

Unit 7: Multiple Linear Regression

Statistics 102 Teaching Team

April 08, 2020

Introduction

Fitting and interpreting a multiple regression model

Evaluating a multiple regression model

Categorical predictors with several levels

Inference for the multiple regression model

Interaction in regression

Model selection for explanatory models

Introduction

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 y and several predictors x_1, x_2, \dots, x_p , where p is the number of predictors.

The statistical model for multiple linear regression is based on

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p.$$

THE MAIN IDEAS. . .

There are several applications of multiple regression.

The two discussed in this course are:

- Estimating an association between a response variable and primary predictor of interest while adjusting for possible confounding variables
- Constructing a model that effectively explains the observed variation in the response variable

Fitting and interpreting a multiple regression model

STATIN USE AND COGNITIVE FUNCTION

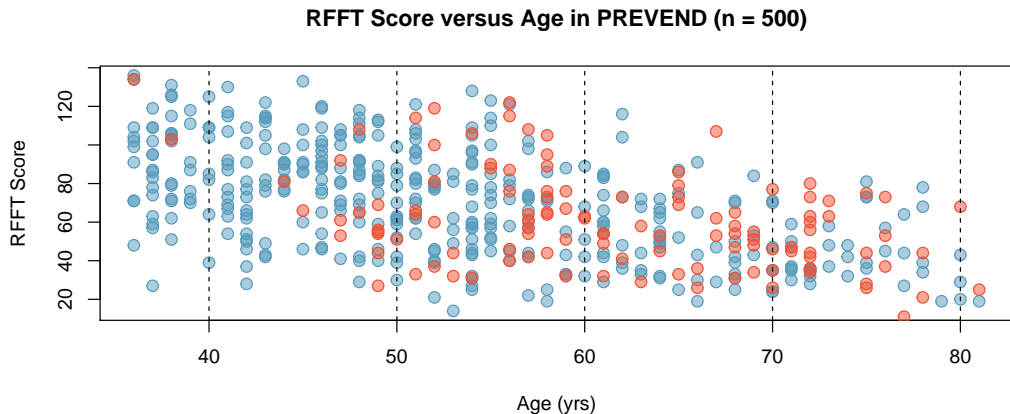
Statins are a class of drugs widely used to lower cholesterol.

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.

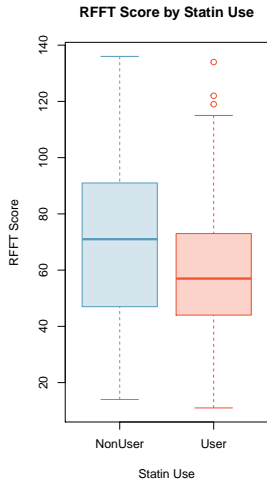
AGE, STATIN USE, AND RFFT SCORE



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  
##          70.71         -10.05
```

INTERPRETATION OF COEFFICIENTS

The statistical model for multiple regression is based on

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p,$$

where p is the number of predictors.

- The coefficient b_j of a predictor x_j is the predicted mean change in y corresponding to a one unit change in x_j , when the values of all other predictors remain constant.

The practical interpretation is that a coefficient in multiple regression estimates the association between a response and that predictor, *after adjusting for the other predictors in the model*.

A specific example helps. . .

RFFT VS. STATIN USE AND AGE

```
#fit the linear model
lm(RFFT ~ Statin + Age, data = prevend.samp)

##
## Call:
## lm(formula = RFFT ~ Statin + Age, data = prevend.samp)
##
## Coefficients:
## (Intercept)  StatinUser      Age
##    137.8822      0.8509    -1.2710
```

RFFT VS. STATIN USE AND AGE...

```
#print the model summary for the coefficients  
summary(lm(RFFT ~ Statin + Age, data = prevend.samp))$coef
```

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	137.8822193	5.12208509	26.919158	4.031155e-99
## StatinUser	0.8508699	2.59570660	0.327799	7.432017e-01
## Age	-1.2709945	0.09430356	-13.477693	1.652922e-35

Evaluating a multiple regression model

ASSUMPTIONS FOR MULTIPLE REGRESSION

Similar to those of simple linear regression. . .

1. Linearity: For each predictor variable x_j , 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 $(y, x_1, x_2, \dots, x_p)$ is independent.
4. Normality of residuals: The residuals are approximately normally distributed.

USING RESIDUAL PLOTS

It is not possible to make a scatterplot of a response against several simultaneous predictors. Instead, use a modified residual plot to assess linearity:

- For each (numerical) predictor, plot the residuals on the y -axis and the predictor values on the x -axis.
- Patterns/curvature are indicative of non-linearity.

Constant variability and normality of residuals can be assessed using the same methods as for simple regression:

- Constant variability: plot the residual values on the y -axis and the predicted values on the x -axis
- Normality of residuals: use normal probability plots

R^2 WITH MULTIPLE REGRESSION

As in simple regression, R^2 represents the proportion of variability in the response variable explained by the model.

As variables are added, R^2 always increases.

In the `summary(lm())` output, Multiple R-squared is R^2 .

```
#extract  $R^2$  of a model  
summary(lm(RFFT ~ Statin + Age, data = prevend.samp))$r.squared
```

```
## [1] 0.2851629
```

ADJUSTED R^2 AS A TOOL FOR MODEL ASSESSMENT

$$R_{adj}^2 = 1 - \left(\frac{\text{Var}(e_i)}{\text{Var}(y_i)} \times \frac{n-1}{n-p-1} \right),$$

where n is the number of cases and p is the number of predictor variables.

Adjusted R^2 incorporates a penalty for including predictors that do not contribute much towards explaining observed variation in the response variable.

- It is often used to balance predictive ability with model complexity.
- Unlike R^2 , R_{adj}^2 does not have an inherent interpretation.

```
#extract adjusted R2 of a model  
summary(lm(RFFT ~ Statin + Age, data = prevend.samp))$adj.r.squared
```

```
## [1] 0.2822863
```

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 `prevend.samp` is

$$\widehat{RFFT} = 70.71 - 10.05(StatinUser)$$

- 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, $70.71 - 10.05 = 60.66$.

WHAT ABOUT MORE THAN TWO LEVELS?

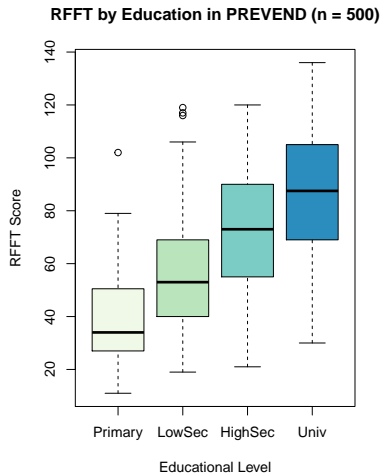
Is RFFT score associated with education?

The variable `Education` indicates the highest level of education an individual completed in the Dutch educational system:

- 0: primary school
- 1: lower secondary school
- 2: higher secondary education
- 3: university education

```
#convert Education to a factor
prevend.samp$Education = factor(prevend.samp$Education,
                                levels = c(0, 1, 2, 3),
                                labels = c("Primary", "LowerSecond",
                                           "HigherSecond", "Univ"))
```

RFFT vs. EDUCATION



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
```

```
#calculate group means
```

```
tapply(prevend.samp$RFFT, prevend.samp$Education, mean)
```

```
##      Primary LowerSecond HigherSecond      Univ
## 40.94118    55.71975    73.07463    85.90506
```

INTERPRETATION OF THE MODEL WITH EDUCATION

The baseline category represents individuals who at most completed primary school (`Education` = 0).

- Intercept is the sample mean RFFT score for these individuals, 40.94 points

The coefficients represent the change in estimated average RFFT relative to the baseline category.

- An increase of 14.78 points is predicted for `LowerSecond`,

$$40.94 + 14.78 = 55.72 \text{ points}$$

- An increase of 32.13 points is predicted for `HigherSecond`,

$$40.94 + 32.13 = 73.07 \text{ points}$$

- An increase of 44.96 points is predicted for `Univ`,

$$40.94 + 44.96 = 85.90 \text{ points}$$

Inference for the multiple regression model

THE MODEL FOR STATISTICAL INFERENCE

The coefficients of a multiple regression model b_0, b_1, \dots, b_p are estimates of the population parameters in the model

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon,$$

where $\epsilon \sim N(0, \sigma)$.

Inference is usually done about the slope parameters: $\beta_1, \beta_2, \dots, \beta_p$

TESTING HYPOTHESES ABOUT A SLOPE COEFFICIENT

Typically, the hypotheses of interest are

- $H_0 : \beta_k = 0$, the variables X_k and Y are not associated
- $H_A : \beta_k \neq 0$, the variables X_k and Y are associated

The t -statistic has $df = n - p - 1$, where n is the number of cases and p is the number of predictors in the model.

$$t = \frac{b_k - \beta_k^0}{\text{s.e.}(b_k)} = \frac{b_k}{\text{s.e.}(b_k)}$$

A 95% confidence interval for the slope β_k is given by

$$b_k \pm (t^* \times \text{s.e.}(b_k)),$$

where t^* is the point on a t -distribution with $n - p - 1$ degrees of freedom and $\alpha/2$ area to the right.

THE F -STATISTIC IN REGRESSION

The F -statistic is used in an overall test of the model to assess whether the predictors in the model, considered as a group, are associated with the response.

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
- H_A : At least one of the slope coefficients is not 0

There is sufficient evidence to reject H_0 if the p -value of the F -statistic is smaller than or equal to α .¹

The F -statistic and its associated p -value are displayed in the output from `summary(lm())`.

¹The formulas associated with calculating the F -statistic are discussed in Section 7.4.2 of *OI Biostat*.

INFERENCE: RFFT VS EDUCATION

```
#model summary
summary(lm(RFFT ~ Education, data = prevend.samp))

##
## Call:
## lm(formula = RFFT ~ Education, data = prevend.samp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -55.905 -15.975  -0.905  16.068  63.280
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      40.941      3.203  12.783 < 2e-16 ***
## EducationLowerSecond  14.779      3.686   4.009 7.04e-05 ***
## EducationHigherSecond 32.133      3.763   8.539 < 2e-16 ***
## EducationUniv       44.964      3.684  12.207 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.87 on 496 degrees of freedom
## Multiple R-squared:  0.3072, Adjusted R-squared:  0.303
## F-statistic: 73.3 on 3 and 496 DF, p-value: < 2.2e-16
```

INFERENCE: RFFT VS EDUCATION...

```
#connection to ANOVA
```

```
summary(aov(prevend.samp$RFFT ~ prevend.samp$Education))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## prevend.samp$Education    3 115041    38347    73.3 <2e-16 ***
## Residuals                496 259469      523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pairwise.t.test(prevend.samp$RFFT, prevend.samp$Education,
                 p.adj = "none")
```

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data:  prevend.samp$RFFT and prevend.samp$Education
##
##           Primary LowerSecond HigherSecond
## LowerSecond 7.0e-05 -                -
## HigherSecond < 2e-16 2.6e-10          -
## Univ         < 2e-16 < 2e-16        2.3e-06
##
## P value adjustment method: none
```

REANALYZING THE PREVENT DATA

```
##
## Call:
## lm(formula = RFFT ~ Statin + Age + Education + CVD, data = prevent.samp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -56.348 -15.586  -0.136  13.795  63.935
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    99.03507     6.33012   15.645 < 2e-16 ***
## StatinUser       4.69045     2.44802    1.916  0.05594 .
## Age            -0.92029     0.09041  -10.179 < 2e-16 ***
## EducationLowerSecond 10.08831     3.37556    2.989  0.00294 **
## EducationHigherSecond 21.30146     3.57768    5.954 4.98e-09 ***
## EducationUniv    33.12464     3.54710    9.339 < 2e-16 ***
## CVDPresent      -7.56655     3.65164   -2.072  0.03878 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.71 on 493 degrees of freedom
## Multiple R-squared:  0.4355, Adjusted R-squared:  0.4286
## F-statistic: 63.38 on 6 and 493 DF,  p-value: < 2.2e-16
```

Interaction in regression

AN IMPORTANT ASSUMPTION

The multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$

assumes that when one of the predictors x_j is changed by 1 unit and the values of the other variables remain constant, the predicted response changes by β_j , *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 x_j 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-variable setting, where one of the predictors is categorical and the other is numerical.

CHOLESTEROL VS. AGE AND DIABETES

Consider a linear model that predicts total cholesterol level (mmol/L) from age (yrs.) and diabetes status.

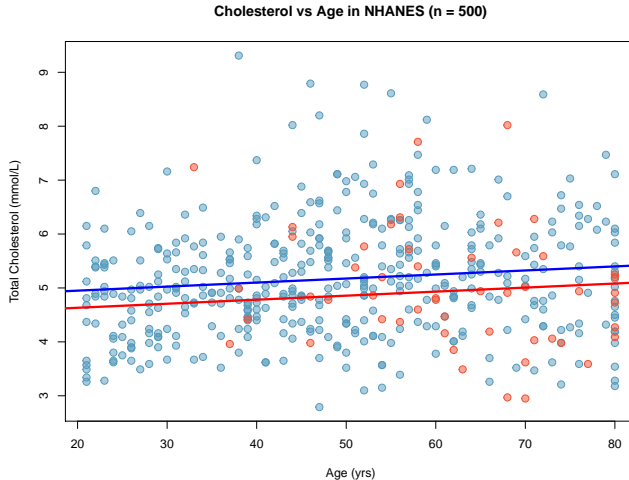
$$E(TotChol) = \beta_0 + \beta_1(Age) + \beta_2(Diabetes)$$

```
#load the data
library(oibistat)
data("nhanes.samp.adult.500")

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

```
## (Intercept)      Age  DiabetesYes
## 4.800011340  0.007491805 -0.317665963
```

CHOLESTEROL VS. AGE AND DIABETES...

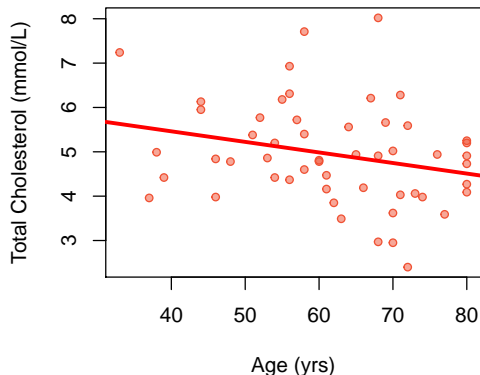


The model equation for non-diabetics is blue; the one for diabetics is red.

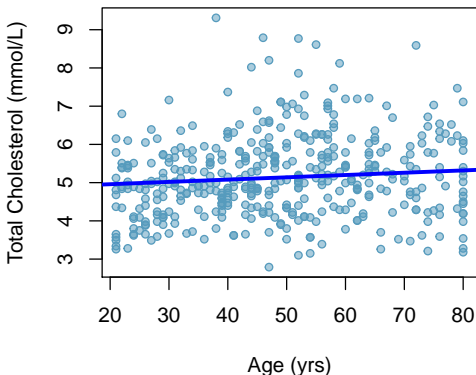
CHOLESTEROL VS. AGE AND DIABETES...

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

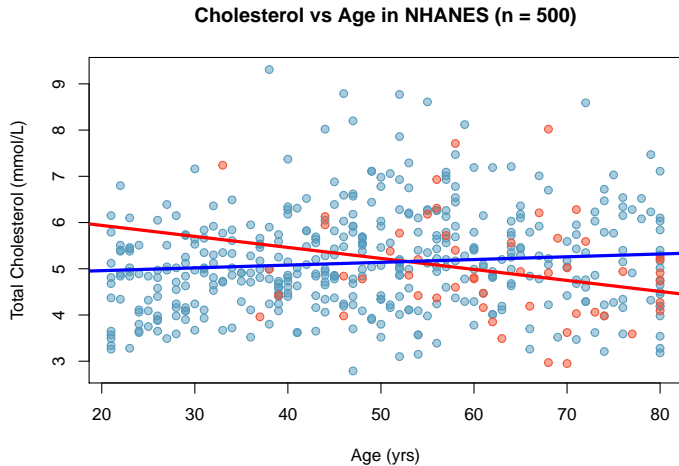
Cholesterol vs Age, Diabetics



Cholesterol vs Age, Non-Diabetics



CHOLESTEROL VS. AGE AND DIABETES...



ADDING AN INTERACTION TERM

Consider the model

$$E(TotChol) = \beta_0 + \beta_1(Age) + \beta_2(Diabetes) + \beta_3(Diabetes \times Age).$$

The term $(Diabetes \times Age)$ is the interaction term between diabetes status and age, and β_3 is the coefficient of the interaction term.

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

```
##      (Intercept)           Age      DiabetesYes Age:DiabetesYes
##      4.695702513      0.009638183      1.718704342      -0.033451562
```

ADDING AN INTERACTION TERM...

$$\widehat{TotChol} = 4.70 + 0.0096(Age) + 0.1.72(DiabetesYes) - 0.033(Age \times DiabetesYes)$$

For diabetics ($DiabetesYes = 1$), the model equation is

$$\begin{aligned}\widehat{TotChol} &= 4.70 + 0.0096(Age) + 1.72(1) - 0.034(Age \times 1) \\ &= 4.70 + 1.72 + (0.0096 - 0.034)(Age) \\ &= 6.42 - 0.024(Age)\end{aligned}$$

For non-diabetics ($DiabeticsYes = 0$), the model equation is

$$\begin{aligned}\widehat{TotChol} &= 4.70 + 0.0096(Age) + 1.72(0) - 0.034(Age \times 0) \\ &= 4.70 + 0.0096(Age)\end{aligned}$$

SIGNIFICANCE OF AN INTERACTION TERM

```
summary(model.interact)
```

```
##
## Call:
## lm(formula = TotChol ~ Age * Diabetes, data = nhanes.samp.adult.500)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3587 -0.7448 -0.0845  0.6307  4.2480
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.695703   0.159691  29.405 < 2e-16 ***
## Age           0.009638   0.003108   3.101  0.00205 **
## DiabetesYes   1.718704   0.763905   2.250  0.02492 *
## Age:DiabetesYes -0.033452   0.012272  -2.726  0.00665 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.061 on 469 degrees of freedom
## (27 observations deleted due to missingness)
## Multiple R-squared:  0.03229,    Adjusted R-squared:  0.0261
## F-statistic: 5.216 on 3 and 469 DF,  p-value: 0.001498
```


Model selection for explanatory models

EXPLANATORY MODELS

In explanatory modeling, the goal is to 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.

- These methods are not discussed in this course.

STEPS FOR MODEL SELECTION

1. *Data exploration.* Examine both the distributions of individual variables and the relationships between variables.
2. *Initial model fitting.* Fit an initial model with the predictors that seem most highly associated with the response.
3. *Model comparison.* Work towards a model with highest adjusted R^2 .
 - Fit new models without predictors that were not statistically significant or only marginally significant and compare R^2_{adj} .
 - Examine whether interaction terms may improve R^2_{adj} .
4. *Model assessment.* Assess the fit of the final model with residual plots.

CASE STUDY: HABITAT FRAGMENTATION

The process of model selection will be illustrated with a case study in which a regression model is built to examine the association between the abundance of forest birds in a habitat patch and features of a patch.

Lab 5 steps through the model selection process.