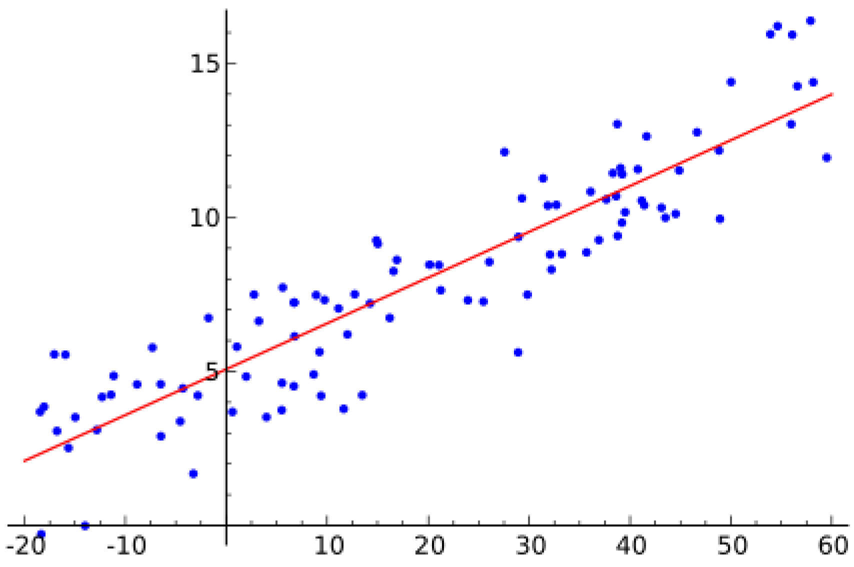
Lecture 11

Correlation and Regression

Overview of the Correlation and Regression Analysis



The Correlation Analysis

- In statistics, **dependence** refers to any statistical relationship between two random variables or two sets of data. **Correlation** refers to any of a broad class of statistical relationships involving dependence.
- Familiar examples of dependent phenomena include the correlation between the physical statures of parents and their offspring, and the correlation between the demand for a product and its price.
- Correlations are useful because they can indicate a predictive relationship that can be exploited in practice.
  - For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. In this example there is a causal relationship, because extreme weather causes people to use more electricity for heating or cooling; however, statistical dependence is not sufficient to demonstrate the presence of such a causal relationship (i.e., Correlation does not imply causation).

Pearson ’s Contribution to Statistics

- Pearson's work was all-embracing in the wide application and development of mathematical statistics, and encompassed the fields of biology, epidemiology, anthropometry, medicine and social history. In 1901, with Weldon and Galton, he founded the journal **Biometrika** whose object was the development of statistical theory.
- **Pearson ’s Correlation coefficient:** defined as the covariance of the two variables divided by the product of their standard deviations.
- **Method of moments:** Pearson introduced moments, a concept borrowed from physics, as descriptive statistics and for the fitting of distributions to samples.
- Foundations of the statistical hypothesis testing theory and the statistical decision theory.
- **Pearson's chi-squared test**: A hypothesis test using normal approximation for discrete data.
- **Principal component analysis:** The method of fitting a linear subspace to multivariate data by minimizing the chi distances.

Karl Pearson (1857-1936)

The Regression Analysis

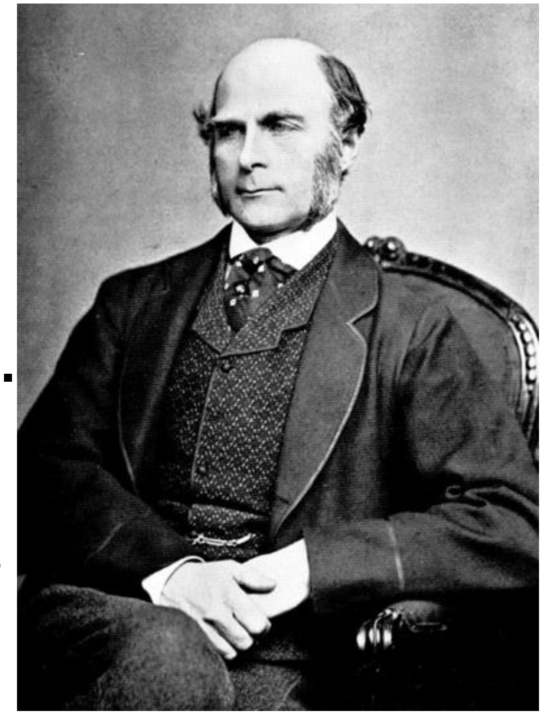
- In statistics, **regression analysis** is a statistical technique for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a **dependent variable** and one or more **independent variables**.
- More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.
- Regression analysis is widely used for **prediction** and **forecasting**, where its use has substantial overlap with the field of **machine learning**.
- Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.

History of Regression

- The earliest form of regression was the **method of least squares**, which was published by Legendre in 1805, and by Gauss in 1809. Gauss published a further development of the theory of least squares in 1821, including a version of the **Gauss–Markov theorem**.
- The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as **regression toward the mean**
- In the 1950s and 1960s, economists used electromechanical desk calculators to calculate regressions. Before 1970, it sometimes took up to 24 hours to receive the result from one regression.
- Regression methods continue to be an area of active research. In recent decades, new methods have been developed for **robust regression**, regression involving correlated responses such as time series and growth curves, regression in which the predictor or response variables are curves, images, graphs, or other complex data objects, regression methods accommodating various types of missing data, nonparametric regression, Bayesian methods for regression, regression in which the predictor variables are measured with error, regression with more predictor variables than observations, and causal inference with regression.

Galton's Contribution to Correlation and Regression

- was an English Victorian polymath: anthropologist, eugenicist, tropical explorer, geographer, inventor, meteorologist, proto-geneticist, psychometrician, and **statistician**.
- Galton produced over 340 papers and books. **He also created the statistical concept of correlation and widely promoted regression toward the mean.** He was the first to apply statistical methods to the study of human differences and inheritance of intelligence, and introduced the use of questionnaires and surveys for collecting data on human communities, which he needed for genealogical and biographical works and for his anthropometric studies.
- He was a pioneer in eugenics, coining the term itself and the phrase "nature versus nurture". His book Hereditary Genius (1869) was the first social scientific attempt to study genius and greatness



Sir Francis Galton (1822-1911)

Hereditary Stature by F. Galton (1886)

HEREDITARY STATURE [1]

IT will perhaps be recollected that, at the meeting last autumn of the British Association in Aberdeen, I chose for my Presidential Address to the Anthropological

[1] Extracts from Mr. F. Galton's Presidential Address to the Anthropological Institute, January 26.

almost absurdly simple, and not only so, but it is explained most easily by a working model that altogether supersedes the trouble of calculation. I exhibit one of these: it is a large card ruled with horizontal lines 1 inch apart, and numbered consecutively in feet and inches, the value of 5 feet 8 inches lying about half way up. A pin-hole is bored near the left-hand margin at a height corresponding to 5 feet $8\frac{1}{4}$ inches. A thread secured at

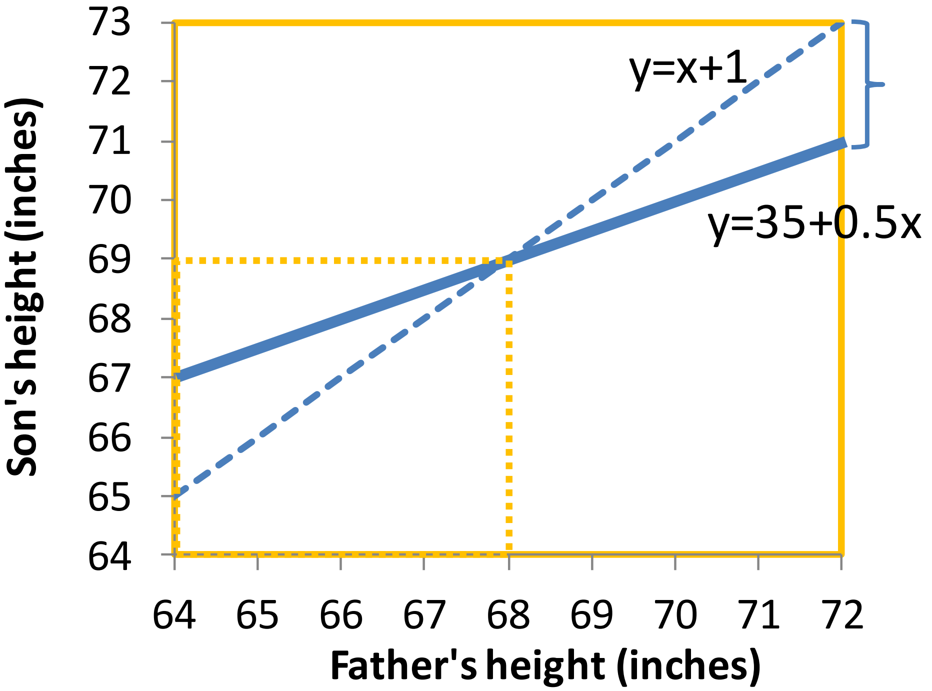
the back of the card is passed through the hole; when it

already explained, we shall see from the divisions on the

- 1078 pairs of son (y) and father (x)
- Average of sons: m(y) = 69 inches
- Average of fathers m(x) = 68 inches
- On average, taller father has taller son
- Can we use y=x+1 to predict son's stature?

Regression of son on father's height

- When grouping on fathers
  - For fathers **x=72** [4 in. taller than m(x)], **y=71** (2 in. shorter than x+1 and 1 in. shorter than x);
  - For fathers **x=64** [4 in. shorter than m(x)], **y=67** (2 in. taller than x+1 and 3 in. taller than x;



Regression of offspring on mid-parent height

- Slope from offspring and mid-parent is higher than slope from son and father!

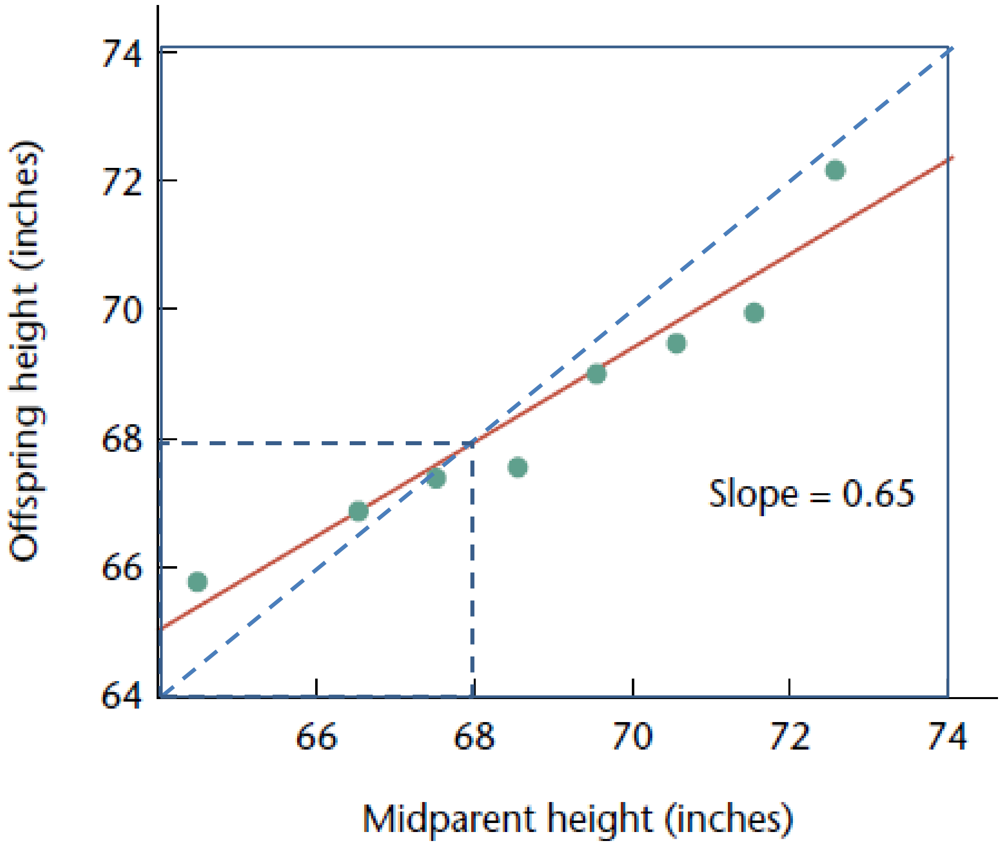


Figure 1 Galton's 1889 plot of average parental height versus average height of offspring.

Galton’s explanation of regression

- Resemblance between offspring and parents
- Regression
  - The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon.
  - The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean).

Correlation Analysis

Correlation analysis

- Correlation Analysis is the study of the relationship between two variables.
  - Scatter Plot
  - Correlation Coefficient

Scatter plot

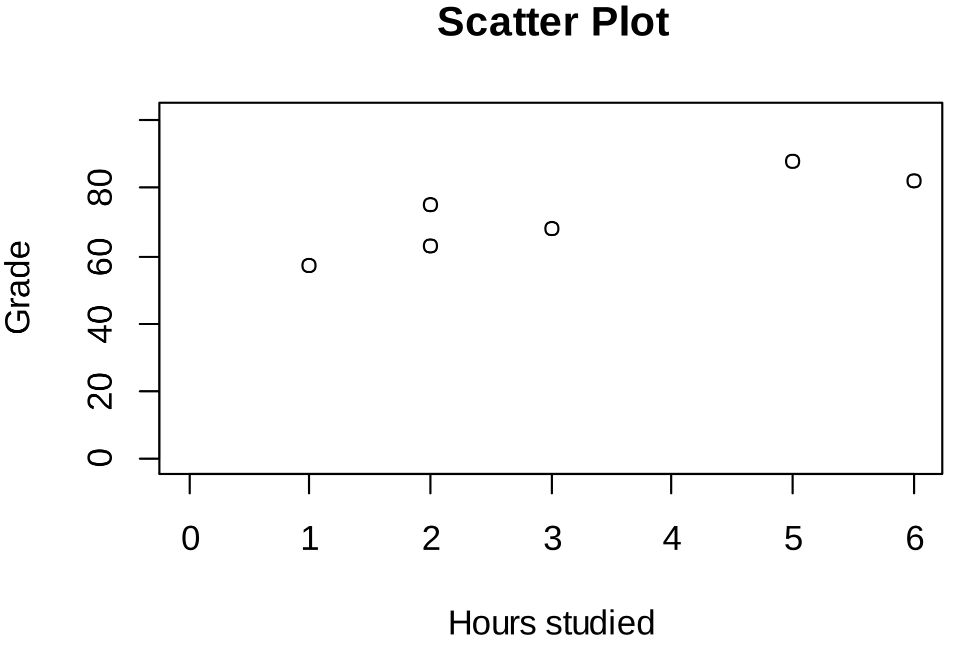
- A scatter plot is a graph of the ordered pairs (X,Y) of numbers consisting of the independent variables X and the dependent variables Y.

- It is usually the first step in correlation analysis.

Scatter plot example

- The plot shows the relationship between the grade and the hours studied of a course of six students

The graph suggests a positive relationship between hours of studies and grades



Correlation coefficient

- Measures the strength and direction of the linear relationship between two variables X and Y
- Population Correlation Coefficient:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{DX}\sqrt{DY}} = \frac{E\left[\left(X - EX\right)\left(Y - EY\right)\right]}{\sqrt{E\left[X - EX\right]^2}\sqrt{E\left[Y - EY\right]^2}}$$

- Sample Correlation Coefficient:

$$r = \frac{s_{xy}^2}{\sqrt{s_x^2 s_y^2}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

Correlation coefficient

- The range of correlation coefficient is -1 to 1.
- If $r<0$, it indicates a positive linear relationship between the two variables. (when one variable increases, the other decreases and vice versa)
- If $r>0$, it indicates a positive linear relationship between the two variables. (both variables increase or decrease at the same time)
- If $r=0$, it indicates the two variables are not related. (not necessarily independent)

Distribution of *r*

- The population correlation coefficient ρ is usually not known. Therefore, the sample statistic r is used to estimate ρ and to carry out tests of hypotheses.
- If the true correlation between X and Y within the general population is ρ=0, and if the size of the sample $n \geq 6$, then

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \rightarrow t(n-2)$$

Example

- The observations of two variables are

X	35.5	34.1	31.7	40.3	36.8	40.2	31.7	39.2	44.2
y	12	16	9	2	7	3	13	9	-1

- Then r=-0.8371. H_0: ρ=0

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.8371\sqrt{9-2}}{\sqrt{1-(-0.8371)^2}} = -4.05$$

$$t_{0.01}(7) = 3.499 < |t|$$

- So at a 99% confidence level, the null hypothesis H_0 of no relationship in the population (ρ=0) is rejected.

Correlation coefficient

- From the example, we find that the t statistic is a function of sample correlation coefficient *r*, sample size *n* and confidence level α.
- For any particular sample size, an observed value of *r* is regarded as statistically significant at the 95% level if and only if its distance from zero is equal to or greater than the distance of the tabled value of *r*.

Correlation coefficient (95% level)

n	±r	n	±r
6	0.73	19	0.39
7	0.67	20	0.38
8	0.62	21	0.37
9	0.58	22	0.36
10	0.55	23	0.35
11	0.52	24	0.34
12	0.5	25	0.34
13	0.48	26	0.33
14	0.46	27	0.32
15	0.44	28	0.32
16	0.43	29	0.31
17	0.41	30	0.31
18	0.4	31	0.3

Linear Regression Analysis

Linear regression

- Linear regression is used to study an outcome as a linear function of one or several predictors.
  - x_i: independent variables (predictors)
  - y: dependent variable (effect)
- Regression analysis with one independent variable is termed simple linear regression.
- Regression analysis with more than one independent variables is termed multiple linear regression.

Linear regression

- Given a data set $\{y_i, x_{i1}, \ldots, x_{ip}\}$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y_i and the p-vector of explanatory variables x_i is linear. This relationship is modeled through a disturbance term or error variable ε_i. Thus the model takes the form

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

Linear regression

- Often these *n* equations are stacked together and written in vector form as

Where $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Linear regression

- y_i is called the response variable or dependent variable
- x_i are called explanatory variables, predictor variables, or independent variables. The matrix is sometimes called the design matrix.
- β is a (p+1)-dimensional parameter vector. Its elements are also called effects, or regression coefficients. β_0 is called intercept.
- ε is called the error term. This variable captures all other factors which influence the dependent variable y_i other than the regressors x_i.

Ordinary least square (OLS)

- Assume the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ satisfies the Gauss-Markov assumptions:

$$E(\varepsilon_i) = 0, \quad Var(\varepsilon_i) = \sigma^2, Cov(\varepsilon_i, \varepsilon_j) = 0$$

(later referred to as **model 11.1**)

- The OLS method minimizes the sum of squared residuals, and leads to a closed-form expression for the estimated value of the unknown parameter **β**:

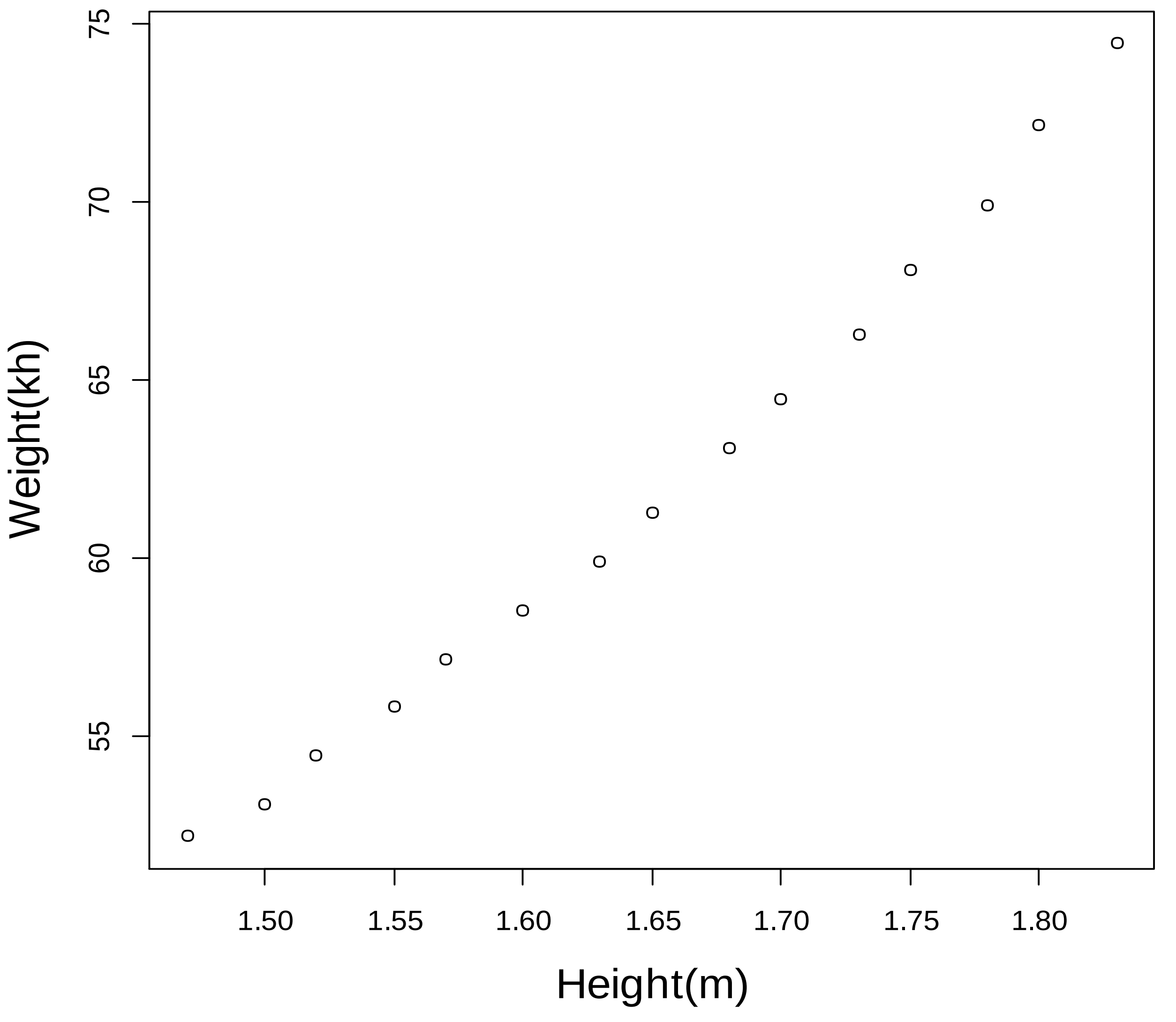
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Example

- The following data set gives average heights and weights for American women aged 30–39

Height (m)	1.47	1.5	1.52	1.55	1.57	1.6	1.63	1.65
Weight (kg)	52.21	53.12	54.48	55.84	57.2	58.57	59.93	61.29
Height (m)	1.68	1.7	1.73	1.75	1.78	1.8	1.83	
Weight (kg)	63.11	64.47	66.28	68.1	69.92	72.19	74.46	

Scatter plot



Example

- The scatter plot suggests that the relationship is strong and can be approximated as a quadratic function.
- OLS can handle non-linear relationships by introducing the regressor $HEIGHT^2$. The regression model then becomes a multiple linear model:

$$w_i = \beta_1 + \beta_2 h_i + \beta_3 h_i^2 + \varepsilon_i$$

Example

- In matrix form: $\mathbf{w} = \mathbf{H\beta} + \boldsymbol{\varepsilon}$

Where

$$\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{15} \end{pmatrix}, \mathbf{H} = \begin{pmatrix} 1 & h_1 & h_1^2 \\ 1 & h_2 & h_2^2 \\ \vdots & \vdots & \vdots \\ 1 & h_{15} & h_{15}^2 \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{15} \end{pmatrix}$$

- The OLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{w} = (128.8128, -143.1620, 61.9603)^T$$

- The relationship between weight and height is $w = 128.8128 - 143.1620 * h + 61.9603 * h^2$

Properties of OLS estimators

- For model 11.1, the Least Square Estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

has the following properties:

1. $E\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$
2. $Cov\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$
3. (Gauss-Markov Theorem) Among any unbiased estimator of $\mathbf{c}^T\boldsymbol{\beta}$, $\mathbf{c}^T\hat{\boldsymbol{\beta}}$ has the minimum variance.

Properties of OLS estimators

1. $$SS_\varepsilon = \mathbf{y}^T\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\mathbf{y}$$

2. $$\hat{\sigma}^2 = \frac{SS_\varepsilon}{n-p-1}$$

- Here $SS_\varepsilon = \hat{\varepsilon}^T\hat{\varepsilon} = \sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2$ is called *residual sum of squares*. Its value reflects the fitness of the regression model.

Centering and scaling

- In application, centering and scaling of data matrix brings convenience.
- Centering:

$$\mathbf{X}_C = \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{np} - \bar{x}_p \end{pmatrix}$$

Centering and scaling

- Scaling: $$\mathbf{Z} = \left(z_{ij}\right)_{n \times p}$$

where $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{s_j}$, $s_j^2 = \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2$

- Since **Z** is centered and scaled, it satisfies:

$$-\mathbf{1}_n^T \mathbf{Z} = \mathbf{0} \qquad \mathbf{R} = \mathbf{Z}^T \mathbf{Z} = \left(r_{ij}\right)_{p \times p}$$

where $$r_{ij} = \frac{\sum_{k=1}^{n} \left(x_{ki} - \bar{x}_i\right)\left(x_{kj} - \bar{x}_j\right)}{s_i s_j}$$

Centering and scaling

- **R** is called the correlation matrix of design matrix **X**. r_{ij} is the correlation coefficient between the ith and jth column of **X**.
- The centered and scaled model takes the form

$$\mathbf{y} = \alpha \mathbf{1}_n + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Correspondingly, the OLS estimator of the unknown parameter is

$$\begin{cases} \hat{\alpha} = \bar{\mathbf{y}} \\ \hat{\boldsymbol{\beta}} = \left(\mathbf{Z}^T \mathbf{Z}\right)^{-1} \mathbf{Z}\mathbf{y} \end{cases}$$

Centering and scaling

- When we have estimated values of intercept and regression parameters (α and $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \cdots, \hat{\beta}_p\right)^T$ respectively) in a centered and scaled model, we can put the regression equation as

$$Y = \hat{\alpha} + \left(\frac{X_1 - \bar{x}_1}{s_1} \right) \hat{\beta}_1 + \cdots + \left(\frac{X_p - \bar{x}_p}{s_p} \right) \hat{\beta}_p$$

$$= \left(\hat{\alpha} - \sum_{i=1}^{p} \frac{\bar{x}_i}{s_i} \hat{\beta}_i \right) + \sum_{i=1}^{p} \frac{\hat{\beta}}{s_i} X_i$$

Multicollinearity

Multicollinearity

- Multicollinearity occurs when there is a linear relationship among several independent variables.
- In the case where we have two independent variables, X_1 and X_2, multicollinearity occurs when $X_{1i}=a+bX_{2i}$,where a and b are constants.
- Intuitively, a problem arises because the inclusion of both X_1 and X_2 adds no more information to the model than the inclusion of just one of them.

Multicollinearity

- For model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, i = 1,2,\cdots,n$

  the variance of, say, β_1 is

$$Var(\beta_1) = \frac{\sigma^2}{\sum (X_{1i} - \bar{X}_1)^2 (1 - r_{12}^2)} = \frac{\sigma^2}{\sum X_{1i}^2 (1 - r_{12}^2)}$$

  where r_{12} is the correlation efficient between X_1 and X_2.

- If X_1 and X_2 are linearly related, then $r_{12}^2 = 1$, and the denominator goes to zero(in the limit), and the variance goes to infinity, which means the estimator is very unstable.

Perfect & near-perfect multicollinearity

- What we have been discussing so far is really perfect multicollinearity.
- Sometimes people use the term multicollinearity to describe situations where there is a *nearly perfect* linear relationship between the independent variables.
- The assumptions of the linear regression model only require that there be no perfect multicollinearity. However, in practice, we almost never face perfect multicollinearity but often encounter near-perfect multicollinearity.

Perfect & near-perfect multicollinearity

- Although the standard errors are technically “correct” and will have minimum variance with near perfect multicollinearity, they will be very, very large.
- The intuition is, again, that the independent variables are not providing much independent information in the model and so out coefficients are not estimated with a lot of certainty.

Detection of multicollinearity

1. Variance Inflation Factor (VIF)

$$VIF = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination of a regression of *j*th independent variable on all the independent variables.

As a rule of thumb, VIF > 10 indicates high multicollinearity.

Detection of multicollinearity

2. Condition Number (k)

$$k = \sqrt{\frac{\lambda_1}{\lambda_m}}$$

where λ_1 and λ_m are the maximum and minimum eigenvalue of the coefficient matrix of design matrix respectively.

- As a rule of thumb, $k>30$ indicates high multicollinearity.

Remedies for multicollinearity

1. Make sure you have not fallen into the *dummy variable trap*; including a dummy variable for every category (e.g., summer, autumn, winter, and spring) and including a constant term in the regression together guarantee perfect multicollinearity.
2. Obtain more data, if possible. This is the preferred solution.

Remedies for multicollinearity

3. Standardize your independent variables. This may help reduce a false flagging of a condition index above 30.
4. Apply a ridge regression or principal component regression.
5. Select a subset of the independent variable(which will be discussed later)

Hypothesis Tests

Hypothesis tests for a single coefficient

- Consider *normal linear regression model*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Where $\varepsilon_i \text{ i.i.d.} \sim N(0, \sigma^2)$

- Suppose that you want to test the hypothesis that the true coefficient β_j takes on some specific value, $\beta_{j,0}$. The null hypothesis and the two-sided alternative hypothesis are

$$H_0 : \beta_j = \beta_{j,0} \text{ vs. } H_1 : \beta_j \neq \beta_{j,0}$$

Hypothesis tests for a single coefficient

- By the property of the OLS estimator, we have

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)$$

- Assume $\mathbf{C}_{p\times p} = \left(c_{ij}\right) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$, we have

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{jj})$$

- So when H_0 is true,

$$\frac{\hat{\beta}_j - \beta_{j,0}}{\sigma\sqrt{c_{jj}}} \sim N(0,1)$$

Hypothesis tests for a single coefficient

- Since in normal linear regression model there exists $\frac{SS_\varepsilon}{\sigma^2} \sim \chi^2_{n-p-1}$ and independent of $\hat{\boldsymbol{\beta}}$, we have

$$t_j = \frac{\hat{\beta}_j - \beta_{j,0}}{\hat{\sigma}\sqrt{c_{jj}}} \sim t_{n-p-1}$$

where $\hat{\sigma}^2 = \frac{SS_\varepsilon}{n-p-1}$.

- With a given confidence level α, when $|t_j| > t_{n-p-1}\left(\frac{\alpha}{2}\right)$

  we can refuse the null hypothesis H_0, otherwise cannot.

Hypothesis tests for a single coefficient

- If the regression model is not normal. By the property of the OLS estimator, we have

$$\sqrt{n}\left(\hat{\beta}_j - \beta_j\right) \rightarrow N\left(0, \sigma^2_{\hat{\beta}_j}\right)$$

- So under H_0, the t statistic

$$t_j = \frac{\hat{\beta}_j - \beta_{j,0}}{SE\left(\hat{\beta}_j\right)} \rightarrow N(0,1)$$

where $SE\left(\hat{\beta}_j\right)$ is the standard error of $\hat{\beta}_j$.

Hypothesis tests for the model

- Consider normal linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Where ε_i i.i.d. $\sim \mathrm{N}(0, \sigma^2)$

- To test hypothesis H_0 on the model: β=0

$$SS_{tot} = \sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2, \; f_T = n - 1$$

$$SS_{reg} = \sum_{i=1}^{n} \left(\hat{y}_i - \overline{y} \right)^2, \; f_M = p$$

$$SS_{err} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2, \; f_M = n - p - 1$$

Hypothesis tests for the model

- Under H_0,

$$F = \frac{SS_{reg}/p}{SS_{err}/(n-p-1)} \sim F(1, n-p-1)$$

Source	D.F.	SS	MS	F
Model	p	SS_{reg}	$MS_{reg}=SS_{reg}/p$	MS_{reg}/MS_{err}
Error	n-p-1	SS_{err}	$MS_{err}=SS_{err}/(n-p-1)$	
Total	n-1	SS_{tot}		

Example

- We also use this data:

X	35.5	34.1	31.7	40.3	36.8	40.2	31.7	39.2	44.2
y	12	16	9	2	7	3	13	9	-1

$$\mathbf{X} = \begin{pmatrix} 1 & 35.5 \\ 1 & 34.1 \\ 1 & 31.7 \\ 1 & 40.3 \\ 1 & 36.8 \\ 1 & 40.2 \\ 1 & 31.7 \\ 1 & 39.2 \\ 1 & 44.2 \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y} = (48.5485, 1.0996)^T$$

$$\left(\mathbf{X}^T \mathbf{X}\right)^{-1} = \begin{pmatrix} 9.616 & -0.256 \\ -0.256 & 0.0069 \end{pmatrix}$$

$$y = 48.5493 - 1.0996x$$

Calculation of SS

$\bar{y} = 7.78$

Observation	Prediction y	Error
12	9.5135	2.4865
16	11.0529	4.9471
9	13.6920	-4.6920
2	4.2354	-2.2354
7	8.0840	-1.0840
3	4.3454	-1.3454
13	13.6920	-0.6920
9	5.4450	3.5550
-1	-0.0530	-0.9470

SS_{Tot}=249.5556, SS_{reg}=174.9935, SS_{err}=74.6679

ANOVA

Source	D.F.	SS	MS	F	P
Model	1	174.99	174.99	16.41**	0.0049
Error	7	74.67	10.67		
Total	8	249.56			

Test for coefficient

- H_0: β=0

$$\hat{\sigma}^2 = \frac{SS_\varepsilon}{n-p-1} = 10.6668$$

$$\mathbf{C}_{p\times p} = (c_{ij}) = (\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 9.616 & -0.256 \\ -0.256 & 0.0069 \end{pmatrix}$$

- So C_{11}=0.007. Under H_0,

$$t = \frac{\hat{\beta}}{\hat{\sigma}\sqrt{c_{jj}}} = \frac{-1.0996}{\sqrt{10.6668 \cdot 0.0069}} = -4.05 \sim t_7$$

- When $\alpha = 0.05, |t| > t_{n-p-1}\left(\frac{\alpha}{2}\right)$, so we reject H_0.

Model Selection in Regression

Model selection

- Model selection consists of two aspects:

    1) linear or non- linear?

    2) which variables to include?

- In this course, we only focus on the second part, the *variable selection* in linear regression.
- There are often

    1) too many variables to choose from

    2) different cost, different power

    3) not an unequivocal “best”

Opposing criteria

- Good fit, good in-sample prediction:
  - Make R^2 large or MSE small
  - Include many variables
- Parsimony:
  - Keep cost of data collection low, interpretation simple, standard errors small
  - Include few variables

Model selection criteria: Coefficient of determination R^2

- Definitions: $$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{err}}{SS_{tot}}$$

  where $$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2, SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

  $$SS_{err} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

- In regression, the R^2 is a statistical measure of how well the regression line approximates the real data points. An R^2 closer to 1 indicates a better fit.
- Adding predictors(independent variables) always increase R^2.

Example

- In previous example:
  - SS_{Tot}=249.5556
  - SS_{reg}=174.9935
  - SS_{err}=74.6679

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = \frac{174.9935}{249.5556} = 0.7012$$

Model selection criteria: Adjusted R^2

- Definitions: $$adj\ R^2 = 1 - \frac{n-1}{n-p}\left(1 - R^2\right)$$

where R^2 is the coefficient of determination.
p is the number of variables of the model

(**including the intercept**).

- adj R^2 will only increase when a predictor has some value, not like R^2.
- Larger adj R^2 (closer to 1) is better.

Model selection criteria: AIC and BIC

- Definition:

  $$\text{AIC} = -2(\text{maximized log-likelihood}) + 2p$$

  $$\text{BIC} = -2(\text{maximized log-likelihood}) + p \log(n)$$

- For linear regression,

  $$-2(\text{maximized log-likelihood}) = n \log(SS_{err}) + C$$

- Smaller value of AIC or BIC is better
- Get a balance between model fit and model size: BIC penalizes larger models more heavily than AIC $\Rightarrow$ BIC tends to prefer smaller models

Model selection criteria: Mallow's C_p

- Definition: $$C_p = \frac{SS_{err}}{\hat{\sigma}^2_{full}} + p - n$$

where $\hat{\sigma}^2_{full}$ estimated from the full model and SS_{err} is obtained from a sub-model of interest.

- Cheap to compute
- Closely related to adj R^2 and AIC, BIC.
- Performs well in predicting.

Variable selection methods: Best subsets selection

- Fit all possible models (all of the various combinations of explanatory variables) and evaluate which fits the data best based on the criteria above(except for R^2).
- Usually takes a long time when dealing with models with many explanatory variables.

Variable selection methods: Forward selection

- Starting with no variables in the model, testing the addition of each variable using a chosen model comparison criterion, adding the variable (if any) that improves the model the most, and repeating this process until none improves the model.

Variable selection methods: Backward selection

- Starting with all candidate variables, testing the deletion of each variable using a chosen model comparison criterion, deleting the variable (if any) that improves the model the most by being deleted, and repeating this process until no further improvement is possible.

Variable selection methods: Stepwise selection

- A combination of the forward selection and the backward selection, testing at each step for variables to be included or excluded.

Model selection example

- We will model a multiple linear regression for a dataset *(Longley's Economic Regression Data*) through different model selection approaches and criteria.
- The dataset shows the relationship between the dependent variable *GNP deflator* and the possible predictor variables.
- The objective is to find out the a subset of all the predictor variables which truly have an significant effect on the dependent variable and to evaluate the effect.

GNP deflator and the possible predictor variables

	GNP Deflator	GNP	Unemployed	Armed Forces	Population	Year	Employed
1947	83	234.289	235.6	159	107.608	1947	60.323
1948	88.5	259.426	232.5	145.6	108.632	1948	61.122
1949	88.2	258.054	368.2	161.6	109.773	1949	60.171
1950	89.5	284.599	335.1	165	110.929	1950	61.187
1951	96.2	328.975	209.9	309.9	112.075	1951	63.221
1952	98.1	346.999	193.2	359.4	113.27	1952	63.639
1953	99	365.385	187	354.7	115.094	1953	64.989
1954	100	363.112	357.8	335	116.219	1954	63.761
1955	101.2	397.469	290.4	304.8	117.388	1955	66.019
1956	104.6	419.18	282.2	285.7	118.734	1956	67.857
1957	108.4	442.769	293.6	279.8	120.445	1957	68.169
1958	110.8	444.546	468.1	263.7	121.95	1958	66.513
1959	112.6	482.704	381.3	255.2	123.366	1959	68.655
1960	114.2	502.601	393.1	251.4	125.368	1960	69.564
1961	115.7	518.173	480.6	257.2	127.852	1961	69.331
1962	116.9	554.894	400.7	282.7	130.081	1962	70.551

Full model

Estimate and significance test of regression parameters				
	Estimate	Std.Error	t value	Pr(>|t|)
(intercept)	2946.85636	5647.97658	0.522	0.6144
GNP	0.26353	0.10815	2.437	0.0376
Unemployed	0.03648	0.03024	1.206	0.2585
Armed Forces	0.1116	0.01545	0.722	0.4885
Population	-1.73703	0.67382	-2.578	0.0298
Year	-1.4188	2.9446	-0.482	0.6414
Employed	0.23129	1.30394	0.177	0.8631
$R^2 = 0.9926$				

Full model

- Not all the predictors have a significant effect on the dependent variable. (the p-value of some regression parameters are no less than 0.05)
- The coefficient of determination R^2 reaches the maximum value (bigger than that of any sub-model).

Best subset selection

- Using the best subset selection with C_p Criterion, we get 3 predictor variables:

GNP	Unemployed	Armed Forces	Population	Year	Employed
TRUE	TRUE	FALSE	TRUE	FALSE	FALSE

- Using the best subset selection with adj R^2 Criterion, we get 4 predictor variables:

GNP	Unemployed	Armed Forces	Population	Year	Employed
TRUE	TRUE	TRUE	TRUE	FALSE	FALSE

Forward/Backward selection

- Using the forward selection with AIC Criterion, we get only one predictor variable:

GNP	Unemployed	Armed Forces	Population	Year	Employed
TRUE	TRUE	FALSE	TRUE	FALSE	FALSE

- Using the backward selection with AIC Criterion, we get 3 predictor variables:

GNP	Unemployed	Armed Forces	Population	Year	Employed
TRUE	TRUE	TRUE	TRUE	FALSE	FALSE

Stepwise selection

- Using the stepwise selection with AIC Criterion, we get 1 predictor variables:

GNP	Unemployed	Armed Forces	Population	Year	Employed
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE

- As we mentioned above, different approaches may yield different selections, there is no unequivocal "best".

Regression in Excel: LINEST(…)

表格 插图 图表

LINEST =LINEST(A3:A18, B3:G18, TRUE, TRUE)

	A	B	C	D	E	F	G
1	Y	X1	X2	X3	X4	X5	X6
2	GNP Defl	GNP	Unemploy	Armed Fo	Populati	Year	Employed
3	83	234.289	235.6	159	107.608	1947	60.323
4	88.5	259.426	232.5	145.6	108.632	1948	61.122
5	88.2	258.054	368.2	161.6	109.773	1949	60.171
6	89.5	284.599	335.1	165	110.929	1950	61.187
7	96.2	328.975	209.9	309.9	112.075	1951	63.221
8	98.1	346.999	193.2	359.4	113.27	1952	63.639
9	99	365.385	187	354.7	115.094	1953	64.989
10	100	363.112	357.8	335	116.219	1954	63.761
11	101.2	397.469	290.4	304.8	117.388	1955	66.019
12	104.6	419.18	282.2	285.7	118.734	1956	67.857
13	108.4	442.769	293.6	279.8	120.445	1957	68.169
14	110.8	444.546	468.1	263.7	121.95	1958	66.513
15	112.6	482.704	381.3	255.2	123.366	1959	68.655
16	114.2	502.601	393.1	251.4	125.368	1960	69.564
17	115.7	518.173	480.6	257.2	127.852	1961	69.331
18	116.9	554.894	400.7	282.7	130.081	1962	70.551
19	b6	b5	b4	b3	b2	b1	b
20	SE(b6)	SE(b5)	SE(b4)	SE(b3)	SE(b2)	SE(b1)	SE(b)
21	R^2	SE(Y)					
22	F	D.F.					
23	SS(Reg)	SS(Resid)					
24	=LINEST(A	-1.4188	-1.73703	0.011161	0.036483	0.263527	2946.856
25	1.303941	2.944602	0.673815	0.015453	0.030245	0.108151	5647.977
26	0.992647	1.194618	#N/A	#N/A	#N/A	#N/A	#N/A
27	202.5094	9	#N/A	#N/A	#N/A	#N/A	#N/A
28	1734.02	12.844	#N/A	#N/A	#N/A	#N/A	#N/A

Exercises with SAS

- Use SAS Proc Regression