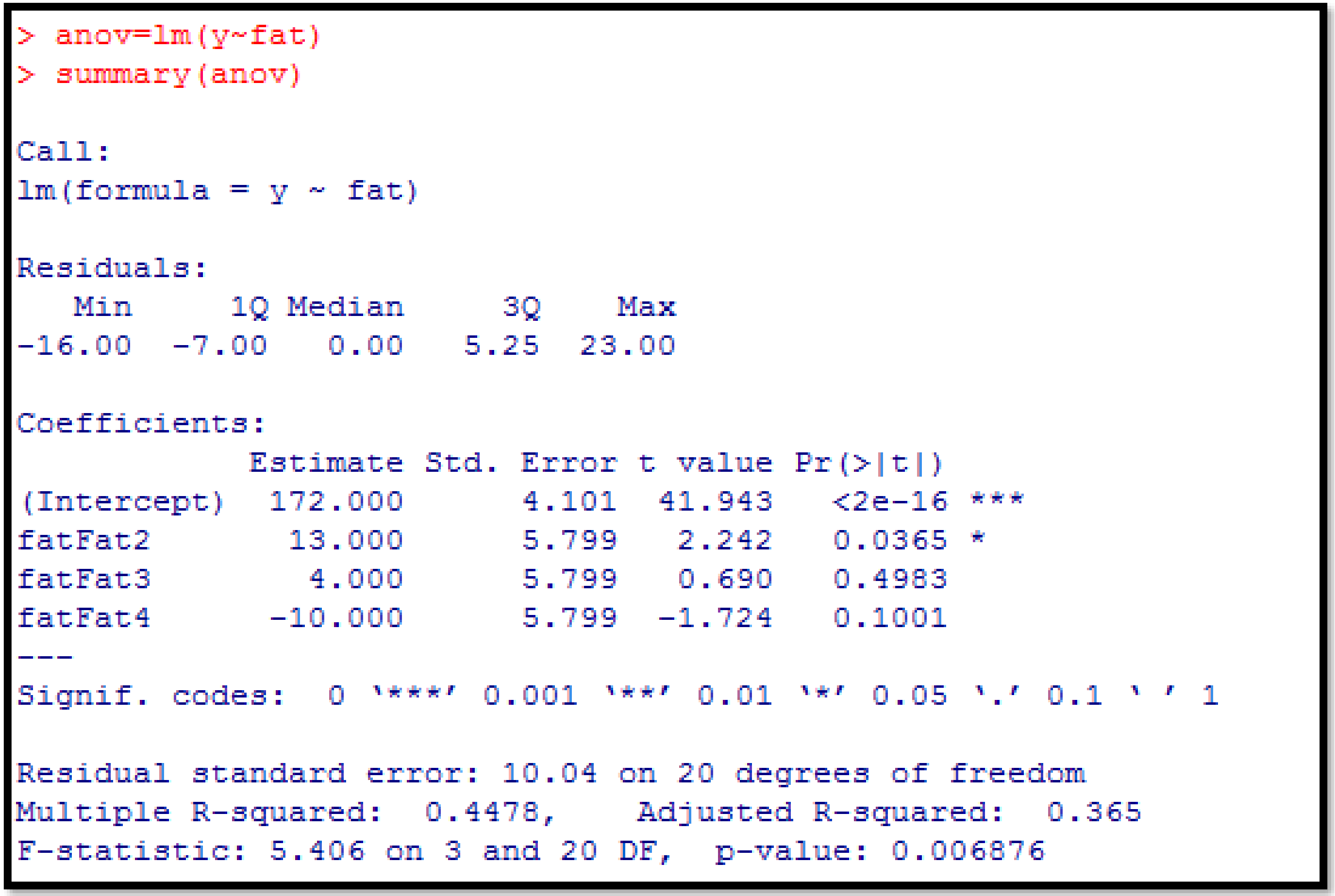


```
> attach(data)
The following objects are masked from data (pos = 3):

    absorbed, Fat

The following objects are masked from data (pos = 4):

    absorbed, Fat

> fat1=data[which(Fat=='Fat1'),]
> fat2=data[which(Fat=='Fat2'),]
> fat3=data[which(Fat=='Fat3'),]
> fat4=data[which(Fat=='Fat4'),]
> fat3
   absorbed  Fat
13      175 Fat3
14      193 Fat3
15      178 Fat3
16      171 Fat3
17      163 Fat3
18      176 Fat3
```



```
> summary(fat1)
    absorbed       Fat
 Min.   :156.0   Fat1:6
 1st Qu.:165.0   Fat2:0
 Median :170.0   Fat3:0
 Mean   :172.0   Fat4:0
 3rd Qu.:175.8
 Max.   :195.0
> summary(fat2)
    absorbed       Fat
 Min.   :177.0   Fat1:0
 1st Qu.:179.0   Fat2:6
 Median :183.5   Fat3:0
 Mean   :185.0   Fat4:0
 3rd Qu.:189.5
 Max.   :197.0
> summary(fat3)
    absorbed       Fat
 Min.   :163.0   Fat1:0
 1st Qu.:172.0   Fat2:0
 Median :175.5   Fat3:6
 Mean   :176.0   Fat4:0
 3rd Qu.:177.5
 Max.   :193.0
> summary(fat4)
    absorbed       Fat
 Min.   :149.0   Fat1:0
 1st Qu.:157.2   Fat2:0
 Median :165.0   Fat3:0
 Mean   :162.0   Fat4:6
 3rd Qu.:167.5
 Max.   :170.0
```

Fitted model – ANOVA table

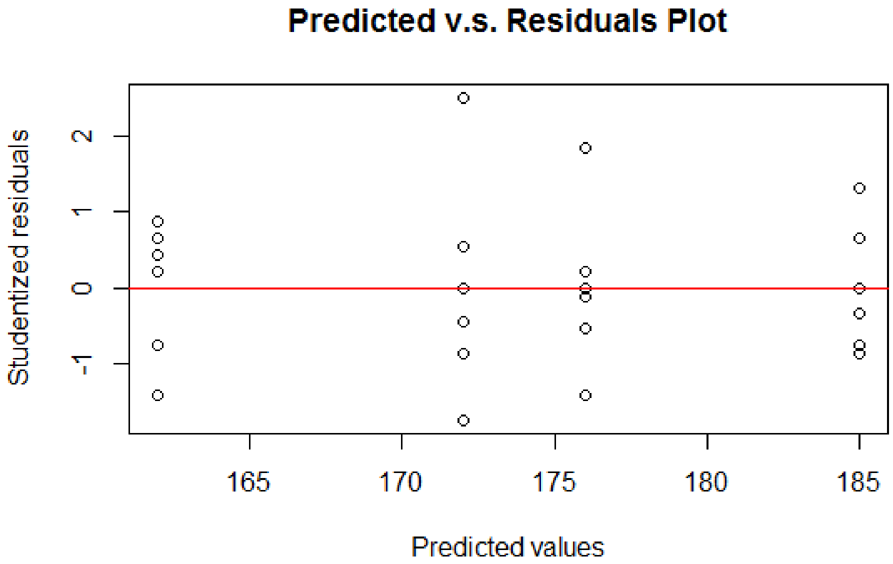


```
> anov=lm(y~fat)
> summary(anov)

Call:
lm(formula = y ~ fat)

Residuals:
   Min     1Q Median     3Q    Max
-16.00  -7.00   0.00   5.25  23.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  172.000      4.101  41.943   <2e-16 ***
fatFat2       13.000      5.799   2.242   0.0365 *
fatFat3        4.000      5.799   0.690   0.4983
fatFat4      -10.000      5.799  -1.724   0.1001
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.04 on 20 degrees of freedom
Multiple R-squared:  0.4478,    Adjusted R-squared:  0.365
F-statistic: 5.406 on 3 and 20 DF,  p-value: 0.006876
```

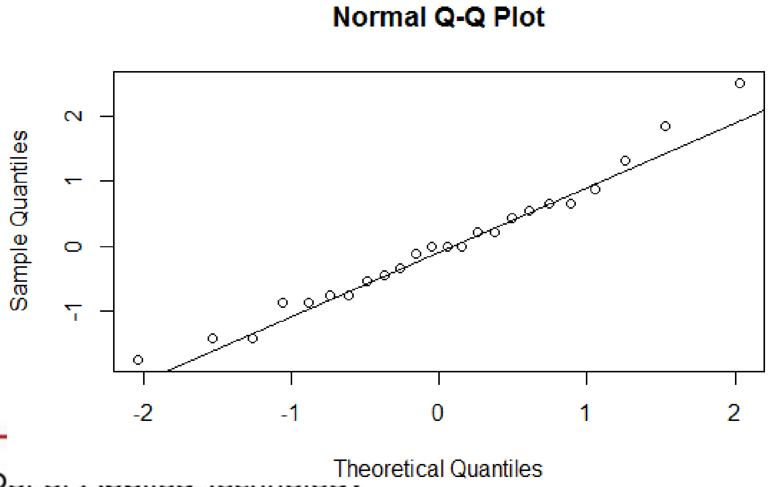
How to Interpret it?

- What about F-test
- What about individual parameter test?
- How to interpret the coefficients?

Check model assumptions: residual analysis



Is the within-groups variability constant?



Is the normality assumption satisfied?

In-Class Practice

- Using data in the case study 1
- Question
  - We have student grades in A, B, C, F
  - Are there differences in age, hours on studying/reading/games/internet among these 4 groups?