Multiple Regression

Berlin Chen

Department of Computer Science & Information Engineering

National Taiwan Normal University

Reference:

1. W. Navidi. *Statistics for Engineering and Scientists*. Chapter 8 (Sec. 8.1-8.2) & Teaching Material

Introduction

- Simple Linear Regression (introduced in Ch. 7)
  - Fit a linear model relating the value of an dependent variable *y* to the value of a single independent variable *x*

- However, there are many situations when a single independent variable is not enough
  - In situations like this, there are several independent variables, $x_1, x_2, \ldots, x_p$, that are related to a dependent variable *y*

Multiple Regression Model

- Assume that we have a sample of *n* items and that on each item we have measured a dependent variable *y* and *p* independent variables, $x_1, x_2, \ldots, x_p$
  - The *i*-th sampled item gives rise to the ordered set $(y_i, x_{1i}, \ldots, x_{pi})$

- We can then fit the multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i$$

Various Multiple Linear Regression Models

- Polynomial regression model (the independent variables are all powers of a single variable)

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i$$

- Quadratic model (polynomial regression of model of degree 2, and powers of several variables)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \beta_4 x_{1i}^2 + \beta_5 x_{2i}^2 + \varepsilon_i$$

  - A variable that is the product of two other variables is called an interaction

- These models are considered linear models, even though they contain nonlinear terms in the independent variables. The reason is that they are linear in the *coefficients* β_i

Estimating the Coefficients

- In any multiple regression model, the estimates $\hat{\beta}_0, \beta_1, ..., \hat{\beta}_p$ are computed by least-squares, just as in simple linear regression
  - The equation

$$\hat{y} = \hat{\beta}_0 + \beta_1 x_1 + \cdots + \hat{\beta}_p x_p$$

    is called the least-squares equation or fitted regression equation
- $\hat{y}_i$ is defined to be the *y* coordinate of the least-squares equation corresponding to the *x* values ($x_{1i}, ..., x_{pi}$)
  - The residuals are the quantities $e_i = y_i - \hat{y}_i$,which are the differences between the observed *y* values and the *y* values given by the equation.
  - We want to compute $\hat{\beta}_0, \beta_1, ..., \hat{\beta}_p$ so as to minimize the sum of the squared residuals E^2 . This is complicated and we rely on computers to calculated them

$$E^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_p x_{1p} \right)^2$$

Sums of Squares

- Much of the analysis in multiple regression is based on three fundamental quantities
  - Regression sum of squares (SSR)
  - Error sum of squares (SSE) $\sum_{i=1}^{n}(y_i - \hat{y})^2$
  - Total sum of squares (SST) $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- The analysis of variance identity is SST = SSR + SSE

Assumptions of the Error Terms

- Assumptions for Errors in Linear Models
  - In the simplest situation, the following assumptions are satisfied:
    1. The errors $\varepsilon_1,\ldots,\varepsilon_n$ are random and independent. In particular, the magnitude of any error ε_i does not influence the value of the next error ε_{i+1}
    2. The errors $\varepsilon_1,\ldots,\varepsilon_n$ all have mean 0
    3. The errors $\varepsilon_1,\ldots,\varepsilon_n$ all have the same variance, which we denote by σ^2
    4. The errors $\varepsilon_1,\ldots,\varepsilon_n$ are normally distributed

Mean and Variance of y_i

- In the multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_I$$

  - Under assumptions 1 through 4, the observations $y_1,\ldots, y_n$ are independent random variables that follow the normal distribution. The mean and variance of y_i are given by

$$\mu_{y_i} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$
$$\sigma^2_{y_i} = \sigma^2$$

  - Each coefficient represents the change in the mean of y associated with an increase of one unit in the value of x_i, when the other x variables are held constant

Statistics (2/3)

- The three statistics most often used in multiple regression are
  - estimated error variance s^2 (an estimate of the error variance σ^2)
  - coefficient of determination R^2
  - F statistic
- Estimated error variance s^2

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p-1} = \frac{\text{SSE}}{n-p-1}$$

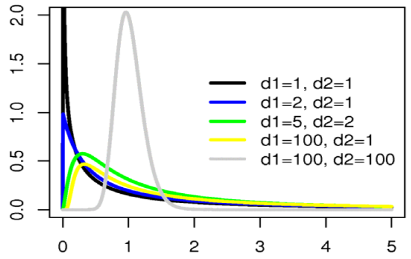
  - We have to adjust the estimated standard deviation since we are estimating $p + 1$ coefficients
  - The estimated variance $s^2_{\hat{\beta}_i}$ of each least-squares coefficient is a complicated calculation and we can find them on a computer based on the value s^2 (multiplying s^2 by a rather complicated function of variables x_{ij})

Statistics (2/3)

- In simple linear regression, the coefficient of determination, r^2, measures the goodness of fit of the linear model. The goodness of fit statistic in multiple regression denoted by R^2 is also called the coefficient of determination
  - The value of R^2 is calculated in the same way as r^2 in simple linear regression. That is, R^2 = SSR/SST

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}}$$

Statistics (3/3)



- Tests of Hypothesis (*F*-test)
  - In simple linear regression, a test of the null hypothesis $\beta_1 = 0$ is almost always made. If this hypothesis is not rejected, then the linear model may not be useful
  - The test is multiple linear regression is $H_0 = \beta_1 = \beta_2 = \ldots = \beta_p = 0$. This is a very strong hypothesis. It says that none of the independent variables has any linear relationship with the dependent variable
  - The test statistic for this hypothesis is

$$F = \frac{\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]/P}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / (n - p - 1)} = \frac{[\text{SST} - \text{SSE}]/p}{\text{SSE}/(n - p - 1)} = \frac{\text{SSR}/p}{\text{SSE}/(n - p - 1)}$$

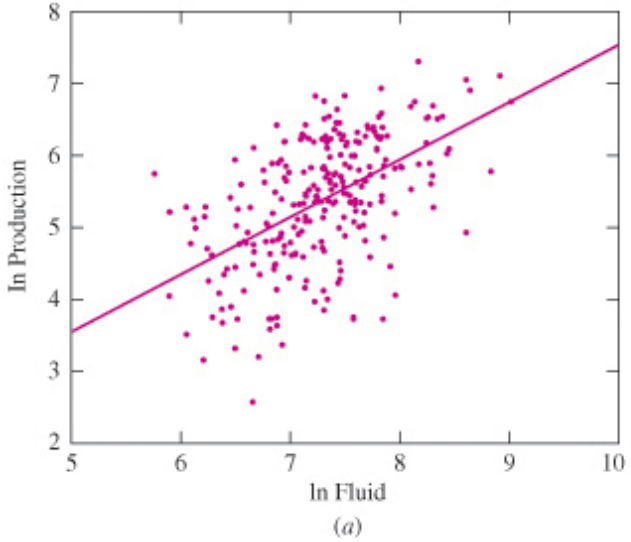
    - This is an *F* statistic and its null distribution is $F_{p,n-p-1}$. Note that the denominator of the *F* statistic is s^2. The subscripts *p*, *n*-*p*-1 are the **degrees of freedom** for the *F* statistic
    - The smaller the test statistic, the more plausible the null hypothesis (H_0)

Confounding (1/3)

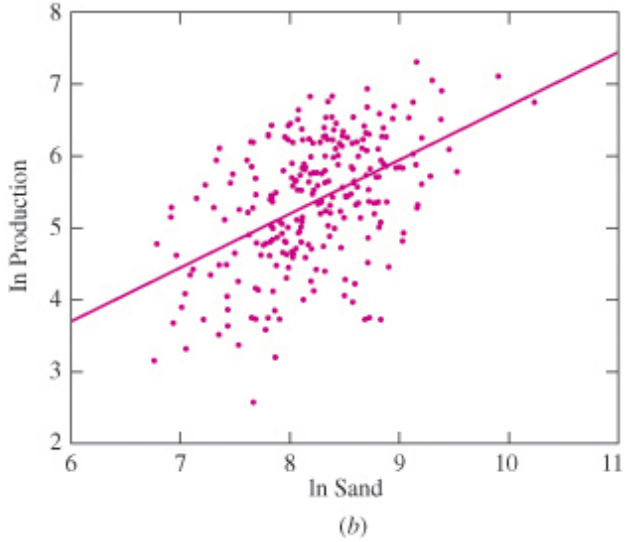
- Fitting separate models to each variable is not the same as fitting the multivariate model
- Example: There are 225 gas wells that received "fracture treatment" in order to increase production. In this treatment, fracture fluid, which consists of fluid mixed with sand, is pumped into the well. The sand holds open the cracks in the rock, thus increasing the flow of gas.
  - Does increasing the volume of fluid pumped increase the production of the well?
  - Does increasing the volume of sand pumped increase the production of the well?

Confounding (3/3)

- We can use sand to predict production or fluid to predict production. If we fit a simple model, then sand and fluid in their models show up as important predictors
  - Scatterplots in "log" domains



$$\ln\_\text{Production} = -0.444 + 0.798\ln\_\text{Fluid}$$
$$P - \text{value} = 0$$

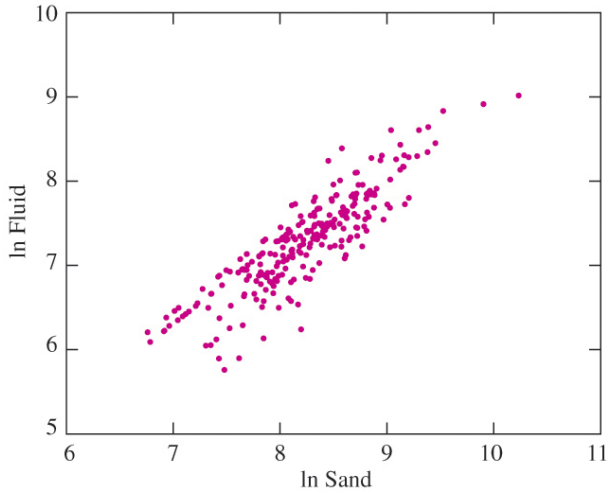


$$\ln\_\text{Production} = -0.778 + 0.748\ln\_\text{Fluid}$$
$$P - \text{value} = 0$$

  - It is tempted to conclude immediately that increasing either the volume of fluid or sand will increase the production

Confounding (3/3)

- However, fluid and sand are pumped in together in a single mixture (the more fluid, the more sand)



- Multiple regression provides a way to solve this issue

$$\ln\_\text{Production} = -0.729 + 0.670\,\ln\_\text{Fluid} + 0.148\,\ln\_\text{Sand}$$
$$P-\text{value} = 0 \qquad P-\text{value} = 0.389$$

  - We can conclude that increasing the amount of fluid tends to increase production, but it is not clear whether sand has such an effect

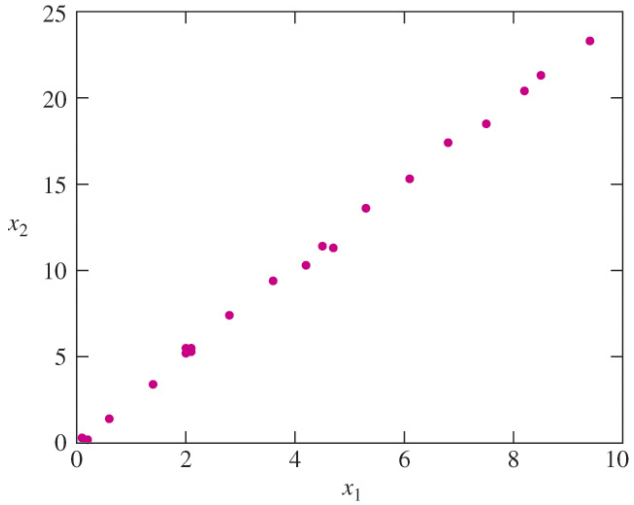
Collinearity (1/2)

- When two independent variables are very strongly correlated, multiple regression may not be able to determine which is the important one
  - In this case, the variables are said to be collinear
  - The word collinear means to lie on the same line, and when two variables are highly correlate, their scatterplot is approximately a straight line
  - The word multicollinearity is sometimes used as well, meaning that multiple variables are highly correlated with each other
  - When collinearity is present, the set of independent variables is sometimes said to be ill-conditioned

Collinearity (2/3)

- Example:

"x_1"	"x_2"	"y"
0.1	0.3	3.6
0.6	1.4	6.0
2.0	5.2	8.4
2.1	5.5	12.7
2.8	7.4	9.9
4.2	10.3	16.3
4.7	11.3	20.2
6.1	15.3	26.6
7.5	18.5	31.0
8.5	21.3	32.4
0.2	0.2	0.3
1.4	3.4	10.6
2.0	5.5	11.8
2.1	5.3	6.8
3.6	9.4	16.7
4.5	11.4	19.9
5.3	13.6	22.9
6.8	17.4	28.1
8.2	20.4	28.8
9.4	23.3	35.0



$$y = 2.90 + 3.53x_1$$

$$P - \text{value} = 0$$

$$y = 2.90 + 1.42x_2$$

$$P - \text{value} = 0$$

$$y = 2.72 - 0.49x_1 + 1.62x_2$$

$$P - \text{value} = 0.914 \quad P - \text{value} = 0.379$$

Larger P-values indicate that the coefficients are plausible to be 0

Collinearity (3/3)

- Sometimes two variables are so correlated that multiple regression cannot determine which is responsible for the linear relationship with *y*
- In general, there is not much that can be done when variables are collinear
- The only way to fix the situation is to collect more data, including some values for the independent variables that are not on a straight line

Summary

- In this chapter, we learned about
  - Multiple regression models
  - Estimating the coefficients
  - Confounding and collinearity