Statistical Data Analysis

Assist. Prof. Dr. Zeyneb KURT

(Slides have been prepared by
Prof. Dr. Nizamettin AYDIN,
updated by Zeyneb KURT)

zeyneb@yildiz.edu.tr
http://avesis.yildiz.edu.tr/zeyneb/

Analysis of Variance (ANOVA)

ANOVA

- The process of evaluating hypotheses regarding the group means of multiple populations is called the Analysis of Variance (ANOVA).
- ANOVA models generalize the *t*-test and are used to compare the means of multiple groups identified by a categorical variable with more than two possible categories.
- Since we are only considering one factor only, this method is specifically called one- way ANOVA.
- An ANOVA with two factors is called a two-way ANOVA.
- In general, the between-groups variation is denoted as SS_B and calculated by

$$SS_B = \sum_{i=1}^{k} n_i \, (\overline{y}_i - \overline{y})^2$$

where k is the number of groups

ANOVA

- The within-groups variation is denoted as SS_W and calculated by

$$SS_w = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{y}_{ij} - \overline{y}_i)^2$$

- We measure the total variation in Y by

$$SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y})^2$$

- The total variation SS is equal to the sum of the between-groups variation SS_B and the within-groups variation SS_W,

$$SS = SS_B + SS_W.$$

- The total variation can be attributed partly to the variation within groups and partly to the variation between groups. SS_B is interpreted as the part of total variation SS that is associated with (and can be explained by) the factor variable X (e.g., syndrome type).
- In contrast, SS_W is regarded as the unexplained part of total variation and is regarded as random.

ANOVA

- Let us denote the overall population mean of Y as μ and group-specific population means as $\mu_1, \ldots, \mu_4$.
- Then we can express the null hypothesis of no difference in means between the groups as

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$$

- The alternative hypothesis H_A is that at least one of the group means μ_i is different from the mean μ.
- The test statistic for examining the null hypothesis is called F-statistic (more specifically, ANOVA F -statistic) and is defined as

$$F = \frac{SS_B/(k-1)}{SS_W/(n-k)}$$

where n is the total sample size, and k is the number of groups.

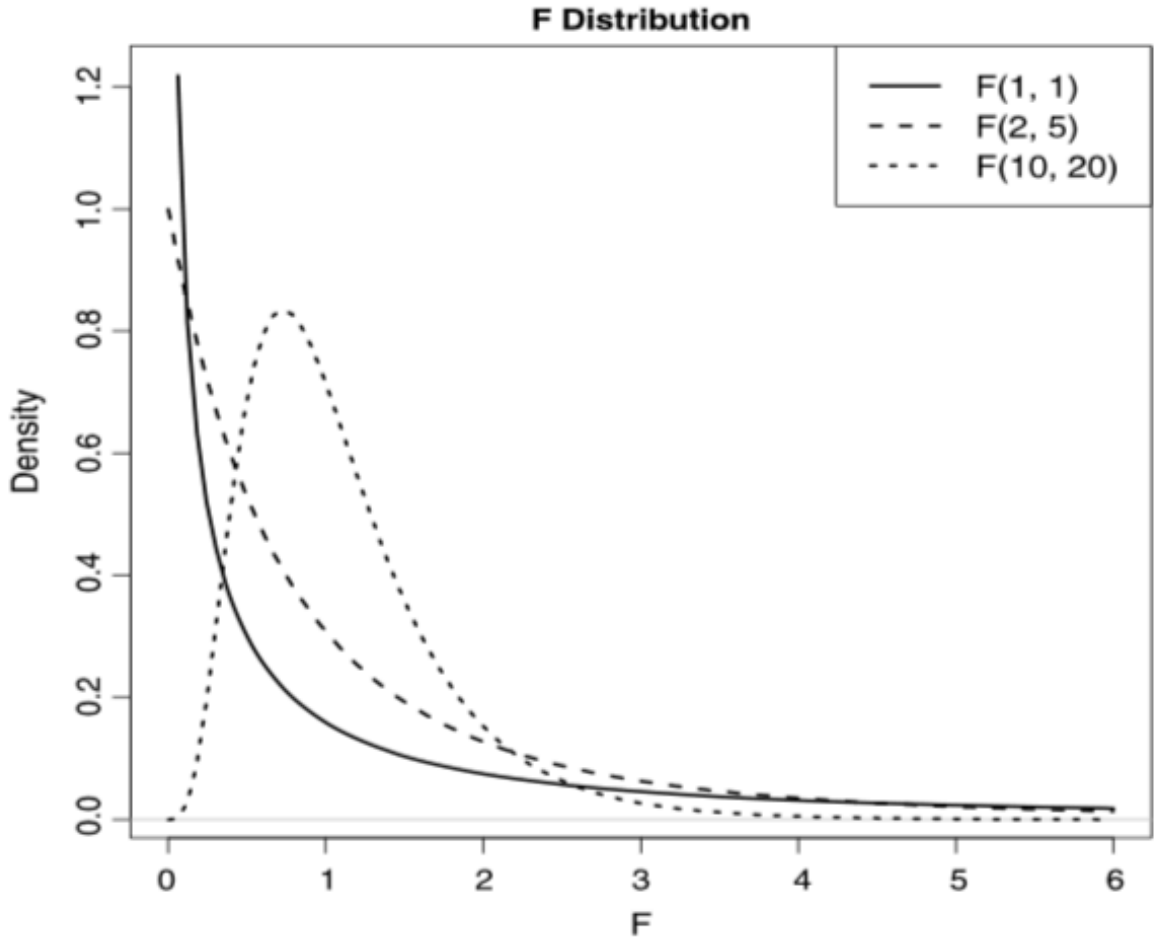
- The numerator is called the mean square for groups, and the denominator is called the mean square error (MSE).

ANOVA

- For the one-way ANOVA, the F-statistic has $F(df_1 = k - 1, df_2 = n - k)$ distribution under the null hypothesis (i.e., assuming that the null hypothesis is true).
- The F-distribution, which is a continuous probability distribution, is very important for hypothesis testing.
- It is specified by two parameters, df_1 and df_2, and is denoted as $F(df_1, df_2)$.
- We refer to df_1 and df_2 as the numerator degrees of freedom and denominator degrees of freedom, respectively.
- Both parameters must be positive.

ANOVA

- The following figure shows the pdf of F-distribution for different values of df_1 and df_2.



Example

- As an example, we analyze the Cushings data set, which is available from the MASS package.
  - Cushing's syndrome is a hormone disorder associated with high level of cortisol secreted by the adrenal gland.
- The *Type* variable in the data set shows the underlying type of syndrome, which can be one of four categories:
  - adenoma (a), bilateral hyperplasia (b), carcinoma (c), and unknown (u).

Example

- Our objective is to find whether the four groups are different with respect to urinary excretion rate of Tetrahydrocortisone.
- We denote by Y the urinary excretion rate of Tetrahydrocortisone and by X the *Type* variable,
  - where $X = 1$ for Type=a, $X = 2$ for Type=b, $X = 3$ for Type=c, and $X = 4$ for Type=u.
- Then, our objective could be defined as investigating whether the *mean* of the response variable Y differs for different values (levels) of the factor X.

Example

- Denote the individual observations as y_{ij} : the urinary excretion rate of Tetrahydrocortisone of the j th individual in group i.
- Total number of observations is $n = 27$,
- The number of observations in each group is

  $n_1 = 6$, $n_2 = 10$, $n_3 = 5$, and $n_4 = 6$.
- The overall (for all groups) observed sample mean for the response variable is $\bar{y} = 10.46$.
- We also find the group specific means, by clicking (in R-Commander) *Statistics→Summaries→Numerical summaries*

  $\bar{y}_1 = 3.0$, $\bar{y}_2 = 8.2$, $\bar{y}_3 = 19.7$, and $\bar{y}_4 = 14.0$.
- The degrees of freedom parameters are

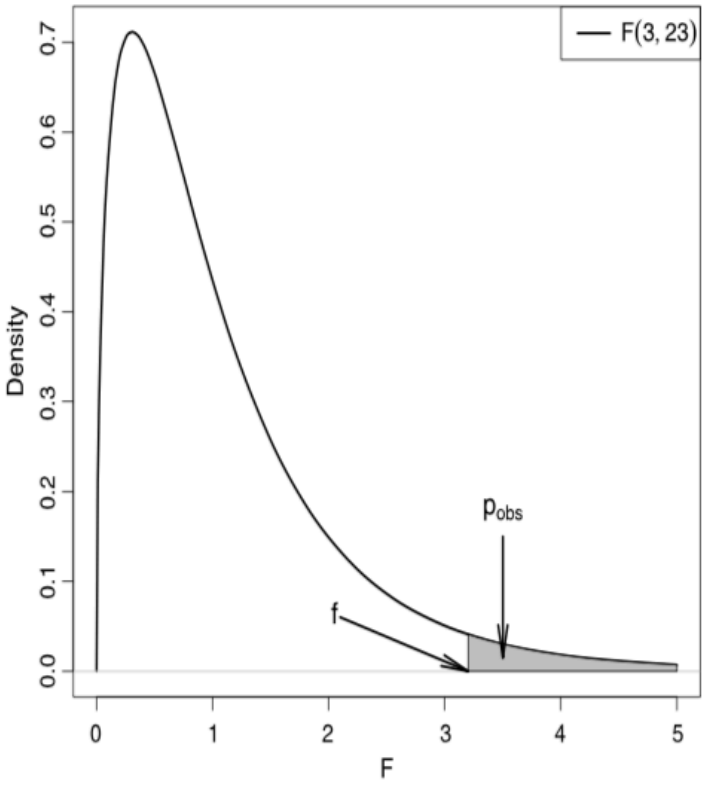
  $df_1 = 4-1 = 3$ and $df_2 = 27 - 4 = 23$.

Example

- $SS_B = 893.5$ and $SS_W = 2123.6$.
- The observed value of F-statistic is $f = 3.2$ given under the column labeled F value.
- The resulting p-value is then 0.04.
- Therefore, we can reject H_0 at 0.05 significance level (but not at 0.01) and conclude that the differences among group means for urinary excretion rate of Tetrahydrocortisone are statistically significant (at 0.05 level).

Example

- For plotting the $F(3, 23)$ distribution using R-Commander, click *Distribution* → *Continuous distributions*→ *F distribution Plot F distribution*.
- Set the *Numerator degrees of freedom* to 3 and the *Denominator degrees of freedom* to 23.



- The density plot of $F(3, 23)$-distribution.
- This is the distribution of F-statistic for the Cushings data assuming that the null hypothesis is true.
- The observed value of the test statistic is $f = 3.2$, and the corresponding p-value is shown as the *shaded area* above 3.2

ANALYSIS OF CATEGORICAL VARIABLES

ANALYSIS OF CATEGORICAL VARIABLES

- Pearson's χ^2 (chi-squared) test is used to test hypotheses regarding the distribution of a categorical variable or the relationship between two categorical variables.
- Pearson's χ^2 test uses a test statistic, which we denote as Q, to measure the discrepancy between the observed data and what we expect to observe under the null hypothesis.
- Higher levels of discrepancy between data and H_0 results in higher values of Q.
- We use q to denote the observed value of Q based on a specific sample of observed data.

Pearson's χ^2 Test for One Categorical Variable

- Let us denote the binary variable of interest as X, based on which we can divide the population into two groups depending on whether $X = 1$ or $X = 0$.
- Further, suppose that the null hypothesis H_0 states that the probability of group 1 is $\boldsymbol{\mu_{01}}$ and the probability of group 2 is $\boldsymbol{\mu_{02}}$.
  - Here $\boldsymbol{\mu_{02}} = 1 - \boldsymbol{\mu_{01}}$.
- If the null hypothesis is true, we expect that, out of n randomly selected individuals, $\boldsymbol{E_1 = n\mu_{01}}$ belong to the first group, and $\boldsymbol{E_2 = n(1 - \mu_{01})}$ belong to the second group.
- We refer to E_1 and E_2 as the expected frequencies under the null.
- We refer to the observed number of people in each group as the observed frequencies and denote them $\boldsymbol{O_1}$ and $\boldsymbol{O_2}$ for group 1 and group 2, respectively.

Pearson's χ^2 Test for One Categorical Variable

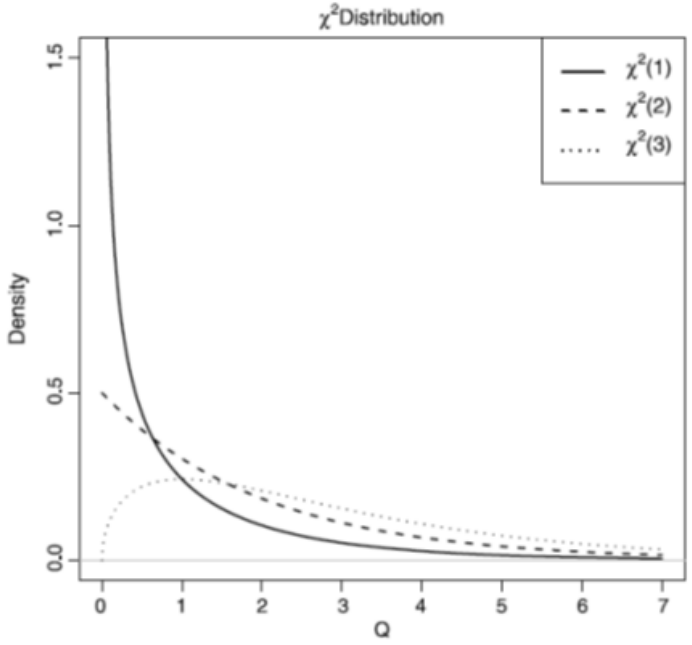
- Pearson's $\chi 2$ test measures the discrepancy between the observed data and the null hypothesis based on the difference between the observed and expected frequencies as follows:

$$Q = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

- The value of Q will be zero only when the observed data matches our expectation under the null exactly.
- When there is some discrepancy between the data and the null hypothesis, Q becomes greater than zero.
- The higher discrepancy between our data and what is expected under H_0, the larger Q and therefore the stronger the evidence against H_0.

Pearson's χ^2 Test for One Categorical Variable

- If the null hypothesis is true, then the approximate distribution of Q is χ^2.
- Like the t-distribution, the χ^2-distribution is commonly used for hypothesis testing and denoted $\chi^2(df)$.
- The plot of the pdf for a χ^2 distribution with various degrees of freedom



- The observed significance level p_{obs} is calculated using the χ^2 distribution with 1 degree of freedom.
- This corresponds to the upper tail probability of q from the $\chi^2(1)$ distribution.

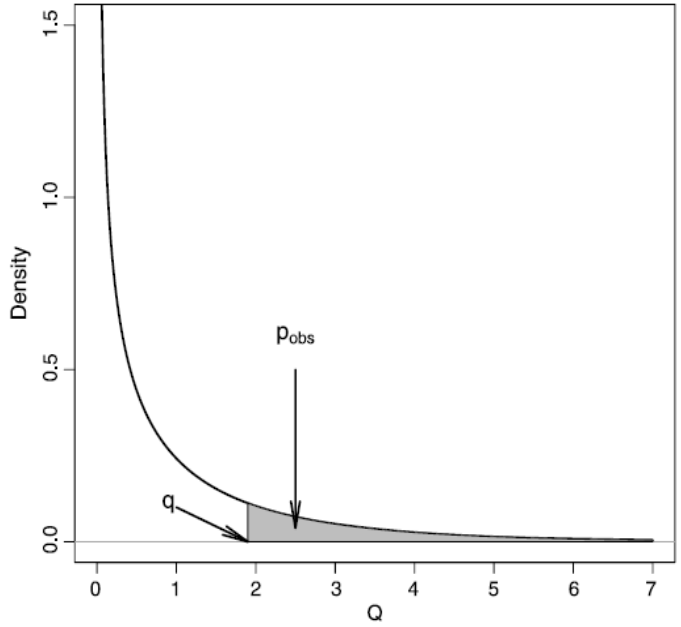
Example

- We use the heart attack survival rate (i.e., the probability of survival after heart attack) within one year after hospitalization.
- Suppose that $\boldsymbol{H_0}$ specifies that the probability of surviving is $\mu_{01} = 0.70$ and the probability of not surviving is $\mu_{02} = 0.30$.
- If we take a random sample of size $n = 40$ from the population, we expect that $E_1 = 0.70 \times 40 = 28$ and $E_2 = 0.30 \times 40 = 12$.
- Now suppose that the observed number of people in each group as the observed frequencies: $O_1 = 24$ and $O_2 = 16$.
- For the heart attack survival example, the observed value of the test statistic is

$$q = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \frac{(24 - 28)^2}{28} + \frac{(16 - 12)^2}{12} = 1.90$$

- The $p_{obs} = P(Q \geq 1.90) = 0.17$ is obtained from a χ^2 distribution with 1 degree of freedom

Example



- The sampling distribution for Q under the null hypothesis: $Q \sim \chi^2(1)$.
- The p-value is the upper tail probability of observing values as extreme or more extreme than $q = 1.90$

- Therefore, the results are not statistically significant, and we cannot reject the null hypothesis at commonly used significance levels (e.g., 0.01, 0.05, and 0.1).
- In this case, we believe that the difference between observed and expected frequencies could be due to chance alone.

Categorical Variables with Multiple Categories

- Pearson's χ^2 test can be generalized to situations where the categorical random variable can take more than two values.
- In general, for a categorical random variable with k possible categories, we calculate the test statistic Q as

$$Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} ,$$

- The approximate distribution of Q is $\chi 2$ with the degrees of freedom equal to $df = k - 1$.
- Therefore, to find p_{obs}, we calculate the upper tail probability of q (the observed value of Q) from the $\chi^2(k-1)$ distribution.

Example

- Suppose that we monitor heart attack patients for one year and divide them into three groups:
  - patients who did not have another heart attack and survived,
  - patients who had another heart attack and survived,
  - patients who did not survive.
- Suppose that $\mu_{01} = 0.5$, $\mu_{02} = 0.2$, and $\mu_{03} = 0.3$.
- The expected frequencies of each category for a sample of $n = 40$:

$$E_1 = 0.5 \times 40 = 20,\ E_2 = 0.2 \times 40 = 8,\ E_3 = 0.3 \times 40 = 12.$$

- This time, suppose that the actual observed frequencies based on a sample of size $n = 40$ for the three groups are

$$O_1 = 13,\ O_2 = 11,\ O_3 = 16.$$

Example

- The amount of discrepancy:

$$Q = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3}$$

- The observed value of this test statistic:

$$q = \frac{(13 - 20)^2}{20} + \frac{(11 - 8)^2}{8} + \frac{(16 - 12)^2}{12} = 4.91$$

- Using R-Commander, we find $p_{obs} = P(Q \geq 4.91) = 0.086$ using the χ^2 distribution with 2 degrees of freedom.
- Therefore, we can reject the null hypothesis at 0.1 level but not at 0.05 level.
- At the 0.1 significance level, we can conclude that the difference between observed and expected frequencies is statistically significant, and it is probably not due to chance alone.

Pearson's χ^2 Test of Independence

- We now discuss the application of Pearson's χ^2 test for evaluating a hypothesis regarding possible relationship between two categorical variables.
- More specifically, we measure the difference between the observed frequencies and expected frequencies under the null.
- The null hypothesis in this case states that the two categorical random variables are independent.
  - For two independent random variables, the joint probability is equal to the product of their individual probabilities.
    - In what follows, we use this rule to find the expected frequencies.

Pearson's χ^2 Test of Independence

- We use the following general form of Pearson's χ^2 test, which summarizes the differences between the expected frequencies (under the null hypothesis) and the observed frequencies over all cells of the contingency table:

$$Q = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} and E_{ij} are the observed and expected values in the ith row and j th column of the contingency table.

- The double sum simply means that we add the individual measures of discrepancies for cells by going through all cells in the contingency table.

Pearson's χ^2 Test of Independence

- As before, higher values of Q provide stronger evidence against H_0.
- For $I \times J$ contingency tables (i.e., I rows and J columns), the Q statistic has approximately the χ^2 distribution with $(I-1)\times(J-1)$ degrees of freedom under the null.
- Therefore, we can calculate the observed significance level by finding the upper tail probability of the observed value for Q, which we denote as q, based on the $\chi 2$ distribution with $(I-1)\times(J-1)$ degrees of freedom.

Example 1

- The probability that the mother is smoker (i.e., smoke =1) and the baby has low birthweight (i.e., low =1) is the product of smoker and low-birthweight probabilities.
- For the baby weight example, we can summarize the observed and expected frequencies in the contingency tables.

	Observed frequency				Expected frequency	
	Normal	Low			Normal	Low
Nonsmoking	86	29		Nonsmoking	79.1	35.9
Smoking	44	30		Smoking	50.9	23.1

Example 1

- Then Pearson's test statistic is

$$Q = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}}$$

$$q = \frac{(86 - 79.1)^2}{79.1} + \frac{(29 - 35.9)^2}{35.9} + \frac{(44 - 50.9)^2}{50.9} + \frac{(30 - 23.1)^2}{23.1} = 4.9$$

- Since the table has $I = 2$ rows and $J = 2$ columns, the approximate null distribution of Q is χ^2 with $(2-1)\times(2-1) = 1$ degrees of freedom.
- Consequently, the observed p-value is the upper tail probability of 4.9 using the $\chi^2(1)$ distribution.
- We find $p_{\text{obs}} = P(Q \geq 4.9) = 0.026$.
- Therefore, at the 0.05 significance level (but not at 0.01 level), we can reject the null hypothesis that the mother's smoking status and the baby's birthweight status are independent.

Example 2

- Suppose that we would like to investigate whether the race of mothers is related to the risk of having babies with low birthweight.
- The race variable can take three values: 1 for white, 2 for African-American, and 3 for others.
- As before, the low variable can take 2 possible values: 1 for babies with birthweight less than 2.5 kg and 0 for other babies.
- Therefore, all possible combinations of race and low can be presented by a 3×2 contingency table.
- The following Table provides the observed frequency of each cell and the expected frequency of each cell if the null hypothesis is true.

Example 2

	Observed frequency				Expected frequency	
Groups	Normal (low=0)	Low (low=1)		Groups	Normal (low=0)	Low (low=1)
1	73	23		1	66	30
2	15	11		2	18	8
3	42	23		3	46	21

- For example, there are 73 babies in the first row and first column.
- This is the number of babies in the intersection of race = 1 (mother is white) and low = 0 (having a baby with normal birthweight).
- If the null hypothesis is true, the expected number of babies in this cell would have been 66.

Example 2

- The observed value of the test statistic Q is obtaned as $q = 5.0$ using the following equations.
- The distribution of Q under the null hypothesis is χ^2 with
  $(3 - 1) \times (2 - 1) = 2$ degrees of freedom.
- To find the corresponding p-value, we need to find the probability of observing values as or more extreme than 5.0.
- This is the upper-tail probability of 5 from the $\chi^2(2)$ distribution: $p_{\text{obs}} = P(Q \geq 5)$.

$$Q = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}} + \frac{(O_{31} - E_{31})^2}{E_{31}} + \frac{(O_{32} - E_{32})^2}{E_{32}}$$

- The value of p_{obs} is 0.08.
- We can reject the null hypothesis at 0.1 level but not at 0.05 level.
- At 0.05 level, the relationship between the two variables (i.e., race of mothers and birthweight status) is not statistically significant.

Regression Analysis

Regression Analysis

- The modeling of the relationship between a response variable and a set of explanatory variables is one of the most widely used of all statistical techniques.
  - We refer to this type of modeling as regression analysis.
- A regression model provides the user with a functional relationship between the response variable and explanatory variables that allows the user to determine which of the explanatory variables have an effect on the response.
  - The regression model allows the user to explore what happens to the response variable for specified changes in the explanatory variables.
    - {For example, financial officers must predict future cash flows based on specified values of interest rates, raw material costs, salary increases, and so on}

Regression Analysis

- The basic idea of regression analysis is to obtain a model for the functional relationship between a response variable (often referred to as the dependent variable) and one or more explanatory variables (often referred to as the independent variables).
- **Regression models have a number of uses**:
  - The model provides a description of the major features of the data set.
    - In some cases, a subset of the explanatory variables will not affect the response variable, and, hence, the researcher will not have to measure or control any of these variables in future studies.
      - This may result in significant savings in future studies or experiments.

Regression Analysis

- The equation relating the response variable to the explanatory variables produced from the regression analysis provides estimates of the response variable for values of the explanatory variables not observed in the study.
  - For example, a clinical trial is designed to study the response of a subject to various dose levels of a new drug.
  - Because of time and budgetary constraints, only a limited number of dose levels are used in the study.
    - The regression equation will provide estimates of the subjects' response for dose levels not included in the study.
- In business applications, the prediction of future sales of a product is crucial to production planning.
  - If the data provide a model that has a good fit in relating current sales to sales in previous months, prediction of sales in future months is possible.

Regression Analysis

- In some applications of regression analysis, the researcher is seeking a model that can accurately estimate the values of a variable that is difficult or expensive to measure using explanatory variables that are inexpensive to measure and obtain.
  - If such a model is obtained, then in future applications it is possible to avoid having to obtain the values of the expensive variable by measuring the values of the inexpensive variables and using the regression equation to estimate the values of the expensive variable.
    - For example, a physical fitness center wants to determine the physical well-being of its new clients. Maximal oxygen uptake is recognized as the single best measure of cardiorespiratory fitness, but its measurement is expensive. Therefore, the director of the fitness center would want a model that provides accurate estimates of maximal oxygen uptake using easily measured variables such as weight, age, heart rate after a 1-mile walk, time needed to walk 1 mile, and so on.

Regression Analysis

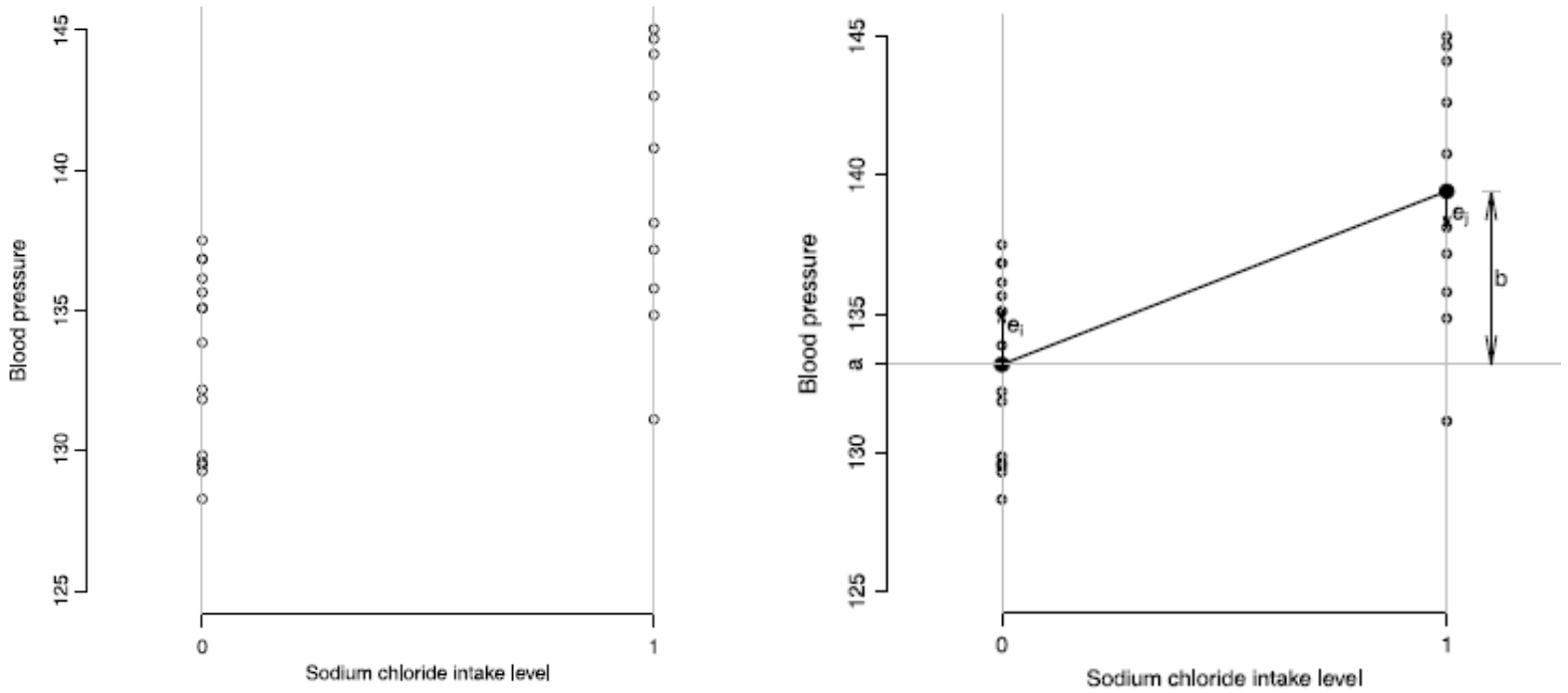
- After this soft introduction, we now discuss linear regression models for either testing a hypothesis regarding the relationship between one or more explanatory variables and a response variable, or predicting unknown values of the response variable using one or more predictors.
  - We use X to denote 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.

Linear Regression Models with 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).
- The next 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.

Linear Regression Models with One Binary Explanatory Variable

- The dot plot for systolic blood pressure for 25 elderly people (left panel), where 15 people follow a low sodium chloride diet ($X = 0$), and 10 people follow a high sodium chloride diet ($X = 1$)
- The dot plot for systolic blood pressure for 25 elderly people (right panel).
  - Here, the sample means among the low and high sodium chloride diet groups are shown as black circles. A straight line connects the sample means. The line intercepts the vertical axis at a = 133.17 and has slope b = 6.25



Linear Regression Models with 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 constant (intercept) term a is interpreted as the predicted value of y when $x = 0$.
  - The slope b of the line is the predicted change in y when there is a one-unit change in x.
- The slope is also known as the regression coefficient of X.
  - For the given 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.

Linear Regression Models with One Binary Explanatory Variable

- For an individual with $x = 0$, the estimate according to the above regression line is

$$\hat{y} = a + b \times 0 = a = \hat{y}_{x=0}$$

which is the sample mean for the first group.

- For an individual with $x = 1$, the estimate according to the above regression line is

$$\hat{y} = a + b \times 1 = a + b = \hat{y}_{x=0} + \hat{y}_{x=1} - \hat{y}_{x=0} = \hat{y}_{x=1}$$

- 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:

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

Linear Regression Models with One Binary Explanatory Variable

- – 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

$$e_i = 135.08 - 133.17 = 1.91$$

- By rearranging the terms in the equation $e_i = y_i - \hat{y}_i$, we can write the observed value y_i in terms of the estimate obtained from the regression line and the corresponding residual,

$$y_i = \hat{y}_i + e_i$$

- For individual i, whose values of the explanatory variable and the response variable are x_i and y_i, respectively, the estimated value of the response variable, denoted as $\hat{y}_i$, is

$$\hat{y}_i = a + bx_i$$

- So, the observed value y_i can be given as

$$y_i = a + bx_i + e_i$$

The linear relationship

- The linear relationship between Y and X in the entire population can be presented in a similar form,

$$Y = \alpha + \beta X + \varepsilon$$

- where α is the intercept, and β is the slope of the regression line , ε is called the **error term**, representing the difference between the estimated and the actual values of Y in the population.
- We refer to the above equation as the **linear regression model**.
  - We refer to α and β as the **regression parameters**.
  - More specifically, β is called the **regression coefficient** for the explanatory variable.
  - The process of finding the regression parameters is called **fitting** a regression model to the data.

Statistical Inference Using Simple Linear Regression Models

- 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 in the previous example to predict the value of blood pressure for a person with high sodium chloride diet (i.e., $x = 1$),

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

Residual sum of squares

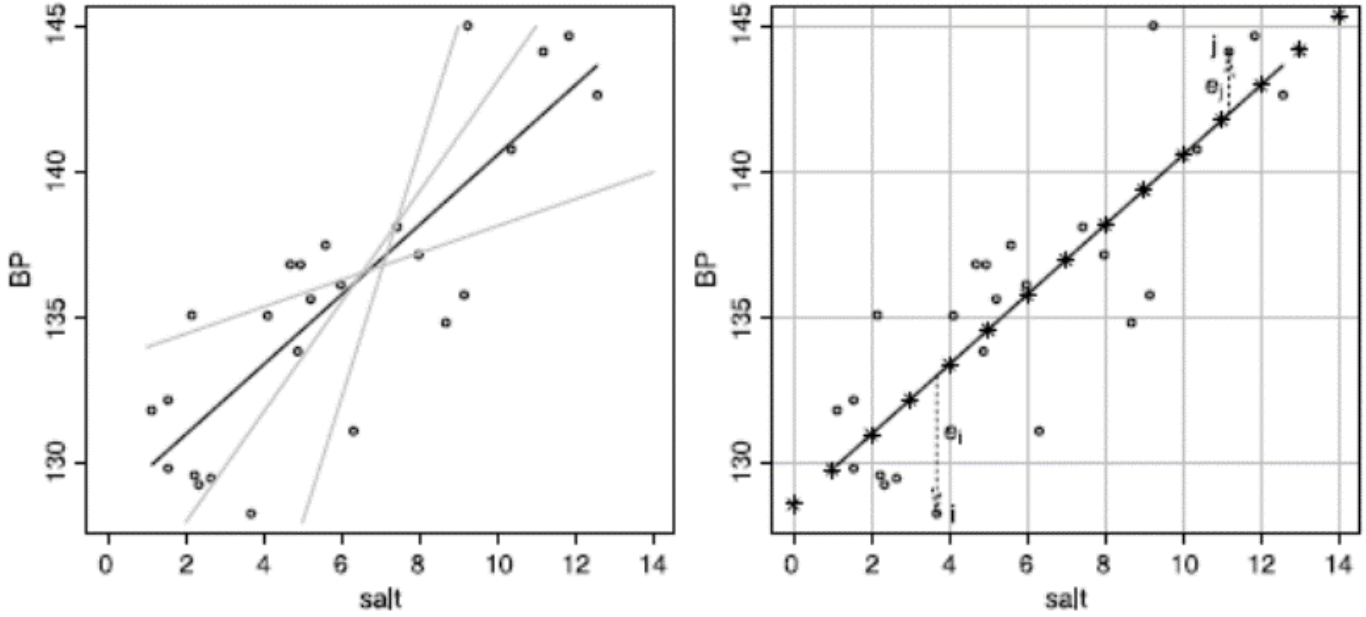
- As a measure of discrepancy between the observed values and those estimated by the line, we calculate the Residual Sum of Squares (*RSS*):

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

- Here, e_i is the residual of the *i*th 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.



- *Left panel*: Scatterplot of blood pressure by daily sodium chloride intake along with some candidate lines for capturing the overall relationship between the two variables.
  - The *black line* is the least-squares regression line.
- *Right panel*: The least-squares regression line for the relationship between blood pressure and sodium chloride intake.
  - The *vertical arrows* show the residuals for two observations.
  - The *stars* are the estimated blood pressure for daily sodium chloride intakes from 0 to 14 grams

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**.
- First, we find the slope of regression line using the sample correlation coefficient r between the response variable Y and the explanatory variable X,

$$b = r\frac{s_x}{s_y}$$

  - Here, s_y is the sample standard deviation of Y, and s_x is the sample standard deviation of X.
    - Note that since s_x and s_y are always positive, the sign of b is the same as the sign of the correlation coefficient: $b > 0$ for positively correlated random variables, and $b < 0$ for negatively correlated variables.

One Numerical Explanatory Variable

- When $r = 0$ (i.e., the two variables are not linearly related), then $b = 0$.
- After finding the slope, we find the intercept as follows:

$$a = \bar{y} - b\bar{x}$$

where $\bar{y}$ and $\bar{x}$ are the sample means for Y and X, respectively.

- Then the least-squares regression line with intercept a and slope b can be expressed as

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

Example

- For the blood pressure example,
  - the sample correlation coefficient is $r = 0.84$;
  - the sample standard deviation of blood pressure is $s_y = 4.94$,
  - the sample standard deviation of sodium chloride intake is $s_x = 3.46$.
- Therefore,

$$b = 0.84 \times 4.94/3.46 = 1.20.$$

- For the observed data,
  - the sample means are $\bar{y} = 135.68$ and $\bar{x} = 5.90$.
- Therefore,

$$a = 135.68 - 1.20 \times 5.90 = 128.60.$$

- The linear regression model can be written as

$$\hat{y} = 128.60 + 1.20x.$$

Example

- We can now use this model to estimate the value of the response variable.
- For the individual i in the right panel of the figure in slide 16,
  - the amount of daily sodium chloride intake is $x_i = 3.68$.
- The estimated value of the blood pressure for this person is

$$\hat{y}_i = 128.60 + 1.20 \times 3.68 = 133.02.$$

- The actual blood pressure for this individual is

$$y_i = 128.3$$

- The residual therefore is

$$e_i = y_i - \hat{y}_i = 128.3 - 133.02 = -4.72$$

Example

- We can also use our model for predicting the unknown values of the response variable (i.e., blood pressure) for all individuals in the target population.
  - For example, if we know the amount of daily sodium chloride intake is $x = 7.81$ for an individual, we can predict her blood pressure as follows:

$$\hat{y} = 128.60 + 1.20 \times 7.81 = 137.97$$

- Of course, the actual value of the blood pressure for this individual would be different from the predicted value.
  - The difference between the actual and predicted values of the response variable is called the model **error** and is denoted as ε.
    - In fact, the residuals are the observed values of ε for the individuals in our sample.

Estimating model parameters

- As an alternative way, the least-squares estimates of slope and intercept can be obtained as follows:

$$\beta = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \alpha = \overline{y} - \beta\overline{x}$$

where

$S_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y})$ and $S_{xx} = \sum_i (x_i - \overline{x})^2$

- Thus, S_{xy} is the sum of x deviations times y deviations and S_{xx} is the sum of x deviations squared.

Example

- In the road resurfacing example
  - Cost y_i (in thousands of dollars): 6.0 14.0 10.0 14.0 26.0
  - Mileage x_i (in miles): 1.0 3.0 4.0 5.0 7.0

- For the road resurfacing data, n =5 and $\sum x_i = 1.0 + 3.0 + 4.0 + 5.0 + 7.0 = 20.0$
- So $\bar{x} = \frac{20.0}{5} = 4.0.$
- Similarly $\sum y_i = 70.0, \bar{y} = \frac{70.0}{5} = 14.0$
- Also,

$$S_{xx} = \sum_i (x_i - \bar{x})^2 = (1.0 - 4.0)^2 + \cdots \ldots + (7.0 - 4.0)^2 = 20.00$$

Example

- And

$$S_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y})$$
$$= (1.0 - 4.0)(6.0 - 14.0) + \cdots \ldots (7.0 - 4.0)(26.0 - 14.0)$$
$$= 60.0$$

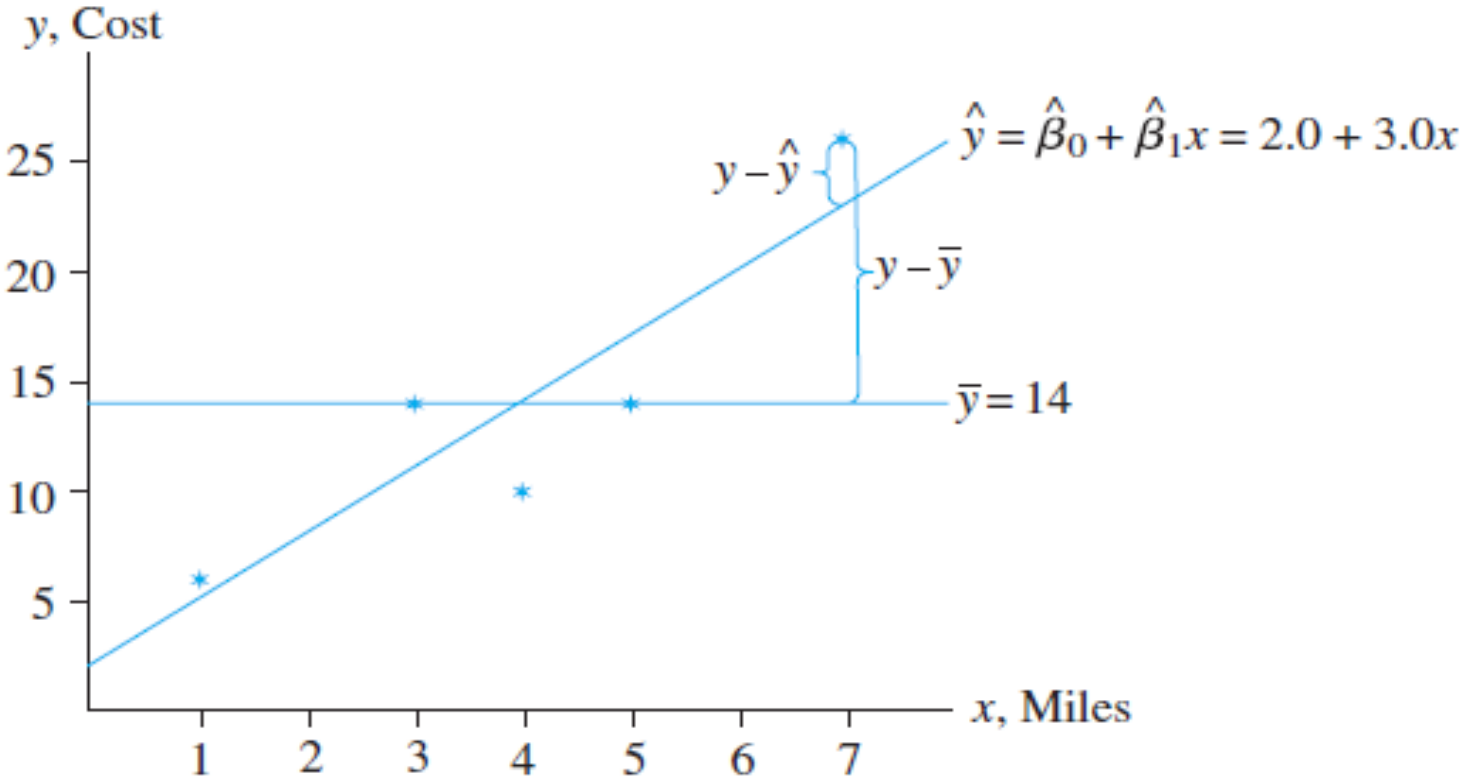
- Thus

$$\beta = \frac{60.0}{20.0} = 3.0$$
$$\alpha = 14.0 - (3.0)(4.0) = 2.0$$

- From the value $= 3.0$, we can conclude that the estimated average increase in cost for each additional mile is \$3,000.

Example

- Deviations from the least-squares line from the mean



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.

Confidence Interval for Regression Coefficients

- 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.

Confidence Interval for Regression Coefficients

- For the blood pressure example,
  - the sample size is $n = 25$.
- Therefore, we use the t -distribution with $25 - 2 = 23$ degrees of freedom.
- If we set the confidence level to 0.95,
  - then $t_{crit} = 2.07$,
    - which is obtained from the t -distribution with 23 degrees of freedom by setting the upper tail probability to $(1-0.95)/2 = 0.025$.
- Therefore,
  - the 95% confidence interval for β is

$$[6.25-2.07\times1.59,\ 6.25+2.07\times1.59] = [2.96,\ 9.55]$$

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.

$$t = \frac{b}{SE_b}$$

- Then, we find the corresponding *p*-value as follows:
  - if $H_A : \beta < 0$, $p_{obs} = P(T \leq t)$,
  - if $H_A : \beta > 0$, $p_{obs} = P(T \geq t)$,
  - if $H_A : \beta \neq 0$, $p_{obs} = 2 \times P(T \geq |t|)$,

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

Hypothesis testing

- In the blood pressure example,
  - the estimate of the regression coefficient was $b = 6.25$,
  - the standard error was $SE_b = 1.59$.
- Therefore,

$$t = b / SE_b = 6.25 / 1.59 = 3.93.$$

- If $H_A : \beta \neq 0$ (which is the common form of the alternative hypothesis),
  - we find the p-value by calculating the upper tail probability of $|3.93| = 3.93$ from the t -distribution with $25 - 2 = 23$ degrees of freedom and multiplying the result by 2.
- For this example,

$$p_{\text{obs}} = 2 \times 0.00033 = 0.00066.$$

- Because p_{obs} for this example is quite small and below any commonly used confidence level (e.g., 0.01, 0.05, 0.1), we can reject the null hypothesis and conclude that blood pressure is related to sodium chloride diet level.

Example

- Data from a sample of 10 pharmacies are used to examine the relation between prescription sales volume and percentage of prescription ingredients purchased directly from the supplier.
- The sample data are shown in the following table

Pharmacy	Sales Volume, y (in \$1,000s)	% of Ingredients Purchased Directly, x
1	25	10
2	55	18
3	50	25
4	75	40
5	110	50
6	138	63
7	90	42
8	60	30
9	10	5
10	100	55

a. Find the least-squares estimates for the regression line $\hat{y} = \alpha + \beta x$

b. Predict sales volume for a pharmacy that purchases 15% of its prescription ingredients directly from the supplier.

c. Plot the (x, y) data and the prediction equation $\hat{y} = \alpha + \beta x$

d. Interpret the value of β in the context of the problem.

Example

a. Least-squares estimates

	y	x	$y - \bar{y}$	$x - \bar{x}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
	25	10	−46.3	−23.8	1,101.94	566.44
	55	18	−16.3	−15.8	257.54	249.64
	50	25	−21.3	−8.8	187.44	77.44
	75	40	3.7	6.2	22.94	38.44
	110	50	38.7	16.2	626.94	262.44
	138	63	66.7	29.2	1,947.64	852.64
	90	42	18.7	8.2	153.34	67.24
	60	30	−11.3	−3.8	42.94	14.44
	10	5	−61.3	−28.8	1,765.44	829.44
	100	55	28.7	21.2	608.44	449.44
Total	713	338	0	0	6,714.60	3,407.60
Mean	71.3	33.8				

$$S_{xx} = \sum (x - \bar{x})^2 = 3{,}407.6 \qquad S_{xy} = \sum (x - \bar{x})(y - \bar{y}) = 6{,}714.6$$

Example

Substituting into the formulas for α and β

$$\beta = \frac{s_{xy}}{s_{xx}} = \frac{6714.6}{3407.6} = 1.97$$

$$\alpha = \bar{y} - \beta\bar{x} = 71.3 - 1.97 \times 33.8 = 4.7$$

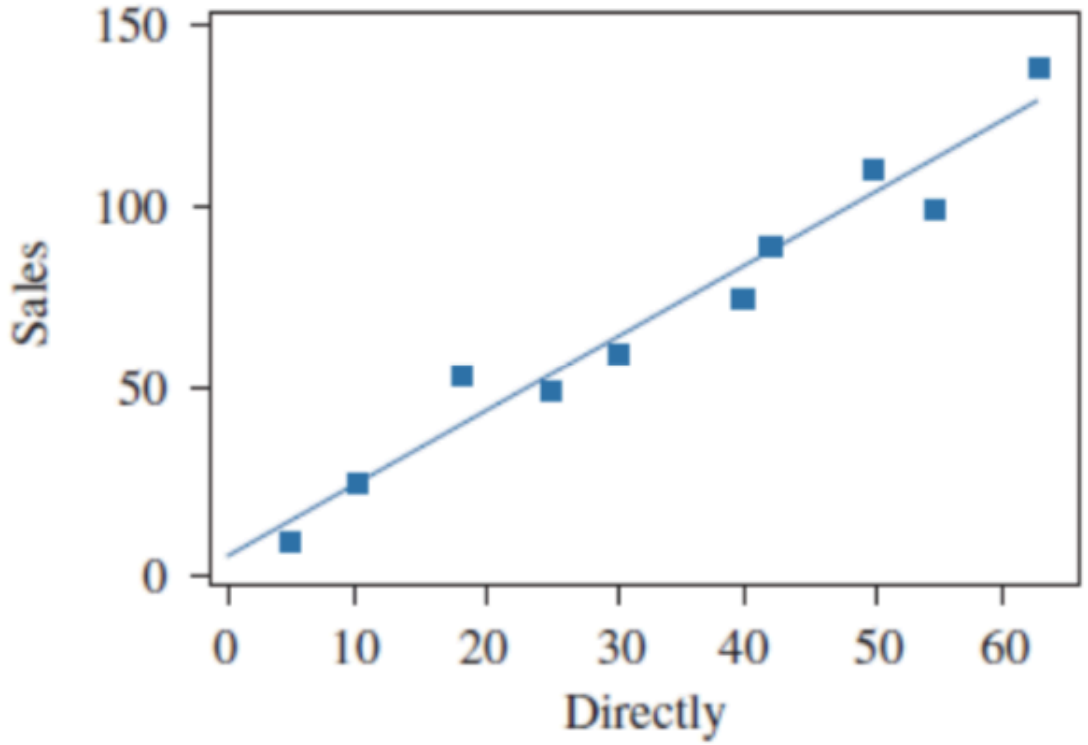
b. When x = 15%, the predicted sales volume is

$$\hat{y} = 4.7 + 1.97 \times 15 = 34.25$$

(that is, \$34,250).

c. The (x, y) data and prediction line are plotted in the next slide:

Example



d. From $\beta = 1.97$, we conclude that if a pharmacy would increase by 1% the percentage of ingredients purchased directly, then the estimated increase in average sales volume would be \$1,970.