Lecture 4
Multiple linear regression

BIOST 515

January 15, 2004

Outline

- Motivation for the multiple regression model
- Multiple regression in matrix notation
- Least squares estimation of model parameters
- Maximum likelihood estimation of model parameters
- Hypothesis testing

Multiple linear regression

We are now considering the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \tag{1}$$

$i = 1, \ldots, n$, $E[\epsilon_i] = 0$, var$(\epsilon_i) = \sigma^2$ and $cov(\epsilon_i, \epsilon_j) = 0$.
We will also require that $p < n$. Additional assumption:

- The predictors $(x_1, \ldots, x_p)$ are fixed and measured without error

Why multiple linear regression?

Previously we've examined the case with one predictor and one outcome (simple linear regression). There are a variety of reasons we may want to include additional predictors in the model.

- Scientific question
- Adjustment for confounding
- Gain precision

Scientific Question

May dictate inclusion of particular predictors

- Predictors of interest
  - ⋆ The scientific factor under investigation can be modeled by multiple predictors (eg - dummy variables, polynomials, etc.)
- Effect modifiers
  - ⋆ The scientific question may relate to detection of effect modification
- Confounders
  - ⋆ The scientific question may have been stated in terms of adjusting for known (or suspected) confounders

Confounding

Sometimes the scientific question of greatest interest is confounded by associations in the data.

From KKMN, pg. 187:
In general, confounding exists if meaningfully different interpretations of the relationship of interest result when an extraneous variable is ignored or included in the analysis.

Precision

Adjusting for an additional covariate changes the standard error of the slope estimate

- Standard error is decreased by having smaller within group variance
- Standard error is increased by having correlations between the predictor of interest and other covariates in the model

General comments on multiple regression

- Can be difficult to choose the "best" model, since many reasonable candidates may exist
- More difficult to visualize the fitted model
- More difficult to interpret the fitted model

The model in matrix notation

$$y = X\beta + \epsilon$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Least Squares Estimates

$$S(\beta) = (y - X\beta)'(y - X\beta)$$

Least squares estimates are obtained by solving

$$\frac{\partial S(\beta)}{\partial \beta} = -2X'y + 2X'X\beta = 0$$

for β. This gives us the *least squares normal equations*:

$$X'X\hat{\beta} = X'y.$$

To get the least squares estimator of β multiply each side by $(X'X)^{-1}$ which gives

$$\hat{\beta} = (X'X)^{-1}X'y$$

provided $(X'X)^{-1}$ exists (which it will if the regressors are linearly independent).

The vector of fitted y-values , $\hat{y}$, corresponding to the observed y-values y is

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy.$$

The $n \times n$ matrix H is often referred to as the hat matrix.

The residuals,

$$e = \hat{y} - y,$$

can also be rewritten as

$$e = y - X\hat{\beta} = y - Hy = (I - H)y.$$

Properties of least squares estimates

$\hat{\beta}$ is an unbiased estimator of β

$$E(\hat{\beta}) = E[(X'X)^{-1}X'y] = (X'X)^{-1}X'X\beta = \beta$$

The variance of $\hat{\beta}$ is expressed by the **covariance matrix**

$$\begin{aligned} \text{cov}(\hat{\beta}) &= E\{[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]'\} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

If we let $C = (X'X)^{-1}$, the variance of $\hat{\beta}_j = \sigma^2 C_{jj}$ and the covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$ is $\sigma^2 C_{ij}$.

Estimation of σ^2

As in simple linear regression, we develop an estimator of σ^2 from SSE (residual sum of squares).

$$SSE = (y - X\hat{\beta})'(y - X\hat{\beta})$$

$$\hat{\sigma^2} = MSE = \frac{SSE}{n - p - 1}$$

As in simple linear regression $\hat{\sigma^2}$ is an unbiased estimator of σ^2.

Example

Cardiovascular Health Study (CHS)

A population-based, longitudinal study of coronary heart disease in people over 65.

Primary aim: To identify risk factors related to the onset and course of coronary heart disease and stroke.
Secondary aim: Describe prevalence and distributions of risk factors.

Scientific question

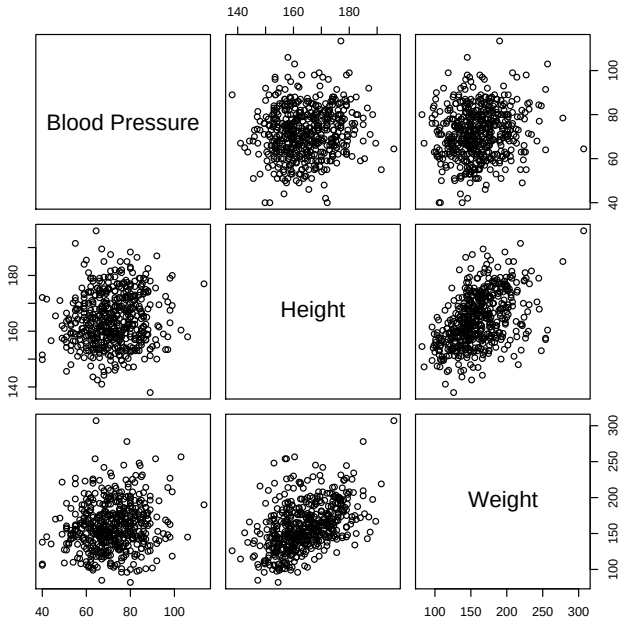
How is a person's weight related to blood pressure?

We might expect people who are more overweight to have higher blood pressure. However, a simple linear regression model with weight as a predictor might be misleading. Why?

We will examine a subset of 500 of these subjects to see how height and weight are related to blood pressure.

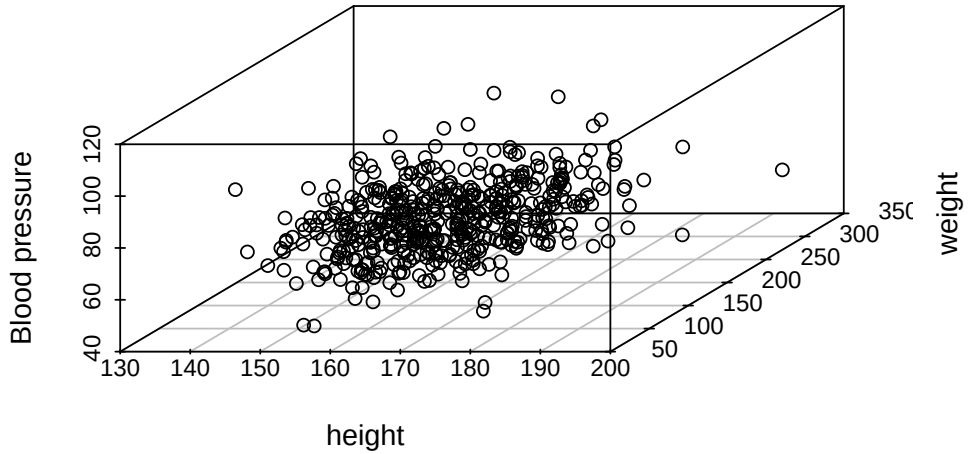
Scatterplot matrix

```
pairs(cbind(chs$DIABP,chs$HEIGHT,chs$WEIGHT),
      labels=c("Blood Pressure", "Height", "Weight"))
```



3d scatterplot

```
library(scatterplot)
scatterplot3d(chs$HEIGHT,chs$WEIGHT,chs$DIABP, xlab="height", ylab="weight",
              zlab="Blood pressure")
```



Regression models we may be interested in:

$$E[BP_i] = \beta_0 + \beta_1 height_i$$

$$E[BP_i] = \beta_0 + \beta_1 weight_i$$

$$E[BP_i] = \beta_0 + \beta_1 height_i + \beta_2 weight_i$$

We've fit models similar to the first 2, but not the 3rd.

Note: 2 observations will be omitted from the analysis because the subjects' weights are missing.

$$BP_i = \beta_0 + \beta_1 height_i + \epsilon_i$$

```
>lmht <- lm(DIABP~HEIGHT,data=chs)
>summary(lmht)

Residuals:
     Min       1Q   Median       3Q      Max
-32.8206  -7.6509  -0.0482   7.4237  40.0141

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 51.56051    8.79011   5.866 8.18e-09 ***
HEIGHT       0.12353    0.05343   2.312   0.0212 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.28 on 496 degrees of freedom
Multiple R-Squared: 0.01066,    Adjusted R-squared: 0.00867
F-statistic: 5.347 on 1 and 496 DF,  p-value: 0.02117
```

$$BP_i = \beta_0 + \beta_1 weight_i + \epsilon_i$$

```
lmwt <- lm(DIABP~WEIGHT,data=chs)
summary(lmwt)

Residuals:
     Min       1Q   Median       3Q      Max
-30.7561  -7.4422  -0.1446   7.3281  40.1285

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.99253    2.50881  25.507  < 2e-16 ***
WEIGHT       0.04905    0.01534   3.198  0.00147 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.23 on 496 degrees of freedom
Multiple R-Squared: 0.0202,     Adjusted R-squared: 0.01822
F-statistic: 10.23 on 1 and 496 DF,  p-value: 0.001474
```

$$BP_i = \beta_0 + \beta_1 height_i + \beta_2 weight_i + \epsilon_i$$

```
lmhtwt <- lm(DIABP~HEIGHT+WEIGHT,data=chs)
summary(lmhtwt)

Residuals:
     Min       1Q   Median       3Q      Max
-31.3833  -7.2260  -0.2881   7.7002  39.6144

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.65777    8.91267   6.245 9.14e-10 ***
HEIGHT       0.05820    0.05972   0.975   0.3302
WEIGHT       0.04140    0.01723   2.403   0.0166 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.23 on 495 degrees of freedom
Multiple R-Squared: 0.02208,    Adjusted R-squared: 0.01812
F-statistic: 5.587 on 2 and 495 DF,  p-value: 0.003987
```

Dummy (indicator) variables

So far we have only dealt with predictors that are continuous. However, we will often have variables that can take on only a small number of levels. Examples:

- Smoking status (current/former/never)
- Race
- Sex (Male/Female)
- Treatment group in a clinical trial

To model the association between these predictors and a response, we assign some sort of numerical scale to them. For example, we could define sex as (0/1), (-1,1), (1,2). Then if

$$y_i = \beta_0 + \beta_1 sex_i + \epsilon_i$$

and $sex = 1$ if female and 0 if male,

$$E(y_i) = \beta_0, \text{ if male}$$

and

$$E(y_i) = \beta_0 + \beta_1, \text{ if female.}$$

What does β_1 represent?

How is being a current smoker related to blood pressure?

$$BP_i = \beta_0 + \beta_1 smoker_i + \epsilon_i$$

```
>lmsmk <- lm(chs$DIABP~smoker)
>summary(lmsmk)

Residuals:
     Min       1Q   Median       3Q      Max
-32.1307  -8.0207  -0.1307   7.8693  41.3093

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.1307     0.5339 135.103   <2e-16 ***
smoker        -2.8344     1.7020  -1.665   0.0965 .

Residual standard error: 11.31 on 496 degrees of freedom
Multiple R-Squared: 0.00556,    Adjusted R-squared: 0.003555
F-statistic: 2.773 on 1 and 496 DF,  p-value: 0.09649
```

We might also want to know if current smoking status is an effect modifier for weight? In other words, is the relationship between blood pressure and weight different for people who smoke than for those who don't smoke?

$$BP_i = \beta_0 + \beta_1 weight_i + \beta_2 smoker_i + \beta_3 weight_i \times smoker_i + \epsilon_i$$

How do we interpret β_0, β_1, β_2, and β_3?

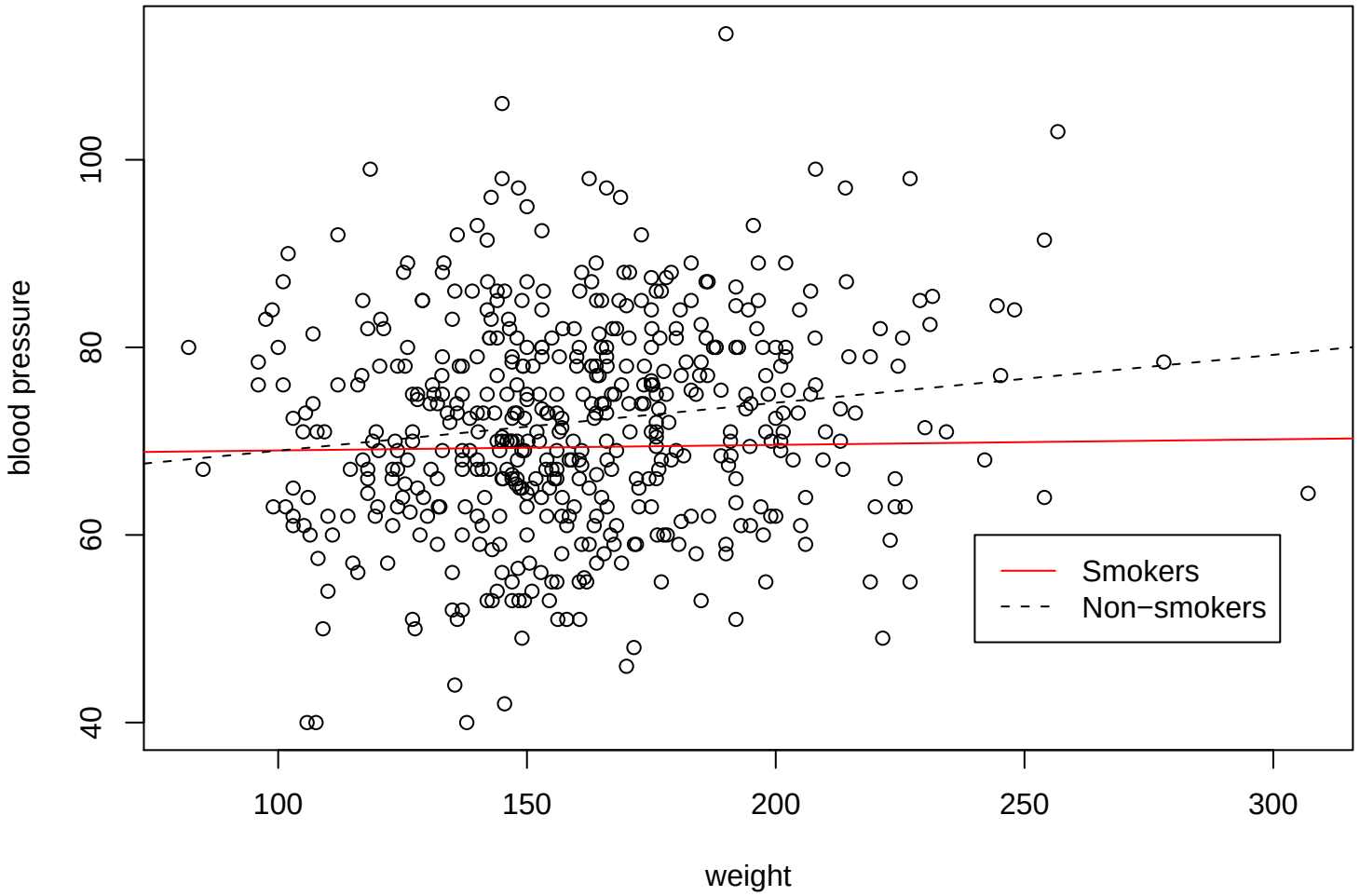
```
> summary(lm(chs$DIABP~smoker*chs$WEIGHT))

Call:
lm(formula = chs$DIABP ~ smoke * chs$WEIGHT)

Residuals:
       Min          1Q      Median          3Q         Max
-3.092e+01 -7.195e+00  4.102e-04  7.586e+00  3.986e+01

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      63.87963    2.67958  23.839  < 2e-16 ***
smoker             4.53954    7.94883   0.571  0.56819
chs$WEIGHT         0.05108    0.01626   3.141  0.00178 **
smoker:chs$WEIGHT -0.04517    0.05187  -0.871  0.38431

Residual standard error: 11.22 on 494 degrees of freedom
Multiple R-Squared: 0.02506,    Adjusted R-squared: 0.01914
```


blood pressure
40
60
80
100
100
150
200
250
300
weight
Smokers
Non−smokers

We know if a subject is a current/former/never smoker. How should we model this relationship with blood pressure?

Option 1:

$$smoke_i = \begin{cases} 1, & never\ smoked \\ 2, & former\ smoker \\ 3, & current\ smoker \end{cases}$$

$$BP_i = \beta_0 + \beta_1 smoke_i + \epsilon,$$

how do we interpret β_1?
$\hat{\beta}_1 = -1.08$ and $\hat{se}(\hat{\beta}_1) = 0.7637$.

Option 2:

Create two dummy variables

$$smoke_{1i} = \begin{cases} 1, & never\ smoked \\ 0, & otherwise \end{cases}$$

and

$$smoke_{2i} = \begin{cases} 1, & former\ smoker \\ 0, & otherwise \end{cases} .$$

$$BP_i = \beta_0 + \beta_1 smoke_{1i} + \beta_2 smoke_{2i} + \epsilon_i$$

How do we interpret β_0, β_1 and β_2?

```
> lmsmk2 <- lm(chs$DIABP~smoke1+smoke2)
> summary(lmsmk2)

Call:
lm(formula = chs$DIABP ~ smoke1 + smoke2)

Residuals:
     Min       1Q   Median       3Q      Max
-32.2824  -7.9008  -0.2824   7.7176  41.5198

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   69.296      1.618  42.839   <2e-16 ***
smoke1TRUE     2.986      1.763   1.694    0.091 .
smoke2TRUE     2.624      1.816   1.445    0.149
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.32 on 495 degrees of freedom
```

Summary

So far, we've discussed

- The motivation for multiple linear regression models
- Matrix notation for multiple linear regression
- Least squares estimates for multiple linear regression
- Recoding factors into dummy variables
- Intrepretation of linear models

Next, we'll discuss hypothesis testing in multiple linear regression

- Overall tests : does the entire set of independent predictors contribute significanly to the prediction of y?
- Test for addition of a single variable: does the addition of one variable significantly improve the prediction of y over other independent predictors already present in the model?
- Test for addition of a group of variables: does the addtion of some group of variables improve the prediction of y over other independent predictors alreay present in the model?