Multiple linear regression analysis: introduction

Tuan V. Nguyen
Professor and NHMRC Senior Research Fellow
Garvan Institute of Medical Research
University of New South Wales
Sydney, Australia

What we are going to learn …

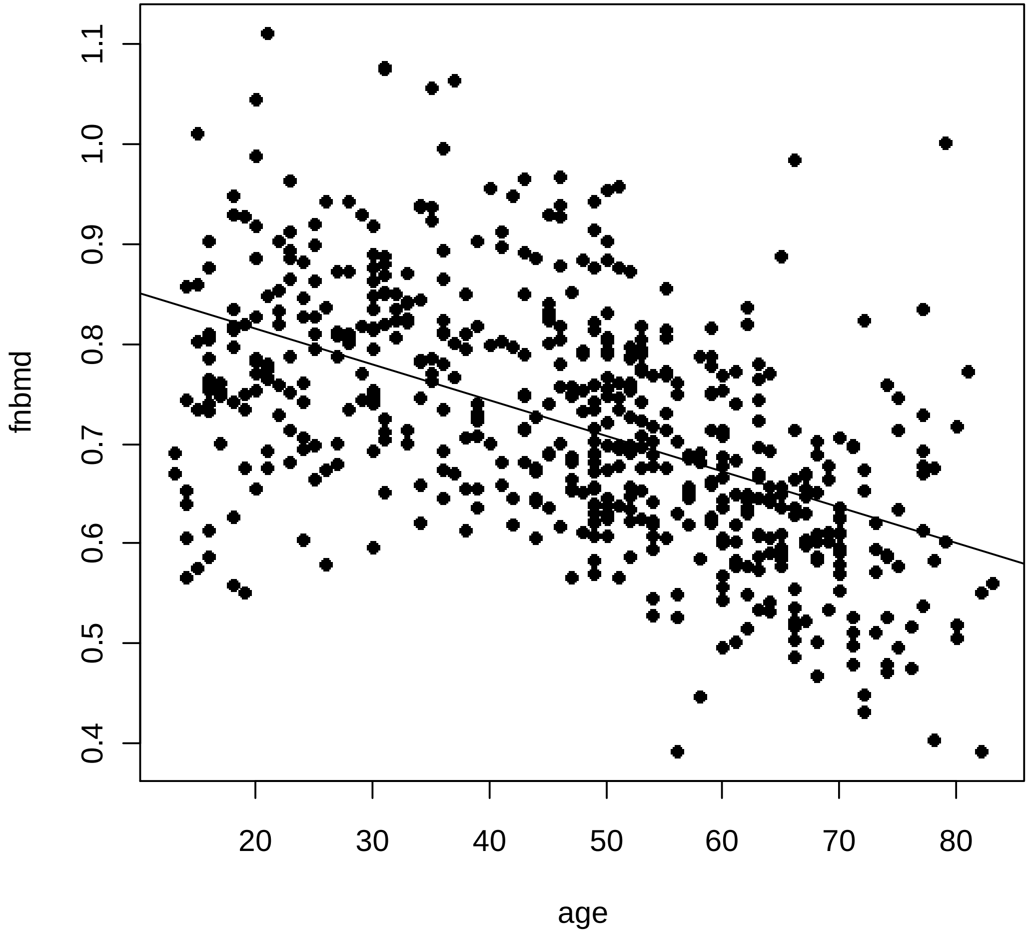
- Examples
- Purposes of MLR
- Questions of interest
- R analysis and Interpretation
- Categorical predictor
- Selection of an "optimal" model

Consider the relationships between

- Femoral neck BMD
- Weight
- and Age

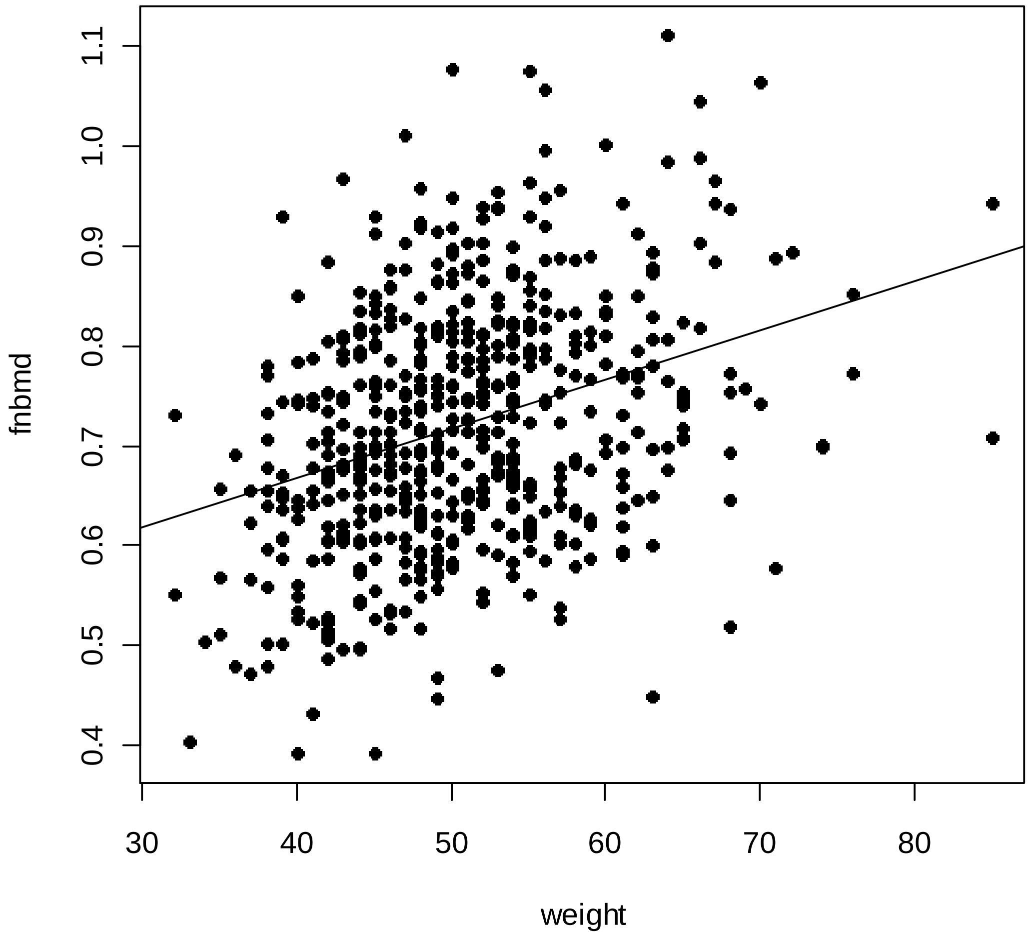
Femoral neck bone density and age

```
women = subset(vd, sex==2)
plot(fnbmd ~ age, pch=16)
abline(lm(fnbmd ~ age))
```

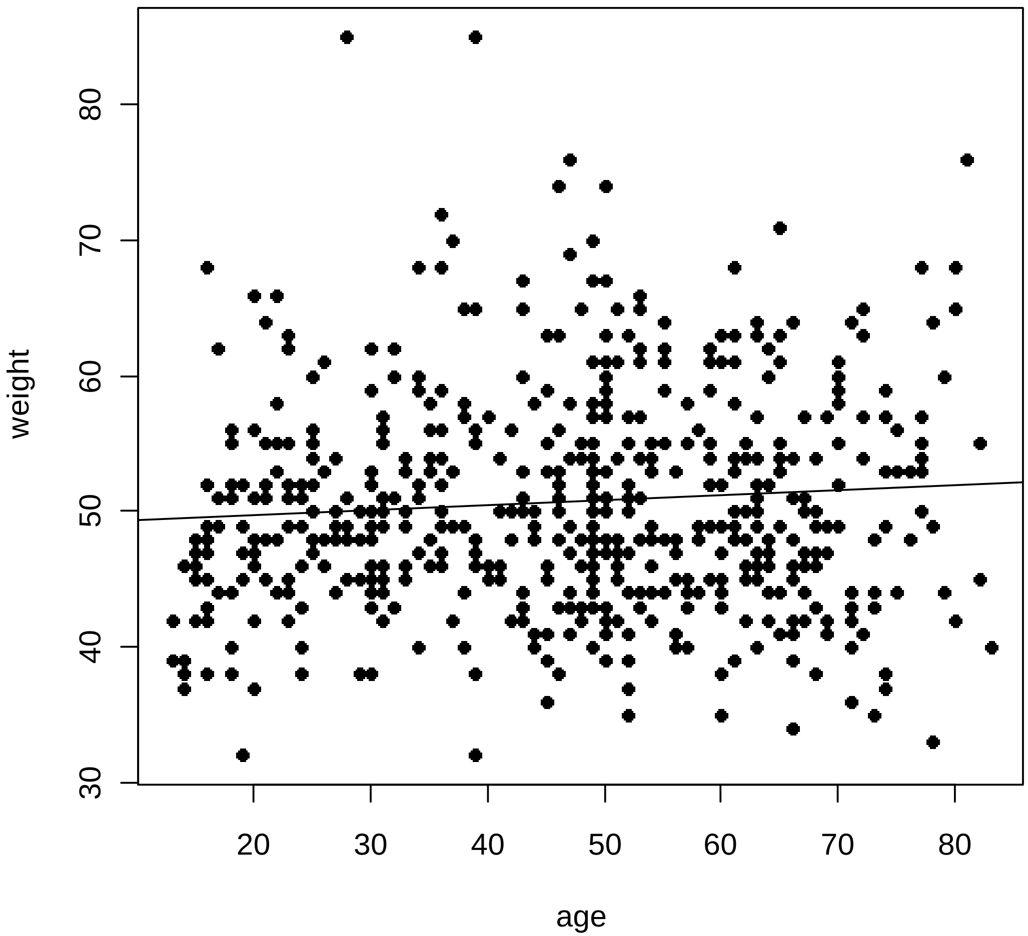


Weight and femoral neck bone density

```
plot(fnbmd ~ weight, pch=16)
abline(lm(fnbmd ~ weight))
```



Relationship between age and weight



Questions of interest

- What are the effects of age and weight on FNBMD?
- Is the effect of age on BMD independent of weight?
- How well age and weight can predict BMD?

Simple and multiple linear regression model

- Simple linear regression model

$$BMD = a + b*weight + e$$

$$BMD = a + b*age + e$$

- Multiple linear regression model

$$BMD = a + b*weight + c*age + e$$

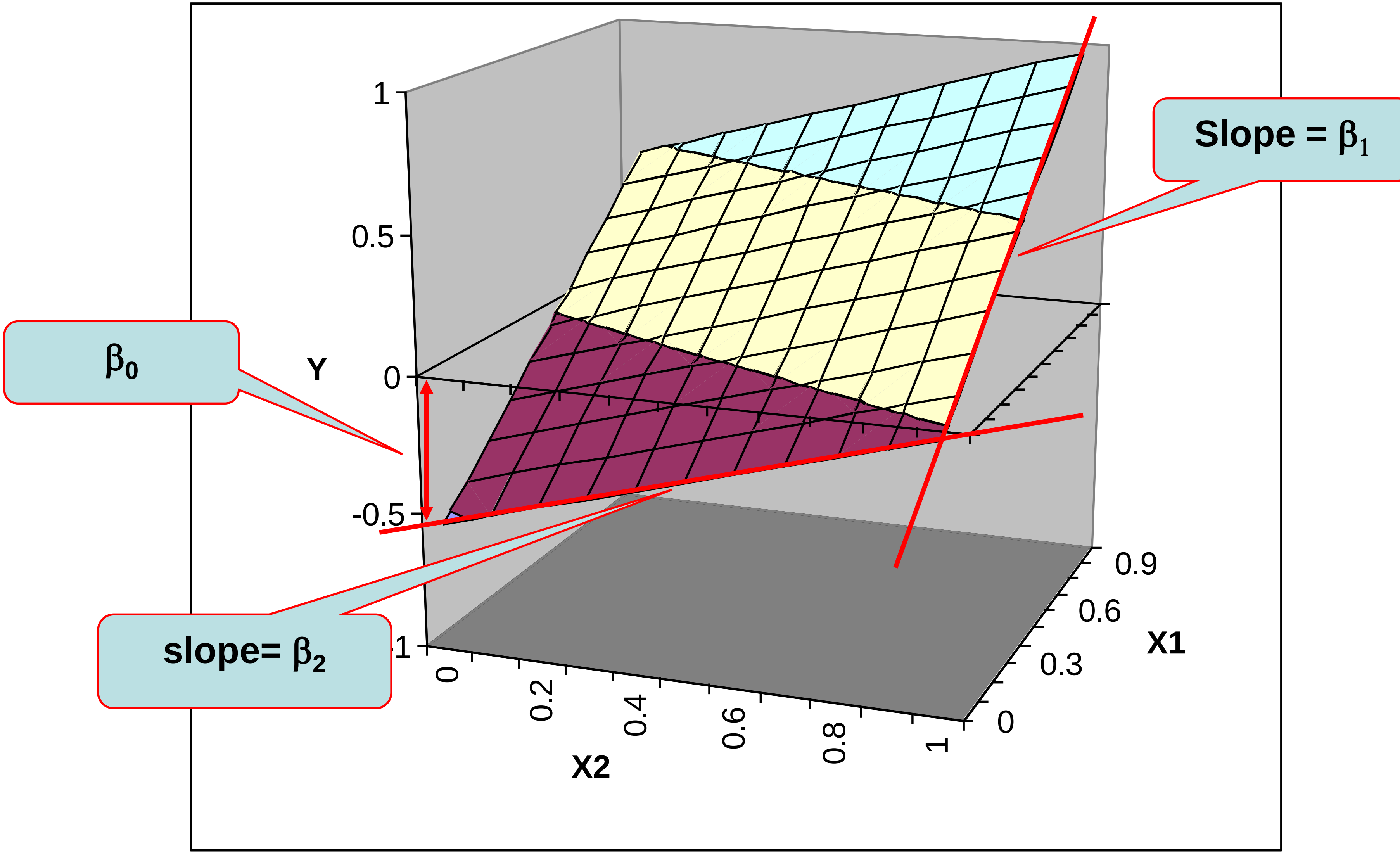
The multiple linear regression model

The model assumes

- The responses are ***normally distributed*** with means μ (each response has a ***different*** mean) and ***constant*** variance σ^2
- The mean response μ of a typical observation depends on the covariates through a ***linear relationship***

$$\mu = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

- The responses are ***independent***



Slope = β_1
β_0
slope= β_2
Y
1
0.5
0
-0.5
-1
X2
0
0.2
0.4
0.6
0.8
1
X1
0
0.3
0.6
0.9

Estimation of the coefficients

- We estimate the (unknown) regression plane by the "least squares plane" (best fitting plane)
- Best fitting plane = plane that minimizes the sum of squared vertical deviations from the plane
- That is, minimize the least squares criterion

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - \ldots - b_k x_{ik})^2$$

Estimation of the coefficients (2)

- The R command `lm` calculates the coefficients of the best fitting plane
- This function solves the *normal equations*, a set of linear equations derived by differentiating the least squares criterion with respect to the coefficients

R analysis

```
m1 = lm(fnbmd ~ weight)
m2 = lm(fnbmd ~ age)
m3 = lm(fnbmd ~ age + weight)
summary(m1); summary(m2); summary(m3)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.4699822  0.0310144   15.15  < 2e-16 ***
weight      0.0049416  0.0006041    8.18 1.95e-15 ***
---
Multiple R-squared: 0.1074,     Adjusted R-squared: 0.1058
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.8871142  0.0123679   71.73   <2e-16 ***
age         -0.0035730  0.0002478  -14.42   <2e-16 ***
---
Multiple R-squared: 0.2721,     Adjusted R-squared: 0.2708
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.6130722  0.0266642   22.99   <2e-16 ***
age         -0.0037703  0.0002243  -16.81   <2e-16 ***
weight       0.0055870  0.0004937   11.32   <2e-16 ***
---
Multiple R-squared: 0.4085,     Adjusted R-squared: 0.4064
```

Let's summarize

Model	Regression coefficient (standard error)		Coefficient of determination (R^2)
	Weight	Age	
1	0.0049 (0.0006)		0.107
2		-0.0036 (0.00025)	0.272
3	0.0056 (0.0005)	-0.0038 (0.0002)	0.408

Our model:

BMD = 0.6131 - 0.0038*age + 0.0056*weight

Interpretation

Model	Regression coefficient (standard error)		Coefficient of determination (R^2)
	Weight	Age	
1	0.0049 (0.0006)		0.107
2		-0.0036 (0.00025)	0.272
3	0.0056 (0.0005)	-0.0038 (0.0002)	0.408

Interpretation: BMD was positively associated with body weight and inversely related to age. Each kg increase in weight was associated with an 0.006 g/cm^2 increase in BMD. Furthermore, each year advancing age was associated with a decline of 0.0038 g/cm^2 in BMD, and the effect was independent of weight. Collectively, age and body weight accounted for approximately 41% of variation in BMD

Let's re-scale the data

$$z = \frac{\text{Individual value} - \text{mean}}{\text{standard deviation}}$$

- Mean of z is always 0
- When z = 0, the individua's value is *equal* to the sample mean
- When z > 0, the individual's value *higher* than the sample mean
- When z < 0, the individual's value *lower* than the sample mean

Re-scale of predictor variables

- Re-scale of predictor variable helps
  - Meaningful interpretation of regression parameter
  - Technical computation (more stable results)

R analysis

```
zage = (age-mean(age)) / sd(age)
zweight = (weight-mean(weight)) / sd(weight)
summary(lm(fnbmd ~ zage+zweight)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.720509   0.003974  181.29   <2e-16 ***
zage        -0.067071   0.003990  -16.81   <2e-16 ***
zweight      0.045150   0.003990   11.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09388 on 555 degrees of freedom
Multiple R-squared: 0.4085,     Adjusted R-squared: 0.4064
F-statistic: 191.7 on 2 and 555 DF,  p-value: < 2.2e-16
```

Presentation

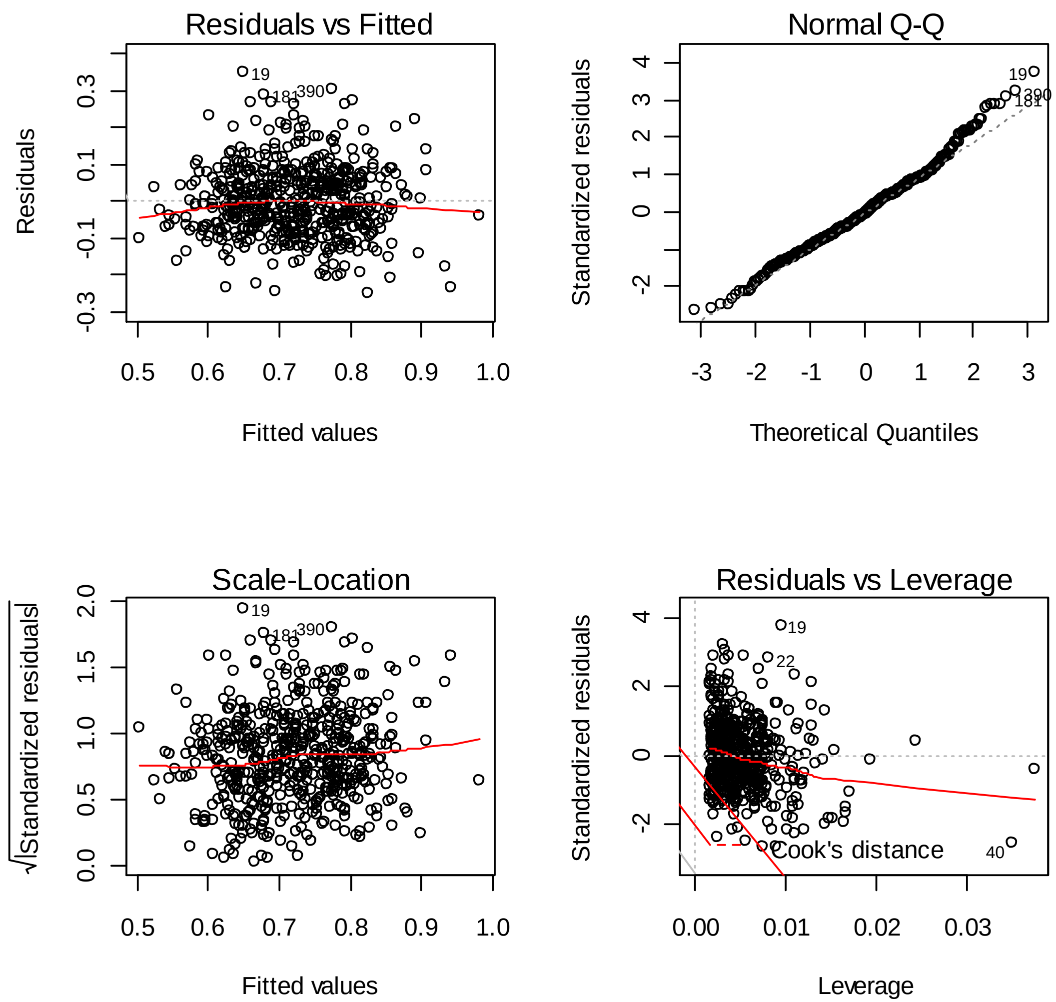
Predictor	Unit of comparison	Regression coefficient and standard error
Age	17.8	-0.067 (0.004)
Weight	8.1	0.045 (0.004)

Which variable is more important ?

Model	Variance of FNBMD	Change in variance
No predictor	0.0148	.
Age	0.0108	-0.004 (down 27%)
Weight	0.0133	-0.001 (down 11%)
Age + Weight	0.0088	-0.006 (down 41%)

Checking model assumptions

```
par(mfrow=c(2,2))
plot(m3)
```



Categorical predictor

Let's look at the data again

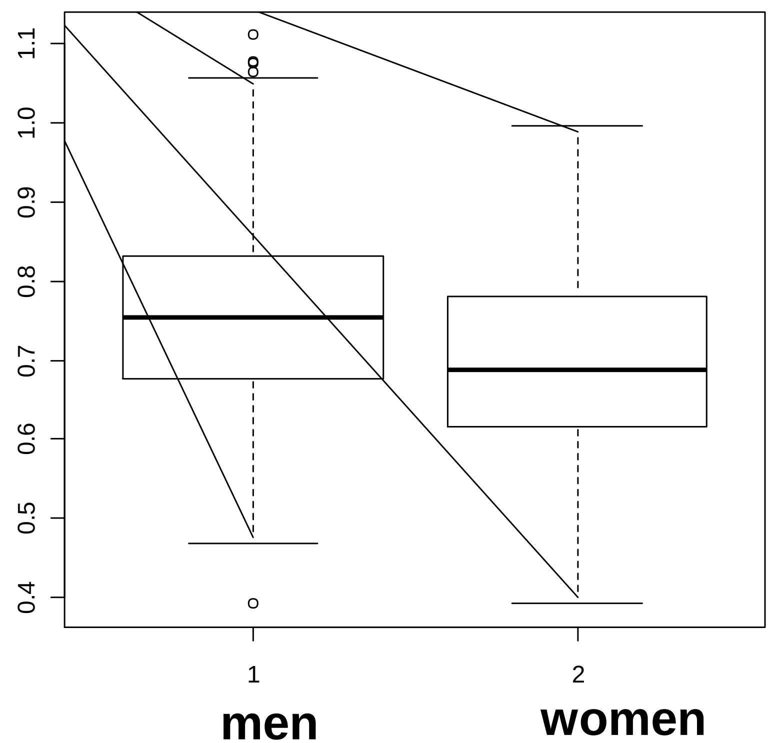
```
setwd("C:/Documents and Settings/Tuan/My Documents/_Current
   Projects/_Vietnam/Huong/Vitamin D")

vd = read.csv("vitaminD.csv", header=T, na.strings=" ")

attach(vd)

names(vd)

boxplot(fnbmd ~ sex)
```

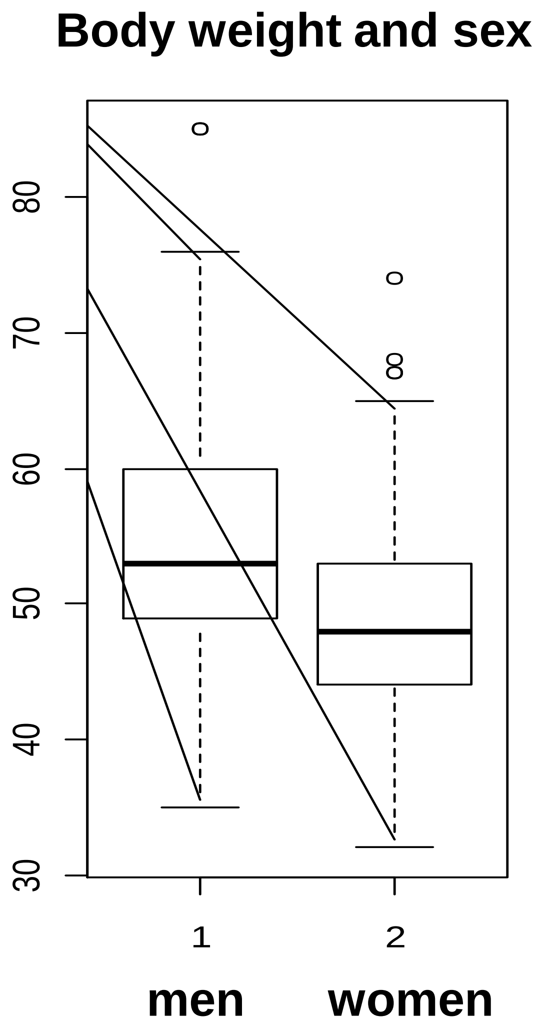


But …

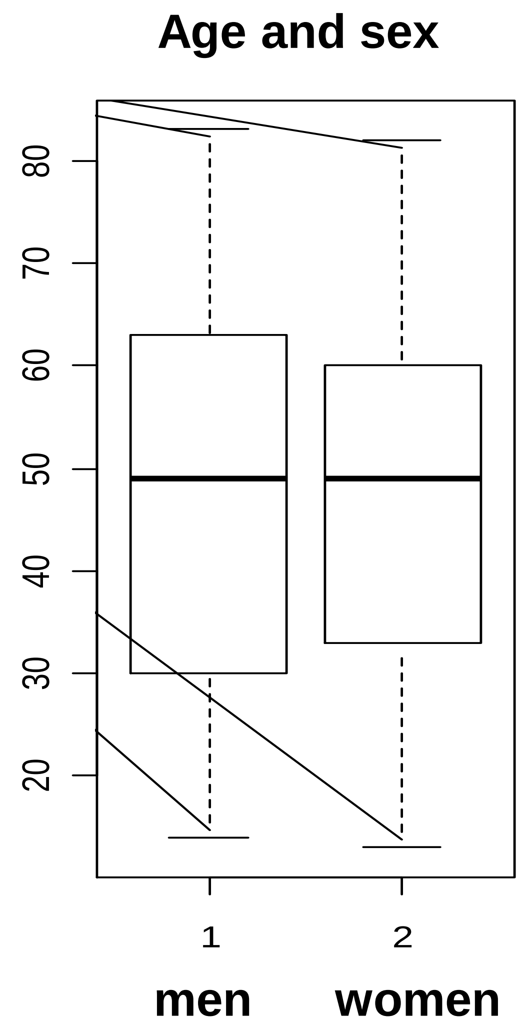
```
> temp=cbind(fnbmd,weight,age)
> describe.by(temp, sex, skew=F)
```

```
INDICES: 1 (men)
       var   n  mean    sd median trimmed   mad   min   max range   se
fnbmd    1 222  0.76  0.12   0.76    0.75  0.12  0.39  1.11  0.72 0.01
weight   2 222 54.46  8.44  53.00   54.06  7.41 35.00 85.00 50.00 0.57
age      3 222 46.67 19.30  49.00   46.57 25.20 14.00 83.00 69.00 1.30

------------------------------------------------------------
INDICES: 2 (women)
       var   n  mean    sd median trimmed   mad   min max range   se
fnbmd    1 336  0.70  0.11   0.69    0.70  0.12  0.39   1   0.6 0.01
weight   2 336 48.21  6.78  48.00   47.94  7.41 32.00  74  42.0 0.37
age      3 336 46.60 16.74  49.00   47.04 19.27 13.00  82  69.0 0.91
```



Body weight and sex
30
40
50
60
70
80
1
2
men
women



Age and sex
20
30
40
50
60
70
80
1
2
men
women

Question of interest

- Is BMD in men higher than women, *after adjusting* for age and weight ?
- Solution: multiple linear regression
- Model:

BMD = a + b*age + c*weight + d*sex

R analysis

```
m4 = lm(fnbmd ~ zage + zweight + sex)
summary(m4)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.769472   0.014466  53.191  < 2e-16 ***
zage        -0.066653   0.003951 -16.868  < 2e-16 ***
zweight      0.039447   0.004270   9.239  < 2e-16 ***
sex         -0.030561   0.008689  -3.517 0.000472 ***
---

Residual standard error: 0.09293 on 554 degrees of freedom
Multiple R-squared: 0.4215,     Adjusted R-squared: 0.4183
F-statistic: 134.5 on 3 and 554 DF,  p-value: < 2.2e-16
```

New model

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.769472   0.014466  53.191  < 2e-16 ***
zage        -0.066653   0.003951 -16.868  < 2e-16 ***
zweight      0.039447   0.004270   9.239  < 2e-16 ***
sex         -0.030561   0.008689  -3.517 0.000472 ***
```

BMD = 0.769 – 0.067*zage + 0.039*zweight – 0.03*sex

Remember sex = 1 (men), 2 (women)

For men:

BMD = 0.769 – 0.067*zage + 0.039*zweight – 0.03*1

= 0.739 – 0.067*zage + 0.039*zweight

For women:

BMD = 0.769 – 0.067*zage + 0.039*zweight – 0.03*2

= 0.709 – 0.067*zage + 0.039*zweight

Interpretation

- After adjusting for age and weight, BMD in women was on average 0.03 g/cm^2 (P = 0.0005) lower than that in men
- Note: before adjusting for age and weight, BMD in men was higher than women by 0.06 g/cm^2 (0.76 vs 0.70)

More analyses: alcohol, tea, coffee …

```
m5 = lm(fnbmd ~ zage + zweight + sex + alcohol + tea +
coffee)
summary(m5)
```

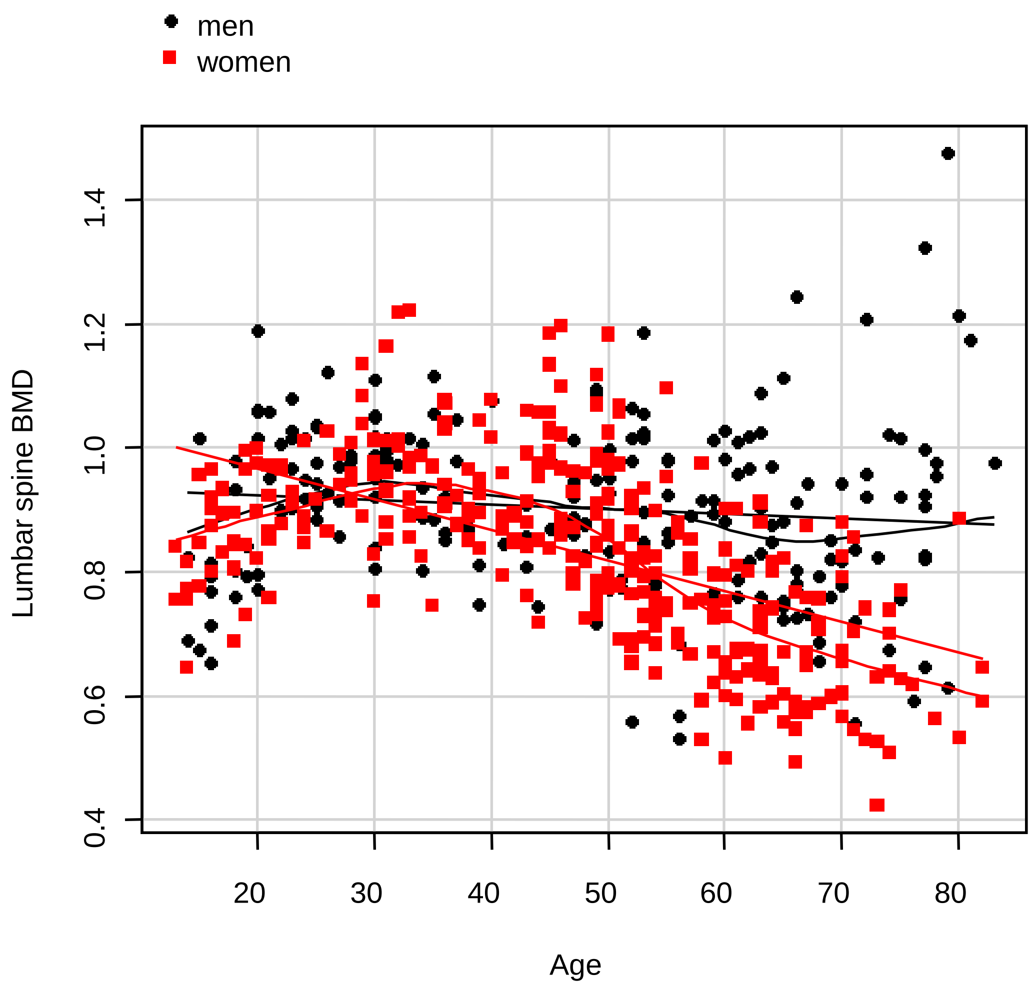
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.724237   0.026137  27.709   <2e-16 ***
zage        -0.067141   0.004061 -16.533   <2e-16 ***
zweight      0.039610   0.004298   9.217   <2e-16 ***
sex         -0.031427   0.011044  -2.846   0.0046 **
alcohol     -0.005053   0.011784  -0.429   0.6682
tea          0.003062   0.009635   0.318   0.7507
coffee       0.027430   0.012176   2.253   0.0247 *
---
Residual standard error: 0.09278 on 545 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared: 0.4184,     Adjusted R-squared: 0.412
F-statistic: 65.35 on 6 and 545 DF,  p-value: < 2.2e-16
```

Non-coffee drinkers had higher BMD than coffee drinkers!

Interaction analysis

Lumbar spine BMD, age, and sex

```
library(car)
scatterplot(lsbmd ~ age | sex, pch=c(16, 15), xlab="Age",
            ylab="Lumbar spine BMD")
```



Observation

It seems …

- **The rate of decline (with age) in lumbar spine is higher in women than in men**
- **The BMD–age relationship is different between men and women**

Interaction

R analysis

```
m6 = lm(lsbmd ~ weight + age + coffee + sex + sex:age)
summary(m6)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3633464  0.0711284   5.108 4.51e-07 ***
weight       0.0061790  0.0007018   8.804  < 2e-16 ***
age          0.0036166  0.0009506   3.805 0.000158 ***
coffee       0.0381405  0.0157832   2.417 0.015999 *
sex          0.1762090  0.0299826   5.877 7.30e-09 ***
age:sex     -0.0045147  0.0005966  -7.567 1.65e-13 ***
---
Residual standard error: 0.1223 on 542 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared: 0.3337,     Adjusted R-squared: 0.3276
F-statistic:  54.3 on 5 and 542 DF,  p-value: < 2.2e-16
```

Model for LSBMD

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.3633464	0.0711284	5.108	4.51e-07	***
weight	0.0061790	0.0007018	8.804	< 2e-16	***
age	0.0036166	0.0009506	3.805	0.000158	***
coffee	0.0381405	0.0157832	2.417	0.015999	*
sex	0.1762090	0.0299826	5.877	7.30e-09	***
age:sex	**-0.0045147**	**0.0005966**	**-7.567**	**1.65e-13**	*****

LSBMD = 0.363 + 0.006*weight + 0.0036*age + 0.038*coffee + 0.176*sex – 0.0045*sex*age

For men:

BMD = 0.363 + 0.006*weight + 0.0036*age + 0.038*coffee + 0.176*1 – 0.0045*1*age

= 0.539 + 0.006*weight - 0.0009*age + 0.038*coffee

For women:

BMD = 0.363 + 0.006*weight + 0.0036*age + 0.038*coffee + 0.176*2 – 0.0045*2*age

= 0.715 + 0.006*weight - 0.0054*age + 0.038*coffee

Summary

- Multiple linear regression is a very useful model for analyzing complex data
- Assumptions: normal distribution, variance is stable across predictor values, independence
- Always check for interaction effects