MATH 189
Multivariate Linear Regression

Wenxin Zhou
UC San Diego

Time: 2:00—3:20 & 3:30—4:50pm TueThur
Location: CENTR 115

Outline

- In the previous lecture, we introduced a classification method, named linear/quadratic discriminant analysis.
  - Bayes' rule
  - Posterior probability
  - Discriminant functions

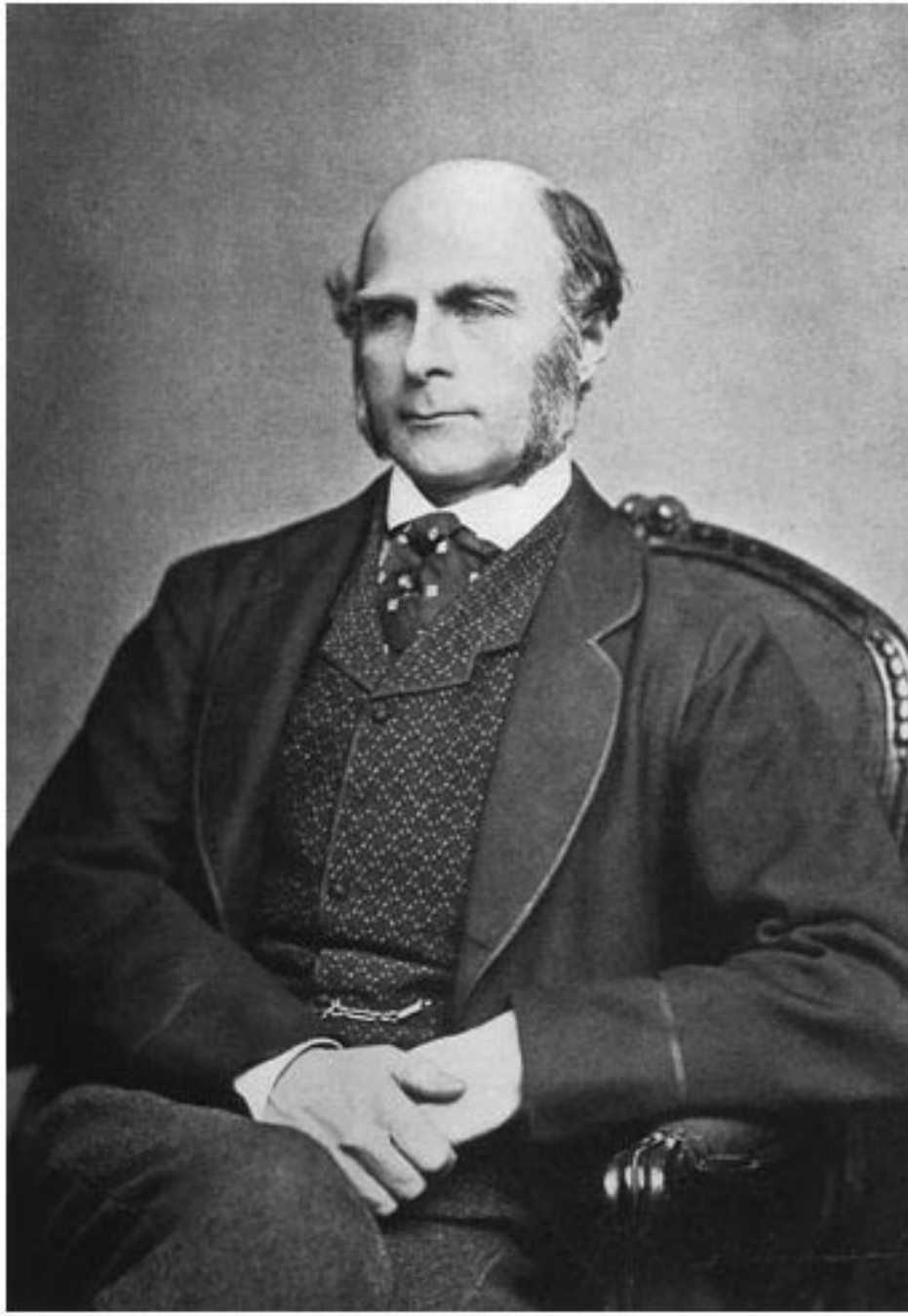
- Today we will introduce the linear regression analysis.
  - Simple linear regression
    - Estimation
    - Inference of parameter
    - Assess model accuracy
  - Multivariate linear regression

Linear Regression Analysis

- Linear regression analysis is one of the most widely used statistical methods: it is the study of linear association and additive effect among several variables.

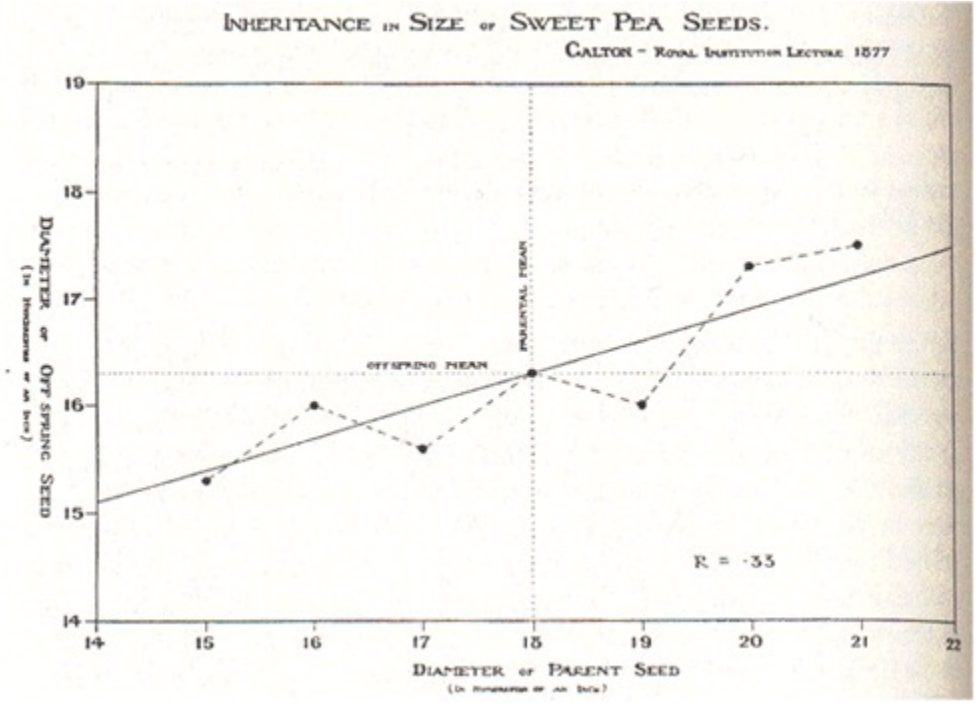
- The first thing you may want to know about linear regression is how the term “regression” was coined. The statistical problem was first studied in depth by Sir Francis Galton (1822-1911), a statistician, progressive, polymath, sociologist, psychologist, etc…

- Galton was a real-life Indiana Jones character. He wrote two best-selling books: “The Art of Travel” and “The Art of Rough Travel” which are still in print. He was also a pioneer in eugenics, and his book “Hereditary Genius” was the first social scientific attempt to study genius and greatness.



Origination of Regression

- Galton was a pioneer in the application of statistical methods to many branches of science.



- In the study on relative sizes of parents and their offspring in various species of plants and animals, he observed the following phenomenon:
- A larger-than-average parent tends to produce a larger-than-average child, but the child is likely to be less larger than the parent in terms of its relative position within its own generation.
- Galton termed this phenomenon "***regression to the mean***".

Regression to the Mean

- Regression to the mean is an important and well observed phenomenon in real life:
  - Your children can be expected to be less exceptional (for better or worse) than you are.
  - Your score on a final exam in a course can be expected to be less bad (or good) than your score on the midterm exam, relative to the rest of the class.
  - A baseball player's batting average in the second half of the season can be expected to be closer to the mean (for all players) than his batting average in the first half of the season.
  - And so on.

- The key word here is "expected". It does not mean it's certain that regression to the mean will occur, but that's the way to bet!

Why Linear Model?

- The next natural question for linear regression is:

  Why should we assume that relationships between variables are linear?

- There are many justifications for this question, here we list a few common answers:
  1. linear relationship is the simplest non-trivial relationship that can be modeled, and hence the easiest to work with.

  2. Many "true" relationships between variables are often at least approximately linear over the range of values that are of interest to us.

  3. Even if they're not, we can often transform the variables in such a way as to linearize the relationships.

Simple Linear Regression

- Simple linear regression lives up to its name: it is a very straightforward approach for predicting a quantitative response Y (not binary/categorical) on the basis of a single regression predictor variable X.

- It assumes an approximately linear relationship between X and Y. Mathematically, we write this linear relationship as

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

  where β_0 and β_1 are two unknown parameters.

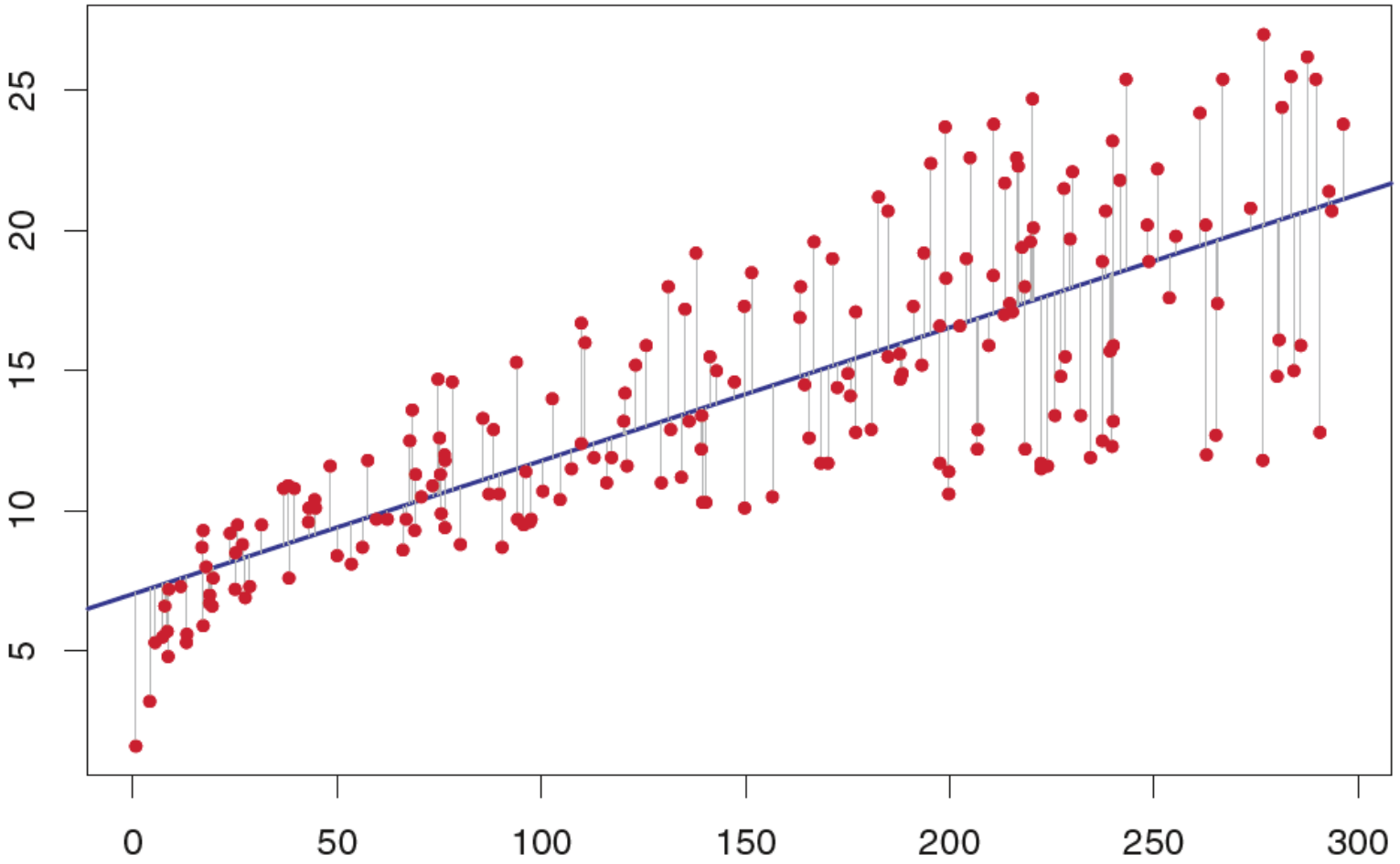
- Usually we call β_0 the intercept and β_1 the slope.
- $\epsilon = Y - \beta_0 + \beta_1 X$ denotes the approximation error of this linear model.

Assumptions of Simple Linear Regression

- The classic simple linear regression should follow the assumptions below:
  1. Linearity: The response variable Y depends (approximately) linearly on the covariate X.
  2. Independence: the observations $y_1, \dots, y_n$ are independent.
  3. Uncorrelated error: X and ϵ are uncorrelated, i.e. $\mathrm{cov}(X, \epsilon) = 0$.
  4. Homoscedasticity: The mean and variance of ϵ_i are the same for all i.
  5. Normality: The error ϵ is normally distributed.

- Assumptions 4 and 5 together can be interpreted as $\epsilon_i' s$ generated from $N(0, \sigma^2)$.

Fit Linear Regression



- The goal of simple linear regression is to find a line which is as as close as possible to the n observations.
- This is equivalent to finding "proper" values of intercept β_0 and slope β_1.

Estimating the Coefficients

- Suppose we observe a random sample of n observations
$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n),$$
each of which consists of a measurement of X and a measurement of Y.
- Our goal is to obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model fits the available data well. How to measure the goodness of the fit?
- In other words, we want to find the values $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the resulting line is as close as possible to the n observations.
- There are a number of ways of measuring closeness. However, by far the most common criterion is based on minimizing the $least\ squares$.
- The method of least squares was first published in 1805 by the French mathematician Adrien-Marie Legendre (1752-1833). He posed LS as a solution to the algebraic problem of solving a system of equations when the number of equations exceeds the number of unknowns.

Residual Sum of Squares

- Given a pair of estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$, we can predict y_i by
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \text{ for } i = 1, \dots, n.$$

- We define the ith residual as the difference between the ith observed response value and the ith fitted response value that is predicted by our linear model:
$$r_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

- Then, we define the *residual sum of squares* (RSS) as
$$\text{RSS} = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2.$$

- RSS is a natural measurement of the "closeness" between the fitted line and n noisy observations.

Least Squares Estimation

- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. This is equivalent to find the parameters that minimize the "closeness" between the fitted line and n observations.

- With some calculations, one can show that the minimizers are

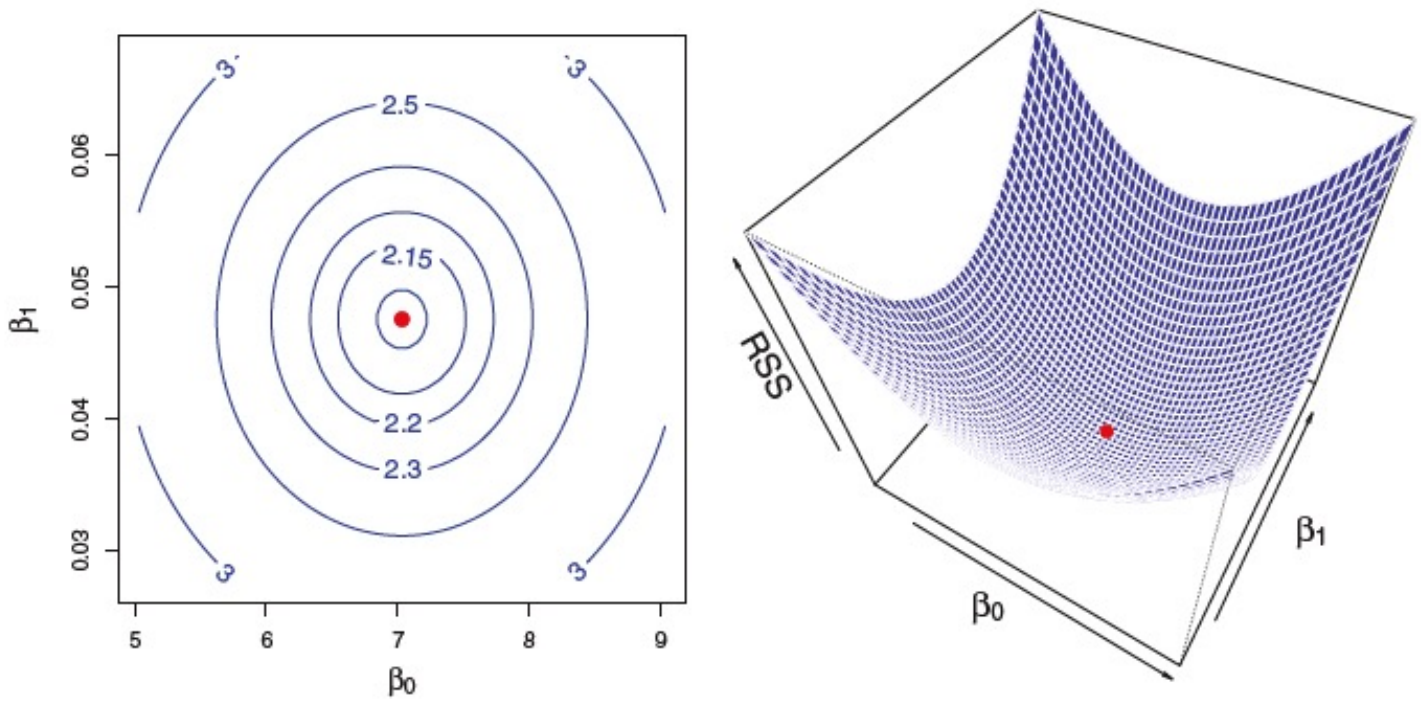
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are sample means of y and x, respectively.

Interpretation of Least Squares Estimator

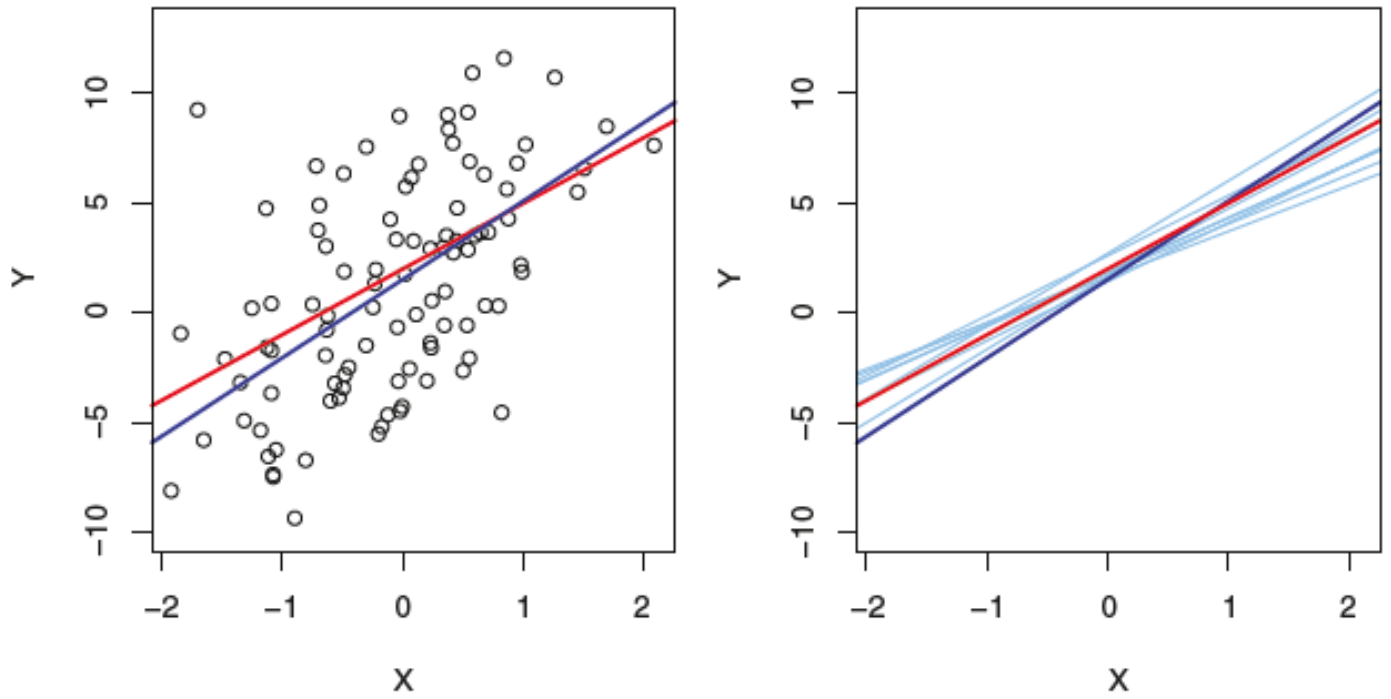
- Least squares estimator is defined as a set of (β_0, β_1) that minimizes RSS. RSS can be interpreted as the overall fitting errors.
- In this sense, the least squares estimator yields the best (in sample) model fitting accuracy.



- **Left Panel**: Contour of RSS in a simple linear regression example. Each blue line represents a set of (β_0, β_1) with the same RSS. The least squares estimator is the red dot (unique!) in the middle.
- **Right Panel**: 3D-surface plot of RSS. The red dot in the middle of the "net" is the least squares estimator.

Variability of Least Square Estimators

- The least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are functions of the observed random sample. Therefore, $\hat{\beta}_0$ and $\hat{\beta}_1$ are two random variables that depend on the data.
- If we observe a different sample from the same population, we may obtain a different pair of $\hat{\beta}_0$ and $\hat{\beta}_1$. We need to quantify the variability of the least squares estimators.



Left panel: The red line is the true relationship $f(X) = 2 + 3X$, the dark blue line is the least squares fitted line.

Right panel: The red and dark blue lines are the same as the left. The 10 light blue lines are least square lines from 10 random samples.

Assessing Accuracy of Coefficient Estimates

- Recall that we assessed the estimation accuracy of sample mean through the analysis of bias and variance. Similarly, least squares estimators can also be analyzed as follows.

- Unbiasedness: with some calculations, we can show that

$$\mathbb{E}(\hat{\beta}_0 \mid x) = \beta_0 \quad \text{and} \quad \mathbb{E}(\hat{\beta}_1 \mid x) = \beta_1.$$

- Standard Error: further, we can show

$$\text{var}(\hat{\beta}_0 \mid x) = \text{se}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right],$$

$$\text{var}(\hat{\beta}_1 \mid x) = \text{se}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{with} \quad \sigma^2 = \text{var}(\epsilon).$$

Estimating σ^2

- The standard errors of least squares estimators depend on the population variance of ϵ, which is in general unknown to us.

- In practice, one can estimate σ^2 by the sample variance of residuals $r_i's$. This estimate is known as the residual standard error, and is given by the formula

$$\hat{\sigma} = \text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}}.$$

- We divide $n-2$ as there are two parameters estimated from the sample. Recall that we divided $n-1$ in the sample variance as it involves 1 estimated parameter - sample mean.

- In the following, with slight abuse of notation, we still use $\text{se}(\hat{\beta}_0)$ and $\text{se}(\hat{\beta}_1)$ to denote the standard errors of the estimated σ^2.

Confidence Interval

- Standard errors can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true (unknown) parameter.

- For linear regression, the 95% confidence interval for β_1 approximately takes the form
$$\hat{\beta}_1 \pm 2\times \mathrm{se}(\hat{\beta}_1).$$

- That is, there is approximately a 95% chance that the interval
$$\left[\hat{\beta}_1 - 2\mathrm{se}(\hat{\beta}_1), \quad \hat{\beta}_1 + 2\mathrm{se}(\hat{\beta}_1)\right]$$
will contain the true β_1.

- Similarly, a confidence interval for β_0 approximately takes the form
$$\hat{\beta}_0 \pm 2\times \mathrm{se}(\hat{\beta}_0).$$

Hypothesis Testing

- Standard errors can also be used to perform hypothesis testing on the coefficients. The most common hypothesis test involves testing the null hypothesis of

  $$H_0: \text{There is no linear relationship between X and Y}$$

  versus the alternative hypothesis

  $$H_1: \text{There is some linear relationship between X and Y}$$

- Mathematically, this corresponds to testing

  $$H_0: \beta_1 = 0 \quad \text{versus} \quad H_1: \beta_1 \neq 0.$$

  If $\beta_1 = 0$, the regression model is reduced to $Y = \beta_0 + \epsilon$ and X is not associated with Y.

- To test the null hypothesis, we need to determine whether $\hat{\beta}_1$, our estimate of β_1, is sufficiently far away from zero. Then, we can be confident that β_1 is non-zero.

Test Statistic

- *How far is far enough?* This of course depends on the accuracy of $\hat{\beta}_1$ which is measured by $\text{se}(\hat{\beta}_1)$, the (estimated) standard error of $\hat{\beta}_1$.

- If $\text{se}(\hat{\beta}_1)$ is small, then even relatively small values of $\hat{\beta}_1$ may provide strong evidence that $\beta_1 \neq 0$. In contrast, if $\text{se}(\hat{\beta}_1)$ is large, then $\hat{\beta}_1$ must be large in absolute value in order for us to reject the null hypothesis.

- In practice, we can compute a t-statistic, given by

$$t = \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)},$$

  which measures the number of standard deviations that $\hat{\beta}_1$ is away from 0.

- Under null hypothesis (no linear relationship between X and Y), we expect this statistic to follow a t-distribution with $n - 2$ degrees of freedom.

Assessing the Accuracy of the Model

- Once we rejected the null hypothesis in favor of the alternative, a follow-up question is: How well can a linear model fit the data?
- The quality of a linear regression fit is typically assessed using two related quantities: the residual standard error (RSE) and the R^2 statistic.
- As introduced in previous slide, RSE measures the standard error of fitted residuals

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n-2}}.$$

- The RSE is considered as a measure of the lack-of-fit of the linear model to the data.
- If $y_i \approx \hat{y}_i$ for $i = 1, \ldots, n$, then RSE will be small, and we can conclude that the model fits the data very well. On the other hand, if $\hat{y}_i$ deviates far from y_i for one or more observations, then the RSE may be quite large, indicating that the model doesn't fit the data well.

Assessing the Accuracy of the Model (cont.)

- The RSE provides an absolute measure of lack of fit of the linear model to the data. The R^2 statistic provides an alternative measure of fit. It takes the form of a proportion.

- To calculate R^2, we use the formula

$$R^2 = \frac{\text{TSS}-\text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}},$$

where $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares which can be regarded as the total amount of variability inherent in the response before the regression is performed.

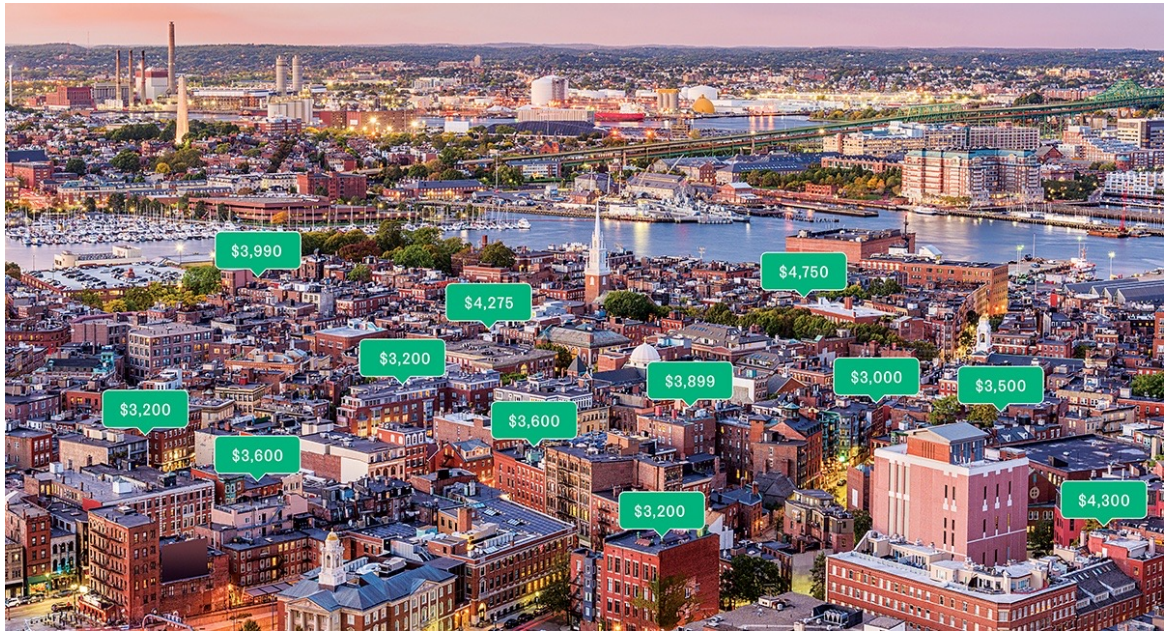
- R^2 measures the proportion of variance explained by the linear model. So, it always takes a value between 0 and 1, and is independent of the scale of Y .

- On can also understand R^2 as a "horse racing" between predicting with least square estimators and predicting with sample mean of response.

Example: Boston Housing Price Data

- The dataset collects median house price and many related variables over 506 neighborhoods around Boston.

- The response variable is the median house price (medv) of each neighborhood. The covariate we use to predict the response is percent of households with low socioeconomic status (lstat)

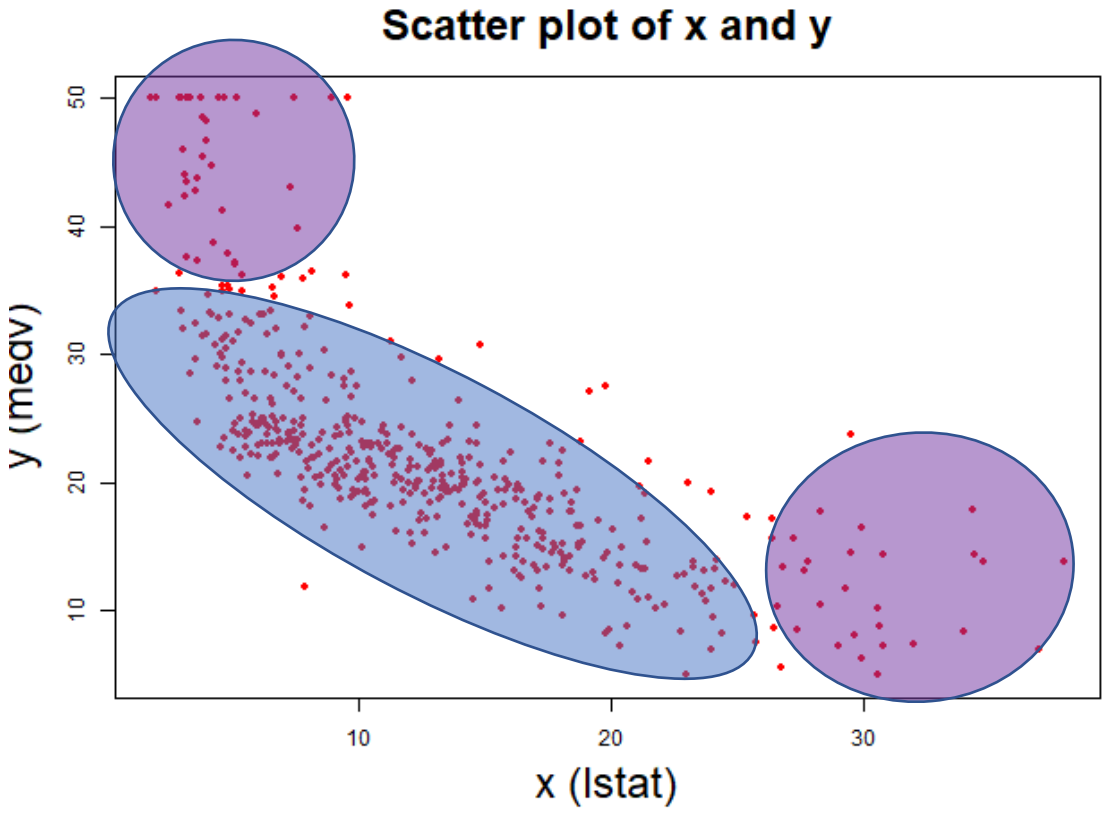


- Fit a linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ for } i = 1, \ldots, 506,$$

where y is the median house price and x is the percent of households with low socioeconomic status

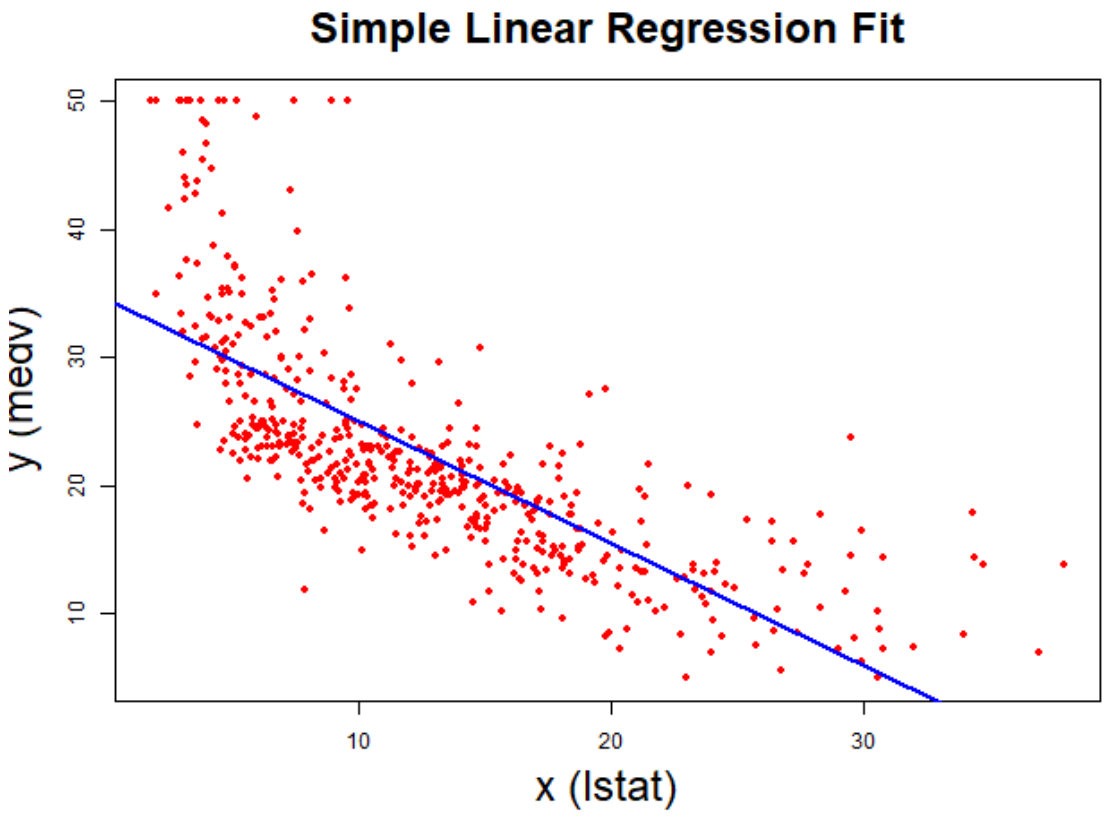
A Peek at Dataset



- We draw a scatter plot between X (low socioeconomic status) and Y (median house price). Each red dot represents a neighborhood in Boston

- We can observe a clear linear pattern in the blue shaded area. However, the linear pattern may not be valid in the purple shaded areas.

Simple Linear Regression Fit



- We fit a simple linear regression model. The estimated regression coefficients are
  $\hat{\beta}_0 = 34.55$ and $\hat{\beta}_1 = -0.95$.

- The estimated linear function is visualized as the blue line in the figure:
  $$Y = 34.55 - 0.95X.$$

Confidence Interval

- We also calculate the standard error of two estimated regression coefficients:

$$\text{se}(\hat{\beta}_0) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] = 0.5626,$$

$$\text{and se}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0.0387,$$

where $\hat{\sigma} = \text{RSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}} = 6.216.$

- The 95% confidence interval for β_0 is

$$\left[\hat{\beta}_0 - 2\text{se}(\hat{\beta}_0), \hat{\beta}_0 + 2\text{se}(\hat{\beta}_0)\right] = [33.42, 35.67].$$

- The 95% confidence interval for β_1 is

$$\left[\hat{\beta}_1 - 2\text{se}(\hat{\beta}_1), \hat{\beta}_1 + 2\text{se}(\hat{\beta}_1)\right] = [-1.03, -0.87\ ].$$

Test the Linear Relationship

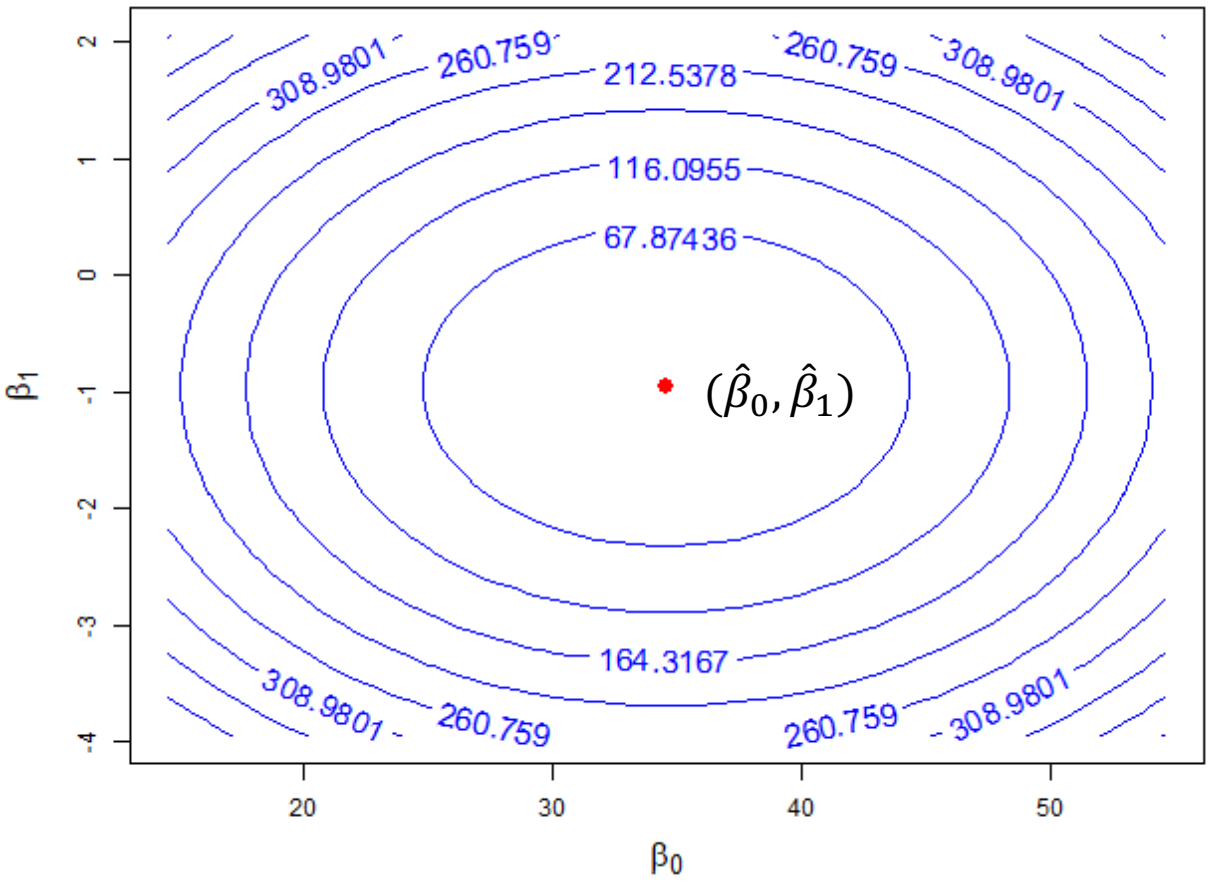
- To check if there is a linear relationship between Y and X, we can test
$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_1: \beta_1 \neq 0,$$

- Compute a t-statistic, given by
$$t = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} = \frac{-0.95}{0.0387} = -24.54.$$

- Under null hypothesis, the t-statistic follows a t-distribution with degrees of freedom 504 (i.e. $n - 2$).

- At a given significance level $\alpha = 0.05$, we reject the null hypothesis as p-value is smaller than 0.00001.

Accuracy of Simple Linear Model



- Based on the estimates, we can calculate the Residual Standard Error of the fitted model

$$\text{RSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n-2}} = 6.216$$

- We also draw the contour plot of RSE. Each blue line represents a set of (β_0, β_1) that has the same model accuracy in terms of RSE.

- The best prediction results (smallest RSE) is obtained at the least squares estimator $\hat{\beta}_0$ and $\hat{\beta}_1$.

Multiple Linear Regression

- Simple linear regression is a useful approach for predicting a response on the basis of a single predictor/covariate/explanatory variable.
- However, in practice we often have more than one predictor.
- For example, in the Boston Housing Price data, we have examined the relationship between median price and the percentage of low socioeconomic status.
- We also have data for the average room per house, the average age of house, crime rate, and so on.
- We want to know whether these variables are also associated with house price.

Problem Setup

- We extend the idea of simple linear regression by assigning each predictor a slope

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

where

- X_j represents the jth predictor.
- β_j quantifies the association between X_j and the response.
- ϵ is the approximation error term (regression error/noise).

- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make predictions using the formula

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

Assumptions of Multivariate Linear Regression

- The assumptions for multivariate linear regression is similar as the simple linear regression case:
  1. Linearity: The response variable Y depends (approximately) linearly on all the covariates $X_1, \dots, X_p$.
  2. Independence: the observations $y_1, \dots, y_n$ are independent.
  3. Uncorrelated error: X_j and ϵ are uncorrelated, i.e. $\text{cov}(X_j, \epsilon) = 0$ for $j = 1, \dots, p$.
  4. Homoscedasticity: The mean and variance of ϵ_i are the same for i.
  5. Normality: The error ϵ is normally distributed.
  6. No multi-collinearity: the covariates $X_1, \dots, X_p$ are not highly correlated.
  7. Low dimension: $n > p$.

- Assumptions 4 and 5 together can be interpreted as having ϵ_i generated from $N(0, \sigma^2)$.
- The importance of new assumptions 6 and 7 will be clear when we review the matrix form of multivariate linear regression.

Estimating the Regression Coefficients

- The parameters can be estimated using the same least squares approach that we used in the context of simple linear regression.

- We choose $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ to minimize the residual sum of squares (RSS)

$$\begin{aligned}\text{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}\right)^2.\end{aligned}$$

- The values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize RSS are called the multiple least squares regression coefficient estimates.

- Unlike the simple linear regression, the multiple regression coefficient estimates have very complex expressions that are most easily represented using matrix algebra.

Matrix Form of Multivariate Linear Model

- Suppose we observe a sample of n observations

$$(y_1, x_{11}, \dots x_{1p}), (y_2, x_{21}, \dots x_{2p}), \dots (y_n, x_{n1}, \dots x_{np})$$

- The matrix form of multivariate linear model is as follows:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \mathbf{e},$$

where

- $\boldsymbol{Y} = (y_1, \dots, y_n)'$ a vector of response variables.
- $\boldsymbol{\beta} = (\beta_0, \beta_1 \dots, \beta_p)'$ is a vector of regression coefficients (one intercept and p slopes).
- $\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \in \mathbb{R}^{n\times(p+1)}$ is called the design matrix.
- $\boldsymbol{e} = (e_1, e_2, \dots, e_n)'$ is a vector of random errors.

Multivariate Least Squares

- With the matrix notation, we can express the multivariate least squares problem in a relative simple form

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2,$$

  where $\|\cdot\|$ is the Euclidean norm (L_2 norm).

- The above minimization problem can be solved by finding the root of the first-order derivative (gradient). This is equivalent to solving the following equation:

$$0 = \frac{\partial \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = -2\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}).$$

- It is easy to check that the solution admits the closed form

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

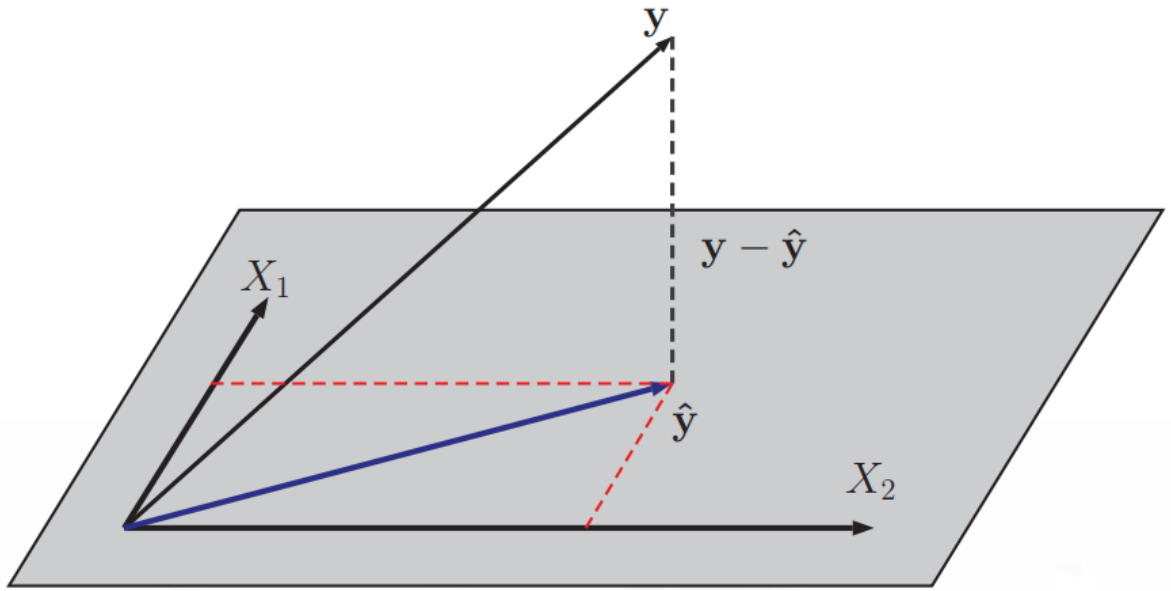
- $\widehat{\boldsymbol{\beta}}$ is usually referred to as the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$.

Interpretation of OLS Estimator

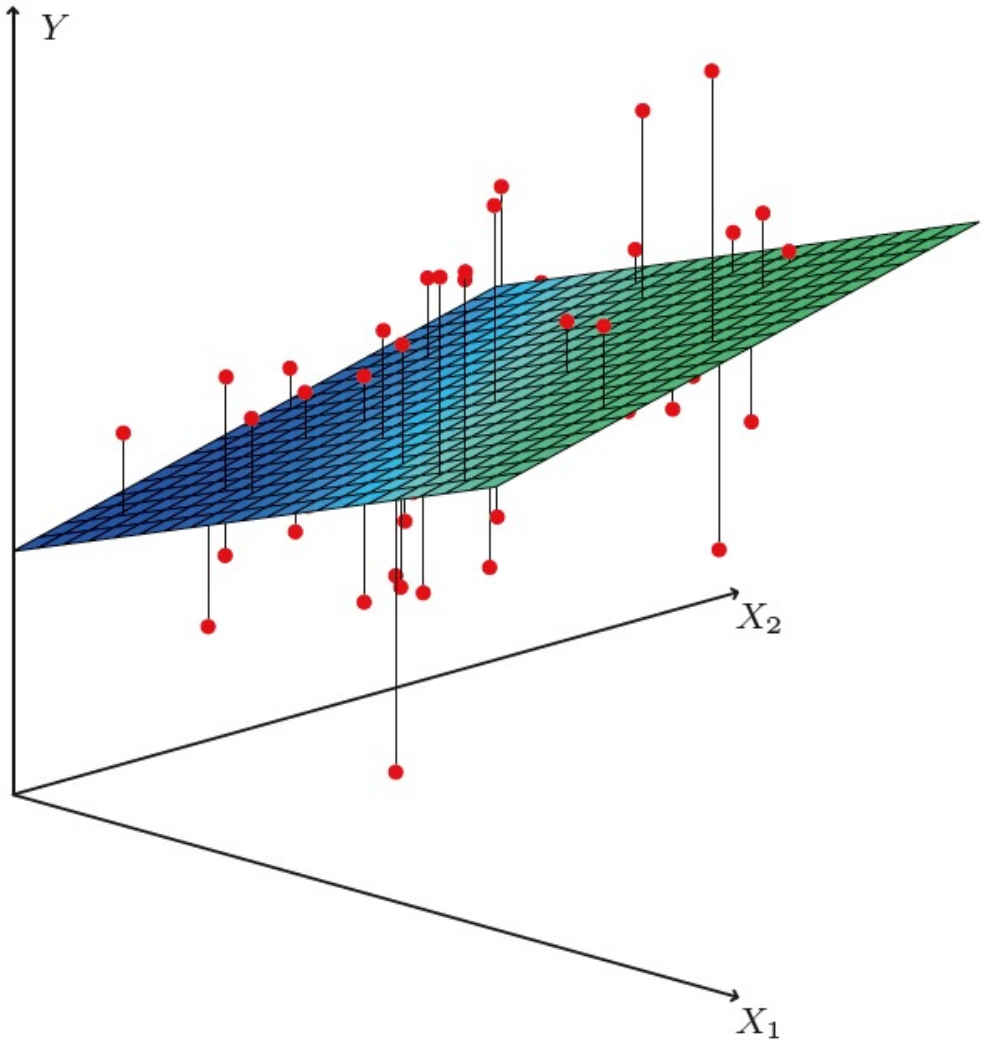
- With the OLS estimator, we can predict the response vector $\boldsymbol{Y}$ by
$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \equiv \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y},$$
where $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the $n \times n$ projection matrix.

- The projection matrix satisfies the following properties:
  1. $\boldsymbol{P}_{\boldsymbol{X}}$ is symmetric.
  2. $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{X}}$.
  3. $\boldsymbol{P}_{\boldsymbol{X}}$ projects any n-vector onto the column space of $\boldsymbol{X}$.

Geometric view of least-squares: The fitted value is the blue arrow, which is the projection of Y on the plane spanned by X_1 and X_2.



Example: OLS Estimator with Two Covariates



- Consider a toy example
  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$
  where we have with two predictors.

- We draw the three dimensional observations (y_i, x_{1i}, x_{2i}), $i = 1, \ldots, n$, in a 3D Scatter plot. Each observation is denoted as a red point.

- In 3-dimensional setting, the least squares regression line becomes a plane.

- The plane is chosen to minimize the sum of the squared vertical distances between each observation and the plane.

Assessing Accuracy of OLS Estimator

- Similarly, we assess the estimation accuracy of ordinary least squares estimators in the multivariate case.

- Unbiasedness: with some calculations, we can show

$$\mathbb{E}(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}) = \boldsymbol{\beta}.$$

- Standard error: moreover,

$$\mathrm{var}(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}) = \mathrm{se}(\widehat{\boldsymbol{\beta}})^2 = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

- Again, notice that the standard error of $\widehat{\boldsymbol{\beta}}$ depends on σ^2, the variance of ϵ_i. In practice we need to estimate σ^2 from the observed sample.

Estimating σ^2 in Multivariate Case

- In most cases σ^2 is an unknown parameter of the model in addition to the regression coefficient vector $\boldsymbol{\beta}$.
- In order to calculate the standard error of $\hat{\boldsymbol{\beta}}$, we first need to find a good estimate of σ^2 .
- RSS in multivariate case can be written as

$$\text{RSS} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}}).$$

- A natural estimator of σ^2 is defined as

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n - p - 1}.$$

- One can show that $\hat{\sigma}^2$ is an unbiased estimator of σ^2, i.e. $\mathbb{E}(\hat{\sigma}^2 | \boldsymbol{X}) = \sigma^2$.

Testing on Marginal Effect

- Similar to the simple linear regression, we can test weather a particular regression coefficient is zero. The null and alternative hypotheses are

$$H_0: \beta_j = 0 \quad \text{versus} \quad H_1: \beta_j \neq 0, \text{ for some } j = 1, \ldots, p.$$

- Under the Gaussian and homoscedastic assumption, the random error follows

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I_n}).$$

- Notice that

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}.$$

- Then we have

$$\widehat{\boldsymbol{\beta}} \sim \boldsymbol{N}(\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2) \quad \text{and} \quad \hat{\beta}_j \sim N(\beta_j, v_j\sigma^2),$$

where v_j is the jth diagonal entry of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Test Statistic

- If we look at each $\hat{\beta}_j$ marginally, then $\hat{\beta}_j \sim N(\beta_j, v_j\sigma^2)$, where v_j is the jth diagonal entry of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

- In addition, $\hat{\sigma}^2$ follows a Chi-squared distribution

$$(n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}.$$

- If we want to test the hypothesis that $\beta_j = 0$, we can use the following t-statistic

$$t = \frac{\hat{\beta}_j}{\sqrt{v_j}\hat{\sigma}}.$$

- Under the null hypothesis, the above t-statistic follows a t-distribution with degrees of freedom $n-p-1$.

Example: Boston Housing Price Data

- Let's revisit the Boston Housing Price Data in a multivariate setting.
- We predict the median house price with three variables: percent of households with low socioeconomic status (lstat), average number of rooms per house (rm) and average age of houses (age).

- Consider a linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i \text{ for } i = 1, \dots, 506,$$

where

- y is the median house price;
- x_1 is the percent of households with low socioeconomic status;
- x_2 is the average number of rooms per house;
- x_3 is the average age of houses;

Example: Estimating Regression Coefficients

- First, we estimate the regression coefficients by the OLS estimator. The results are reported in the following table.

$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
-1.175	-0.669	5.019	0.009

- The standard errors of estimated coefficients are reported as follows:

$\mathrm{se}(\widehat{\beta}_0)$	$\mathrm{se}(\widehat{\beta}_1)$	$\mathrm{se}(\widehat{\beta}_2)$	$\mathrm{se}(\widehat{\beta}_3)$
3.182	0.054	0.454	0.011

- The standard deviation σ of ϵ is estimated by the residual standard error:

$$\hat{\sigma} = \mathrm{RSE} = 5.542.$$

Example: Testing on Marginal Effect

- To certify the importance of each predictor, we test the marginal effects.
- Consider null and alternative hypotheses:

$$H_0: \beta_j = 0 \quad \text{versus} \quad H_1: \beta_j \neq 0, \text{ for } j = 1, \dots, 3.$$

- Calculate t-statistic for each coefficient as

$$t_j = \frac{\widehat{\beta}_j}{\sqrt{v_j}\widehat{\sigma}}, \text{ for } j = 1, \dots, 3.$$

- The test statistics, p-values and testing results are summarized below.

Estimator	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
t-statistic	-12.298	11.048	0.009
p-value	<0.0001	<0.0001	0.418
Decision ($\alpha = 0.05$)	**Reject**	**Reject**	**Not Reject**

Example: Compare with Simple Linear Regression

- We compare the multivariate linear regression fitting results with the previous simple linear regression fitting results.

Simple Linear Regression

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.
- Residual Sum of Squares: 19,473.
- R^2 statistic: 0.544.

Multivariate Linear Regression

- $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$.
- Residual Sum of Squares: 15,418.
- R^2 statistic: 0.639.

- Linear model with three variables (lstat, rm and age) yields smaller residual standard error and larger R^2 statistic than the simple linear model with only one variable (lstat).

Can we say the right model is more adequate than the left?

Testing on Model Adequacy

- In many applications the null hypothesis is that a subset of the covariates have zero regression coefficients. This subset of covariates can be deleted from the regression model: they are unrelated to the response variable given the remaining variables.

- Consider two competitive linear models:

$$\text{Null: } Y = \beta_0 + \sum_{j \in S} \beta_j X_j + \epsilon \quad \text{versus} \quad \text{Full: } Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon,$$

  where S is a subset of $\{1, \ldots, p\}$ and contains p_0 elements.

- Denote the RSS_0 and RSS_1 the residual sum of squares based on the null model and full model, respectively.
- If the null hypothesis is true, then these two RSS should be close. This leads to the F-statistic

$$F = \frac{(\text{RSS}_0 - \text{RSS}_1)/(p - p_0)}{\text{RSS}_1/(n - p - 1)}.$$

- Under the null hypothesis, the test statistic follows $F_{p-p_0, n-p-1}$.

Example: Testing on Model Adequacy

- Consider two competitive linear models:

$$\text{null:} \quad y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

$$\text{and alternative:} \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i.$$

- We can test the model adequacy by calculating the F-statistic

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p - p_0)}{\mathrm{RSS}_1/(n - p - 1)}.$$

- From previous calculations, we have

$$\mathrm{RSS}_0 = 19473, \mathrm{RSS}_1 = 15418, p_0 = 1, p = 3 \text{ and } n = 506.$$

- The F-statistic takes value 66.01.
- At a significance level $\alpha = 0.05$, the p-value is smaller than 0.0001. Therefore, we reject the null hypothesis.