Basics of Statistics II

Lecture # 5

Linear Regression

In linear regression, the dependent variable, y_i, is a linear combination of the parameters (but need not be linear in the independent variables). For modeling n data points there is one independent variable xi, and two parameters, β_o and β_1

straight line: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n.$

In multiple linear regression, there are several independent variables or functions of independent variables

parabola: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \ i = 1, \ldots, n.$

ε_i is an error term with i indexes for a particular observation.

For a random sample we estimate the population parameters and obtain the sample linear regression model:

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i.$$

The residual corresponding to the difference between the value of the dependent variable predicted by the model and the true value of the dependent variable is

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

We estimate the parameters so that the sum of the squares of the error is minimized

$$SSR = \sum_{i=1}^{n} e_i^2.$$

By minimizing this relation $\widehat{\beta}_0, \widehat{\beta}_1$. are estimated.

The formula for the least squares solution is:

$$\widehat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \text{ and } \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

Where $\bar{x}$ is the mean of the x values and $\bar{y}$ the mean of the y values.

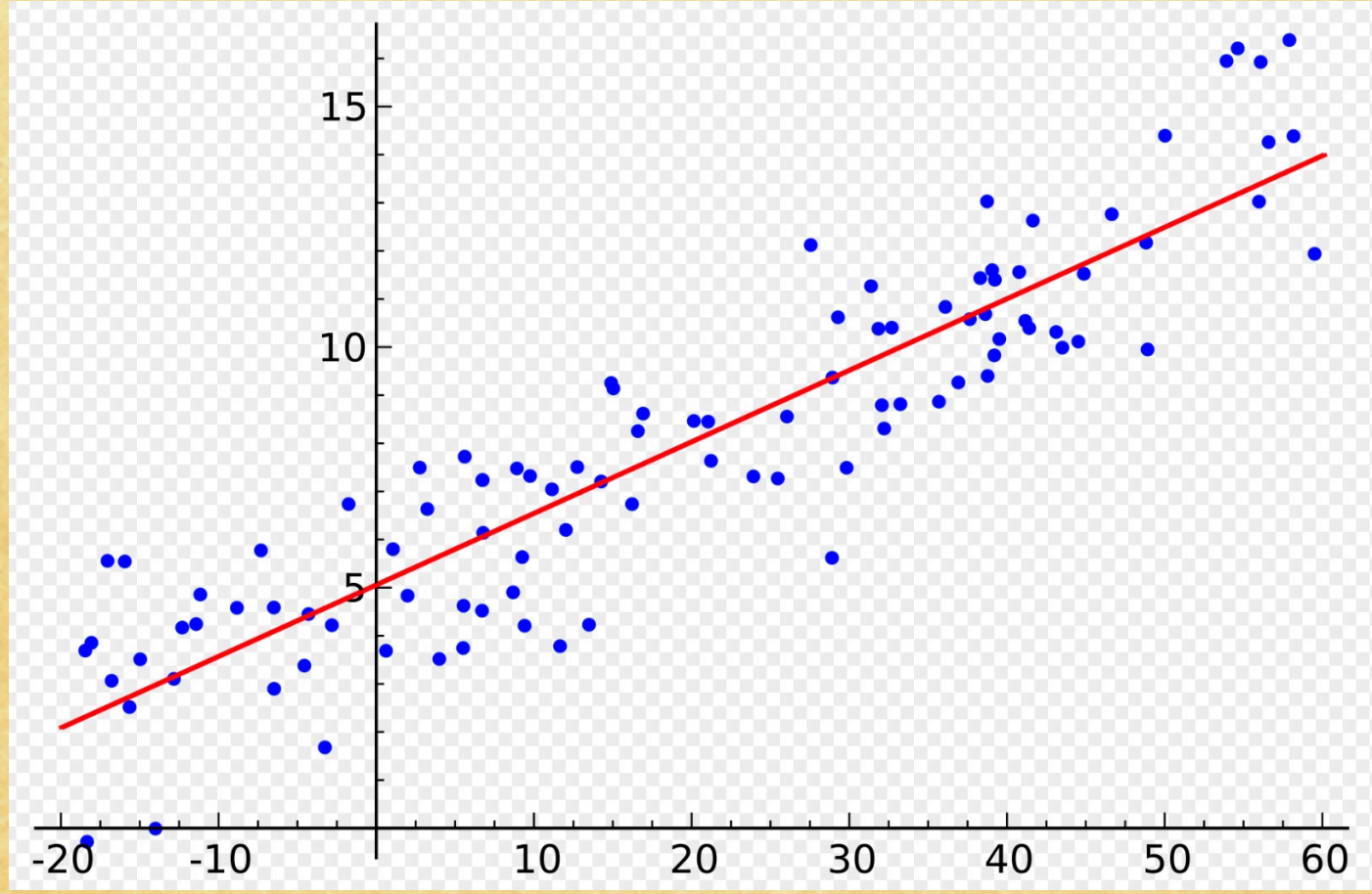
The standard errors of the parameters are:

$$\hat{\sigma}_{\beta_1} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}$$

$$\hat{\sigma}_{\beta_0} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}} = \hat{\sigma}_{\beta_1} \sqrt{\frac{\sum x_i^2}{n}}.$$

With the variance $$\hat{\sigma}_\varepsilon^2 = \frac{SSR}{n-2}.$$

Regression



Multivariate Regression

In multiple regression models where there are p independent variables

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

Where x_{ij} is the i^{th} observation on the j^{th} independent variable. If the first independent variable takes the value 1 for all I, $x_{i1} = 1$, then β_1 is called the regression intercept. The least squares parameter estimates are obtained from *p* normal equations with the residual as

$$\varepsilon_i = y_i - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}.$$

The equations are then

$$\sum_{i=1}^{n} \sum_{k=1}^{p} x_{ij} x_{ik} \hat{\beta}_k = \sum_{i=1}^{n} x_{ij} y_i, \ j = 1, \ldots, p.$$

In matrix notation:

$$(\mathbf{X}^\top \mathbf{X}) \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{Y},$$

Where the ij elements of x are x_{ij}, the ith element of column vector Y is y_i. X is *nxp*; Y is nx1. The solution is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Method of the Least Squares

The least squares method is a standard approach in regression analysis to approximate the solution of overdetermined systems. This is a set of equations in which there are more equations than unknowns. The most important application of this is in data fitting. The best fit minimizes the sum of squared residuals (a residual being the difference between an observed value and the fitted value provided by a model.

Objective: adjusting the parameters of a model function to best fit a data set. A dataset consists of (x_i,y_i), i=1,…n where x_i is the independent variable and yi is a dependent variable whose values are found by observation. The model function has the form f (x,b) where m adjustable parameters are held in the vector b. The aim here is to find the parameters that best fit the data. This is to minimize the residual

$$r_i = y_i - f(x_i, \boldsymbol{\beta}).$$

The least squares method finds the optimal parameters by minimizing

$$S = \sum_{i=1}^{n} {r_i}^2.$$

The y-intercept b_0 and slope b_1 need to be calculated for the model function

$$f(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x.$$

The minimum of S is found by setting its gradient to zero

$$\frac{\partial S}{\partial \beta_j} = 2\sum_i r_i \frac{\partial r_i}{\partial \beta_j} = 0,\ j = 1, \ldots, m,$$

and since $r_i = y_i - f(x_i, \boldsymbol{\beta})$, the gradient equations become

$$-2\sum_i r_i \frac{\partial f(x_i, \boldsymbol{\beta})}{\partial \beta_j} = 0,\ j = 1, \ldots, m.$$

intercept can be found by solving this equation system.

Bivariate Data and Pearson Correlation

The Pearson correlation coefficient is a measure of the strength of the relation between two variables. It is referred to as Pearson correlation or correlation coefficient. If the relation is not linear then the coefficient does not adequately represent the strength of the correlation. Between two variables x and y the correlation coefficient is

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

Where r changes between 0 and 1.

Limitations

Regression for prediction: a model is fitted to provide a prediction rule for application in a similar situation to which the data used for fitting apply. Here the errors in the dependent variable are subject to the same types of observational errors as the data used for fitting.

Regression for fitting a "true relationship": in standard regression analysis there is the implicit assumption that errors in the independent variable are zero or well controlled. When this is not the case, models of measurement errors can be used which would lead to parameter estimates, hypothesis testing and confidence intervals.

Chi-squared (χ^2) Distribution

A standard normal deviate is a random sample from the standard normal distribution. The Chi Square distribution is the distribution of the sum of squared standard normal deviates.

The degrees of freedom of the distribution is equal to the number of standard normal deviates being summed. Therefore, Chi Square with one degree of freedom, written as $\chi 2(1)$, is simply the distribution of a single normal deviate squared.

Example: you sample two scores from a standard normal distribution, square each score, and sum the squares. What is the probability that the sum of these two squares will be six or higher?

Since two scores are sampled, the answer can be found using the Chi Square distribution with two degrees of freedom. A Chi Square calculator can be used to find that the probability of a χ^2 (with 2 df) being six or higher is 0.050.

χ^2 distributions are positively skewed, with the degree of skew decreasing with increasing degrees of freedom. As the degrees of freedom increases, the χ^2 distribution approaches a normal distribution.

Two of the more common tests using the χ^2 distribution are tests of deviations of differences between theoretically expected and observed frequencies (one-way tables) and the relationship between categorical variables (contingency tables).

The χ^2 distribution can be used to test whether observed data differ significantly from theoretical expectations.

Example

We first conduct the significant test to compute the expected frequency for each outcome given that the null hypothesis is true. The expected frequency of each of the numbers is: (1/6) (36) = 6 - given that the probability of each number is 1/6 and we drew 36 rolls of the die. If E is the expected frequency of an outcome and O be the observed frequency of that outcome, we compute $(E - O)^2/E$ for each outcome. We then add up all the values

$$\sum \frac{(E-O)^2}{E} = 5.33$$

The LHS is approximately distributed as a Chi square with k-1 degrees of freedom where k is the number of categories. The χ^2 with 5 degrees of freedom is 5.33.

The probability of a c2 of 5.33 or larger is 0.377 (from c2 table). Therefore the probability that the die is fair cannot be rejected.

Freq of die

Outcome	Frequency
1	8
2	5
3	9
4	2
5	7
6	5

The data in the table were obtained by rolling a six-sided die 36 times. Some outcomes occurred more frequently than others- e.g. a “3” came up nine times, whereas a “4” came up only two times. Are these data consistent with the hypothesis that the die is a fair die?

We need to conduct a significance test. The null hypothesis is that the die is fair. This hypothesis is tested by computing the probability of obtaining frequencies as discrepant or more discrepant from a uniform distribution of frequencies as obtained in the sample. If this probability is sufficiently low, then the null hypothesis that the die is fair can be rejected.

Joint Probability Distribution

Given random variables X,Y,…, the joint probability distribution for X,Y,… is a probability distribution that gives the probability that each of X, Y,… falls in any particular range or discrete set of values specified for that variable. In case of only two random variables, this is called bivariate distribution. The generalization of this to any number of random variables gives a multivariate distribution.

Conditional probability distribution gives the probabilities for any subset of the variables conditional on particular values of the remaining variables.

Example: Consider the flip of two coins with A and B being the discrete random variables associated with the outcomes of the coins. Each coin flip is a Bernoulli trial with a Bernoulli distribution. If a coin displays "heads" then the associated random variable takes the value 1, and it takes the value 0 otherwise. The probability of each of these outcomes is ½. The marginal (unconditional) density functions are

$P(A)=1/2$ for $A \in \{0,1\}$ $P(B) = 1/2$ for $B \in \{0,1\}$

The joint probability density function of A and B defines probabilities for each pair of outcomes

(A=0, B=0), (A=0, B=1), (A=1, B=0), (A=1, B=1)

Since each outcome is equally likely, the joint probability density function becomes $P(A,B) = ¼$ for $A, B \in \{0,1\}$

Since the coin flips are independent, the joint probability density function is the product of the marginal

$P(A,B) = P(A)P(B)$ for $A,B \in \{0,1\}$

Null Hypothesis and Significant Level

Null Hypothesis:

The hypothesis that an apparent effect is due to chance is called the null hypothesis.

A low probability value casts doubt on the null hypothesis. How low must the probability value be in order to conclude that the null hypothesis is false? Although there is clearly no right or wrong answer to this question, it is conventional to conclude the null hypothesis is false if the probability value is less than 0.05. More conservative researchers conclude the null hypothesis is false only if the probability value is less than 0.01.

When a researcher concludes that the null hypothesis is false, the researcher is said to have rejected the null hypothesis. The probability value below which the null hypothesis is rejected is called the α level or simply α. It is also called the significance level.

When the null hypothesis is rejected, the effect is said to be statistically significant.

Covariance

Covariance is a measure of the joint variability of two random variables. If the greater value of one variable corresponds to the greater value of the other variable, and the same holds for the lesser values, the covariance is positive.

The covariance between two jointly distributed real-valued random variables X and Y is defined as the expected product of their deviations from their individual expected values

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

Where E[X] is the expected value of X also known as the mean of X.

$$Cov(X,Y) = E[XY - X E[Y] - E[X] Y + E[X] E[Y]] = E[XY] - E[X] E[Y] - E[X] E[Y] + E[X] E[Y] =$$

$$E[XY] - E[X] E[Y]$$

$$Cov(X,X) = var(X) = s2(X)$$

Covariance of Linear Combinations

$$
\begin{aligned}
\operatorname{cov}(X, a) &= 0 \\
\operatorname{cov}(X, X) &= \operatorname{var}(X) \\
\operatorname{cov}(X, Y) &= \operatorname{cov}(Y, X) \\
\operatorname{cov}(aX, bY) &= ab\ \operatorname{cov}(X, Y) \\
\operatorname{cov}(X + a, Y + b) &= \operatorname{cov}(X, Y) \\
\operatorname{cov}(aX + bY, cW + dV) &= ac\ \operatorname{cov}(X, W) + ad\ \operatorname{cov}(X, V) + bc\ \operatorname{cov}(Y, W) + bd\ \operatorname{cov}(Y, V)
\end{aligned}
$$

Covariance Matrix

$$\mathrm{K}_{\mathbf{XX}} = \begin{bmatrix} \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_1 - \mathrm{E}[X_1])] & \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_2 - \mathrm{E}[X_2])] & \cdots & \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_n - \mathrm{E}[X_n])] \\ \mathrm{E}[(X_2 - \mathrm{E}[X_2])(X_1 - \mathrm{E}[X_1])] & \mathrm{E}[(X_2 - \mathrm{E}[X_2])(X_2 - \mathrm{E}[X_2])] & \cdots & \mathrm{E}[(X_2 - \mathrm{E}[X_2])(X_n - \mathrm{E}[X_n])] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mathrm{E}[X_n])(X_1 - \mathrm{E}[X_1])] & \mathrm{E}[(X_n - \mathrm{E}[X_n])(X_2 - \mathrm{E}[X_2])] & \cdots & \mathrm{E}[(X_n - \mathrm{E}[X_n])(X_n - \mathrm{E}[X_n])] \end{bmatrix}$$

The definition above is equivalent to the matrix equality

$$\mathrm{K}_{\mathbf{XX}} = \mathrm{cov}[\mathbf{X}, \mathbf{X}] = \mathrm{E}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^{\mathrm{T}}] = \mathrm{E}[\mathbf{X}\mathbf{X}^T] - \mu_{\mathbf{X}}\mu_{\mathbf{X}}{}^T \quad \textit{(Eq.1)}$$

where $\mu_{\mathbf{X}} = \mathrm{E}[\mathbf{X}]$.

Conditional Probability

Previously, we defined the *conditional probability* of two events A and B as follows:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Let these events be described by the random variable $X = x$ and $Y = y$. Then we can write:

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f(x, y)}{h(y)}$$

where $f(x, y)$ is the joint probability distribution of X and Y and $h(y)$ is the marginal marginal distribution of y.

Joint Probability Distribution

From Miles Osborne (School of Informatics, Edinburgh University)

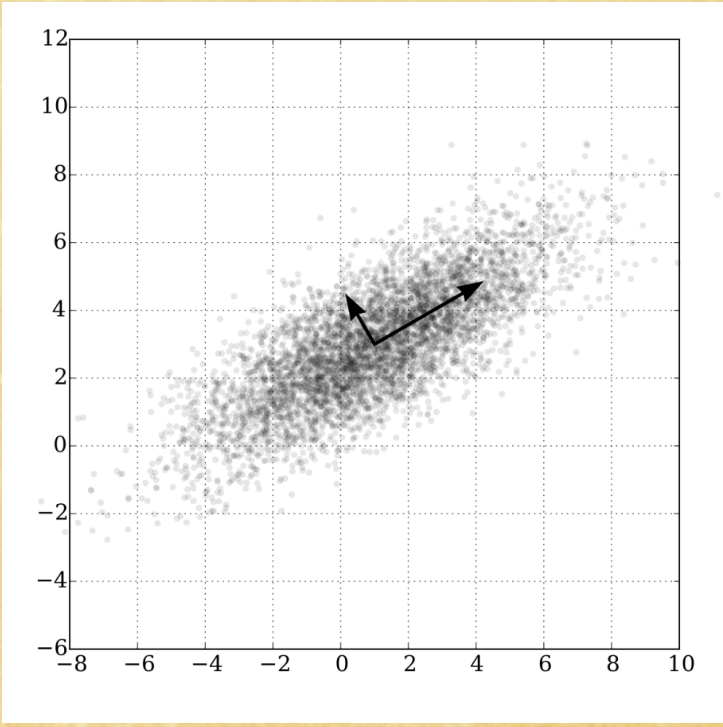
If $f(x, y)$ is the value of the joint probability distribution of the discrete random variables X and Y at (x, y) and $h(y)$ is the value of the marginal distributions of Y at y, and $g(x)$ is the value of the marginal distributions of X at x, then:

$$f(x|y) = \frac{f(x,y)}{h(y)} \quad \text{and} \quad w(y|x) = \frac{f(x,y)}{g(x)}$$

are the conditional distributions of X given $Y = y$, and of Y given $X = x$, respectively (for $h(y) \neq 0$ and $g(x) \neq 0$).

Bivariate Normal Distribution

Because the x and y components co-vary, the variances of x and y do not fully describe the distribution. A 2 x 2 matrix is needed. The direction of the arrows correspond to the eigenvectors of this covariance matrix and their length to the square root of the eigenvalues.

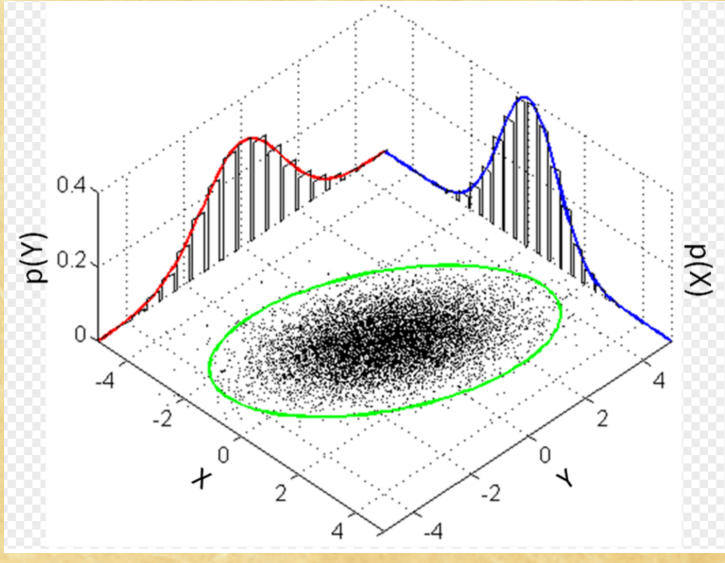


Multivariate Normal Distribution

Definition

The multivariate normal distribution is a generalization of the one-dimensional normal distribution to higher dimensions. A random vector is k-variate normally distributed if every linear combination of its k components has a univariate normal distribution. The multi-variate normal distribution is often used to describe any set of correlated real random variables each of which clustered around a mean value

Probability Density Function



Bayes Theorem

Bayes' theorem is the probability of an event, based on prior knowledge of conditions that might be related to the event. It is stated as

$P(A \mid B) = P(B \mid A)\, P(A) / P(B)$

Where A and B are events and P(B)

P(A | B) is conditional probability: the likelihood of event A occurring given that B is true.

P(B | A) is conditional probability: the likelihood of event B occurring given that A is true

P(A) and P(B) are the probabilities of observing A and B independently of each other. This is known as marginal probability