Statistical Data Analysis

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

1

Regression Analysis

2

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}

3

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.

4

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.

5

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.

6

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.

7

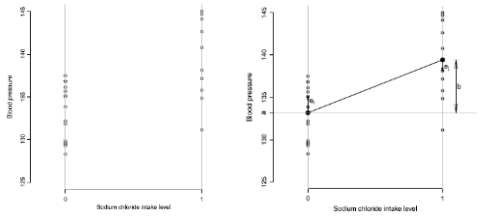
Linear Regression Models with One Binary Explanatory Variable

- 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.

8

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 $b = 6.25$



9

Linear Regression Models with One Binary Explanatory Variable

- Using the intercept $\boldsymbol{a}$ and slope $\boldsymbol{b}$, 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.

10

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:

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

11

Linear Regression Models with One Binary Explanatory Variable

  - For instance, if someone belongs to the first group, her estimated blood pressure is

$$\hat{y}_i = \text{a} = 133.17$$

  - 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

$$y_i = a + bx_i + e_i$$

12

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., $x = 1$),

$$\hat{y} = 133.17 + 6.25 \times x$$
$$= 133.17 + 6.25 \times 1$$
$$= 139.42$$

Residual sum of squares

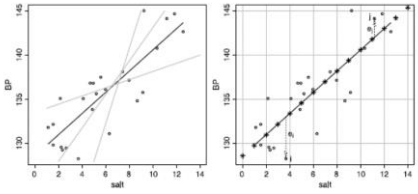
- As a measure of discrepancy between the observed values and those estimated by the line, we calculate the **R**esidual **S**um of **S**quares (*RSS*):

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

- 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.



- *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,
      - $b < 0$ 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:

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

  where $\overline{y}$ and $\overline{x}$ 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 $\overline{y} = 135.68$ and $\overline{x} = 5.90$.
- 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 $\overline{x} = \frac{20.0}{5} = 4.0$.
- Similarly $\sum y_i = 70.0, \overline{y} = \frac{70.0}{5} = 14.0$
- Also,

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

Example

- and

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

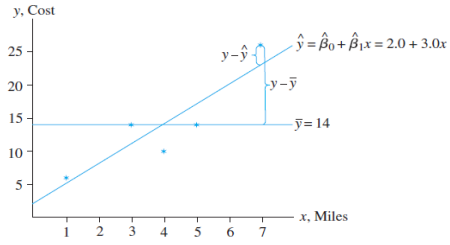
- Thus

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

- From the value $\beta = 3.0$, 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 $t_{\text{crit}} = 2.07$,
    - 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

$$[6.25 - 2.07 \times 1.59,\ 6.25 + 2.07 \times 1.59] = [2.96,\ 9.55]$$

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 $H_A : \beta < 0$, $p_{\text{obs}} = P(T \le t)$,
  - if $H_A : \beta > 0$, $p_{\text{obs}} = P(T \ge t)$,
  - if $H_A : \beta \ne 0$, $p_{\text{obs}} = 2 \times P(T \ge |t|)$,

  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 $H_A : \beta \ne 0$ (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

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

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

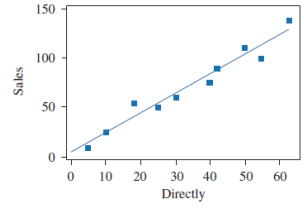
b. When x = 15%, the predicted sales volume is

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

(that is, $34,250).

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

Example



d. From β = 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.