Second course in Statistics. Lecture 23.

Multiple linear regression

- Model specification and estimation
- Statistical inference
- Residual analysis
- Comparing models: analysis of variance

Multiple linear regression

The statistical model for multiple linear regression is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad \text{for} \quad i = 1, 2, \cdots, n$$

The mean response μ_y is a linear function of the explanatory variables:

$$\hat{Y} = \mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

This equation describes how the mean of y varies with the x's.

ε_i are assumed to be independent and normally distributed with mean 0 and standard deviation σ.

The parameters of the model are $\beta_0, \beta_1, \cdots \beta_p$, and σ^2.

Estimation

Least squares estimation

The method of least squares chooses the values of $\hat{\beta}_i$ that make the sum of squares of the residuals as small as possible. In other words the parameter estimates $\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_p$ minimized the quantity

$$\sum\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip} \right)^2$$

The derivation is much more complicated than in simple linear regression.

Estimator of σ^2 is $s^2 = \dfrac{\sum\left(y_i - \hat{y}_i \right)^2}{n - p - 1}$

Statistical inference

$100(1-\alpha)\%$ Confidence interval for β_i is

$$\hat{\beta}_i \pm t_{\alpha/2}(n-p-1)SE_{b_j}$$

Where SE_{b_j} is the standard error of $\hat{\beta}_i$

To test the hypothesis:

$$H_0 : \beta_j = 0 \qquad j = 1, \ldots p$$

t statistic is

$$t = \hat{\beta}_j / SE_{\hat{\beta}_j}$$

Reject H_0 if $|t| > t_{\alpha/2}(n-p-1)$

Analysis of variance

For multiple regression, the analysis of variance is very rich technique that is used to divide variability and to compare models that include different sets of variables.

In the overall analysis of variance, the full model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

is compared to the model with no x variables,

$$y_i = \beta_0 + \varepsilon_i$$

Equivalently:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

$H_0 :$ at least one of β_j is not equal 0

Analysis of variance

Source	d.f.	SS	MS
Variance due to regression model	p	$SSM = \sum\left(\hat{y}_i - \bar{y}\right)^2$	$MSM = \dfrac{SSM}{p}$
Error (residual)	$n - p - 1$	$SSE = \sum\left(y_i - \hat{y}_i\right)^2$	$MSE = \dfrac{SSE}{n - p - 1}$
Total	$n - 1$	$SST = \sum\left(y_i - \bar{y}\right)^2$	

Test statistic: $F = \dfrac{MSM}{MSE}$ with $\nu_1 = p$ and $\nu_2 = n - p - 1$

R.R. Reject H_0 if $F > F_\alpha(\nu_1, \nu_2)$

Coefficient of determination

Coefficient of determination (a.k.a. squared multiple correlation) is defined

$$R^2 = \frac{SSM}{SST} = \frac{\sum\left(\hat{y}_i - \overline{y}\right)^2}{\sum\left(y_i - \overline{y}\right)^2}$$

This statistic is the proportion of the variation of the response variable y that is explained by the explanatory variables $x_1, x_2, \ldots x_p$ in a multiple linear regression.

Example

Suppose we had also recorded the age of each student in the sample. Since a company may reward some experience that an older graduate might have, it is possible that the age of a graduate might influence the average starting salary. The data is augmented as the following table:

Salary	18,5	20	21,1	22,4	21,2	15	18	18,8	15,7	14,4	15,5	17,2	19	17,2	16,8
GPA	2,95	3,2	3,4	3,6	3,2	2,85	3,1	2,85	3,05	2,7	2,75	3,1	3,15	2,95	2,75
Age	22	23	23	23	27	22	25	28	23	22	28	22	26	23	26

Example

We include the linear effect of x_2 in the regression model and fit:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i .$$

Where Y_i is starting salary, x_1 is GPA and x_2 is age.

LSE equation: $\hat{Y}_i = -16.88 + 8.74 x_{i1} + 0.338 x_{i2}$

(5.476) (1.221) (0.137)

The values in bracket are the standard error for the estimated statistics.

Test null hypothesis that

$$\beta_0 = 0, \ \beta_1 = 0, \ \beta_2 = 0$$

Respectively.

Example

According to the estimated equation

(a) Calculate sum of squares due to linear regression $\sum\left(\hat{y}_i - \bar{y}\right)^2$

(b) Calculate sum of squares residual $\sum\left(y_i - \hat{y}_i\right)^2$

(c) Estimate variance of random error σ^2.

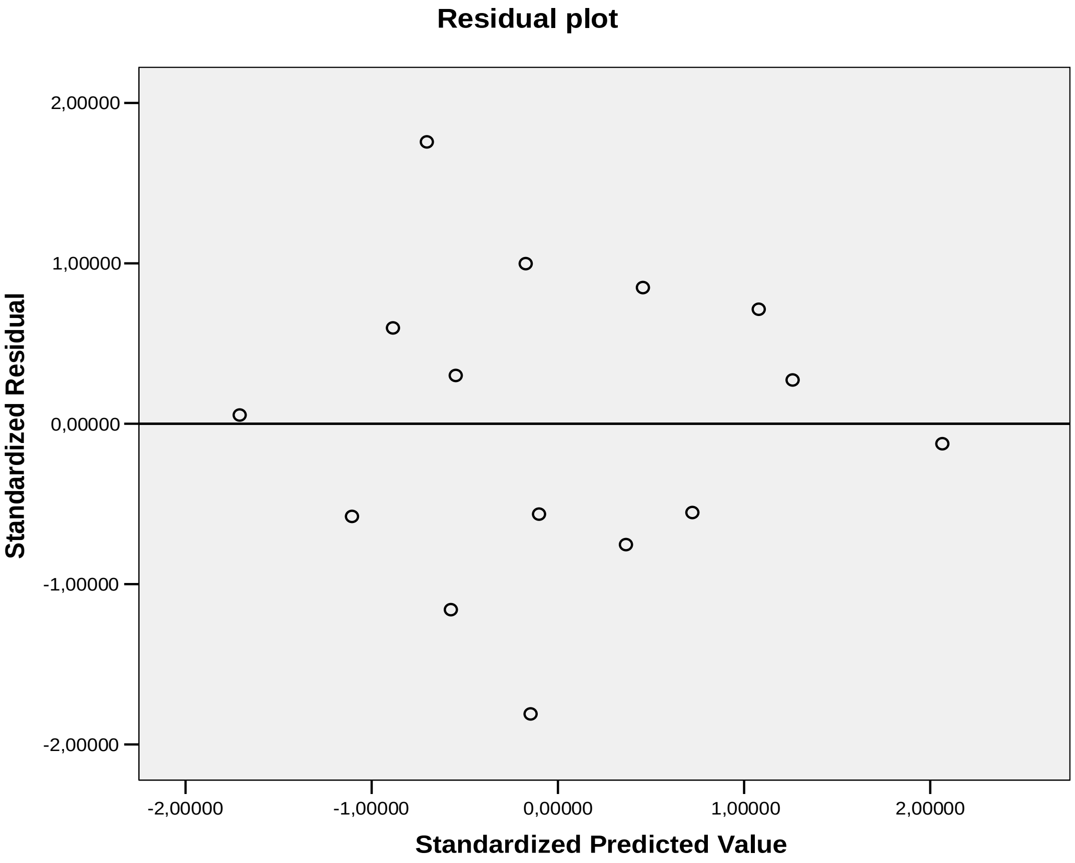
ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	66,099	2	33,050	26,130	,000[a]
	Residual	15,178	12	1,265		
	Total	81,277	14			

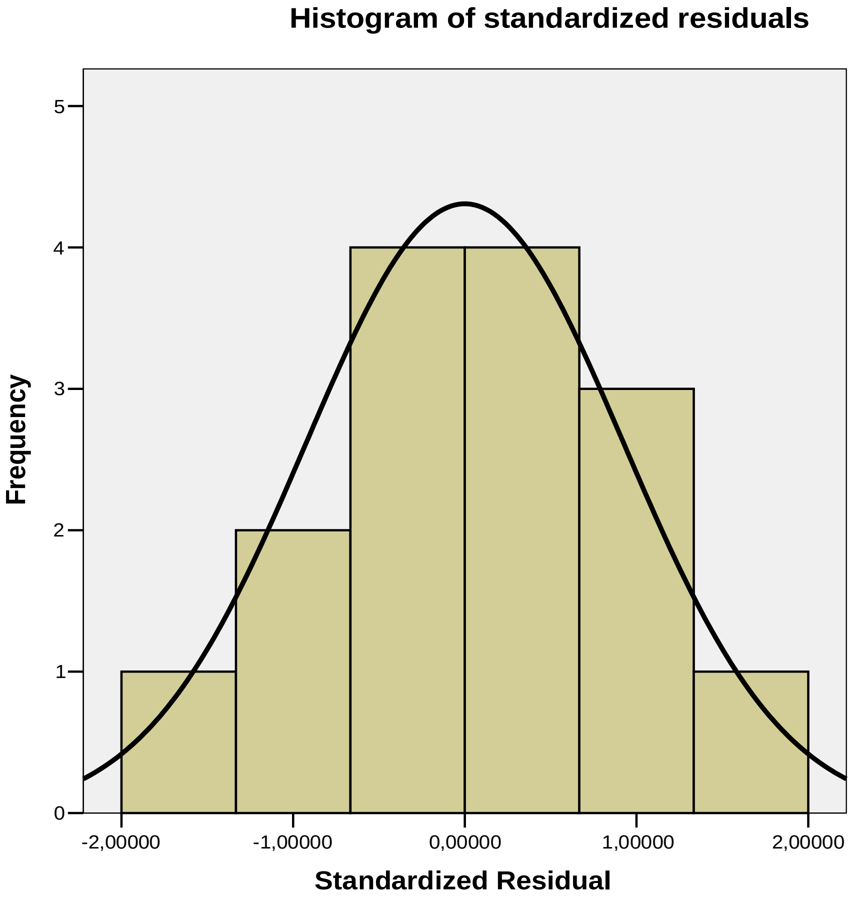
a. Predictors: (Constant), age, GPA

b. Dependent Variable: Salary

Residual analysis



Residual analysis



Example

Test the null hypothesis that $H_0 : \beta_1 = \beta_2 = 0$ in regression

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i .$$

Calculate coefficient of determination under the model

Analyze which one of $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ and $Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$ is better by

(a) comparing estimate of variance of random error

(b) comparing coefficient of determination

(c) ANOVA

(d) analyzing the residuals