2. Simple and Multiple Linear Regression

1. Overview and models representation.
2. Point estimation, hypothesis tests and CI's.
3. Goodness of fit
4. ANOVA F-test for nested models and variable selection
5. The log-transformation
6. Collinearity
7. Residuals checks and outliers.
8. Matrix representation and properties of E(Y) and Var(Y)
9. Estimation via Maximum Likelihood

2.1. Overview, Models representation

Specific learning objectives:

1. Write the models in mathematical form.
2. State the underlying distributional assumptions.

Regression models

- Regression models are equations that consider sources of variation that help understand the relationship between variables a response (Y) and a covariate (X).

- When main interest lies in the response (Y) by itself:
  - if other measurements (X) are not available, sample mean and variance may be our best guess.

  - if X is available, we can estimate the population attributes with greater precision, X is called **covariate**.

Regression models

- When the main interest lies in the relationship of Y and X, X is called **main predictor.**

- When model includes:
  - one covariate or predictor = **simple linear regression**
  - more than one covariate or predictor = **multiple linear regression.**

- X may be continuous or categorical.

Mathematical form

Simple case (k=1, one covariate):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

Simplest case: relationship between Y and X is described by a straight line, β_0 and β_1 are the intercept and slope

Multiple case (k≥2): Described by a hyper-plane in a k-dimensional space

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

where i indexes subjects, $i = 1,\ldots,n$, and

Y_i is the response or dependent variable

X_i's are the independent variables (regressors)

$\beta_0, \beta_1, \ldots, \beta_k$ are the regression coefficients

ε_i random error for subject i

In PK analysis, the response is CL, AUC, etc. and covariates are usually demographic (age, sex, weight, race, smoking status, etc.), continuous or categorical.

Model Assumptions

Linear: derivatives do not depend on the model coefficients

$$\frac{\partial Y_i}{\partial \beta_0} = 1, \quad \frac{\partial Y_i}{\partial \beta_1} = X_{1i}, \ \ldots, \ \frac{\partial Y_i}{\partial \beta_k} = X_{ki}.$$

Note here we have k regressors and p parameters including the intercept (p=k+1)

Model Assumptions:

1. Linearity of Y vs. X
2. ε_i are iid $\sim \mathrm{N}(0, \sigma^2)$

Verifiable mostly via post-modeling graphical assessment (more on this later)

This is, the random errors

i) Are normally distributed

ii) Independent

iii) Have constant variance

Residual normality, independence and homoscedasticity.

2.2. Point estimation, hypothesis tests and CI's

Specific learning objectives:

1. Identify the regression coefficients for intercept and slope as well as varying intercepts and slopes.
2. Calculate hypothesis tests and CI's for individual model coefficients.
3. Fit and Identify the regression coefficient estimates from an R output.

Estimation goal, simple case (k=1)

Find a line that lies closest to the data points.

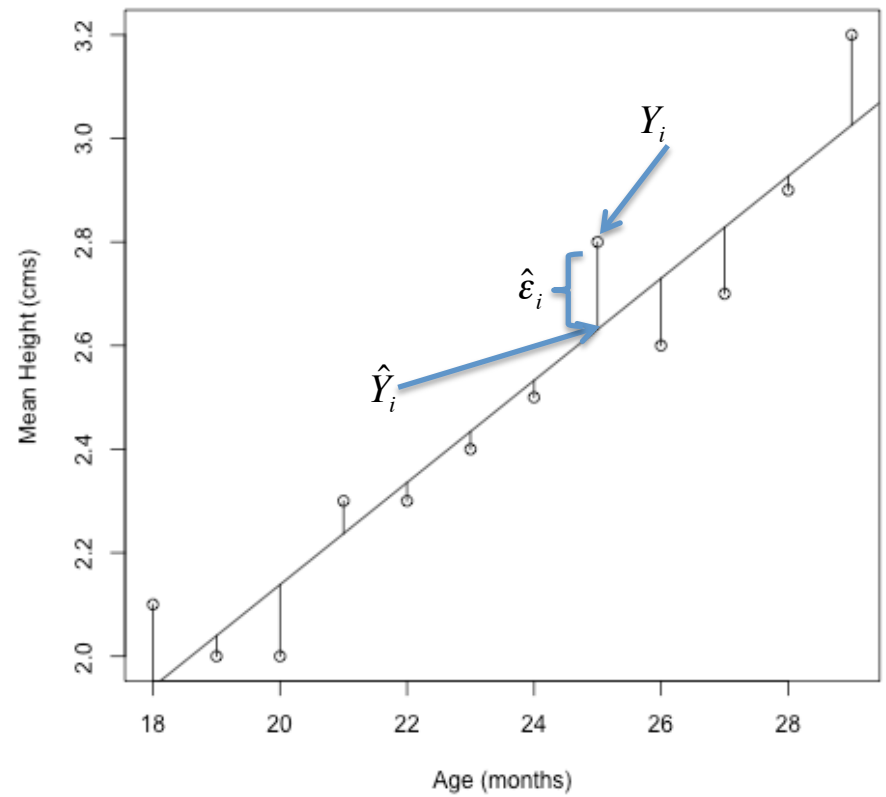
The Ordinary Least Squares (OLS) estimation method:
to obtain estimates of β_0, β_1 by minimizing the estimated residual sum of squares:

$$\hat{\varepsilon}_1^{\,2} + \hat{\varepsilon}_2^{\,2} + \cdots + \hat{\varepsilon}_n^{\,2}$$

The estimate of σ is:

$$\hat{\sigma} = \sqrt{\frac{\hat{\varepsilon}_1^{\,2} + \hat{\varepsilon}_2^{\,2} + \cdots + \hat{\varepsilon}_n^{\,2}}{n-p}}$$

Number of model parameters, p=2



? why p =2?

Because belta 0 and belta 1, p is the number of parameters

Interpretation of coefficients and estimates
Simple case (k=1)

β_0 intercept of the regression line, and the mean of the Y when $X = 0$.
In cases where $X = 0$ does not make sense, β_0 has no physical interpretation.

β_1 slope of the regression line, indicates the average change in Y for a one unit change in X.

$\hat{\beta}_0, \hat{\beta}_1$ estimates of the regression coefficients,

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i}$ predicted or estimated mean for subject i given a value of X_{1i},

$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ predicted or estimated value of the residual for subject i.

Usually of interest: explain the behavior of the variables, predict current or future observations.

Note the new estimate for σ has a form similar to that of the sample s.

$$\hat{\sigma} = \sqrt{\frac{\hat{\varepsilon}_1^{\,2} + \hat{\varepsilon}_2^{\,2} + \cdots + \hat{\varepsilon}_n^{\,2}}{n-p}} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y})^2}{n-p}}$$

More on the interpretation of the intercept β_0

"β_0 is the intercept of the regression line and the mean of Y when X=-."

- What if X=- makes no physical sense?

We can center it to its mean: shift the origin of the X's from zero to $\overline{X}$.

$$Y_i = \beta_0' + \beta_1 (X_i - \overline{X})$$

The value of the estimated slope will remain unchanged.

β_0' is the mean of Y when X = $\overline{X}$

- What if a zero intercept makes no sense? E.g. Y=AUC vs. X=Dose

It is recommended include the intercept in the model anyway.

  - The estimate of β_0 will be very small (close to zero) anyway and will give small sampling error.
  - Quantities of interest such as R^2 lack of meaning under a non-intercept model and the slope estimator will be biased.

Hypothesis tests and CI's for individual coefficients

Hypothesis test

$$H_0 : \beta_j = 0 \qquad \text{vs.} \qquad H_1 : \beta_j \neq 0$$

$$T = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim \text{Student's } t\ (n-p)$$

Reject H_0 at an α significance level if

$$|T_0| > t_{1-\alpha/2,n-p}, \quad \text{or}$$

$$\text{p - value} = P(|T| > T_0) \leq \alpha$$

Note p=#parameters in model including intercept = k+1

Observed value of T

Confidence interval

$$(1-\alpha)100\% \text{ CI for } \beta_j \text{ is } \left\{\hat{\beta}_j \pm t_{1-\alpha/2,n-p}\ SE(\hat{\beta}_j)\right\}$$

Both the CI and the T-test approaches produce equivalent results:

At an α level, CI includes zero $\longleftrightarrow$ H_0 is not rejected

Example simple case: Body composition measurements (% Fat)

25 normal adults, men and women, between 23 and 61 years old.

(Altman, 1991).

- Variables: body fat percentage, age, sex
- Sample size: 10 men, 15 women, n=25
- Age range: 23 - 61 years.

First few rows of data set in R:

```
> head(agefat)
  age  fat    sex
1  24 15.5   male
2  37 20.9   male
3  41 18.6   male
4  60 28.0   male
5  31 34.7 female
6  39 30.2 female
```

Example: Male Body Composition vs. Age

The model:

$$Y_i = \beta_0 + \beta_1 Age_i + \varepsilon_i,$$

where

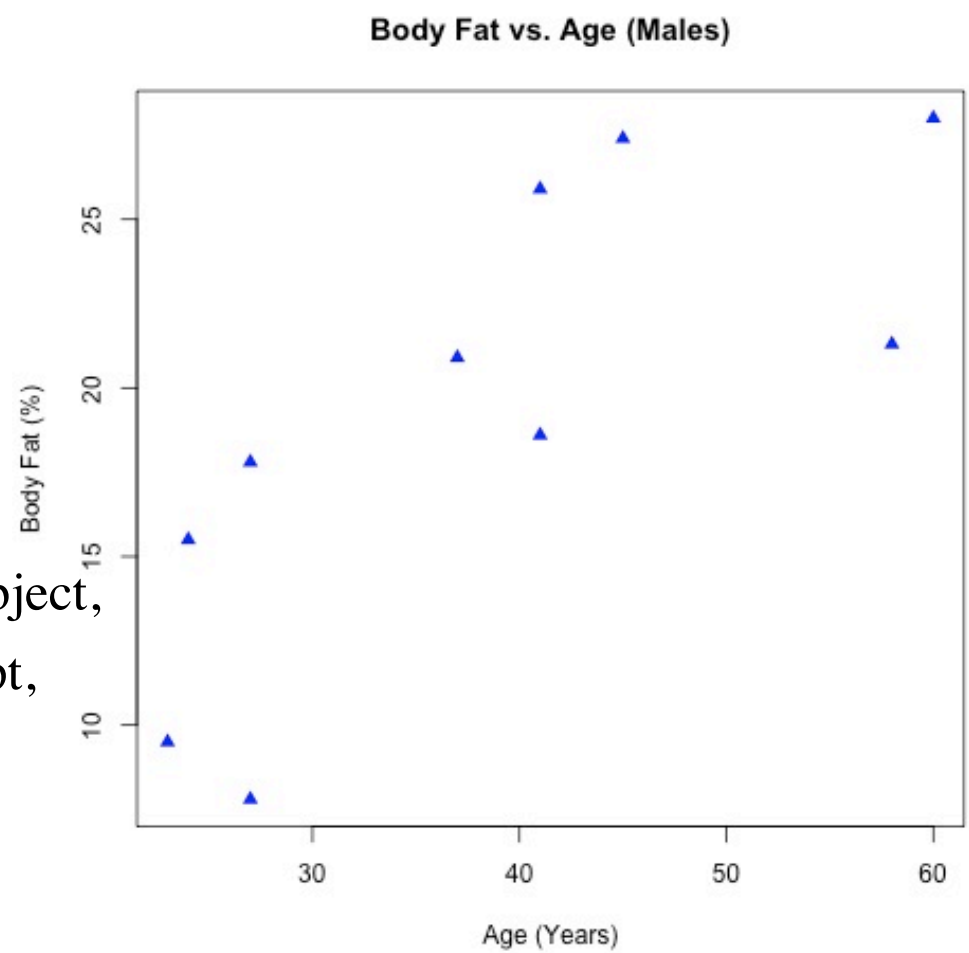
$i = 1,2,...,10$ indexes the subjects,

Y_i is the %Fat in $i-th$ subject,

Age is the predictor (years) of $i-th$ subject,

β_0, β_1 are the unkown slope and intercept,

ε_i error for $i-th$ subject.



Interpretation of slope: The change in mean %Fat for a one unit increase in Age

```
plot(fat~age,data=males, col="blue",main="Body Fat vs. Age
                    (Males)",xlab="Age (Years)",ylab="Body Fat (%)")
```

R code and output: Males Body Composition vs. Age

```
fit <- lm(fat~age,data=subset(agefat,agefat$sex=="male"))
summary(fit)
```

R output (portion):

Tests $H_0: \text{Estimate} = 0$ vs. $H_1: \text{Estimate} \neq 0$

$\hat{\beta}_0$

$\hat{\beta}_1$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.9061     4.7263   0.826  0.43250
age           0.4011     0.1171   3.425  0.00902 **
---
Signif. codes:  - '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.706 on 8 degrees of freedom
Multiple R-squared:  0.5945,  Adjusted R-squared:  0.5438
F-statistic: 11.73 on 1 and 8 DF,  p-value: 0.009022
```

The estimated model for males results suggest that the coefficient for Age is highly statistically different from zero (p-value=0.009).

A 95% CI for β_1 is $\left\{\hat{\beta}_1 \pm 2.31\ \text{SE}(\beta_1)\right\} = \{0.131,\ 0.671\}$

Note this CI does not include zero

```
> c(.4011-qt(.975,8)*.1171,.4011+qt(.975,8)*.1171)
```

The estimated model:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \, Age_i$$
$$= 3.906 + 0.401 \, Age_i$$

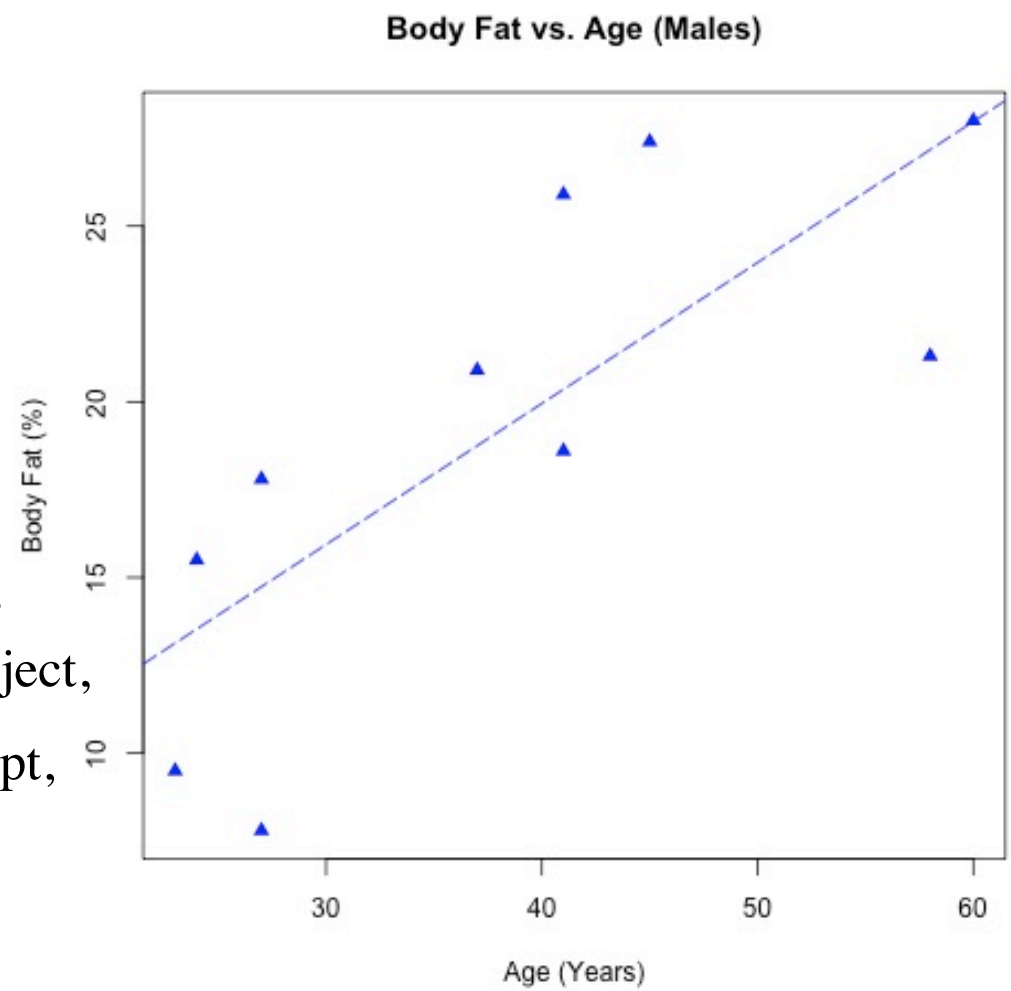
where

$i = 1,2,\ldots,10$ indexes the subjects,

$\hat{Y}_i$ is the predicted %Fat in $i-th$ subject,

Age is the regressor (years) of $i-th$ subject,

$\hat{\beta}_0, \hat{\beta}_1$ are the estimated slope and intercept,



Interpretation of estimated slope:

For each 1 unit increase in Age (year), there is statistically significant 0.401 increase in mean %Fat (p-val=0.009).

Estimation goal, interpretation of coefficients Multiple case (k≥2)

Case k=2: find a plane or surface that lies closest to the data points.

- β_0 is the intercept of the plane and the mean of Y when X_1=- and X_2=- .
- β_0 has no physical interpretation when the values of the predictors cannot be zero.
- β_1 indicates the mean change in Y per unit change in X_1 when X_2 is held constant.
- Similarly, β_2 indicates the mean change in Y per unit change in X_2 when X_1 is held constant.

Estimation goal, interpretation of coefficients Multiple case (k≥2)

Case k>2: find a hyper-plane in the k-dimensional space of the covariates

The parameter β_k represents the expected change in Y per unit change in X_k when the remaining covariates are held constant

Example Multiple Case, k=2
Body Composition vs. Sex & Age

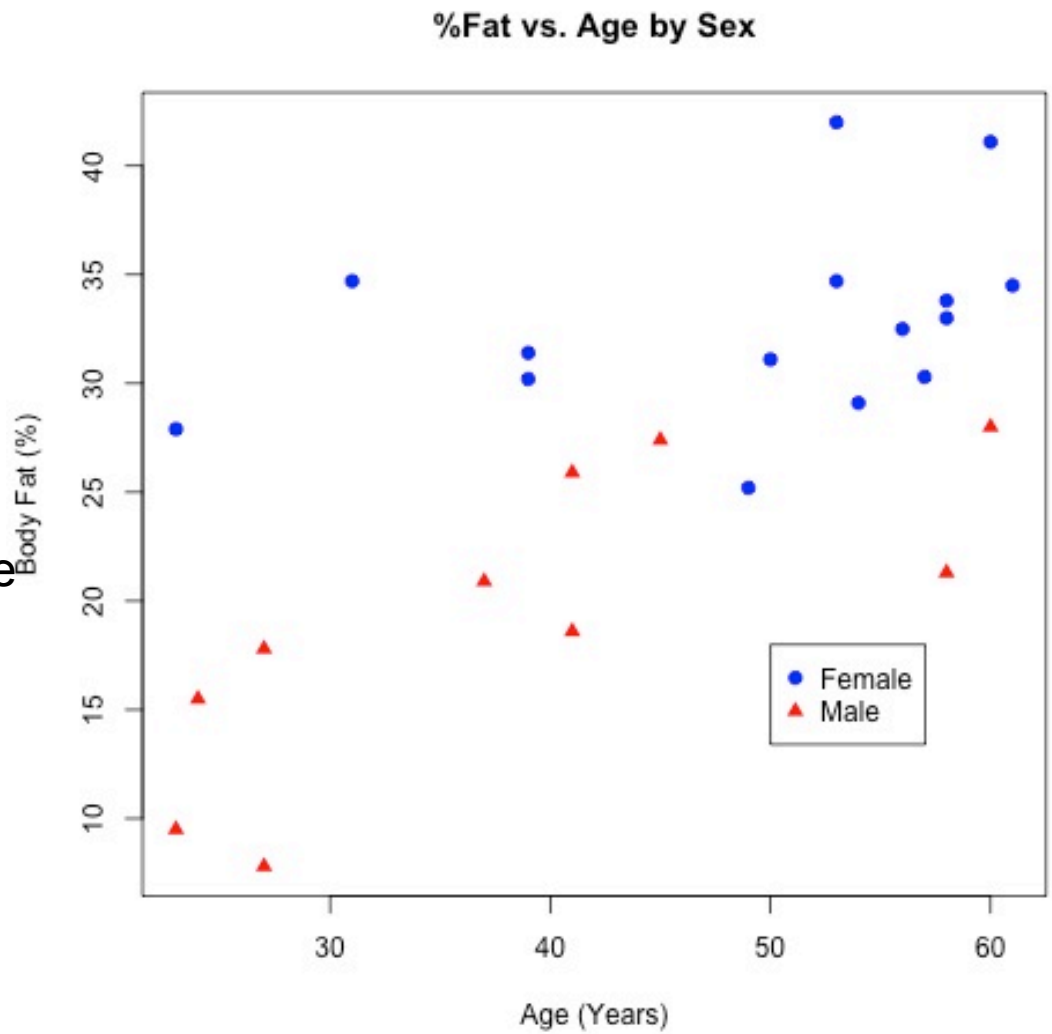
- Body fat seems to increase more steeply for men with age.

- Body Fat appears to have mostly higher values for women than men.

Fitting a simple linear model will enable us to:

- Test whether the relationship of Body Fat and Age is linear for both genders.

- Assess whether gender and/or age have a significant effect on body fat.



R Code for plot %Fat vs. Age and Sex

```
range.plot <- range(agefat$fat)

plot(fat~age,data=agefat,ylim=range.plot,type="n",
     main="%Fat vs. Age by Sex",
     xlab="Age (Years)", ylab="Body Fat (%)")

points(fat~age,pch=19,col="blue",data=subset(agefat,sex=="female"))

points(fat~age,pch=17,col="red",data=subset(agefat,sex=="male"))

legend(50,18,c("Female","Male"),pch=c(19,17),col=c("blue","red"))
```

Note: the type="n" option in the plot() function results in not plotting the data points at all. The points are added in the subsequent lines.

Body composition: Multiple regression model with k=2 regressors

Common slope by sex

$$Y_i = \beta_0 + \beta_1 \, Sex_i + \beta_2 \, Age_i + \varepsilon_i; \;\; i = 1,\ldots,25, \;\; \varepsilon_i \sim N(0,\sigma^2).$$

A dummy variable (Sex) represents a shift in a regression through its effect on the intercept.

where

$Sex_i = 1$ if subject i is female, $Sex_i = 0$ if male,

Age_i is the Age in years of subject i.

Males intercept

$\beta_0 + \beta_2 \, Age_i$ is the mean %Fat for males

$(\beta_0 + \beta_1) + \beta_2 \, Age_i$ is the mean %Fat for females

Females intercept

Testing for significance :

β_1 : is mean %Fat different for men vs. women? i.e., H_0:β_1=-

β_2: is age a significant factor in mean %Fat? i.e., H_0:β_2=-

Fit for common slopes in R

```
fit <- lm(fat~age+sex,data=agefat)
summary(fit)
```

R output (portion):

Tests H_0 : Estimate = 0 vs. H_1 : Estimate ≠ 0

$\hat{\beta}_0$

$\hat{\beta}_1$

$\hat{\beta}_2$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.09894    3.38163   2.691  0.01335 *
age          0.26556    0.07953   3.339  0.00297 **
sexfemale   10.54892    2.09140   5.044 4.74e-05 ***
---
Signif. codes:  - '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.644 on 22 degrees of freedom
Multiple R-squared:  0.7375,  Adjusted R-squared:  0.7137
F-statistic: 30.91 on 2 and 22 DF,  p-value: 4.069e-07
```

$\hat{\sigma}$

Est. Mean %Fat for females = (9.099+10.549) + 0.265 Age
Est. Mean %Fat for males= 9.099 + 0.265 Age

According to this model, both age and gender are statistically significant for percent body fat.

Body composition: Multiple regression model with k=3 regressors

Different slopes by sex (interaction term)

$$Y_i = \beta_0 + \beta_1 \, Sex_i + \beta_2 \, Age_i + \beta_3 \left(Sex_i \times Age_i \right) + \varepsilon_i;$$

$$i = 1,\ldots,25, \quad \varepsilon_i \sim N(0,\sigma^2).$$

$Sex_i = 1$ if subject i is female, $Sex_i = 0$ if male

Age_i is the Age in years of subject i.

$Sex \times Age_i$ is the interaction between Sex and Age,

and $Sex \times Age_i = Age_i$ when subject i is female.

Testing for significance :

β_1 : is %Fat different for men vs. women? i.e., H_0:β_1=-

β_2: is age a significant factor in %Fat? i.e., H_0:β_2=-

β_3: is the rate of %Fat increase with age the same between men and women? i.e., H_0:β_3=-

Body composition: Multiple regression model with k=3 regressors
Different slopes by sex (interaction term)

Regression equations by gender:

Males intercept

Males slope

$$\beta_0 + \beta_2 Age_i \quad \text{is the mean \%Fat for males}$$

$$(\beta_0 + \beta_1) + (\beta_2 + \beta_3) Age_i \quad \text{is the mean \%Fat for females}$$

Females intercept

Females slope

Testing for significance :

β_1 : is mean %Fat different for men vs. women? i.e., H_0:β_1=-

β_2: is age a significant factor in mean %Fat? i.e., H_0:β_2=-

β_3: is the rate of mean %Fat increase with age the same between men and women? i.e., H_0:β_3=-

R code and output: Body Composition Example, different slopes

```
Fit.full <- lm(fat~age+sex+age:sex,data=agefat)
summary(fit.full)
```

Age and sex are highly significant, interaction is not at a 5% level (p-val=0.107>0.05).

R output (portion):

$\hat{\beta}_0$

$\hat{\beta}_1$

$\hat{\beta}_2$

$\hat{\beta}_3$

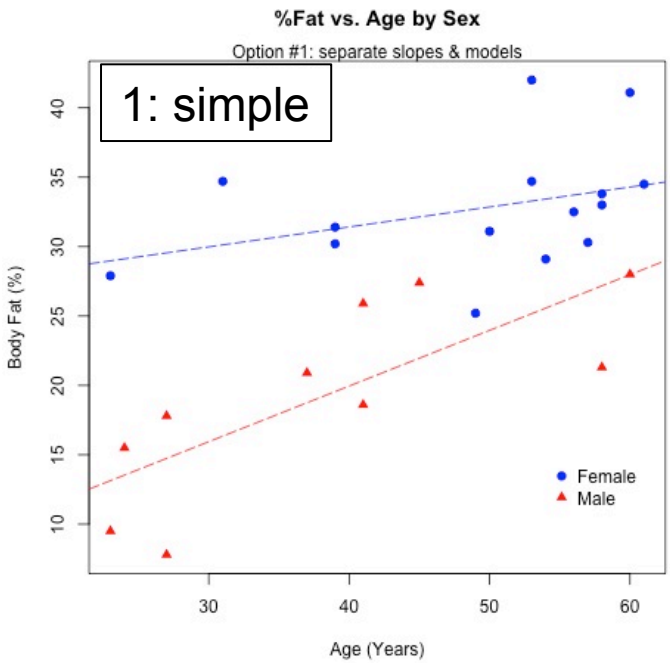
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.9061     4.4818   0.872  0.39331
age             0.4011     0.1111   3.612  0.00164 **
sexfemale      21.7625     6.9625   3.126  0.00511 **
age:sexfemale  -0.2575     0.1531  -1.682  0.10735
---
Signif. codes:  - '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.462 on 21 degrees of freedom
Multiple R-squared:  0.7687,    Adjusted R-squared:  0.7357
F-statistic: 23.27 on 3 and 21 DF,  p-value: 7.081e-07
```

Est. Mean %Fat for females = (3.906+21.762) +(0.401-0.257) Age
Est. Mean %Fat for males= 3.906 + 0.401 Age

According to this model, both age and gender are statistically significant for mean %Fat, but the interaction is not.

Better model may be one without the interaction term.

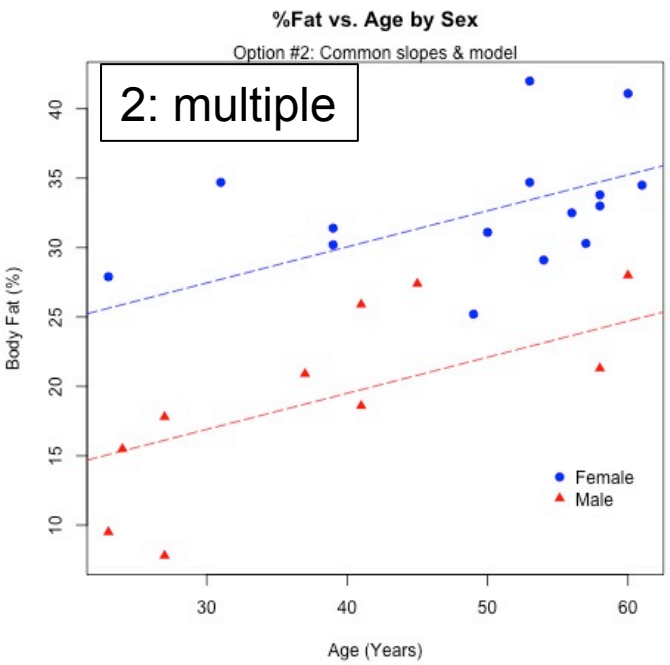


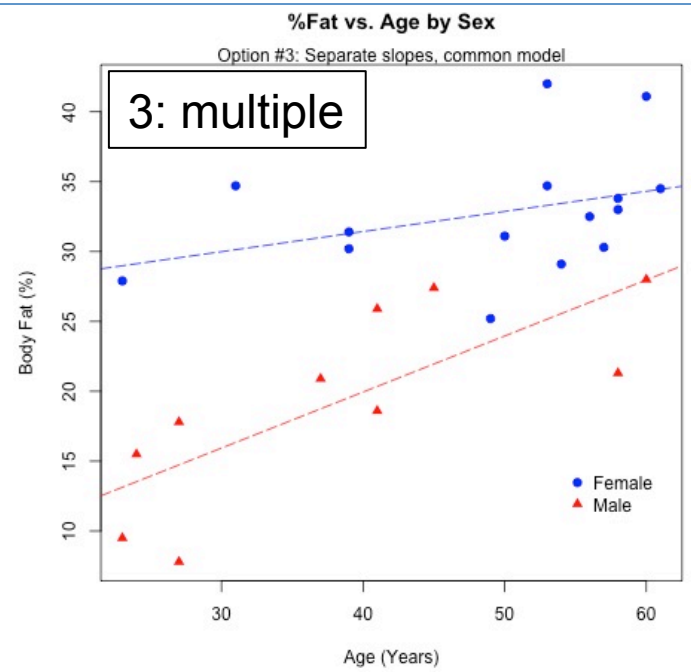
Summary of modeling options, Body Composition (%Fat) example.

Assuming intercepts are different for men and women:

1. Separate fits by gender
2. Same fit for both genders, common slope
3. Same fit for both genders, different slope

Options 1&3 give the same values for slopes and intercepts, options 2&3 allow to test whether slopes between genders are significantly different.





A note on Correlation vs. Causality

- Just because a regression has indicated a strong relationship between two variables, this does not imply that the variables are related in any causal sense.
- Causality implies correlation, correlation does not imply causality.
- Regression analysis can only address the issues on correlation.

2.3. Goodness of fit

Specific learning objectives:

Identify in R output and interpret the following:

2.3.1. ANOVA Global F-Test for Goodness of Fit.
2.3.2. R^2 and R^2_{adj}.

Goodness of Fit or Model Adequacy

Once we have estimated the parameters in the model we face two immediate questions:

1. What is the overall adequacy of the model (goodness of fit)?

   a) ANOVA F-Test

   b) Coefficient of determination R^2 and R^2_{adj}

2. When $k \geq 2$, which specific regressors seem important?

   Hypothesis tests and Confidence Intervals on individual regression coefficients and also subsets of coefficients.

ANOVA F-Test

Assessment of the general significance of the model.

Analysis of variance

Partition of variability (Sums of squares) :

Total = Explained + Unexplained

Related to the sample variance*

Given by the independent variables <u>in</u> the regression line (in simple case)

Error = Variability <u>around</u> the regression line (in the simple case)

$$\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$$

$$\text{SST} = \text{SSR} + \text{SSE}$$

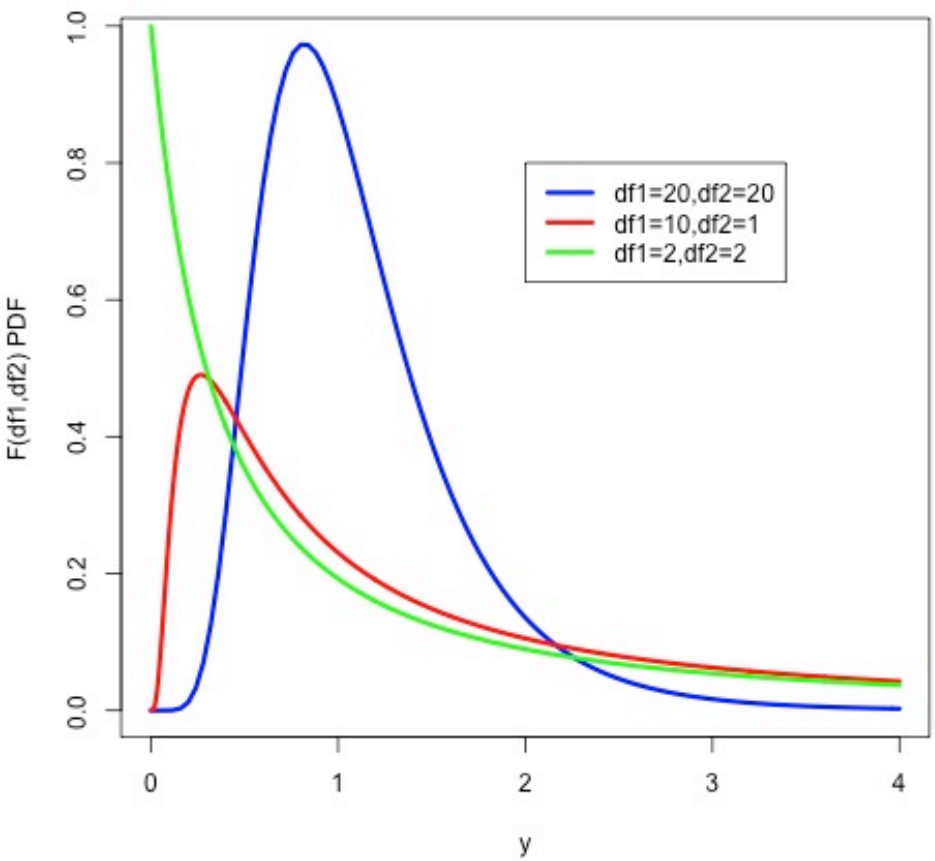
Ideally, SSR >>> SSE

*Recall

$$s^2 = \frac{\left(Y_1-\bar{Y}\right)^2+\cdots+\left(Y_n-\bar{Y}\right)^2}{n-1}$$

How do we test whether SSR >>> SSE?

F-Distribution is used to compare SSR vs. SSE



- Right-skewed.
- Ranges in [0, ∞).
- Used most commonly in Analysis of Variance.
- Results from the ratio of two squared random quantities (e.g., SST,SSR,SSE).
- Has two parameters called degrees of freedom: df of numerator, df of denominator.
- The numerator df is always given first, as switching the order changes the distribution (e.g., $F_{(10,12)}$ does not equal $F_{(12,10)}$).

AKA Fisher–Snedecor distribution, after Biologist & Statistician Ronald Fisher and Mathematician & Statistician George W. Snedecor, mid 1900's

R functions for the F(df1,df2) Distribution
E.g., df1=df2=20

```
rf(100,df1=20,df=20)        # simulates 100 observations

pf(1,df1=20,df2=20)         # gives the probability below the value of 1

df(1,df1=20,df2=20)         # gives the value of the PDF evaluated at 1

qf(.25,df1=20,df2=20)       # gives the Q1
qf(c(.25,.75))              # gives the Q1 and Q3
qf(.85,df1=20,df2=20)       # gives the 85th percentile
qf(.10,df1=20,df2=20)       # gives the 1st decile
```

"r" in rf for random
"p" in pf for probability
"d" in df for density
"q" in qf for quantile

How do we test whether SSR >>> SSE?

ANOVA TABLE
Global F-Test for Goodness of Fit

Used to assess how big SSR is with respect to SSE on average

The **linear** relationship of Y vs. covariates is assessed through:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0 \text{ for at least one } j$$

AKA "Omnibus Test"

Source	Degrees of freedom	Sum of Squares	Mean Squares	F value	P-Value
Model	*k*	*SSR*	*MSR = SSR / k*	*F = MSR/MSE*	$P(F>F_{k,n-k-1,\alpha})$
Residual	*n-k-1*	*SSE*	*MSE = SSE / (n-k-1)*		
Total	*n-1*	*SST*	*MST = SST/ (n-1)*		

$\hat{\sigma}^2$

(Note: under the regression model, MST is still the sample variance but not the best estimate for σ^2 anymore.)

k=1

Failing to reject H_0 implies that there is no **linear** relationship between X and Y, note there still could be a relationship (e.g. quadratic) or no relationship at all.

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

k≥2

Rejection of H_0 implies that at least one of the covariates contributes significantly to the model and has a linear relationship with Y

Another Goodness of Fit measure…

Coefficient of determination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Develops from the analysis of variance of the model.
- Measures the amount of variability in Y remaining after X has been considered.
- In the simple linear model (k=1), it is equivalent to the square of the sample correlation.
- Ranges within [0,1] inclusive, values close to 1 imply that most of the variability is explained by X.
- Does not necessarily imply that the regression model is an accurate predictor, so always interpret along with a scatter plot of Y vs. X.

Adjusted coefficient of determination (relevant in the multiple case, k≥2)

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

- p is the number of parameters including β_0, $p = k+1$
- R^2 never decreases when a regressor is added to the model regardless of the value of the contribution of that variable, so R^2_{adj} is sometimes preferred.
- In the variable selection stage, R^2_{adj} penalizes for adding terms that are not necessary and so it is helpful in evaluating and comparing candidate regression models.
- Ranges within [0,1] inclusive, values close to 1 imply that most of the variability is explained by the regressors.

R code and output: Body Composition Example, different slopes Illustrates Goodness of fit and coefficient significance

```
fit.full <- lm(fat~age+sex+age:sex,data=agefat)
summary(fit.full)
```

R output (portion):

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.9061     4.4818   0.872  0.39331
age             0.4011     0.1111   3.612  0.00164 **
sexfemale      21.7625     6.9625   3.126  0.00511 **
age:sexfemale  -0.2575     0.1531  -1.682  0.10735
---
Signif. codes:  - '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.462 on 21 degrees of freedom
Multiple R-squared:  0.7687,  Adjusted R-squared:  0.7357
F-statistic: 23.27 on 3 and 21 DF,  p-value: 7.081e-07
```

ANOVA F-test shows high global significance of the model

```
> 1-pf(23.27,3,21)
[1] 7.07056e-07
```

About R^2_{adj}=74% of the total variation is explained by this model, adjusting for the number of parameters.

The Global significance F-test Rejects the null hypothesis in favor of the alternative:
H_0: all coefficients are zero vs. H_1: at least one coefficient is not equal to zero.

2.4. ANOVA F-Test for nested models and Variable selection

Specific learning objectives:

2.4.1. Implement the ANOVA F-Test for nested models in R and interpret results.
2.4.2. Implement variable selection methods step by step in R.

ANOVA F-test for nested models

- Used for joint hypothesis testing, e.g.,

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{vs.} \quad H_1 : \text{ at least one } \beta_j \neq 0, \quad j = 1,2,3.$$

Accumulated Type I error for individual tests ≥ Type I error for joint tests

$$\left\{ \begin{array}{lcl} H_0 : \beta_1 = 0 & \text{vs.} & H_1 : \beta_1 \neq 0 \\ H_0 : \beta_2 = 0 & \text{vs.} & H_1 : \beta_2 \neq 0 \\ H_0 : \beta_3 = 0 & \text{vs.} & H_1 : \beta_3 \neq 0 \end{array} \right.$$

When to combine hypothesis? *:

- Set of dummy variables
- Combined effects across independent variables
- Polynomials

*Veazie, P.J.(2006) "When to Combine Hypotheses and Adjust for Multiple Tests." http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1713204 (accessed Jan 21 2016)

When to combine hypothesis

Set of Dummy Variables, Example

Veazie, P.J.(2006)

Suppose we have the following two hypotheses:

1. The utilization of emergency services is not greater for Blacks than Whites.
2. Utilization is not greater for Native Americans than Whites.

$$Y_i = \beta_0 + \beta_1 Black_i + \beta_2 NativeA_i + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 + \varepsilon_i, \quad \text{if } i \text{ is Black,}$$

$$Y_i = \beta_0 + \beta_2 + \varepsilon_i, \quad \text{if } i \text{ is Native American,}$$

$$Y_i = \beta_0 + \varepsilon_i, \quad \text{if } i \text{ is White.}$$

- If our interest in each minority group is independent of the other:

$$H_0 : \beta_1 = 0 \quad \text{or} \quad H_0 : \beta_2 = 0$$

to reduce type I error

- If interest is in both groups combined: $H_0 : \beta_1 = \beta_2 = 0$

A claim that "Blacks and Native Americans both do not differ from Whites in their utilization" makes sense only if both coefficients are simultaneously zero.

Combined effects across independent variables, Example

Veazie, P.J.(2006)

Suppose we reject the two following hypotheses:

1. Age does not differentiate health care utilization.
2. Wealth does not differentiate health care utilization.

These individual hypothesis tests do not warrant claims regarding wealthy elderly, poor youth, or other combinations.

The coefficients for the age and wealth variables must both be nonzero, if such claims are to be made.

When to combine hypothesis

Polynomials

- Regression model of a dependent variable on a second order polynomial, a practice used to capture nonlinear relationships:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^{\ 2} + \varepsilon_i$$

- If the null hypothesis for each coefficient of the polynomial is rejected according to its individual *t*-statistic,

$$H_0 : \beta_1 = 0 \quad \text{and} \quad H_0 : \beta_2 = 0$$

- It could be (erroneously) concluded that the explanatory variable has a parabolic relationship with the dependent variable, suggesting the hypotheses that *both* coefficients were simultaneously zero was rejected:

$$H_0 : \beta_1 = \beta_2 = 0$$

- Testing second-order nonlinearity (not a parabolic shape) implies testing only

$$H_0 : \beta_2 = 0$$

ANOVA F-Test for nested models Procedure

- Consider models 1 and 2, model 1 is 'nested' within model 2.

Model 1: “Restricted”,” reduced”

$$\text{E.g.,} \quad Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i;$$

Model 2: “Unrestricted”, “full”

$$\text{E.g.,} \quad Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i.$$

- Both are estimated using the same sample of size n.
- Model 1 has $p1$ (e.g. 2) parameters, and Model 2 has p_2 (e.g. 4) parameters, where $p_2 > p_1$.

Hypothesis statement

H_0: The omitted coefficients in Model1 are all zero
vs.
H_1: At least one of the omitted coefficients is zero

H_0: $\beta_2=\beta_3=-$ zero
vs.
H_1: at least one $\beta_j \neq -$, j=2,3.

H_0: Model 2 does not provide a better fit than model 1
vs.
H_1: Model 2 does provide a better fit than model 1.

ANOVA F-Test for nested models

Test statistic

$$F = \frac{\left(\frac{SSE_1 - SSE_2}{df_1 - df_2}\right)}{\left(\frac{SSE_2}{df_2}\right)} \sim F - Distribution(df_1 - df_2, df_2)$$

Decision rule

Reject H_0 at an α level of significance if

$$F_0 > f_{(1-\alpha, df1-df2, df2)} \quad \text{or} \quad P(F>F_0) < \alpha,$$

where:

- F_0 is the observed value of F and
- $f_{(1-\alpha, df1-df2, df2)}$ is the quantile to the *(1-α)x100-th* percentile of the F-Distribution with parameters (df_1-df_2, df_2).

Example: Laboratory analysis results on calcium (mm/l) on 176 elderly subjects

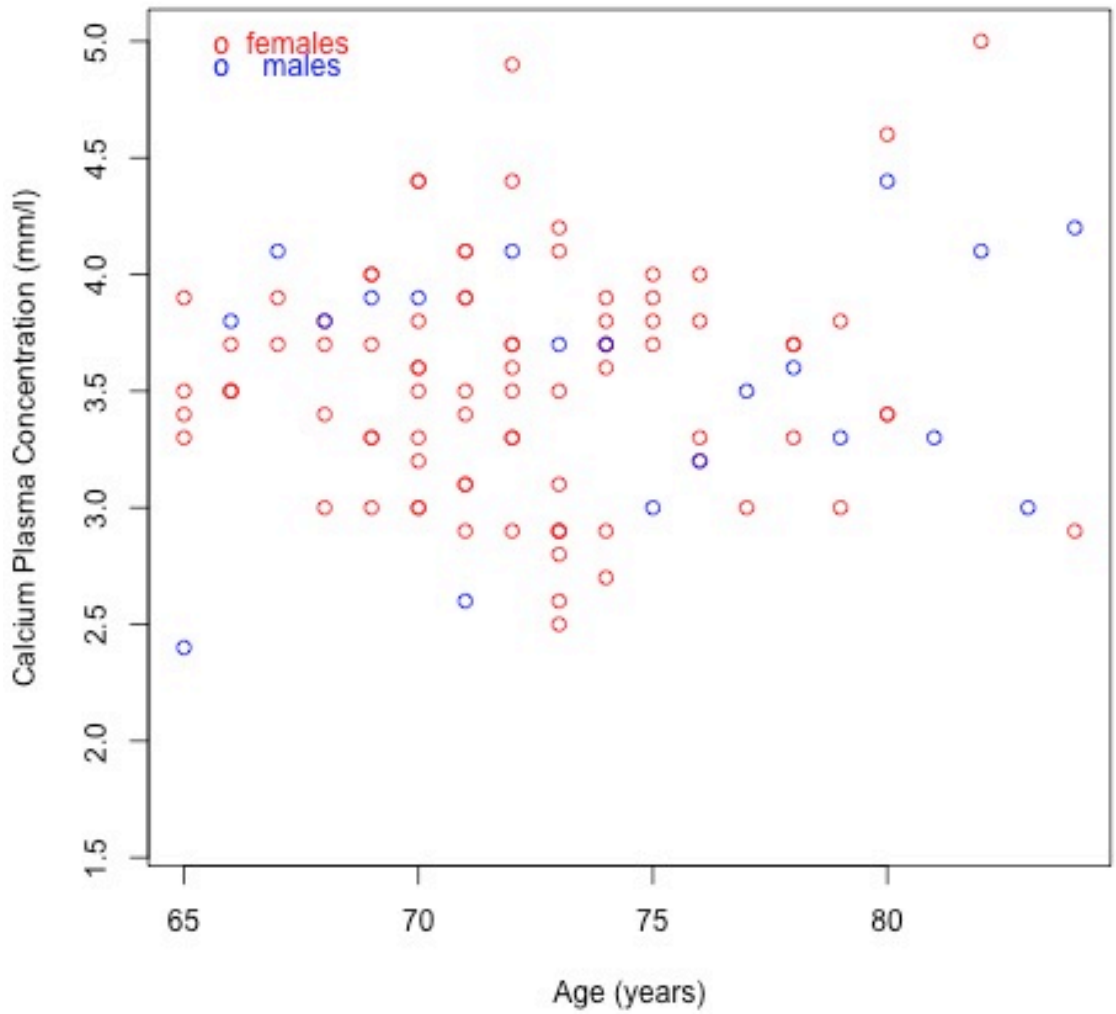
Boyd et al. (1998)*

- Results of laboratory analyses of plasma concentration of calcium (mm/l), inorganic phosphorous (mm/l), and alkaline phosphatase (iu/l).
- 176 subjects (91 male, 87 female) aged 65 – 89 years.
- Primary purpose: to determine if significant gender differences exist in the mean values of calcium in subjects over age 65.
- Secondary purpose: to determine if analytical variation between six laboratories would affect the mean values of the study variable.

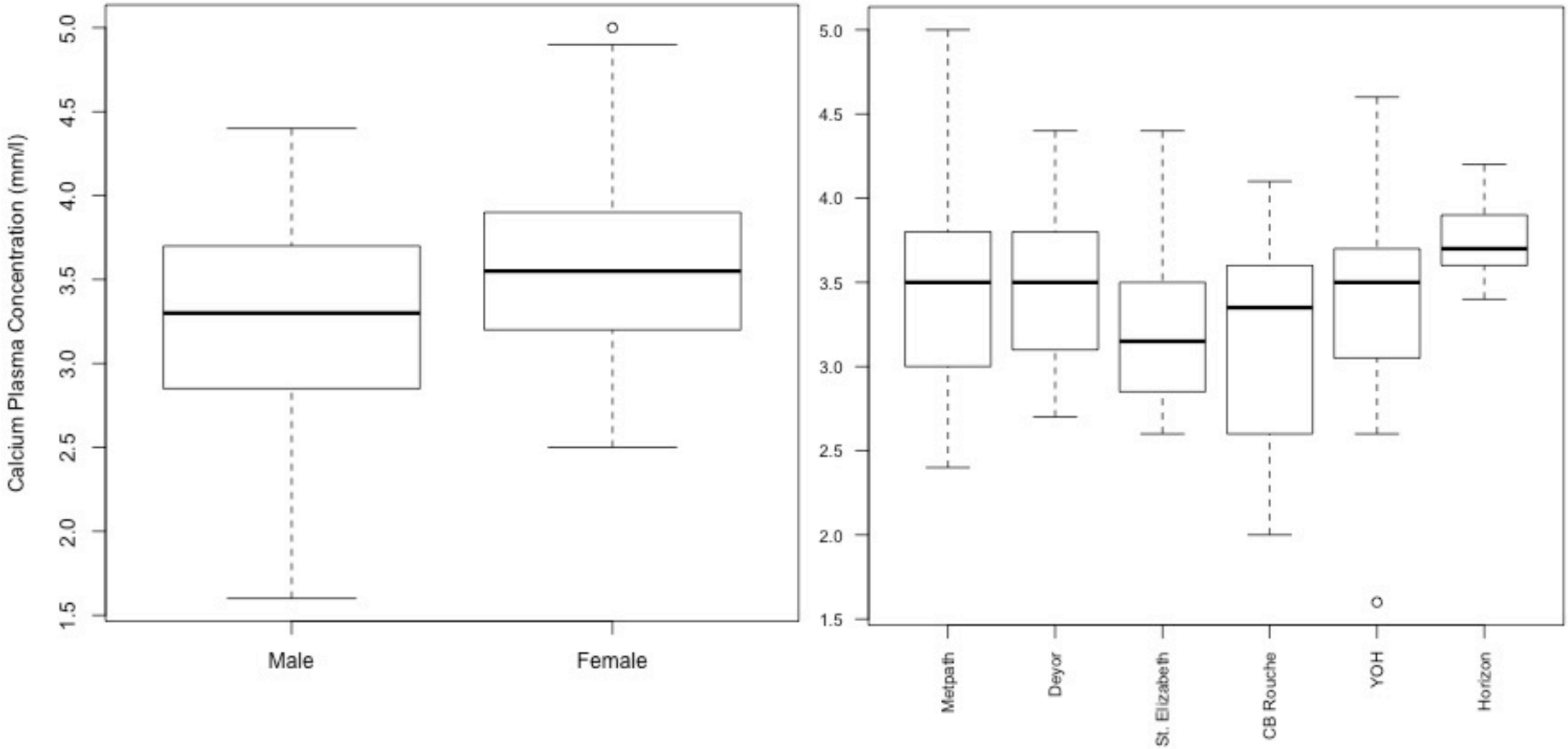
Laboratories labels and names in Youngston, OH, US.
1=Metpath; 2=Deyor; 3=St. Elizabeth's
4=CB Rouche; 5=YOH; 6=Horizon

* Boyd, J., Delost, M., and Holcomb, J., (1998). "Calcium, phosphorus, and alkaline phosphatase laboratory values of elderly subjects," Clinical Laboratory Science, 11, 223-227.
Data accessed Jan 22, 2016:
http://www.amstat.org/publications/jse/jse_data_archive.htm

Example: Laboratory analysis results on calcium (mm/l) on 176 elderly subjects. Boyd et al. (1998)*

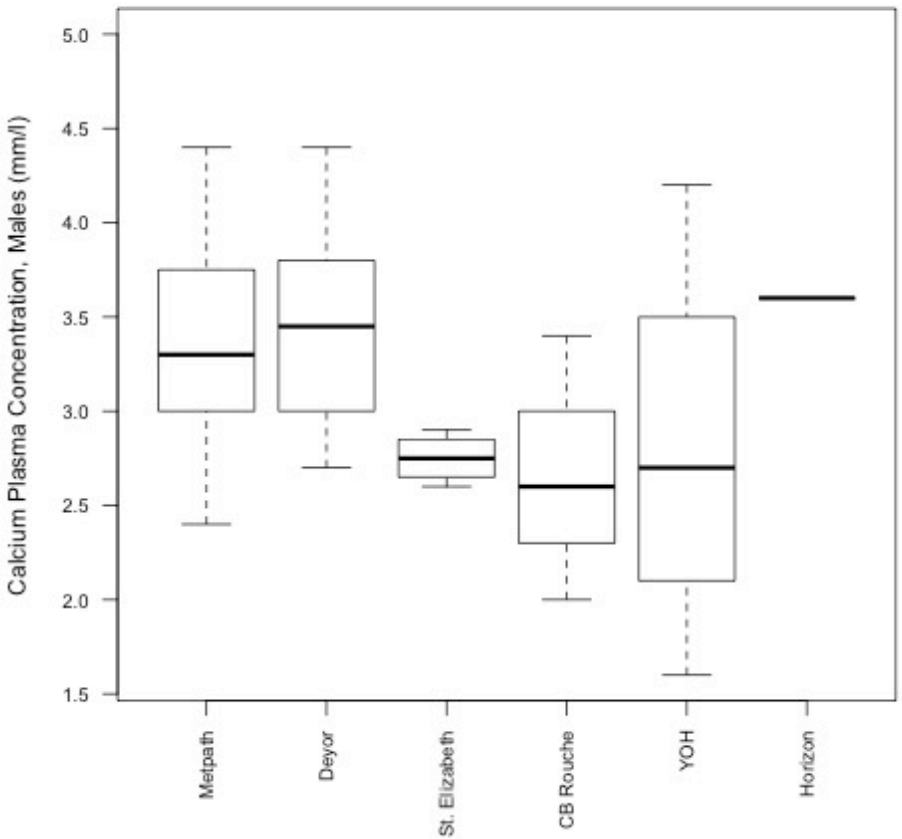


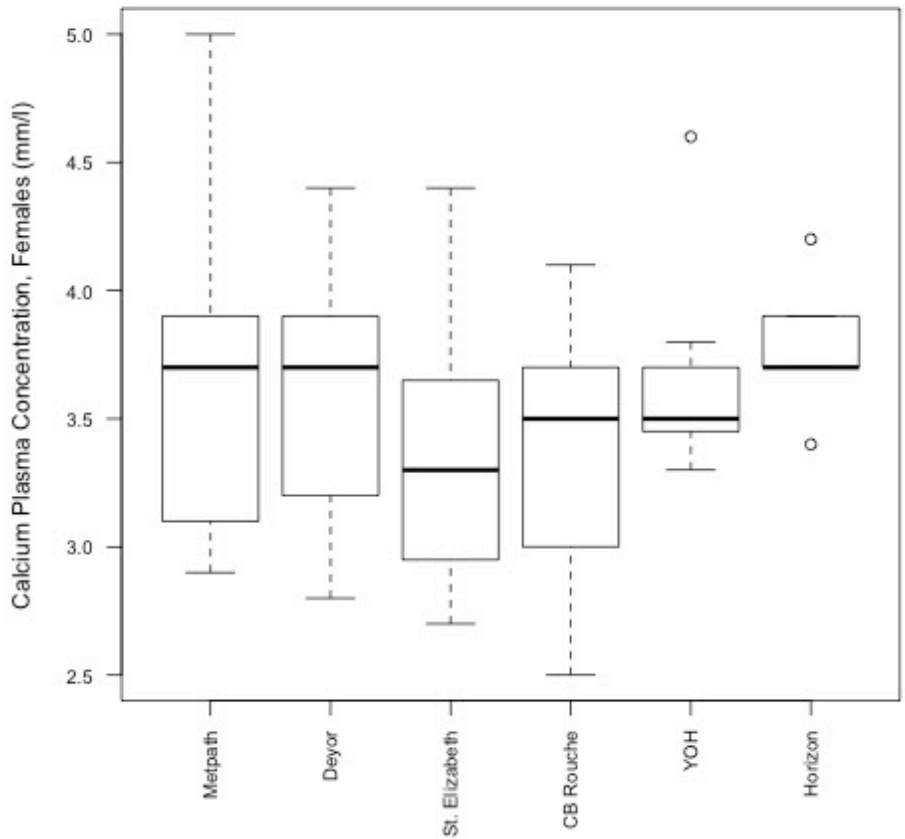
Example: Laboratory analysis results on calcium (mm/l) on 176 elderly subjects. Boyd et al. (1998)*



Example: Laboratory analysis results on calcium (mm/l) on 176 elderly subjects. Boyd et al. (1998)*

```
> table(Sex,Lab)
   Lab
Sex            1  2  3  4  5  6
  (Males)  1  59 20  4  3  4  1
  (Females)2  29 21 12 11  7  5
```





Example: Laboratory analysis results on calcium (mm/l) on 176 elderly subjects
Boyd et al. (1998)

Restricted Model (Model 1)

$$Y_i = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \varepsilon_i$$

Males $Y_i = \beta_0 + \beta_1 Age_i + \varepsilon_i$

Females $Y_i = (\beta_0 + \beta_2) + \beta_1 Age_i + \varepsilon_i$

Unrestricted Model (Model 2)

$$Y_i = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \beta_3 Lab2 + \beta_4 Lab3 + \beta_5 Lab4 + \beta_6 Lab5 + \beta_7 Lab6 + \varepsilon_i$$

	Males	Females
For Lab1 = Metpath	$Y_i = \beta_0 + \beta_1 Age_i + \varepsilon_i$	$Y_i = (\beta_0 + \beta_2) + \beta_1 Age_i + \varepsilon_i$
For Lab2 = Deyor	$Y_i = (\beta_0 + \beta_3) + \beta_1 Age_i + \varepsilon_i$	$Y_i = (\beta_0 + \beta_2 + \beta_3) + \beta_1 Age_i + \varepsilon_i$
$\vdots$	$\vdots$	$\vdots$
For Lab6 = Horizon	$Y_i = (\beta_0 + \beta_7) + \beta_1 Age_i + \varepsilon_i$	$Y_i = (\beta_0 + \beta_2 + \beta_7) + \beta_1 Age_i + \varepsilon_i$

Example: Laboratory analysis results on calcium (mm/l) on 176 elderly subjects
Boyd et al. (1998)

```
> fit.red <-lm(Cammol~ Age + Sex,data=cal2)
> summary(fit.red)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.380600   0.600126   7.299 1.02e-11 ***
Age          -0.015166   0.008244  -1.840  0.06755 .
factor(Sex)2  0.254820   0.080002   3.185  0.00172 **
---
Signif. codes:  - '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5284 on 172 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.07522, Adjusted R-squared:  0.06446
F-statistic: 6.995 on 2 and 172 DF,  p-value: 0.001201
```

Example: Laboratory analysis results on calcium (mm/l) on 176 elderly subjects
Boyd et al. (1998)

```
> fit.full<-lm(Cammol ~ Age + Sex + Lab,data=cal2)
> summary(fit.full)
...

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.173664   0.594798   7.017 5.41e-11 ***
Age          -0.012015   0.008222  -1.461 0.145813
factor(Sex)2  0.316782   0.084544   3.747      246 ***
factor(Lab)2  0.037967   0.100092   0.379 0.704932
factor(Lab)3 -0.327141   0.146384  -2.235 0.026756 *
factor(Lab)4 -0.333440   0.156083  -2.136 0.034112 *
factor(Lab)5 -0.145783   0.168492  -0.865 0.388160
factor(Lab)6  0.181450   0.223297   0.813 0.417608
---
Signif. codes:  - '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5186 on 167 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1351,  Adjusted R-squared:  0.09883
F-statistic: 3.726 on 7 and 167 DF,  p-value:      8962
```

Example: Laboratory analysis results on calcium (mm/l) on 176 elderly subjects
Boyd et al. (1998)

```
> anova(fit.red,fit.full)
Analysis of Variance Table

Model 1: Cammol ~ Age + Sex
Model 2: Cammol ~ Age + Sex + factor(Lab)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    172 48.018
2    167 44.909  5    3.1084 2.3118 0.04618 *
---
Signif. codes:  - '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P(F>2.213)= 0.0462

SSE1 – SSE2
= 48.018 – 44.909
= 3.108

(3.108/5) / (44.909/167) = 2.312

df1-df2=172-167=5

df2=167

```
> 1-pf(2.31,5,167)
[1] 0.04632661
```

Example: Laboratory analysis results on calcium (mm/l) on 176 elderly subjects
Boyd et al. (1998)

Decision Rule

$F \sim F(df_1\text{-}df_2, df_2)$

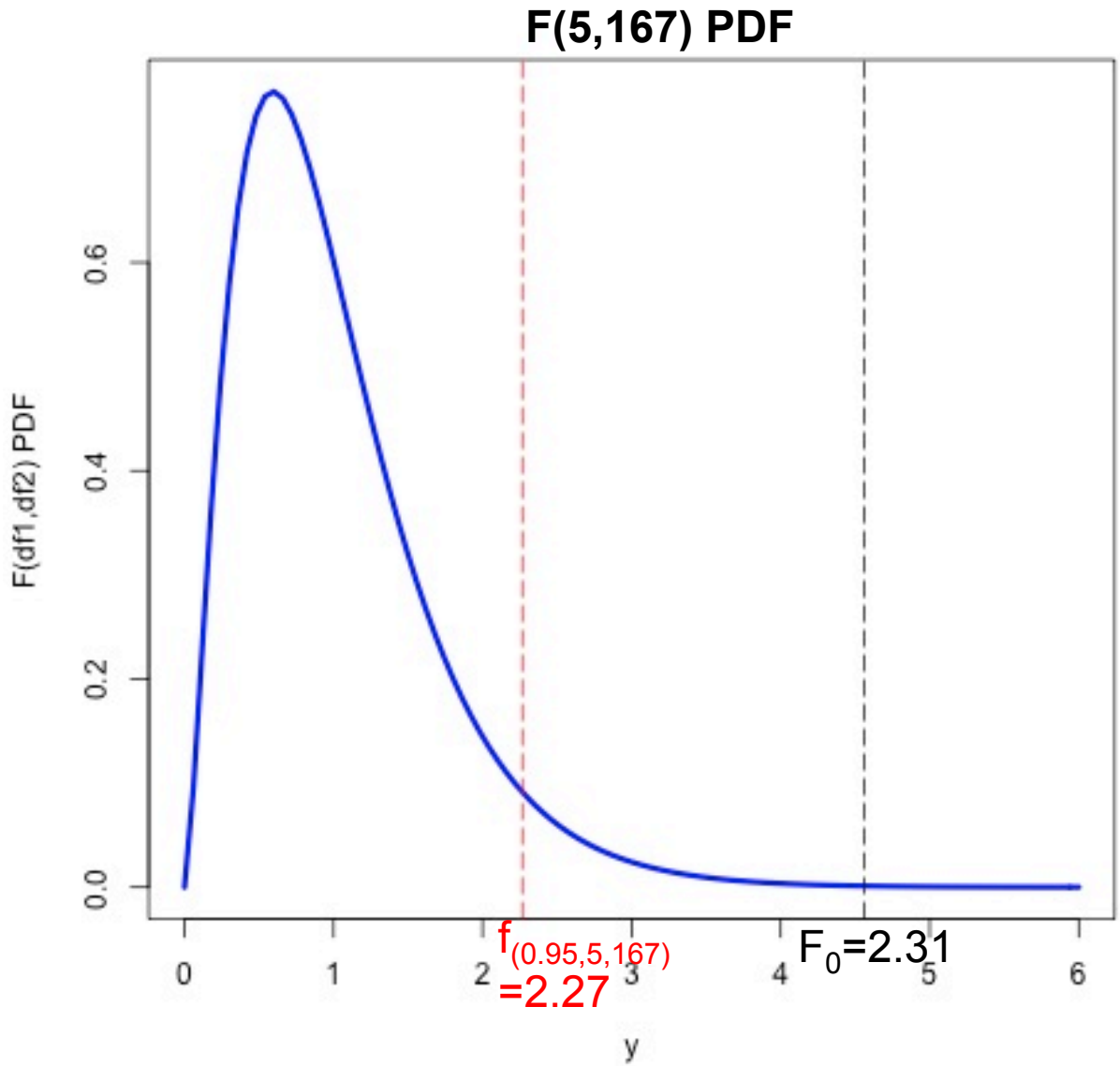
$df_1\text{-}df_2 = 172 - 167 = 5$
$df_2 = 167$

$F(df_1\text{-}df_2, df_2) = F(5, 167)$

1. P-value
   $P(F > F_0) \leq \alpha$
   $F_0 = 2.31$
   $\alpha = 0.05$

$P(F > F_0) = 0.046 \leq \alpha$

2. Critical Region
   $F_0 > f_{(1-\alpha, df1-df2, df22)}$
   $F_0 > f_{(0.95,5,167)}$




```
> qf(.95,df1=5,df2=167)
[1] 2.268267
```

```
> 1-pf(2.31,5,167)
[1] 0.04632661
```

Variable selection

- Find an appropriate subset of regressors for the model among a pool of candidates.

- A “best” regression equation compromises between:

  - Having enough number of regressors (information) that explains the response variable,

  - but not as many that interpretation becomes too complex and the prediction too uncertain (i.e., as the variance of the prediction of Y increases with the number of regressors).

- “In general, we would like to describe the system with as few regressors as possible while simultaneously explaining the substantial portion of the variability in Y.”

Variable selection methods

Compare all possible models:

- By comparing R^2, R^2_{adj} : Not used much in practice.

- Automated sequential algorithms: **Should be used with caution**
  - Forward Selection
  - Backwards Elimination,
  - Stepwise Regression.

  Risks: explanatory covariates may not necessarily have physiological or physical sense: scientific judgment can only be introduced by the user.

Suggestions:

- Narrow down the number of covariates through exploratory analyses and simple regressions,
- Assess collinearity,
- Perform the algorithms manually with a few potential regressors,

All this while objectively interpreting the resulting models (sets of covariates) in each step.

Forward Selection Method

Begins with no covariates in the model.

Step 1. Find the single variable that has the strongest association with the response and enter it to the model.

Step 2. Find one of the remaining variables that when added to the model, explains the largest amount of the remaining variability and/or has the highest significance.

Step 3. Repeat step (2) until the addition of an extra variable is not statistically significant at some chosen significance level.

Forward selection
Example, Cystic Fibrosis Data

Step 1. Find the single variable that has the strongest association with the response and enter it to the model.

Results of separately regressing Pemax on each explanatory variable

	Estimate	Std. Error	t value	Pr(>|t|)
age	4.055	1.088	3.726	0.001
sex	-19.045	13.176	-1.445	0.162
height	0.932	0.260	3.590	0.002
weight	**1.187**	**0.301**	**3.944**	**0.001***
bmp	0.639	0.565	1.131	0.270
fev1	1.354	0.555	2.439	0.023
rv	-0.123	0.077	-1.595	0.124
frc	-0.319	0.145	-2.202	0.038
tlc	-0.358	0.404	-0.886	0.38

* Most significant slope or highest correlation with Pemax

Correlation Matrix of Pemax vs. (and between) potential predictors

	age	sex	height	weight	bmp	fev1	rv	frc	tlc	pemax
age	1	-0.167	**0.926**	**0.906**	0.378	0.294	-0.552	-0.639	-0.469	0.613
sex	-	1.000	-0.168	-0.190	-0.138	-0.528	0.271	0.184	0.024	-0.289
height	-	-	1.000	**0.921**	0.441	0.317	-0.570	-0.624	-0.457	0.599
weight	-	-	-	1.000	0.673	0.449	-0.622	-0.617	-0.418	**0.635**
bmp	-	-	-	-	1.000	0.546	-0.582	-0.434	-0.365	0.230
fev1	-	-	-	-	-	1.000	-0.666	-0.665	-0.443	0.453
rv	-	-	-	-	-	-	1.000	**0.911**	0.589	-0.316
frc	-	-	-	-	-	-	-	1.000	0.704	-0.417
tlc	-	-	-	-	-	-	-	-	1.000	-0.182
pemax	-	-	-	-	-	-	-	-	-	1.000

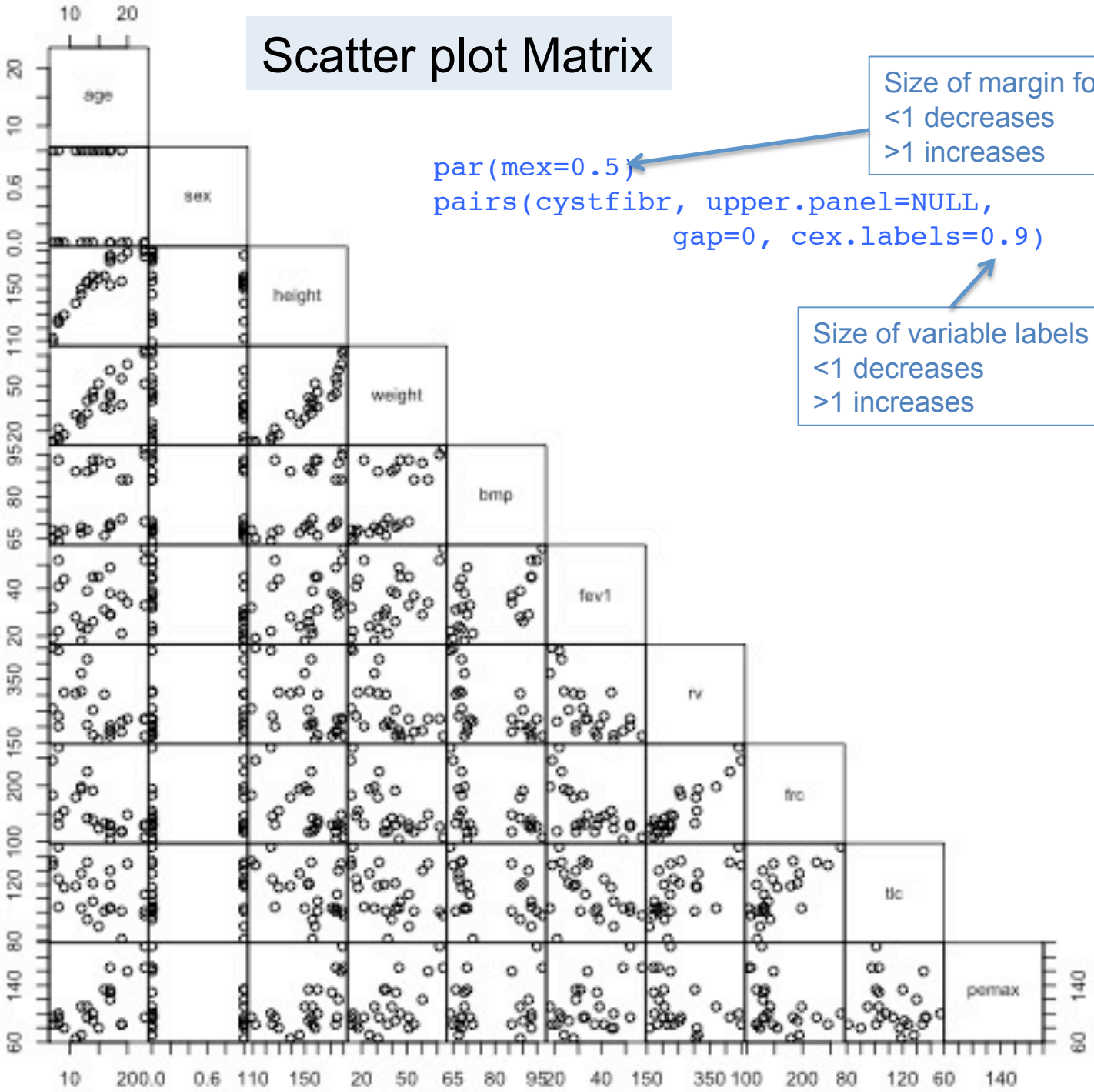
```
> mat <- as.matrix(cystfibr[,c("weight","pemax")])
> rcorr(mat, type="pearson")
       weight pemax
weight   1.00  0.64
pemax    0.64  1.00

n= 25

P
       weight pemax
weight        6e-04
pemax  6e-04
```

```
# calculating correlation matrix and removing lower triangular values
cor.mat <- round(cor(cystfibr),3)
cor.mat[lower.tri(cor.mat)] <- NA
```

Scatter plot Matrix




```
par(mex=0.5)
pairs(cystfibr, upper.panel=NULL,
                gap=0, cex.labels=0.9)
```

Size of margin fonts
<1 decreases
>1 increases

Size of variable labels
<1 decreases
>1 increases

Step 2. Find one of the remaining variables that when added to the model, explains the largest amount of the remaining variability and/or has the highest significance.

Results of adding one variable to the model with Weight

	Estimate	Std. Error	t value	Pr(>|t|)	R2adj
age	1.402	2.552	0.549	0.588	0.358
sex	-11.478	10.796	-1.063	0.299	0.381
height	0.147	0.655	0.224	0.825	0.351
bmp	**-1.005**	**0.581**	**-1.729**	**0.098**	**0.427**
fev1	0.629	0.534	1.179	0.251	0.388
rv	0.050	0.081	0.620	0.542	0.360
frc	-0.031	0.160	-0.194	0.848	0.350
tlc	0.201	0.355	0.567	0.576	0.359

Final model found by Forward Selection Algorithm

If the level of significance is set at α=0.1.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 124.8297    37.4786   3.331 0.003033 **
weight        1.6403     0.3900   4.206 0.000365 ***
bmp          -1.0054     0.5814  -1.729 0.097797 .
---
Signif. codes:  - '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.31 on 22 degrees of freedom
Multiple R-squared:  0.4749,  Adjusted R-squared:  0.4271
F-statistic: 9.947 on 2 and 22 DF,  p-value:   0.0008374
```

Backwards Elimination Method

Initial model contains all the covariates.

Step 1. Fit a full model that includes all covariates of interest.

Step 2. Remove unimportant variables until all those remaining in the model contribute significantly.

For more details, see Bonate

Step 1. Fit a full model that includes all covariates of interest.

```
> fit.full <- lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc,data=dat)
> summary(fit.full)
```

```
...

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 176.0582   225.8912   0.779    0.448
age          -2.5420     4.8017  -0.529    0.604
sex          -3.7368    15.4598  -0.242    0.812
height       -0.4463     0.9034  -0.494    0.628
weight        2.9928     2.0080   1.490    0.157
bmp          -1.7449     1.1552  -1.510    0.152
fev1          1.0807     1.0809   1.000    0.333
rv            0.1970     0.1962   1.004    0.331
frc          -0.3084     0.4924  -0.626    0.540
tlc           0.1886     0.4997   0.377    0.711

Residual standard error: 25.47 on 15 degrees of freedom
Multiple R-squared:  0.6373,    Adjusted R-squared:  0.4197
F-statistic: 2.929 on 9 and 15 DF,  p-value: 0.03195
```

Sex is the least significant with p-value=0.812

Step 2. Remove unimportant variables until all those remaining in the model contribute significantly.

从里面减去 sex

```
> fit.red1 <- update(fit.full,.~. -sex)
> summary(fit.red1)

. . .

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 153.0385   198.7149   0.770    0.452
age          -2.1145     4.3308  -0.488    0.632
height       -0.3948     0.8517  -0.464    0.649
weight        2.8349     1.8420   1.539    0.143
bmp          -1.7416     1.1207  -1.554    0.140
fev1          1.2651     0.7429   1.703    0.108
rv            0.1779     0.1743   1.021    0.323
frc          -0.2483     0.4123  -0.602    0.555
tlc           0.2084     0.4782   0.436    0.669

Residual standard error: 24.71 on 16 degrees of freedom
Multiple R-squared:  0.6359,  Adjusted R-squared:  0.4539
F-statistic: 3.493 on 8 and 16 DF,  p-value: 0.0159
```

tlc (total lung capacity) is the least significant with p-value=0.669

Step 2. Remove unimportant variables until all those remaining in the model contribute significantly.

```
> fit.red2 <- update(fit.red1,.~. -tlc)
> summary(fit.red2)
. . .
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 198.2942   165.3311   1.199   0.2468
age          -2.6632     4.0438  -0.659   0.5190
height       -0.4896     0.8037  -0.609   0.5505
weight        3.1557     1.6478   1.915   0.0725 .
bmp          -1.9625     0.9753  -2.012   0.0603 .
fev1          1.2479     0.7240   1.724   0.1029
rv            0.1596     0.1651   0.967   0.3472
frc          -0.1765     0.3687  -0.479   0.6384
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.11 on 17 degrees of freedom
Multiple R-squared:  0.6316,  Adjusted R-squared:  0.4799
F-statistic: 4.164 on 7 and 17 DF,  p-value: 0.007668
```

frc (functional residual capacity) is the least significant with p-value=0.479

Step 2. Remove unimportant variables until all those remaining in the model contribute significantly.

```
> fit.red4 <- update(fit.red3,.~. -age)
> summary(fit.red4)
. . .
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 137.09584  133.85586   1.024   0.3186
height       -0.44853    0.75059  -0.598   0.5572
weight        2.33869    1.06009   2.206   0.0399 *
bmp          -1.64100    0.72460  -2.265   0.0354 *
fev1          1.47177    0.60072   2.450   0.0241 *
rv            0.11012    0.08845   1.245   0.2283
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.13 on 19 degrees of freedom
Multiple R-squared:  0.6212,  Adjusted R-squared:  0.5215
F-statistic: 6.232 on 5 and 19 DF,  p-value: 0.001396
```

height is the least significant with p-value=0.5572

Step 2. Remove unimportant variables until all those remaining in the model contribute significantly.

```
> fit.red5 <- update(fit.red4,.~. -height)
> summary(fit.red5)
. . .
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.94669   53.27673   1.200 0.244057
weight       1.74891    0.38063   4.595 0.000175 ***
bmp         -1.37724    0.56534  -2.436 0.024322 *
fev1         1.54770    0.57761   2.679 0.014410 *
rv           0.12572    0.08315   1.512 0.146178
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.75 on 20 degrees of freedom
Multiple R-squared:  0.6141,  Adjusted R-squared:  0.5369
F-statistic: 7.957 on 4 and 20 DF,  p-value: 0.000523
```

rv (residual volume) is the least significant with p-value=0.1461

Step 2. Remove unimportant variables until all those remaining in the model contribute significantly.

Final model

```
> fit.red6 <- update(fit.red5,.~. -rv)
> summary(fit.red6)
. . .
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 126.3336    34.7199   3.639 0.001536 **
weight        1.5365     0.3644   4.216 0.000387 ***
bmp          -1.4654     0.5793  -2.530 0.019486 *
fev1          1.1086     0.5144   2.155 0.042893 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.44 on 21 degrees of freedom
Multiple R-squared:   0.57,   Adjusted R-squared:  0.5086
F-statistic: 9.279 on 3 and 21 DF,  p-value: 0.000418
```

bmp := body mass percent
fev1 := forced expiratory volume

Final model via Forward Selection

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 124.8297    37.4786   3.331 0.003033 **
weight        1.6403     0.3900   4.206      365 ***
bmp          -1.0054     0.5814  -1.729 0.097797 .
---
Signif. codes:  - '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.31 on 22 degrees of freedom
Multiple R-squared:  0.4749,    Adjusted R-squared:  0.4271
F-statistic: 9.947 on 2 and 22 DF,  p-value:   0.0008374
```

Final model via Backwards Elimination

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 126.3336    34.7199   3.639 0.001536 **
weight        1.5365     0.3644   4.216 0.000387 ***
bmp          -1.4654     0.5793  -2.530 0.019486 *
fev1          1.1086     0.5144   2.155 0.042893 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.44 on 21 degrees of freedom
Multiple R-squared:   0.57, Adjusted R-squared:  0.5086
F-statistic: 9.279 on 3 and 21 DF,  p-value: 0.000418
```

Forward Selection vs. Backwards Elimination

- Based on comments by Altman, 1991. -

- The two methods often yield the same model, but differences are not uncommon.

- Neither approach is more correct than the other.

- We might choose the larger model as it includes three Variables that are significant at the 5% level.

- On the other hand, it is peculiar to include both weight and BMP in the model.

- This example shows that p-values alone cannot choose an appropriate model.