

# Introduction

**Before:** **Simple** regression model: the relationship between a response variable  $y$  and **ONE** predictor variable  $x$ :

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (1)$$

**Now:** **Multiple** regression model: the relationship between a response variable  $y$  and **MORE THAN ONE** predictor variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{p-1} X_{p-1} + \epsilon \quad (2)$$

# Introduction

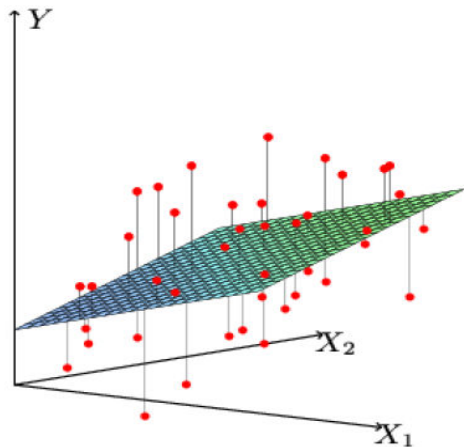
- ▶ **The good news!**– Everything we learned about the simple linear regression model extends (with minor modification) to multiple linear regression.
- ▶ **The even better news!**– We will learn new stuff! Because now we have a more complicated model. We need to consider not only the relation between predictors and response, but also relations among predictors.

## Example

Suppose that a researcher is studying factors that might affect blood pressures for women aged 45 to 65 years old. The  $Y$ -variable is blood pressure. Suppose that two predictor variables ( $X$ -variables) of interest are age and body mass index (calculated as  $weight/height^2$ ). The general structure of a linear multiple regression model for this situation would be

$$\text{Blood Pressure} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Body Mass} + \text{Error}$$

- ▶ The equation  $\beta_0 + \beta_1 \text{Age} + \beta_2 \text{Body Mass}$  describes the mean value of Blood Pressure for specific values of Age and Body Mass.
- ▶ The Error term describes the characteristics of the differences between individual values of blood pressure and the mean blood pressure =  $\beta_0 + \beta_1 \text{Age} + \beta_2 \text{Body Mass}$ .



# Notation for Multiple Regression Model

## Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{p-1} X_{p-1} + \epsilon.$$

or

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{i,p-1} + \epsilon_i.$$

**Assumption:** The linear model above is reasonable, and  $\epsilon_i$  are independent and follows  $N(0, \sigma^2)$  distribution.

**Note:** The subscript  $i$  refers to the  $i$ th individual or unit. In the notation for the  $x$ -variables, the subscript following  $i$  indicates which  $x$ -variable it is. e.g.  $x_{i2}$  represents the 2nd  $x$ -variable for subject  $i$ .

# Matrix Notation for Multiple Regression Model

## Goal:

represent the above regression models in terms of **matrices, vectors and their operations**.

## Reasons:

1. Neat, clean.
2. Statistical softwares basically realize all the calculations for estimations etc using matrix/vector representations.

# Matrix Notation for Multiple Regression Model

Consider the multiple regression models for all the individuals in the sample (size  $n$ ), with intercept and  $p - 1$  predictor variables:

$$y_1 = \beta_0 + \beta_1 x_{1,1} + \beta_2 x_{1,2} + \dots + \beta_{p-1} x_{1,p-1} + \epsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{2,1} + \beta_2 x_{2,2} + \dots + \beta_{p-1} x_{2,p-1} + \epsilon_2$$

...

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

...

$$y_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \dots + \beta_{p-1} x_{n,p-1} + \epsilon_n$$

# Matrix Notation for Multiple Regression Model

And the corresponding regression equations:

$$E(y_1) = \beta_0 + \beta_1 x_{1,1} + \beta_2 x_{1,2} + \dots + \beta_{p-1} x_{1,p-1}$$

$$E(y_2) = \beta_0 + \beta_1 x_{2,1} + \beta_2 x_{2,2} + \dots + \beta_{p-1} x_{2,p-1}$$

...

$$E(y_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_{p-1} x_{i,p-1}$$

...

$$E(y_n) = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \dots + \beta_{p-1} x_{n,p-1}$$



# Matrix Notation for Multiple Regression Model

Then, we have:

$$y_1 = E(y_1) + \epsilon_1$$

$$y_2 = E(y_2) + \epsilon_2$$

...

$$y_i = E(y_i) + \epsilon_i$$

...

$$y_n = E(y_n) + \epsilon_n$$

# Matrix Notation for Multiple Regression Model

We can denote all the observed response as a vector of length  $n$ , all the error terms as a vector of length  $n$ , all the mean responses as a vector of length  $n$ . Then, we have:

$$Y = E(Y) + \epsilon$$

$$\text{where } Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}, E(Y) = \begin{pmatrix} E(y_1) \\ E(y_2) \\ \ddots \\ E(y_i) \\ \ddots \\ E(y_n) \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{pmatrix}. \text{ Hence,}$$

we only need to focus on the matrix representation of  $E(Y)$ .

# Matrix Notation for Multiple Regression Model

Now we write  $E(Y)$  as a design matrix (or say  $X$  matrix)  $X$  postmultiplied by a vector of  $\beta$  coefficients. Notice that for each subject  $i$ ,

$$\begin{aligned}E(y_i) &= \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_{p-1} x_{i,p-1} \\&= 1 \cdot \beta_0 + x_{i,1} \cdot \beta_1 + x_{i,2} \cdot \beta_2 + \dots + x_{i,p-1} \cdot \beta_{p-1} \\&= (1 \ x_{i,1} \ x_{i,2} \dots x_{i,p-1}) \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix} .\end{aligned}$$

# Matrix Notation for Multiple Regression Model

$$E(y_1) = 1 \cdot \beta_0 + x_{1,1} \cdot \beta_1 + x_{1,2} \cdot \beta_2 + \dots + x_{1,p-1} \cdot \beta_{p-1}$$

$$\begin{pmatrix} E(y_1) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$$

# Matrix Notation for Multiple Regression Model

$$E(y_i) = 1 \cdot \beta_0 + x_{i,1} \cdot \beta_1 + x_{i,2} \cdot \beta_2 + \dots + x_{i,p-1} \cdot \beta_{p-1}; i = 1, 2$$

$$\begin{pmatrix} E(y_1) \\ E(y_2) \\ \vdots \\ \vdots \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p-1} \\ & \cdot & \cdot & & \cdot \\ & \cdot & \cdot & & \cdot \\ & \cdot & \cdot & & \cdot \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$$

# Matrix Notation for Multiple Regression Model

$$E(y_i) = 1 \cdot \beta_0 + x_{i,1} \cdot \beta_1 + x_{i,2} \cdot \beta_2 + \dots + x_{i,p-1} \cdot \beta_{p-1}; i = 1, 2, 3$$

$$\begin{pmatrix} E(y_1) \\ E(y_2) \\ E(y_3) \\ \vdots \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p-1} \\ 1 & x_{3,1} & x_{3,2} & \dots & x_{3,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$$

# Matrix Notation for Multiple Regression Model

$$E(y_i) = 1 \cdot \beta_0 + x_{i,1} \cdot \beta_1 + x_{i,2} \cdot \beta_2 + \dots + x_{i,p-1} \cdot \beta_{p-1}; i = 1, 2, 3, \dots, n$$

$$E(Y) = \begin{pmatrix} E(y_1) \\ E(y_2) \\ E(y_3) \\ \vdots \\ E(y_n) \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p-1} \\ 1 & x_{3,1} & x_{3,2} & \dots & x_{3,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p-1} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$$

# Matrix Notation for Multiple Regression Model

## [IMPORTANT]

Regression equation:  $E(Y) = X\beta$ ;

Regression model:  $Y = E(Y) + \epsilon$  or  $Y = X\beta + \epsilon$ , where

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p-1} \\ 1 & x_{3,1} & x_{3,2} & \dots & x_{3,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p-1} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$



# Matrix Notation for Multiple Regression Model

$X$  is called **design matrix**. Each row of  $X$  corresponds to one subject/unit, each column (except the 1st) corresponds to a predictor. The first column is the intercept. And, its form is the same as the way we input data in R.

# Matrix Notation for Multiple Regression Model: Example

Consider a toy example with a data set of size 4, we have response  $Y$  and two predictors  $X_1, X_2$  and consider fitting a multiple regression model. Based on the data given in the following table, how can we represent the models using matrix notations?

$Y$	12	17	15	11
$X_1$	3	5	4	2
$X_2$	1	1	2	2

In other words, for the matrix notation representation  $Y = X\beta + \epsilon$ , what are the exact forms of the design matrix and vectors based on the data?

**How to estimate the coefficients  $\beta_0, \beta_1, \dots, \beta_{p-1}$ , or in other words, the coefficient vector  $\beta$ ?**

# Estimates (Least Square Criterion)

- The estimates of the  $\beta$  coefficients minimize the **sum of squared distances** from the observation points to the regression hyperplane.

$$\min \sum_i (y_i - \beta_0 - \beta_1 x_{i,1} - \dots - \beta_{p-1} x_{i,p-1})^2$$

- The letter  $b$  is used to represent a sample estimate of a  $\beta$  coefficient.

$$(b_0, b_1, \dots, b_{p-1}) = \operatorname{argmin}_{\beta_0, \dots, \beta_{p-1}} \sum_i (y_i - \beta_0 - \dots - \beta_{p-1} x_{i,p-1})^2$$

# Interpretation

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ \vdots \\ b_{p-1} \end{pmatrix}.$$

- ▶  $b_0$ : The estimated mean  $y$  when **all  $x$  variables are 0**.
- ▶  $b_j, j = 1, \dots, p - 1$ : The estimated mean change in  $y$  when  $x_j$  increases one unit **when other  $x$ -variables remain the same**.

# Estimation (in matrix form)

- In matrix form:

$$\begin{aligned} \mathbf{b} \quad & \text{minimize} \quad \sum_i (y_i - \beta_0 - \beta_1 x_{i,1} - \dots - \beta_{p-1} x_{i,p-1})^2 \\ \Leftrightarrow \quad & \text{minimize} \quad (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

- Solution:

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \dots \\ \dots \\ b_{p-1} \end{pmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

# Sampling Distribution of $\mathbf{b}$

Recall: How do we get a sampling distribution.

Now

- Under normality assumption about random errors, we have:

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \sim \text{MVN}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$$

- Each  $b_i$  follows a normal distribution with mean  $\beta_i$  and variance the  $i$ th diagonal elements of  $(\mathbf{X}^T \mathbf{X})^{-1}$

# Estimation

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

- **Keep in mind:** Whenever some of the columns in design matrix  $\mathbf{X}$  are highly correlated (i.e. some of the predictor variables are linear related), there is no unique solution for inverse of  $\mathbf{X}^T \mathbf{X}$ . That is, in such a case,  $(\mathbf{X}^T \mathbf{X})^{-1}$  does not exist uniquely. This is the problem of **multicollinearity**.

Consequences of **high** multicollinearity

- Often confusing and misleading results: interpretation.
- Unstable estimates



# Model fitting

## ► Predicted/Fitted value:

$$\hat{y}_i = b_0 + b_1x_{i,1} + b_2x_{i,2} + \dots + b_{p-1}x_{i,p-1}, \quad i = 1, \dots, n;$$

In matrix notation:  $\hat{Y} = Xb$ .

## ► Residual:

$$e_i = y_i - \hat{y}_i, \quad i = 1, \dots, n;$$

In matrix notation:  $e = Y - \hat{Y}$ , where  $e = (e_1, \dots, e_n)^T$ .

## ► SSE and MSE:

$$SSE = \sum_i (y_i - b_0 - b_1x_{i,1} - \dots - b_{p-1}x_{i,p-1})^2$$

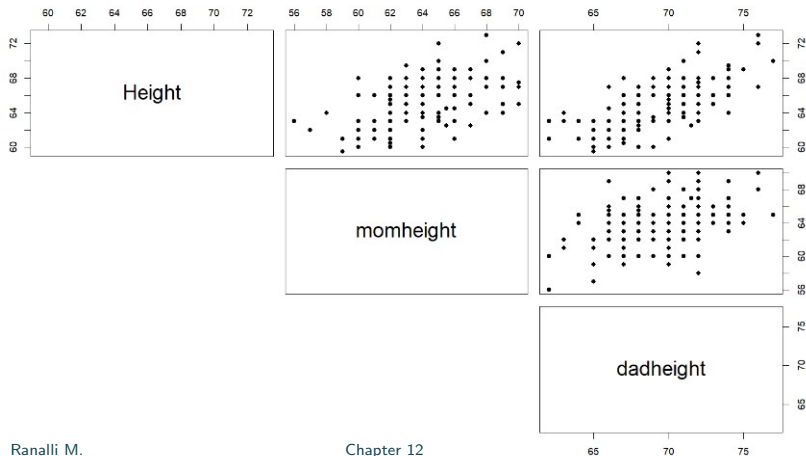
In matrix notation:  $SSE = (Y - Xb)^T(Y - Xb) = e^T e$ . And  $MSE = \frac{SSE}{n-p}$  estimates  $\sigma^2$ ,  $S = \sqrt{MSE}$  estimate  $\sigma$ .

# Validate Model Assumptions

- ▶ **Residual plot (versus the fit):**  
Still a 2-D plot (  $e_i$  versus  $\hat{y}_i$  )! Ideally should resemble a horizontal random band.
- ▶ **Normal probability plot:**  
Expect the normal probability plot to be a straight line and R-J test to have p-value larger than 0.05.

## R Example

The sample is from  $n = 124$  girls at University of California at Davis.  $y$  = student's self-reported height,  $x_1$  = her mother's height, and  $x_2$  = her father's height. All heights are in inches. The following are scatter plots of between each pair of variables.



# R Example

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.30733    4.69649   3.046  0.00284 **
momheight    0.26191    0.07133   3.672  0.00036 ***
dadheight    0.48875    0.06602   7.404  1.94e-11 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.002 on 121 degrees of freedom
Multiple R-squared:  0.5001,    Adjusted R-squared:  0.4918
F-statistic: 60.51 on 2 and 121 DF,  p-value: < 2.2e-16
```

- We can interpret the “slopes” in the same way as the simple linear model, but we have to add the constraint that values of other variables remain constant.
  1. **When father’s height is held constant**, the *average* student height increases 0.262 inches with one-inch increase in mother’s height.
  2. **When mother’s height is held constant**, the *average* student height increases 0.489 inches with one-inch increase in father’s height.

# R Example

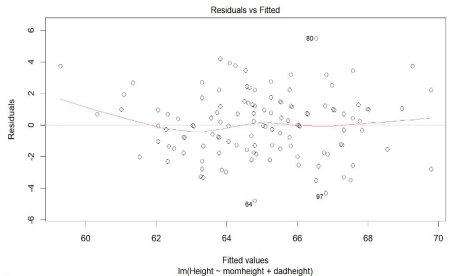
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.30733    4.69649   3.046  0.00284 **
momheight    0.26191    0.07133   3.672  0.00036 ***
dadheight    0.48875    0.06602   7.404  1.94e-11 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.002 on 121 degrees of freedom
Multiple R-squared:  0.5001,    Adjusted R-squared:  0.4918
F-statistic: 60.51 on 2 and 121 DF,  p-value: < 2.2e-16
```

- ▶ The sample regression equation is  
Fitted/Predicted student height =  
 $14.3 + 0.262 \times \text{Mother's height} + 0.489 \times \text{Father's height}$
- ▶ The value of  $R^2 = 50.0\%$  means that the model (the two x-variables) explains 50.0% of the observed variation in student heights (...but look at the Adjusted  $R^2$ ).
- ▶ The value  $S = 2.002$  is the estimated standard deviation of the errors. Roughly, it is the average absolute size of a residual.  
 $S = \sqrt{MSE}$ .

# Previous R Example

## Residual Plot

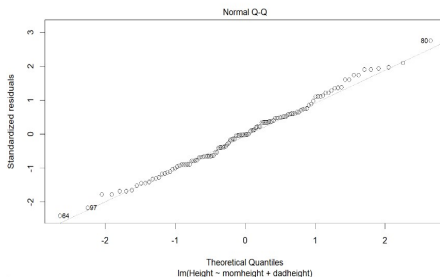


Just as in simple regression, we can use a plot of residuals versus fits to evaluate the validity of assumptions. The residual plot for these data is shown above. Roughly, it looks about as it should, although there may be a bit of increasing (vertical) variance as we move across.

# Previous R Example

## Normal Probability Plot of the residuals and Shapiro test

The straight line pattern indicates normality as does the  $p$ -value of the test ( $p - value = 0.82 > 0.05$ ). This means that it's reasonable to assume that the errors follow a normal distribution.



```
> shapiro.test(out$residuals)
```

Shapiro-wilk normality test

```
data: out$residuals  
W = 0.99327, p-value = 0.8196
```

# Inference: Testing Significance of Each $\beta$ Coefficient

We may want to assess whether a particular  $x$ -variable is making a useful contribution to the model. That is, **given the presence of the other  $x$ -variables in the model**, does a particular  $x$ -variable help us to explain more about the  $y$ -variable?

As an example, suppose that we have three  $x$ -variables in the model. The general structure of the model could be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$



# Inference: Testing Significance of Each $\beta$ Coefficient

To determine whether variable  $x_1$  is a useful predictor variable in this model, we could test

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

If the null hypothesis above were the case,  $y$  and  $x_1$  are not significantly related, or  $x_1$  is not important **when  $x_2$  and  $x_3$  are in the model.**

**NOTE:** If the null were true, we would still be left with variables  $x_2$  and  $x_3$  being present in the model. So when we cannot reject the null hypothesis above, we should say that do not need variable  $x_1$  in the model given that variables  $x_2$  and  $x_3$  will remain in the model.

# Inference: Testing Significance of Each $\beta$ Coefficient

Carry out the test:

$$T = \frac{\text{sample coefficient}}{\text{standard error of the coefficient}} \sim t(n - p) \text{ under } H_0.$$

1. **critical value approach:** if  $|T_0| > t_{\alpha/2}(n - p)$ , reject  $H_0$ .
2. **p-value approach:** if  $p\text{-value} = P(|T| > |T_0|) < \alpha$ , reject  $H_0$ .

where  $T_0$  is the observed value of  $T$  using the given sample.

# Previous R Example

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.30733    4.69649   3.046  0.00284 **
momheight    0.26191    0.07133   3.672  0.00036 ***
dadheight    0.48875    0.06602   7.404  1.94e-11 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.002 on 121 degrees of freedom
Multiple R-squared:  0.5001,    Adjusted R-squared:  0.4918
F-statistic: 60.51 on 2 and 121 DF,  p-value: < 2.2e-16
```

- The  $p$ -values given for the two  $x$ -variables tell us that student height is significantly related to both predictors.

## Previous R Example

Does father's height has more impact on the average heights of children than mother's height?

# Other inference tools

- ▶ Confidence interval for  $E(Y)$  given  $x = (x_1, x_2, \dots, x_{p-1})$
  
  
  
  
  
  
  
  
  
  
- ▶ Prediction interval for  $y$  given  $x_h = (x_{h,1}, x_{h,2}, \dots, x_{h,p-1})$

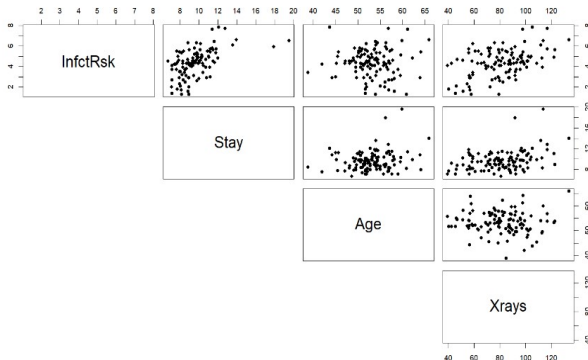
## R Example

Data from  $n = 113$  hospitals in the United States are used to assess factors related to the likelihood that a hospital patients acquires an infection while hospitalized. The variables here are  $y$  =infection risk,  $x_1$  =average length of patient stay,  $x_2$  =average patient age,  $x_3$  =measure of how many x-rays are given in the hospital.

**Note:** sample size  $n = 113$ , number of predictors= 3 (so, if we include all of them, the regression model will have 4  $\beta$  coefficients, i.e.  $p = 4$ ).

# R Example

Step 1. Check individual scatter plots of  $y$  versus  $x_i$ ,  $i = 1, 2, 3$ .



It seems that  $y$  has obvious linear relations with both  $x_1$  and  $x_3$  but mild linear relation with  $x_2$ . And relation between  $x_1$ ,  $x_2$ , and  $x_3$  are not significantly patterned.

# R Example

Step 2. Now, simply include all of the three predictor variables into the model and fit the multiple regression model.

```
Call:
lm(formula = InfctRsk ~ Stay + Age + Xrays, data = Senic)

Residuals:
    Min       1Q   Median       3Q      Max
-2.77320 -0.73779 -0.03345  0.73308  2.56331

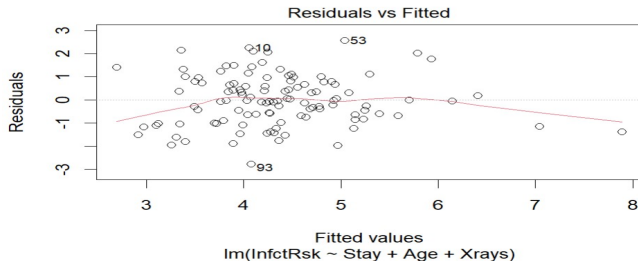
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.001162   1.314724   0.761 0.448003
Stay         0.308181   0.059396   5.189 9.88e-07 ***
Age        -0.023005   0.023516  -0.978 0.330098
Xrays       0.019661   0.005759   3.414 0.000899 ***
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.085 on 109 degrees of freedom
Multiple R-squared:  0.363,    Adjusted R-squared:  0.3455
F-statistic: 20.7 on 3 and 109 DF,  p-value: 1.087e-10
```



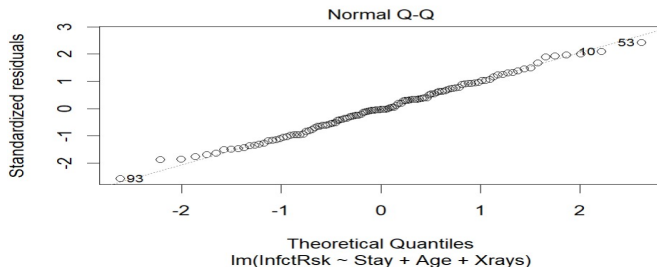
# R Example

Step 3. Check validity of assumptions for multiple regression model.



# R Example

Step 3. Check validity of assumptions for multiple regression model.



shapiro-wilk normality test

```
data: out$residuals  
w = 0.9939, p-value = 0.9037
```

# R Example

Step 4. Inference. The hypothesis testing for  $\beta_1$ :

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

- ▶  $T_0 = \frac{b_1}{s.e.(b_1)} = 5.189 \Rightarrow |T_0| > t_{0.025}(113 - 4) = 1.98$ , so reject  $H_0$ .
- ▶ p-value =  $Prob(|T| \geq 5.189) \approx 0$ , so reject  $H_0$ .

# R Example

Step 4. Inference. The hypothesis testing for  $\beta_2$ :

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

- ▶  $T_0 = \frac{b_2}{\text{s.e.}(b_2)} = -0.978 \Rightarrow |T_0| < t_{0.025}(113 - 4) = 1.98$ , so fail to reject  $H_0$ .
- ▶ p-value =  $\text{Prob}(|T| \geq 0.98) \approx 0.33$ , so fail to reject  $H_0$ .

Thus we cannot reject the null hypothesis  $H_0 : \beta_2 = 0$ , so “Age” is not a useful predictor **within this model** (i.e. **given the presence of the other two predictors**).

# R Example

Step 4. Inference. The hypothesis testing for  $\beta_3$ :

$$H_0 : \beta_3 = 0$$

$$H_1 : \beta_3 \neq 0$$

In a similar fashion, we can also obtain the p-value =  $0.001 < 0.05$ . Therefore, X-rays is useful for predicting  $y$  with the other two variables in the model.

**Note:** Usually, we don't worry about the p-value for “Constant”. It only refer to the “intercept” of the model and doesn't give us information about how changing an x-variables might change the mean of response  $y$ .

# Multiple Regression: Analysis of Variance (ANOVA) Table

**Recall:** In simple linear regression, we can obtain two types of information in a ANOVA table

- ▶ A decomposition of variance ( $SSTO=SSR+SSE$ )
- ▶ A significance test (F-test)

For multiple regression, it is very similar.

# 1. A decomposition of variance

ANOVA table displays quantities that measure how much of the variability in the  $y$ -variable is explained and how much is not explained by the  $y$ -variables relationship with the  $x$ -variables.

Recall (**Basic Idea**):

Overall variation in  $y$  = variation explained by regression + error variation

$$SST = SSR + SSE$$

# 1. A decomposition of variance

- ▶ **Sums of squares for total:**  $SST = \sum_{i=1}^n (y_i - \bar{y})^2$ .
  - ▶ Total degrees of freedom =  $n - 1$ .
  - ▶  $SST$  is a measure of the overall variation in the  $y$ -variables.
- ▶ **Sums of squared errors:**  $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ .
  - ▶ Error degrees of freedom =  $n - p$ .  
 $p = \# \beta$  coefficients in the model (including)  $\beta_0$ .
  - ▶  $MSE = \frac{SSE}{n-p}$  is the mean squared error.
- ▶ **Sums of squares due to Regression:**  $SSR = SST - SSE$ .
  - ▶ Regression degrees of freedom  
= total df – error df  
=  $(n - 1) - (n - p) = p - 1$ .
  - ▶  $MSR = \frac{SSR}{p-1}$  is the mean square for the regression.



# 1. A decomposition of variance

Source	DF	SS	MS	F
Regression	$p - 1$	SST-SSE	MSR	MSR/MSE
Error	$n - p$	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	MSE	
Total	$n - 1$	$\sum_{i=1}^n (y_i - \bar{y})^2$		

**The computation of the table is identical with the simple regression model.**

## 2. Significance test: F test

- The  $F$  statistic in the analysis of variance can be used to test whether the  $y$ -variable is related to at least one  $x$ -variables in the model. Specifically,

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$$

$$H_a : \text{at least one of the } \beta_i \neq 0, \text{ for } i = 1, \dots, p-1.$$

1. The null hypothesis means that the  **$y$ -variable is not related to any of the  $x$ -variables in the model.**

The alternative hypothesis means that the  **$y$ -variable is related to one or more of the  $x$ -variables in the model.**

## 2. Significance test: F test

- $F = MSR/MSE \sim F(p - 1, n - p)$  under  $H_0$ .  
If  $F_0 > F_\alpha(p - 1, n - p)$ , reject  $H_0$ .
- Statistical software reports a p-value for this test statistic:  
 $p\text{-value} = P(F > F_0)$ . Usually, if  $p\text{-value} \leq 0.05$ , reject the null hypothesis, and conclude that  $y$  is related to **at least one** of the  $x$ -variables in the model.
- Cautious! In multiple linear regression, the T-test and F-test are different!**  
T-test tests linear relation between  $y$  and a certain  $x_i$  while all other  $x$ -variables are in the model, while F-test test the linear relation between  $y$  and all  $x$ -variables together!

## Other Uses of ANOVA Table: $MSE$ and $R^2$

- ▶  $MSE$  is the estimate of the error variance. Thus  $S = \sqrt{MSE}$  estimates the standard deviation of the errors.
- ▶ The Total and Error lines give the  $SS$  values used in the calculation of  $R^2 = \frac{SST - SSE}{SST} = \frac{SSR}{SSE}$ .

# R Example

**Example** For a sample of individuals, we have measurements of  $y$  = body fat,  $x_1$  = triceps skinfold thickness,  $x_2$  = thigh circumference, and  $x_3$  = midarm circumference. Some results for a multiple regression with these variables are as follows:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	117.085	99.782	1.173	0.258
Triceps	4.334	3.016	1.437	0.170
Thigh	-2.857	2.582	-1.106	0.285
Midarm	-2.186	1.595	-1.370	0.190

Residual standard error: 2.48 on 16 degrees of freedom

Multiple R-squared: 0.8014, Adjusted R-squared: 0.7641

F-statistic: 21.52 on 3 and 16 DF, p-value: 7.343e-06

```
> anova(out)
```

Analysis of Variance Table

Response: Bodyfat

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Triceps	1	352.27	352.27	57.2768	1.131e-06 ***
Thigh	1	33.17	33.17	5.3931	0.03373 *
Midarm	1	11.55	11.55	1.8773	0.18956
Residuals	16	98.40	6.15		

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> |
```

## R Example

- ▶ The number of beta coefficients in the model is  $p = 4$ .
- ▶ The value of  $F$  involved in this test was

$$F = 21.52 = MSR/MSE = 132.33/6.15,$$

and the  $p$ -value is given as 0.000. This means that at least one of the three  $x$ -variables is a useful predictor of the  $y$ -variable.

- ▶ The value of  $R^2 = \frac{SST - SSE}{SST} = \frac{495.39 - 94.40}{495.39} = 0.801$ , or 80.1%. The model explains 80.1% of the observed variation in body fat.
- ▶ The estimated standard deviation of the errors is  $S = \sqrt{MSE} = 2.48$ .

# Practice

1. Suppose we are interested in students' final score for Statistics course by considering two predictors: average hours of studying per day and average score of two midterms. By fitting a multiple linear regression model, we have the following fitted model:

$$(\text{estimated}) \text{ Final Score} = b_0 + b_1 \cdot \text{Study Hour} + b_2 \cdot \text{Midterm}$$

where  $b_0, b_1, b_2$  are the estimated coefficients (based on Least SSE criterion) with some positive real values. Please write down the interpretation of each of the three estimated coefficients in this context.

# Practice

2. Complete the following ANOVA table for a multiple linear regression model if the number of x-variables is 3, and the number of observations in the sample is 35.

Source	DF	SS	MS	F
Regression		322.1		
Error				NA
Total		547.6	NA	NA