Group 3: Data Analysis

Nick Cummings, Liza Nicoll, Emily Ramos, and Yiding Zhang

University of Massachusetts, Amherst

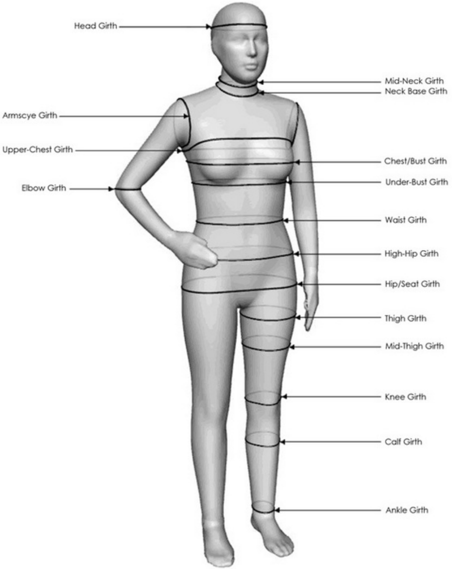
April 28, 2014

Overview

1. Group Analysis
   - Introduction
   - Description of Variables
   - Outcome of Interest
   - Initial Model
2. Individual Analysis
   - Model Diagnostics
   - Model Selection
   - Logistic Regression and Ada-boosting
   - Regression Trees

Introduction

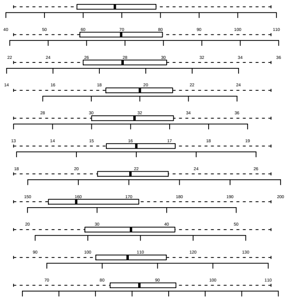
- The dataset, "Exploring Relationships in Body Dimensions", contains 25 variables: 21 body dimension measurements as well as age, weight, height, and gender for 507 physically active, young individuals.
- Of the body dimension measurements, 9 were skeletal/diameter measurements and 12 were girth/circumference measurements.
- Of the 507 individual observations, there are 247 men and 260 women.
- No missing values. Measurements made with metric scale.



Description of Variables

Summary table(left) Boxplot(right)

variable	Min.	Median	Mean	Max.
weight	42.00	68.20	69.15	116.40
chest.diam	22.20	27.80	27.97	35.60
chest.dep	14.30	19.00	19.23	27.50
bitro.diam	24.70	32.00	31.98	38.00
wrist.min	13.00	16.10	16.10	19.60
ankle.min	16.40	22.00	22.16	29.30
height	147.20	170.30	171.10	198.10
age	18.00	27.00	30.18	67.00
shoulder	85.90	108.20	108.20	134.80
navel	64.00	84.60	85.65	121.10
hip	78.80	96.00	96.68	128.30

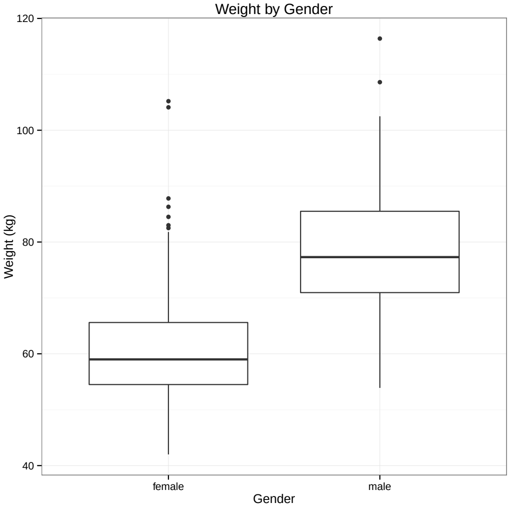


Outcome of Interest: Weight

Applications of Data:

- investigate correspondence of frame size, girths, and weight of young, athletic people
- estimate ideal weight
- inform predictions of lean/fat body compositions
- identify gender in forensic science
- design appropriate clothing (law enforcement/military)

Note: Outliers in the distribution of weight may be present because some of the individuals had unusually high muscle mass due to their high level of physical fitness.



Multiple Linear Regression Model

Initial Model

$$\begin{aligned}\text{weight}_i &= \beta_0 + \beta_1 \text{chest.diam}_i + \beta_2 \text{chest.dep}_i + \beta_3 \text{bitro.diam}_i \\ &+ \beta_4 \text{wrist.min}_i + \beta_5 \text{ankle.min}_i + \beta_6 \text{height}_i\end{aligned}$$

R Output:

(Intercept)	chest.diam	chest.dep	bitro.diam	wrist.min	ankle.min	height	R^2
-109.89	1.34	1.54	1.20	1.11	1.15	0.18	0.8882

- This model was chosen by the authors of the dataset based on the idea that these measurements remain constant after physical maturation.
- It seems that, in our model, chest depth has the largest impact on weight.
- Model seems like a good fit, but can we find a better one?

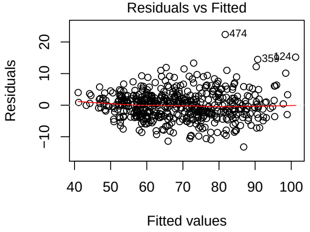
Model Diagnostics (Nick)

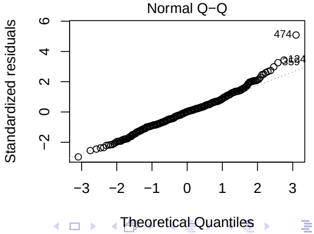
Model Diagnostics are important to fitting an appropriate model

- Outliers: extreme points that are distant form the other observations, measured by residuals
- Leverage: an extreme point, measured by the hat matrix. An "outlier on the X axis"
- Influential Points: points that have a large effect on the slope of the model
- Influence = Leverage + "Outlyingness"

Outliers (Nick)

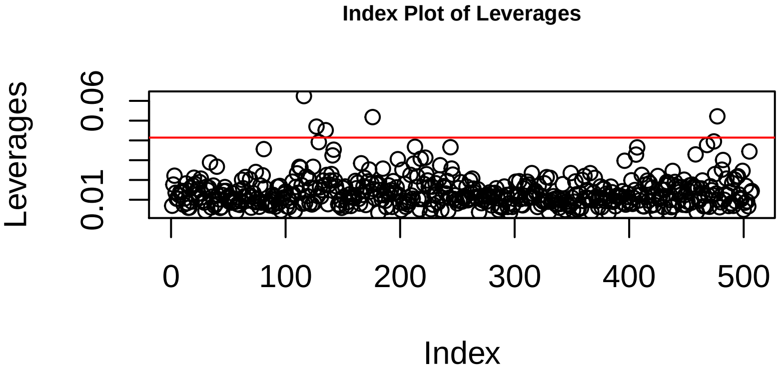
- Assumption violations can lead to biased or faulty results
- Outliers are commonly measurement errors, or are points indicative of a population that has a heavy tailed distribution
- Outliers are present in the model (points 124, 359, 474)





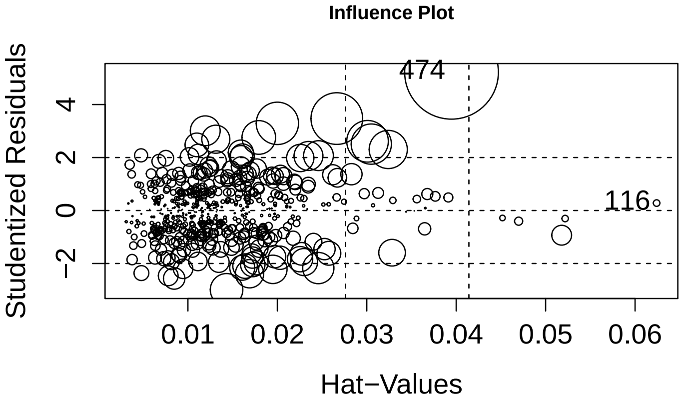
Leverage (Nick)

A "rule of thumb" is that leverages of more than $\frac{2p}{n}$ or $\frac{3p}{n}$ should be looked at more closely. Points of concern are located above the red line.



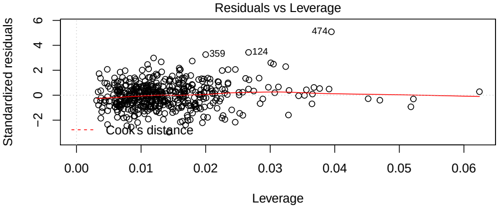
Influence (Nick)

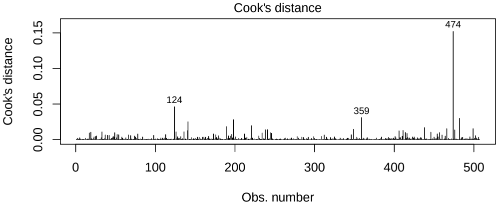
An **influential point** is one whose removal from the dataset causes a drastic change in the fit. An influential point will either be an outlier in the data, will have high leverage, or will have both.



Cook's Distance (Nick)

- We can look at leverage, outlyingness, and influence altogether
- The Cook's Distance plot shows the values of these problematic points
- When we remove these points, does the model improve much?





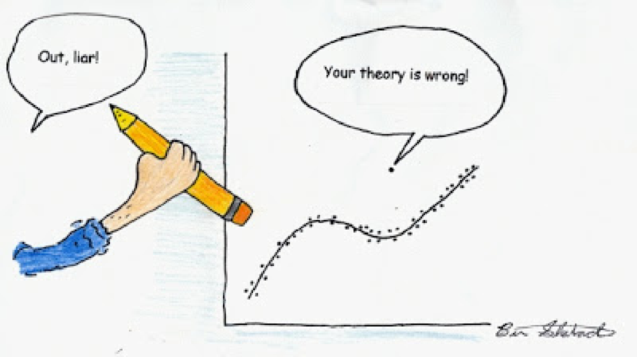
Better Fit? (Nick)

Initial Model:

(Inter.)	chest.diam	chest.dep	bitro.diam	wrist.min	ankle.min	height	Adj. R^2
-109.89	1.34	1.54	1.20	1.11	1.15	0.18	0.887

New model with influential points removed:

(Inter.)	chest.diam	chest.dep	bitro.diam	wrist.min	ankle.min	height	Adj. R^2
-109.11	1.38	1.55	1.10	0.97	1.17	0.19	0.891



Model Selection Criteria and Methods (Emily)

Goal: Find the “best” method for model selection.

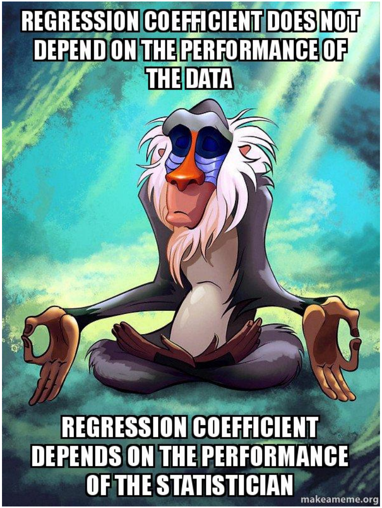
The methods used to create ten different models were:

- include all variables (1)
- suggested by paper (2)
- my selection (1)
- stepAIC (1)
- leaps (2)
- combinations of R functions and human intuition (3).

These models were compared using:

- AIC/ BIC: measure goodness-of-fit through residual sum of squares and penalizes for adding more predictors; the smaller the better.
- Adjusted R^2: adjusts R^2 so that the model is penalized for adding more predictors; the larger the better.
- PRESS is a summary measure focused on prediction; the smaller the better.

My Selection (Emily)



- **Predictors needed:** age, height and gender (since these contribute significantly to weight)
- **Predictors we will allow:** the predictors used in the inital model (chest diameter, chest depth and bitro diameter), pelvic bredth, shoulder, chest, waist, hip and thigh (since these are directly associated with weight)
- The model I decided was the "best" includes the predictors chest.dep, chest.diam, shoulder, waist, hip, thigh and height^2, based on AIC and BIC.

R function: stepAIC() (Emily)

To demonstrate these two R functions, we will begin with a simple model:

```
MLRex <- lm(weight   height + wrist.min + ankle.min + chest, data = body)
```

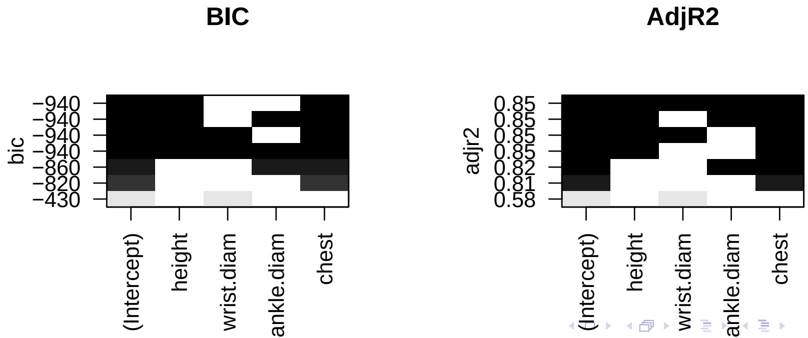
The R function found in the package “MASS” called “stepAIC()” performs stepwise model selection by AIC. This function allows you to indicate the direction of the search: forward, backward or both.

```
step <- stepAIC(MLRex, direction = "both")
head(stepanova)
```

```
Initial Model:  weight   height + wrist.diam + ankle.diam + chest
Final Model:  weight   height + ankle.diam + chest
             Step  Df   Deviance  Resid. Df  Resid. Dev       AIC
 1                                      502    13294.14  1666.150
 2   - wrist.diam   1   47.90427        503    13342.05  1665.974
```

R package: leaps (Emily)

```
leaps <- regsubsets(weight   height + wrist.diam + ankle.diam + chest, nbest = 2,
data = body)
par(mfrow = c(1, 2))
plot(leaps, main = "BIC")
plot(leaps, scale = "adjr2", main = "AdjR2")
```



Summary (Emily)

Criteria Summary for each model

MLR#	AIC	BIC	PRESS	Adjusted R^2	Method
all	2216	2326	2384	0.9753	all variables from dataset used
i	2970	3004	10405	0.8869	suggested by paper
1	2256	2319	2560	0.9727	suggested by paper
2	2402	2441	3408	0.9632	my model
3	2206	2282	2329	0.9754	stepAIC
4	2195	2271	2281	0.9759	stepAIC and adjustments
5	2207	2292	2335	0.9755	leaps (adj R^2)
6	2189	2278	2255	0.9764	leaps(adj R^2)and adjustments
7	2213	2272	2353	0.9749	leaps(BIC)
8	2205	2264	2316	0.9753	leaps(BIC) and adjustments

Red corresponds to the best value for that criteria, blue is the second best.

- Model Selection truely is an art form.
- R can mechanically run through steps, interactions, combinations, etc.
- R cannot subjectively look at the variables to determine the absolute best model.
- To acheive the model of "best" fit, it is best to utilize a combination of R functions, criteria methods, and your own adjustments/ intuition.

Model Selection (Yiding)

Logistic Model: $logit(p_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_m X_{im}$

Model selection criteria:

- *AIC* $= nlog(RRS/n) + 2(p+1)$
  in R: *step*()

AIC criteria model selection

Step #	Model	AIC
1	*logi*(SEX)=WT+CDM+CDP+BDM+WR+ANK+HT+AGE+SHD+NAV+HIP	110.35
2	*logi*(SEX)=WT+CDP+BDM+WR+ANK+HT+AGE+SHD+NAV+HIP	108.36
3	*logi*(SEX)=WT+CDP+BDM+WR+HT+AGE+SHD+NAV+HIP	106.45
4	*logi*(SEX)=CDP+BDM+WR+HT+AGE+SHD+NAV+HIP	104.55
5	*logi*(SEX)=CDP+WR+HT+AGE+SHD+NAV+HIP	102.69
6	*logi*(SEX)=CDP+WR+HT+AGE+SHD+HIP	101.05

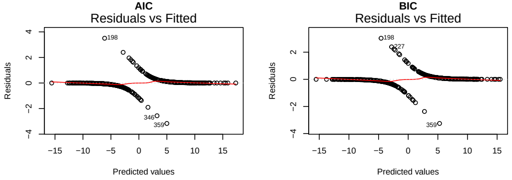
- *BIC* $= nlog(RSS/n) + (p+1)log(n)$
  &
  Posterior Probability
  $= p(\theta \mid x) = \frac{p(x|\theta)p(\theta)}{p(x)}$
  in R: *BMA* packages *bic.glm*()

BIC criteria model selection

Model #	Model	BIC	Posterior prob
1	*logi*(SEX)=WR+HT+SHD+HIP	-3033	0.453
2	*logi*(SEX)=WR+HT+AGE+SHD+HIP	-3031	0.157
3	*logi*(SEX)=CDP+WR+HT+AGE+SHD+HIP	-3031	0.151
4	*logi*(SEX)=WR+HT+SHD+NAV+HIP	-3031	0.142
5	*logi*(SEX)=WR+HT+SHD+HIP	-3029	0.047

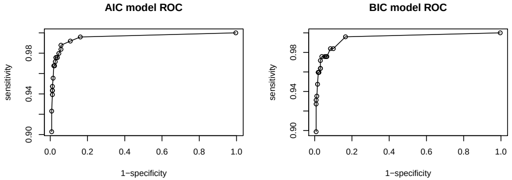
Check the Goodness of Model (Yiding)

- **Linearity**
  Residuals vs Fitted Values
  Pearson residuals:
  $$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}(y_i)}} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$



- **Predictive ability**
  ROC curve:
  sinsitivity *vs* 1-specificity
  Somer's rank correlation:
  $D_{xy} = 2(c - 0.5)$
  c is the area under the ROC curve.

  Hmisc package: *somers*2()



Model	Area ROC curve	Somers' D_{xy}
AIC	0.9944721	0.9889443
BIC	0.9941607	0.9883214

Ada-boosting (Yiding)

Algorithem:

Given: $(x_1, y_1), \dots, (x_m, y_m)$ where

$x_i \in X, y_i \in Y = \{-1, +1\}$ Initialize

$D_1(i) = 1/m$.

For $t = 1, \dots, T$:

- Train weak learner using distribution D_t.
- Get weak hypothesis $h_t : X \rightarrow \{-1, +1\}$ with

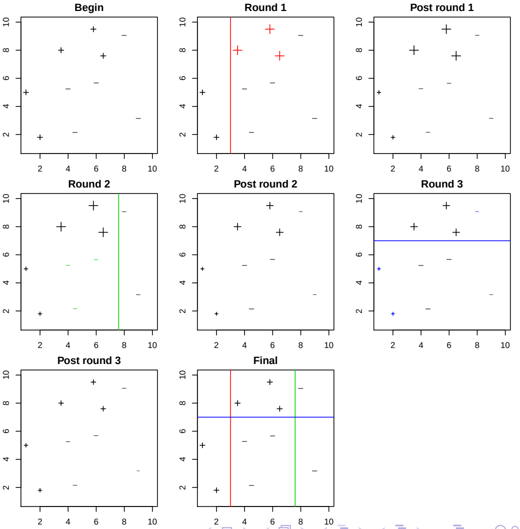
error $\epsilon_t = Pr_{i \sim D_t}[h_t(x_i) \neq y_i]$

- Choose $\alpha_t = \frac{1}{2} ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$
- Update:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases}$$

$= \frac{D_i(i) exp(\alpha_t y_i h_t(x_i))}{Z_t}$

Final hypothesis: $H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$



Logistic Regression vs Ada-boosting (Yiding)

Logistic Regression
Predictive ability depends on cutoff point

Cutoff point	Sensitivity	Specificity
0.00	1.000	0.004
0.05	0.996	0.838
0.10	0.992	0.892
0.15	0.988	0.942
0.20	0.984	0.954
...	...	...
0.85	0.923	0.992
0.90	0.903	0.992

Ada-boosting
Predictive ability depends on variables

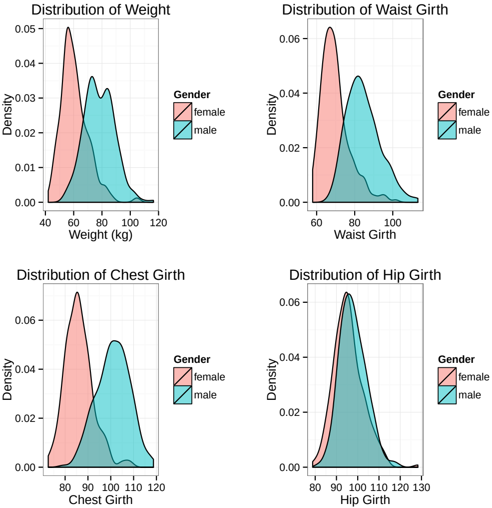
Round #	Original data Sensitiviy; Specificity; Accuracy	AIC data Sensitiviy; Specificity; Accuracy	BIC data Sensitiviy; Specificity; Accuracy
1	0.975; 0.925; 0.949	0.967; 0.962; 0.965	0.967; 0.947; 0.957
2	0.953; 0.936; 0.945	0.953; 0.920; 0.937	0.961; 0.936; 0.949
3	0.949; 0.951; 0.949	0.954; 0.959; 0.957	0.954; 0.959; 0.957
4	0.983; 0.932; 0.957	0.983; 0.955; 0.961	0.983; 0.962; 0.972
5	0.983; 0.924; 0.953	0.983; 0.931; 0.957	0.983; 0.947; 0.965
Avarage	0.969; 0.934; 0.951	0.968; 0.945; 0.955	0.970; 0.950; 0.960

- **Inference**
  Logistic Regression $\surd$

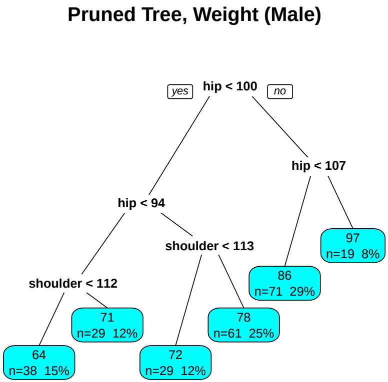
- **Prediction**
  Ada-boosting $\surd$

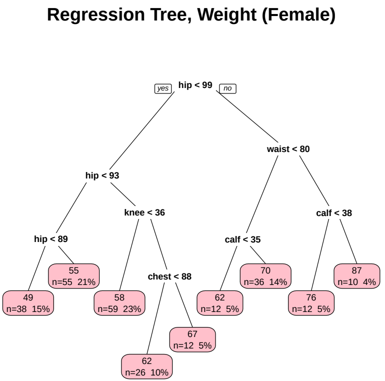
Differences between Males and Females (Liza)

- Are there significant differences in the body measurements most useful for predicting weight in males and females?
- Is one regression formula appropriate for predicting weight for both genders?
- Can we use regression trees to help explore these questions?



Regression Trees (Liza)



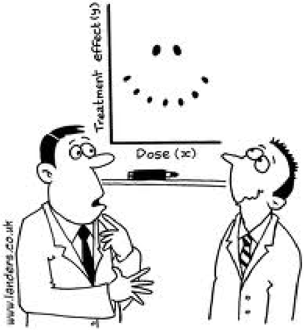


- Variables used in male tree: hip shoulder girths
- Variable used in female tree: hip, knee, chest, waist and calf girths

Conclusions (Liza)

- Regression trees are useful for exploring data and provide a useful alternative to parametric regression methods, though are not intended for making predictions.
- Results here suggest that separate models for males and females might be appropriate.
- Model fitting and selection exercises could test this hypothesis.

Group Conclusions



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

What have we learned?

- Always check your assumptions, even low influence outliers can change your model fit
- With model selection, a combination of selection criteria, R functions and intuition are needed to create the model of "best" fit.
- Inference choose Logistic Regression; prediction choose Ada-boosting.
- Regression trees are a powerful, yet simple, non-parametric method for exploring data.
- If time allowed, a combination of methods shown may produce and even better fitting model.