STATS 8: Introduction to Biostatistics

Regression Analysis

Babak Shahbaba
UCI, Spring of 2012

Introduction

- We now discuss **linear regression models** for either testing a hypothesis regarding the relationship between one or more *explanatory variables* and a response variable, or **predicting** unknown values of the response variable using one or more *predictors*.

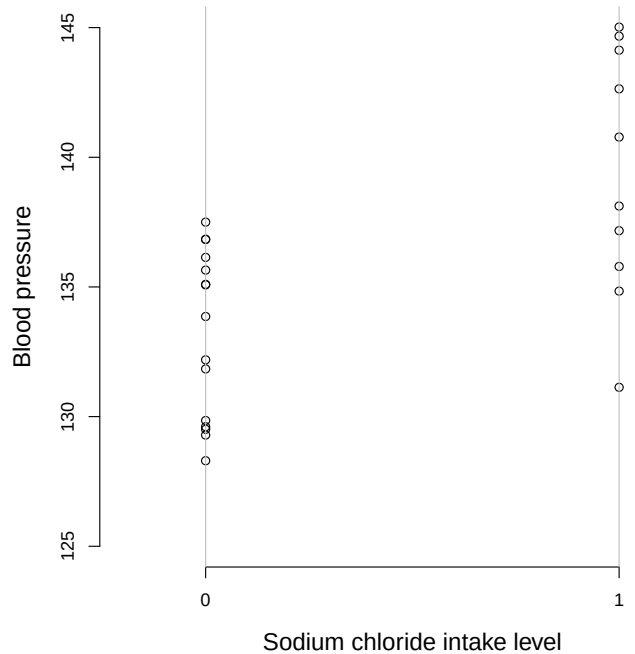
- We use X to denote explanatory variables and Y to denote response variables.

- We start by focusing on problems where the explanatory variable is binary. As before, the binary variable X can be either 0 or 1.

- We then continue our discussion for situations where the explanatory variable is numerical.

One Binary Explanatory Variable

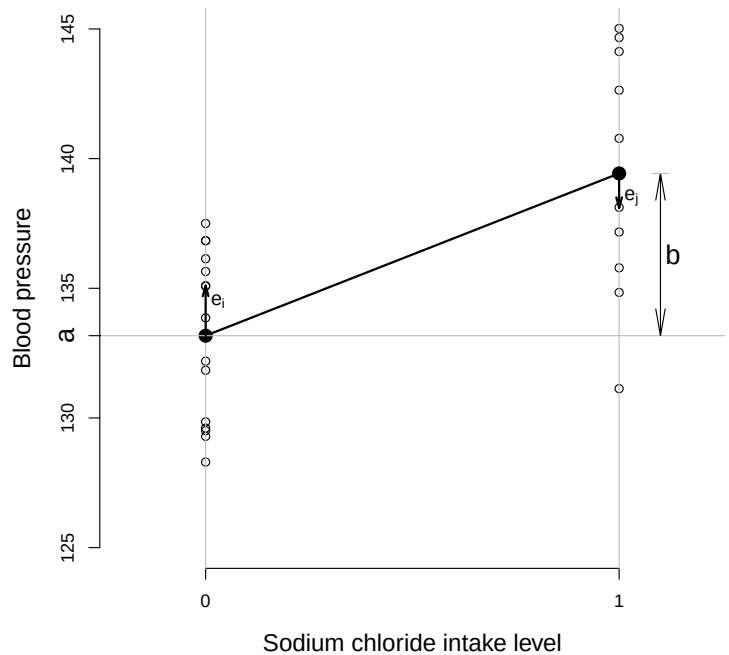
- Suppose that we want to investigate the relationship between sodium chloride (salt) consumption (low vs. high consumption) and blood pressure among elderly people (e.g., above 65 years old).



One Binary Explanatory Variable

- The following figure shows the dot plot along with sample means, shown as black circles, for each group.

- We connect the two sample means to show the overall pattern for how blood pressure changes from one group to another.



One Binary Explanatory Variable

- Using the intercept a and slope b, we can write the equation for the straight line that connects the estimates of the response variable for different values of X as follows:

$$\hat{y} = a + bx.$$

- The slope is also knows as the **regression coefficient** of X.

- For this example,

$$\hat{y} = 133.17 + 6.25x.$$

- We expect that on average the blood pressure increases by 6.25 units for one unit increase in X.

- In this case, one unit increase in X from 0 to 1 means moving from low to high sodium chloride diet group.

One Binary Explanatory Variable

- For an individual with $x = 0$ (i.e., low sodium chloride diet), the estimate according to the above regression line is

$$\begin{aligned} \hat{y} &= a + b \times 0 = a \\ &= \hat{y}_{x=0}, \end{aligned}$$

  which is the sample mean for the first group.

- For an individual with $x = 1$ (i.e., high sodium chloride diet), the estimate according to the above regression line is

$$\begin{aligned} \hat{y} &= a + b \times 1 = a + b \\ &= \hat{y}_{x=0} + \hat{y}_{x=1} - \hat{y}_{x=0} \\ &= \hat{y}_{x=1}. \end{aligned}$$

One Binary Explanatory Variable

- We refer to the difference between the observed and estimated values of the response variable as the **residual**.

- For individual i, we denote the residual e_i and calculate it as follows:

$$e_i = y_i - \hat{y}_i.$$

- For instance, if someone belongs to the first group, her estimated blood pressure is $\hat{y}_i = a = 133.17$.

- Now if the observed value of her blood pressure is $y_i = 135.08$, then the residual is

$$e_i = 135.08 - 133.17 = 1.91.$$

Residual sum of squares

- As a measure of discrepancy between the observed values and those estimated by the line, we calculate the **Residual Sum of Squares** (RSS):

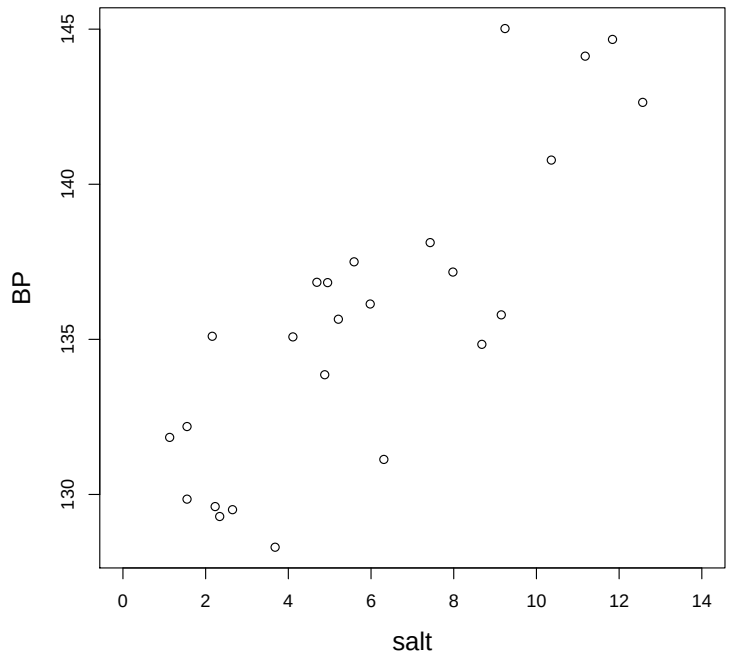
$$RSS = \sum_{i}^{n} e_i^2.$$

- Here, e_i is the residual of the ith observation, and n is the sample size.

- The square of each residual is used so that its sign becomes irrelevant.

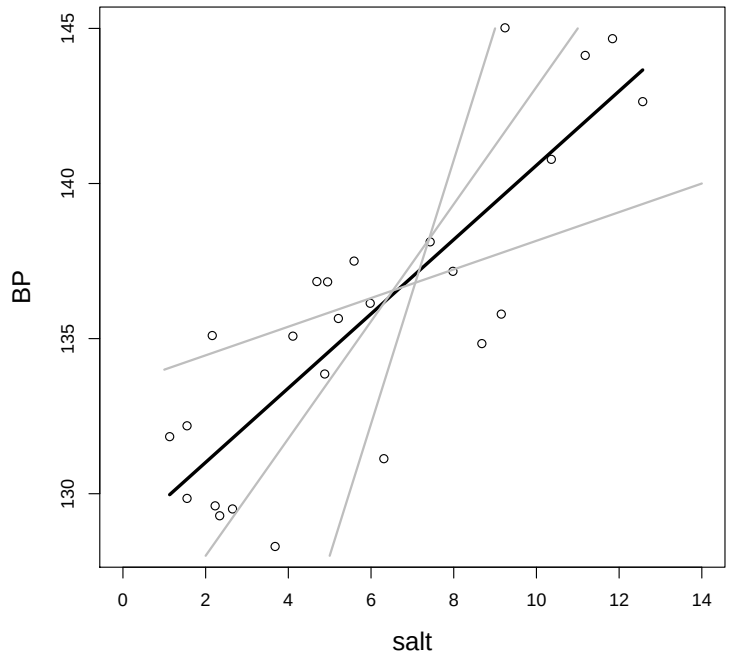
One Numerical Explanatory Variable

- We now discuss simple linear regression models (i.e., linear regression with only one explanatory variable), where the explanatory variable is numerical.



One Numerical Explanatory Variable

- Among all possible lines we can pass through the data, we choose the one with the smallest sum of squared residuals.



- The resulting line is called the **least-squares regression line**

Statistical inference using regression models

- We can use R or R-Commander to find the least-squares regression line.
- The slope of the regression line plays an important role in evaluating the relationship between the response variable and explanatory variable(s).
- We can also use this regression line to predict the unknown value of the response variable.

Prediction

- Using the regression line, we can estimate the unknown value of the response variable for members of the population who did not participate in our study.

- In this case, we refer to our estimates as **predictions**.

- For example, we can use the linear regression model we built previously to predict the value of blood pressure for a person with high sodium chloride diet (i.e., $x = 1$),

$$\begin{aligned} \hat{y} &= 133.17 + 6.25x \\ &= 133.17 + 6.25 \times 1 \\ &= 139.42. \end{aligned}$$

Confidence interval

- We can find the confidence interval for the population regression coefficient as follows:

$$[b - t_{\text{crit}} \times SE_b, b + t_{\text{crit}} \times SE_b].$$

- For simple (i.e., one predictor) linear regression models, SE_b is obtained as follows:

$$SE_b = \frac{\sqrt{RSS/(n-2)}}{\sqrt{\sum_i (x_i - \bar{x})^2}}.$$

- The corresponding t_{crit} is obtained from the t-distribution with $n - 2$ degrees of freedom.

Hypothesis testing

- To assess the null hypothesis that the population regression coefficient is zero, which is interpreted as no linear relationship between the response variable and the explanatory variable, we first calculate the t-score.

$$t = b/SE_b$$

- Then, we find the corresponding p-value as follows:

$$\begin{aligned}
&\text{if } H_A : \beta < 0, \quad p_{\text{obs}} = P(T \leq t),\\
&\text{if } H_A : \beta > 0, \quad p_{\text{obs}} = P(T \geq t),\\
&\text{if } H_A : \beta \neq 0, \quad p_{\text{obs}} = 2 \times P\big(T \geq |t|\big),
\end{aligned}$$

where T has the t-distribution with $n - 2$ degrees of freedom.