Correlations + Simple Linear Regression + Multiple Linear Regression

Statistics 102 Teaching Team

CORREATION

```
#load the data
library(oibiostat)
library(openintro)
```

```
## Loading required package:
airports
```

```
## Loading required package:
cherryblossom
```

```
## Loading required package: usdata
```

```
data("prevend.samp")
data("famuss")
```

Relationships between two variables

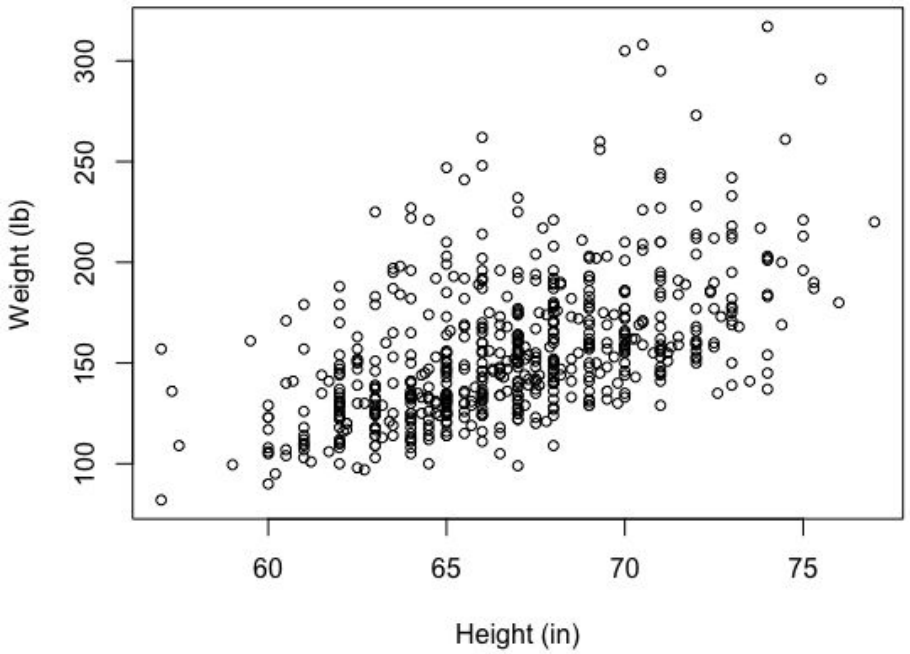
Summarizing relationships between two variables

Approaches for summarizing relationships between two variables vary depending on variable types…

Two numerical variables

Two categorical variables

One numerical variable and one categorical variable



Two numerical variables

Two variables and are

positively associated if increases as increases.

negatively associated if decreases as increases.

Two numerical variables

Correlation is a numerical summary that measures the strength of a linear relationship between two variables.

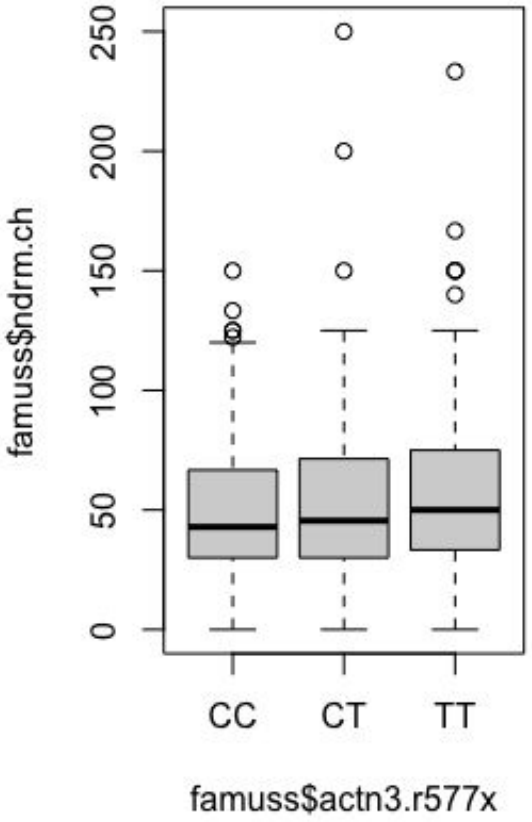
Introduced in *OI Biostat* Section 1.6.1; details in Ch. 6.

The correlation coefficient takes on values between -1 and 1.

The closer is to , the stronger the linear association.

```
cor(famuss$height,
famuss$weight)
```

```
## [1] 0.5308787
```



_____

SIMPLE LINEAR REGRESSION

The main ideas

Linear regression provides methods for examining the association between a quantitative response variable and a set of possible predictor variables.

- Linear regression should only be used with data that exhibit linear or approximately linear relationships.

Simple linear regression is used to estimate the linear relationship between a response variable and a single predictor .

- The response variable can be referred to as the *dependent* variable, and the predictor variable, the

Examining scatterplots

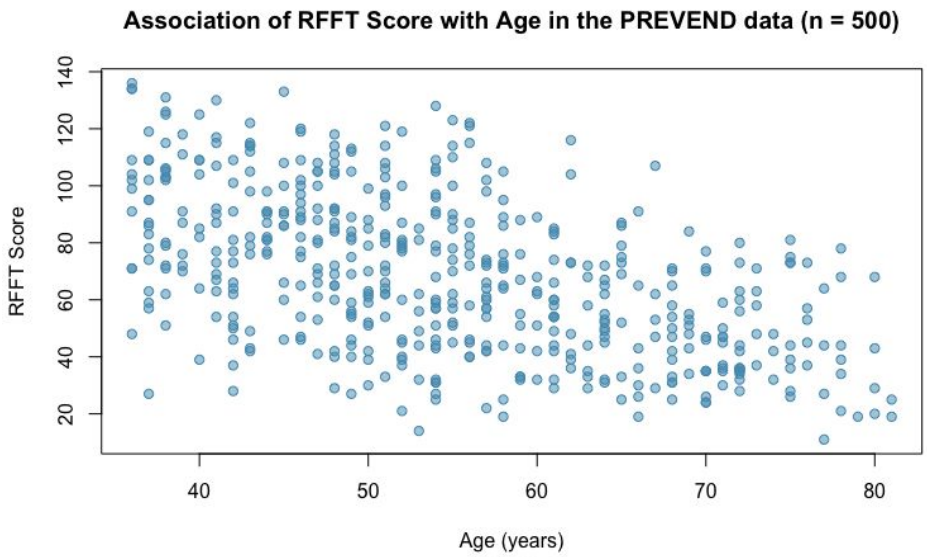
The PREVEND study

As adults age, cognitive function changes over time; largely due to various cerebrovascular and neurodegenerative changes.

The Prevention of REnal and Vascular END-stage Disease (PREVEND) study measured various clinical and demographic data for participants in a series of surveys between 1997 - 2006.

Data from 4,095 participants are in the dataset in the package.

Cognitive function was assessed with the Ruff Figural Fluency Test (RFFT), which provides information about cognitive abilities such as planning and the ability to switch between



Age vs RFFT in …

The relationship between age and RFFT score appears linear. A line might provide a useful summary of this association.

Lab 1 steps through fitting and interpreting a line as well as evaluating whether the assumptions for linear regression are satisfied.

Assumptions for linear regression

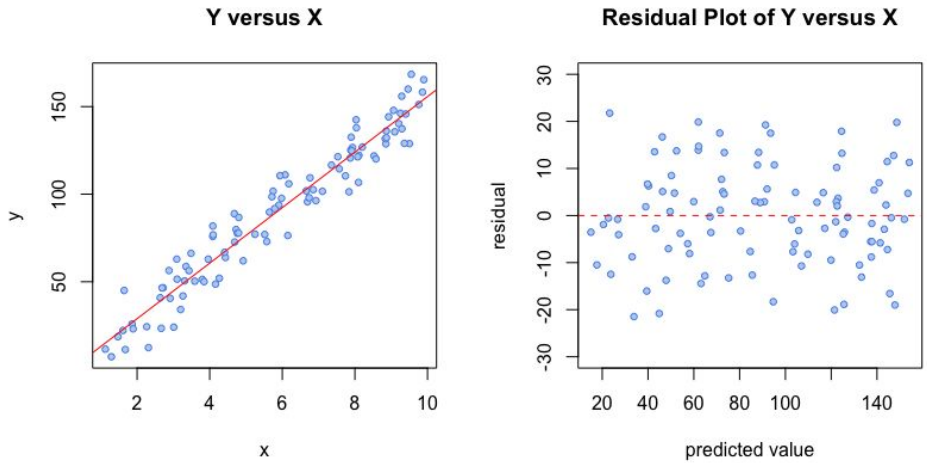
There are 4 assumptions that should be satisfied for

Least squares regression

Residuals in linear regression

The vertical distance between a point in the scatterplot and the predicted value on the regression line is the **residual** for the point.

For an observation , where is the predicted value according to the line , the residual is the value

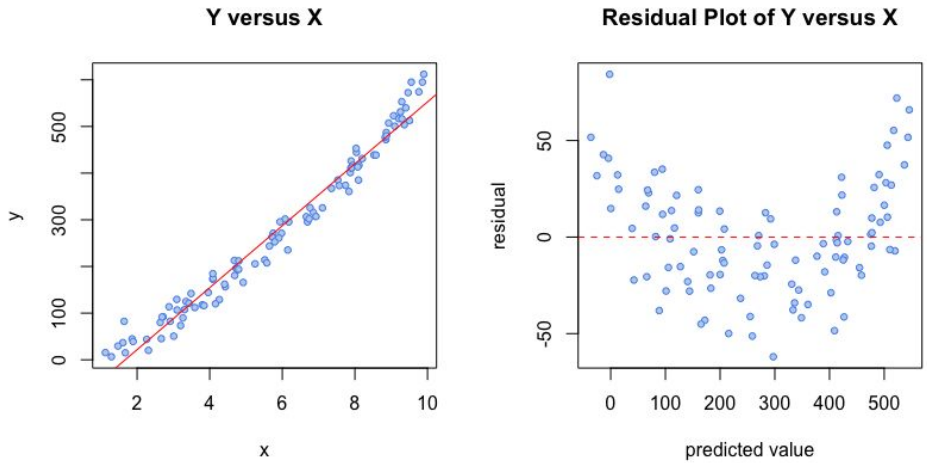


Residuals in linear regression...

Estimating a line using least squares

The least squares regression line is the line which minimizes the sum of the squared residuals for all the points in the plot.

Checking linearity and constant variability...



Checking normality of the residuals

Normality of the residuals is fundamental to the underlying model

since is assumed to be nomally distributed with mean 0 and standard deviation .

Normality of the residuals is checked using normal probability plots.

- These plots were used to check the normality

Interpreting a linear model

Categorical predictors with two levels

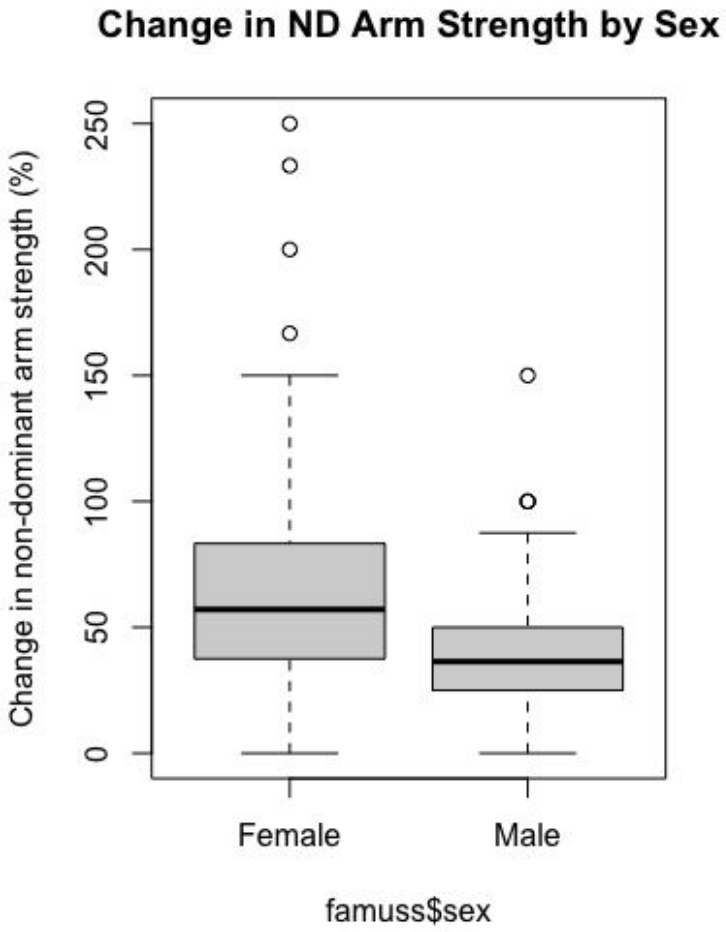
Categorical predictors with two levels

Although the response variable in linear regression is necessarily numerical, the predictor may be either numerical or categorical.

Simple linear regression only allows for categorical predictor variables with two levels.

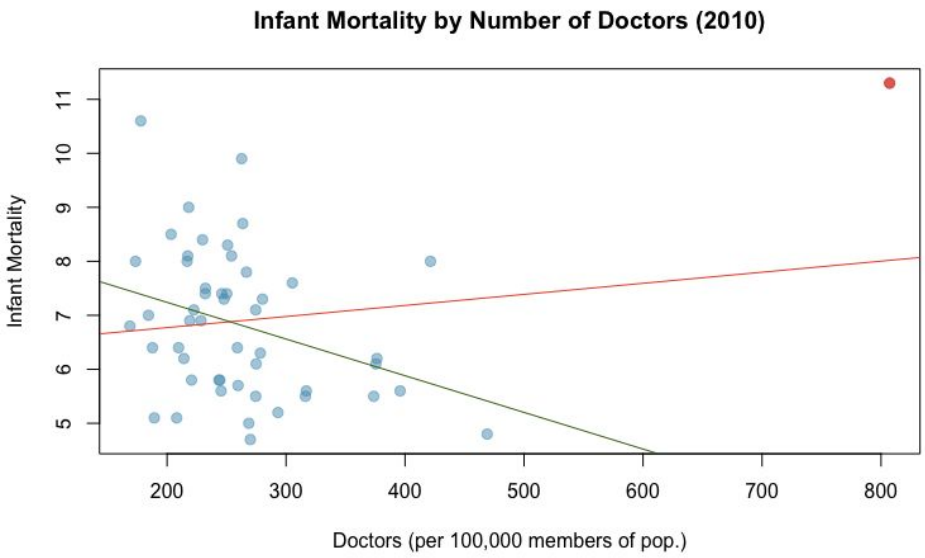
Examining categorical predictors with more than two levels requires multiple linear regression.

Fitting a simple linear regression model with a two-level categorical predictor is analogous



FAMuSS: comparing ndrm.ch by sex…

```
#calculate mean
ndrm.ch in each group
tapply(famuss$ndrm.ch,
famuss$sex, mean)

##   Female     Male

## 62.92720 39.23512

#fit a linear model of
ndrm.ch by sex
lm(famuss$ndrm.ch ~
famuss$sex)$coef

##    (Intercept)
famuss$sexMale

##       62.92720
-23.69207
```



The line is the model fit to all 51 observations.

The line is the model fit to 50 observations, excluding the red point.

Infant mortality and number of doctors...

```
#identify the influential point
census.2010$state[census.2010$doctors > 700]
```

```
## [1] "District of Columbia"
```

The point marked in red corresponds to the District

Statistical inference in regression

The model for statistical inference

The observed data are assumed to have been randomly sampled from a population where the explanatory variable and the response variable follow a population model

where .

Under this assumption, the slope and intercept of the regression line, and , are estimates of the population parameters and .

Hypothesis testing in regression

MULTIPLE LINEAR REGRESSION

The main ideas

In most practical settings, more than one explanatory variable is likely to be associated with a response.

Multiple linear regression is used to estimate the linear relationship between a response variable and several predictors , where is the number of predictors.

The statistical model for multiple linear regression is based on

The main ideas

Fitting and interpreting a multiple regression model

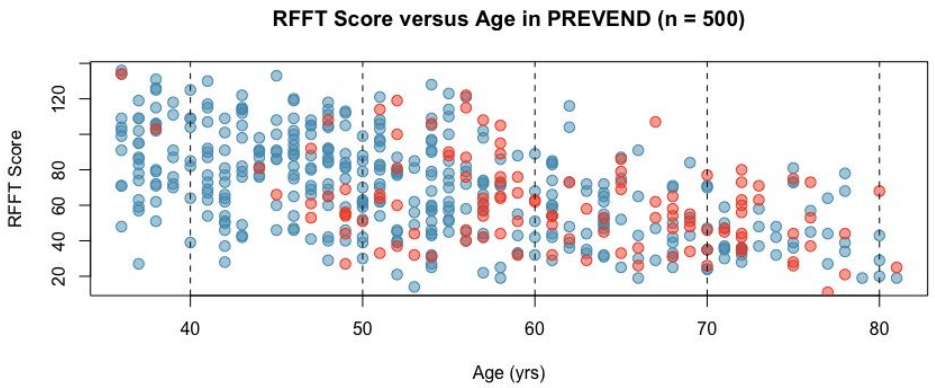
Statin use and cognitive function

Statins are a class of drugs widely used to lower cholesterol.

If followed, recent guidelines for prescribing statins would lead to statin use in almost half of Americans between 40 - 75 years of age and nearly all men over 60.

A few small studies have suggested that statins may be associated with lower cognitive ability.

The PREVEND study collected data on statin use as well as other demographic factors.

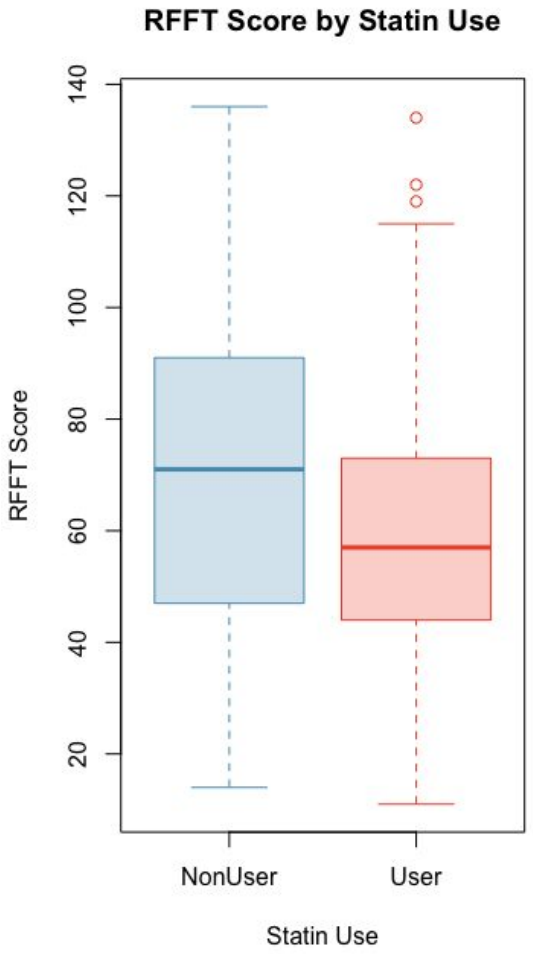


Age, statin use, and RFFT score

Red dots represent statin users; blue dots represent non-users.

Lab 1 examines the association between cognitive function and statin use after adjusting for age as a potential confounder.

RFFT score vs. statin use



RFFT score vs. statin use...

```
#fit the linear model
lm(RFFT ~ Statin,
data=prevend.samp)
##
## Call:
## lm(formula = RFFT ~ Statin, data
= prevend.samp)
##
## Coefficients:
## (Intercept)   StatinUser
```

Evaluating a multiple regression model

Assumptions for multiple regression

Similar to those of simple linear regression…

1. Linearity: For each predictor variable , change in the predictor is linearly related to change in the response variable when the value of all other predictors is held constant.
2. Constant variability: The residuals have approximately constant variance.
3. Independent observations: Each set of observations is independent.
4. Normality of residuals: The residuals are

Categorical predictors with several levels

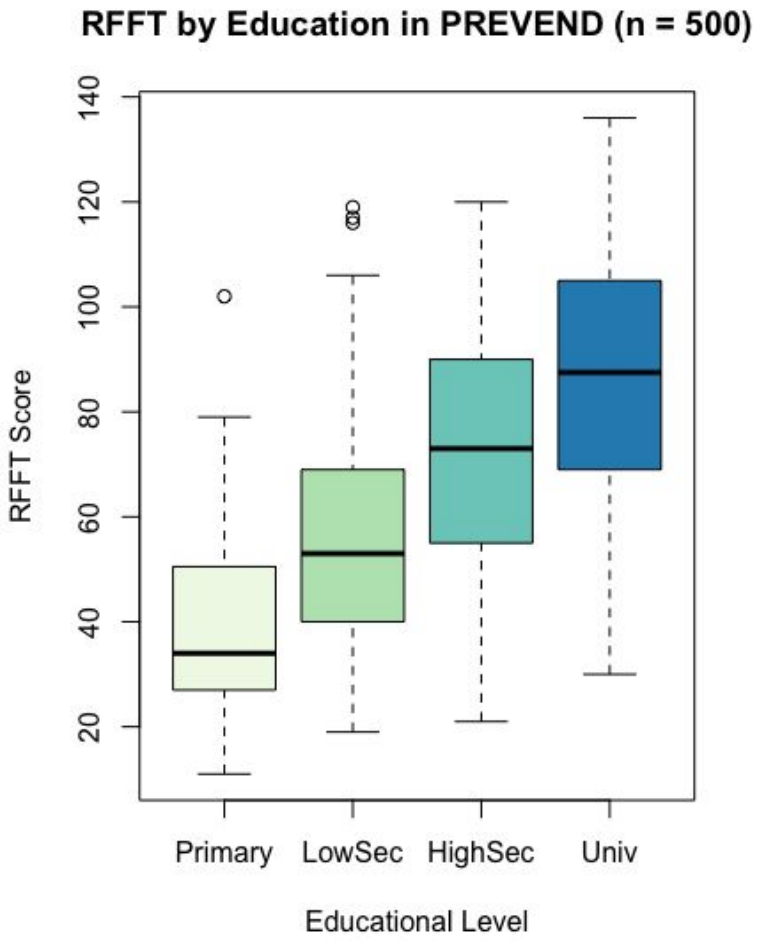
Categorical predictor with two levels

In the setting of a binary predictor, a linear regression estimates the difference in the mean response between the groups defined by the levels.

For example, the equation for the linear model predicting RFFT score from statin use based on the data in is

Mean RFFT score for individuals not using statins is 70.71.

Mean RFFT score for individuals using statins is 10.05 points lower than non-users, .



RFFT vs. Education

```
#fit a model
lm(RFFT ~ Education, data =
prevend.samp)$coef
##             (Intercept)
EducationLowerSecond
EducationHigherSecond
##                40.94118
  14.77857                  32.13345
##          EducationUniv
##                44.96389
```

Inference for the multiple regression model

The model for statistical inference

The coefficients of a multiple regression model are estimates of the population parameters in the model

where .

Inference is usually done about the slope parameters:

Testing hypotheses about a slope coefficient

Typically, the hypotheses of interest are

- , the variables and are not associated

Interaction in regression

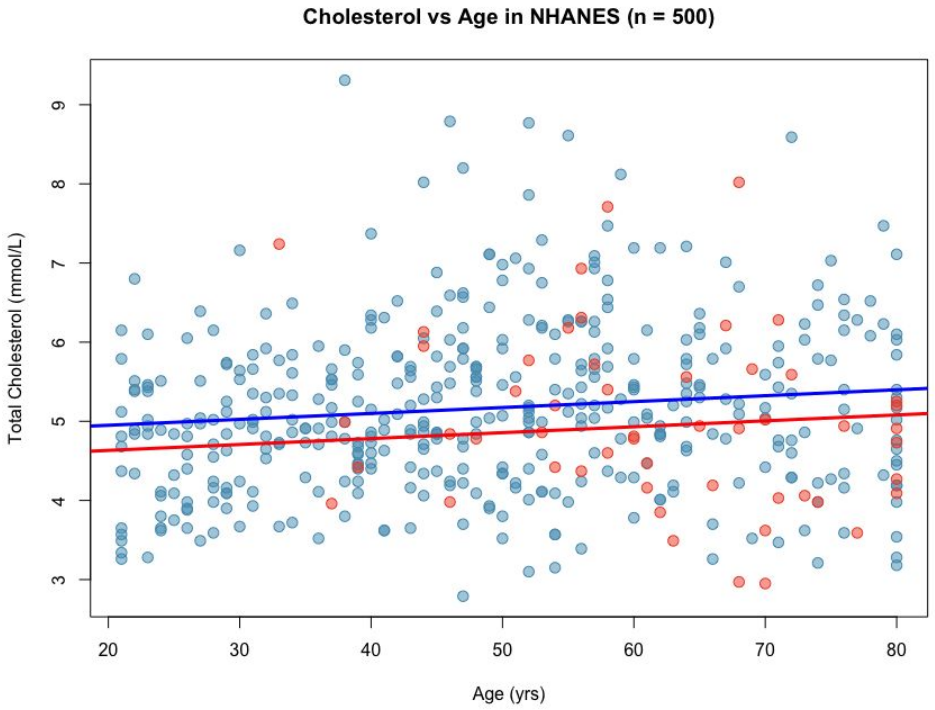
An important assumption

The multiple regression model

assumes that when one of the predictors is changed by 1 unit and the values of the other variables remain constant, the predicted response changes by , *regardless of the values of the other variables*.

A statistical **interaction** occurs when this assumption is not true, such that the effect of one explanatory variable on the response depends on the particular value(s) of one or more other explanatory variables.

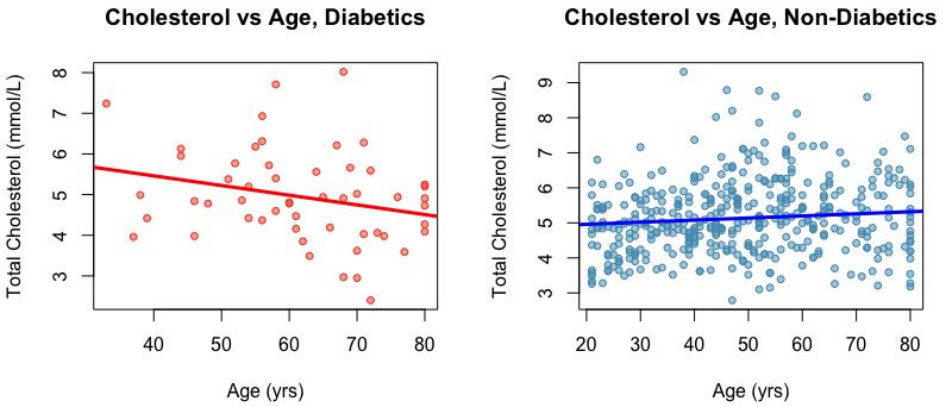
In this course, we specifically examine interaction in a two-



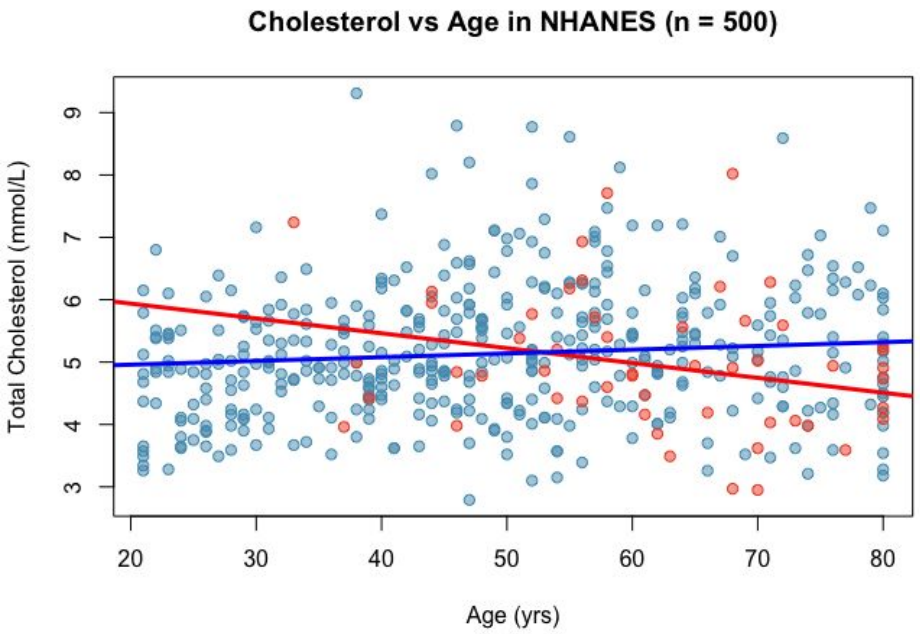
The model equation for non-diabetics is ; the one for diabetics is .

Cholesterol vs. Age and Diabetes...

Suppose two separate models were fit for the relationship between total cholesterol and age; one in diabetic individuals and one in non-diabetic individuals.



Cholesterol vs. Age and Diabetes…



Adding an interaction term

Consider the model

The term is the interaction term between diabetes status and age, and is the coefficient of the interaction term.

```
#fit the model
model.interact = lm(TotChol ~
Age*Diabetes, data =
nhanes.samp.adult.500)
coef(model.interact)
```

Model selection for explanatory models

Explanatory models

In explanatory modeling, the goal is to a construct a model that explains the observed variation in the response variable.

- Typically desirable to have a *parsimonious model*, a model which explains variation in the response using as few predictors as possible

This course discusses model selection in the context of a small set of potential predictors.

There exist purely algorithmic methods that screen a large set of predictors and choose a final model by optimizing a numerical criterion.