Linear Regression Models

Paweł Polak

Septmeber 6, 2017

Linear Regression Models - Lecture 1

Course Description

Meetings Time & Location

Monday and Wednesday 2:40 PM - 3:55 PM, 501 Northwest Corner Building

- Instructor: Paweł Polak
  - Office Building: 1255 Amsterdam Ave, Room 928 (SSW, 9th floor)
  - Office Hours: 4:30 PM - 6:00 PM, Monday
    (please send me an email if you plan to come)
  - E-mail: pp2501@columbia.edu
    (please start the title of the email with [GU4205] or [GR5205])
- Teaching Assistant: Tong Li
  - Office Hours: 13:00-14:30 on Thursdays at the lounge room of the Stat Department (10th floor in School of Social Work).
  - E-mail: tl2794@columbia.edu
    (please start the title of the email with [GU4205] or [GR5205] and Cc me)

Course Description

Course content:

- Theory and practice of regression analysis.
- Simple and multiple regression: estimation, testing, and confidence procedures.
- Modeling, regression diagnostics and plots.
- Polynomial regression.
- Collinearity and confounding.
- Model selection.
- Geometry of least squares.
- Shrinkage and Selection Methods (Ridge, LASSO, Elastic Net).
- Introduction to GLM, and PCA.

Course Description

Materials:

- Slides from the lecture & homework materials.
- Textbook:
  - *Applied Linear Regression Models* (**ALRM**) (4th Ed.) by Kutner, Nachtsheim, and Neter.
- Additional Readings:
  - *Statistical Inference* by George Casella and Roger L. Berger;
  - *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* by Trevor Hastie, Robert Tibshirani, and Jerome Friedman (the book is available here: `http://statweb.stanford.edu/~tibs/ElemStatLearn/`).

Course Outline

- Single variable linear regression:
  - Least squares
  - Maximum likelihood, normal model
  - Tests / inferences
  - ANOVA
  - Diagnostics
  - Remedial Measures

Course Outline

- Multiple linear regression and other related topics:
    - Multiple linear Regression
    - Linear algebra review
    - Matrix approach to linear regression
    - Multiple predictor variables
    - Diagnostics
    - Tests
    - Model selection
    - Shrinkage and Selection Methods for Linear Regression (Ridge, LASSO and Elastic Net)
    - Principle Component Analysis
    - Generalized Linear Models

Requirements

- Calculus
  - Derivatives, gradients, convexity
- Linear Algebra
  - Matrix notation, inversion, eigenvectors, eigenvalues, rank
- Probability and Statistics (Appendix A in ALRM book)
  - Random variable
  - Expectation, variance
  - Estimation
  - Bias/Variance
  - Basic probability distributions
  - Hypothesis Testing
  - Confidence Interval.

Software

- R: The R Project for Statistical Computing.
- MATLAB: The Language of Technical Computing.
- Python: High-level, Interpreted, Dynamic Programming Language
- All the examples in the lectures will be made in R or MATLAB.
- For homework you can use R, MATLAB or Python depending on your preference.

Homeworks

- Homeworks (30%)
  - There will be 4-6 HW assignments.
  - Collaboration is allowed in solving the problems, but each student should hand in her or his own independently written solutions.
  - DUE: one week time
  - Homework must be submitted in class.
  - HW cannot be submitted to your TA by e-mail.
  - Please do not contact the TA or the grader directly concerning your grades.
  - Please write [GU4205] or [GR5205] in the subject heading of all e-mail correspondence with instructor/TA. (This is in general effective in weeding out spam email.)
  - No late homework accepted.
  - Lowest score will be dropped.

Grading: 30/25/45

- 30% Homeworks.
- 25% midterm exam (in class):
  - TBA
- 45% final (in class):
  - TBA, (Consult Student Services Online for Final Exam Schedule).
- Exams are closed-book, closed-notes. One double-sided cheat sheet is allowed for each exam.
- An Important Note: no make-up exams will be given.
- The final letter grade depends on your performance in homeworks, midterm, and final exam.

Simple Linear Regression

Why (Linear) Regression?

- Suppose we observe N values of two quantities (e.g. weights and hights of a group of people):
  - $\mathbf{Y} = (Y_1, \ldots, Y_N)$ - the **dependent** variable, the **regressand**, the **response** variable, the **output** variable, **predicted** variable, and
  - $\mathbf{X} = (X_1, \ldots, X_N)$ - the **independent** variable, the **regressor**, the **explanatory** variable, the **input** variable, **predictor** variable, the **exogenous** variable, the **covariate**.
- The observed values of the pair $(\mathbf{Y}, \mathbf{X})$ are called the **sample** or the **data**.
- If we know a function relation between $\mathbf{Y}$, and $\mathbf{X}$, then we can write that

$$\mathbf{Y} = f(\mathbf{X}),$$

  e.g., $\mathbf{X}$ is a number of units sold, $\mathbf{Y}$ is dollar sales, and the price of the product is fixed at p, then

$$\mathbf{Y} = p\mathbf{X} \text{ for given } p > 0.$$

Why (Linear) Regression?

- But (i) the real world is noisy, (ii) perhaps there are other unobserved variables which influence $\mathbf{Y}$, and (iii) the relation between the variables might not be known exactly, i.e., how do you determine f? (e.g., think about the relation between the weight and the hight).
- We have two goals in mind:
  (1) Estimation: Understanding the relationship between the predictor variable $\mathbf{X}$, and the response variable $\mathbf{Y}$.
  (2) Prediction: Predicting the future response given the new observed predictors.
- A **model** is a set of restrictions $\mathcal{R}$ on the joint distribution of the data - dependent and independent variables

$$(\mathbf{Y}, \mathbf{X}) \sim f_{(\mathbf{Y},\mathbf{X})} \in \mathcal{R}.$$

Why (Linear) Regression?

- Linear regression is a model which restricts the joint distribution $f_{(\mathbf{Y},\mathbf{X})}$ by imposing a linear relationship between $\mathbf{Y} = (Y_1, \ldots, Y_N)$ and $\mathbf{X} = (X_1, \ldots, X_N)$, i.e.,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, N,$$

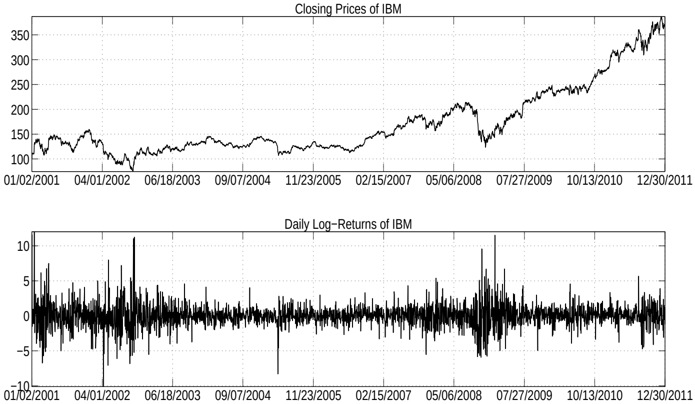
  where β_0 and β_1 are unknown parameters to be estimated, and ε_i is the *unobserved* random error term.
- It is called a **Simple Linear Regression** model because there is only one explanatory variable $\mathbf{X}$.
- For more than one explanatory variable, e.g., $\mathbf{X}_1, \ldots, \mathbf{X}_K$, the model is called **Multiple Linear Regression**

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_K X_{i,K} + \varepsilon_i, \quad \text{for } i = 1, \ldots, N.$$

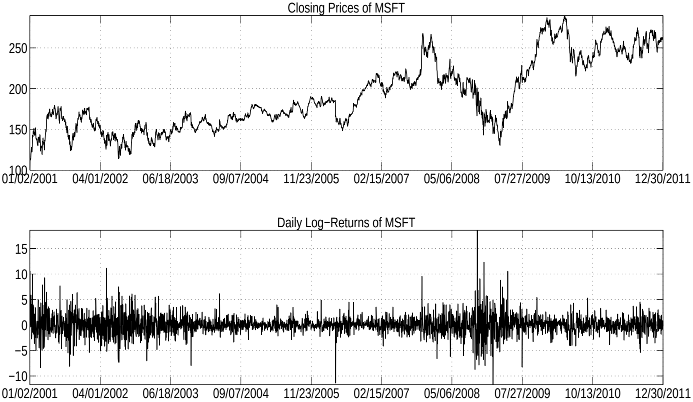
History

- Authors who made substantial contributions are (among others):
  - Adrien-Marie Legendre (1805) and later Carl Friedrich Gauss (1809) developed the least-squares method;
  - The term "regression" was coined by Sir Francis Galton in the late 19th century to describe a biological phenomenon;
  - Cauchy introduced the idea of orthogonality;
  - Chebyshev applied it to polynomial models;
  - Pizzetti found the distribution of the sum of squares of the residuals on the Normal assumption;
  - Karl Pearson (1897), (1903) linked the model with the multivariate Normal thereby broadening the field of applications; and
  - R. A. Fisher (1922) and (1925), extended the orthogonality to qualitative comparisons, and laid the foundations of the modern theory of experimental design; and many others.
- Computational aspect: Before 1970, it sometimes took up to 24 hours to receive the result from one regression.

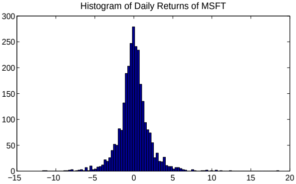
Example 1: Stock Prices

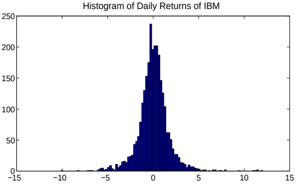


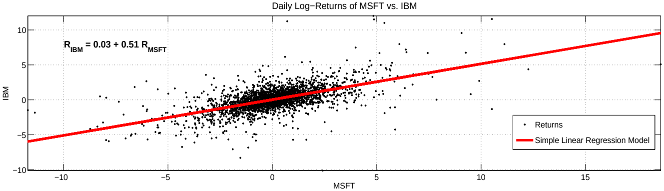
Example 1: Stock Prices



Example 1: Stock Prices

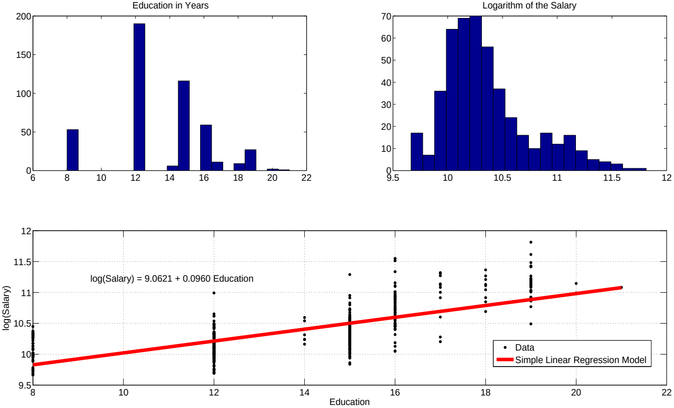






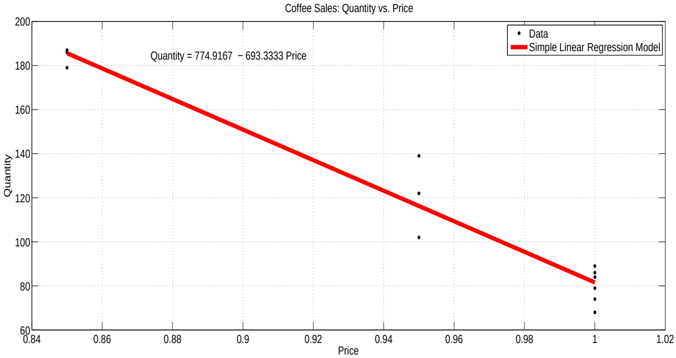
- mean: 0.0470, 0.0348; median: 0.0344, 0; std. dev.: 1.6813, 1.9755,
- skewness: 0.4763, 0.4708; kurtosis: 9.5307, 10.1892.

Example 2: Education vs. Wage in a Bank



- 474 observations on education (in terms of finished years of education) and salary (in logarithms),
- each point in the scatter plot corresponds to the education and salary of an employee,
- on average salaries are higher for higher educated people,
- however, for fixed level of education there remains much variation in salaries.

Example 3: Coffee Sales



- 12 observations on price and quantity sold of a brand of coffee,
- the data were obtained from a controlled marketing experiment in stores in Paris,
- the price is index with value 1 for a usual price, two price actions are investigated, with reduction 5% or 15% of the usual price,
- the quantity sold is in units of coffee per week,
- clearly, lower prices result in higher sales,
- further for a fixed price there remains variation in sales (different values on the vertical axis).

Example 4: Hight vs. Weight

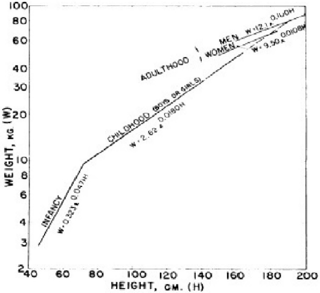


Figure from *Weight-Height Relationship of Young Men and Women* by D. W. Sargent, American Journal of Clinical Nutritions (1963).

- example of *piecewise regression model*, the average relation between height and weight from birth to maturity for men and women.
- in each segment the relation is estimated by a linear regression model.
- each segment has a different constant and relative rate of increase.
- what if we would fit a linear regression model to the whole set of data?

Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, N,$$

where

- Y_i value of the dependent variable for $i = 1, \ldots, N$,
- X_i value of the explanatory variable for $i = 1, \ldots, N$,
- β_0 and β_1 are unknown parameters to be estimated, and
- ε_i is the *unobserved* random error term with mean $\mathbb{E}(\varepsilon_i) = 0$ and variance $\mathrm{Var}(\varepsilon_i) = \sigma^2$,
- ε_i and ε_j are uncorrelated, for $i \neq j$, $i, j = 1, \ldots, N$.

Properties

- The expected value of the predicted variable is

$$\begin{aligned}\mathbb{E}(Y_i) = \mathbb{E}(\beta_0 + \beta_1 X_i + \varepsilon_i) \\ = \beta_0 + \beta_1 X_i + \mathbb{E}(\varepsilon_i) \\ = \beta_0 + \beta_1 X_i,\end{aligned}$$

since $\mathbb{E}(\varepsilon_i) = 0$.

Expectation Review

- Let X be a random variable with probability density function $f(x)$, if $\int |x| f(x) \, \mathrm{d}x < \infty$, then the expected value of X is defined as

$$\mathbb{E}(X) = \int x f(x) \, \mathrm{d}x.$$

- Expected value is linear, i.e.,

  (i) $\mathbb{E}(aX) = a\mathbb{E}(X)$ for any $a \in \mathbb{R}$

  (ii) $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ for any $a, b \in \mathbb{R}$

Example: Expectation Derivation

Suppose p.d.f. of X is $f(x) = 2x, 0 \leq x \leq 1$, then

$$\begin{aligned}
\mathbb{E}(X) &= \int_0^1 xf(x)\,\mathrm{d}x \\
&= \int_0^1 2x^2\mathrm{d}x \\
&= \frac{2}{3}x^3 \mid_0^1 \\
&= \frac{2}{3}.
\end{aligned}$$

Example: Expectation Derivation for Normal Distribution

Suppose $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$, then

$$
\begin{aligned}
\mathbb{E}(X) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \mathrm{d}x \quad (\text{setting } z = x - \mu) \\
&= \underbrace{\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z}_{\text{expected value of } \mathrm{N}\left(0,\sigma^2\right)} + \mu \underbrace{\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z}_{1} \\
&= 0 + \mu = \mu.
\end{aligned}
$$

Expectation of a Product of Random Variables

If X, Y are random variables with joint density function $f(x, y)$, then the expectation of the product is given by

$$\mathbb{E}(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \; f(x, y) \, \mathrm{d}x\mathrm{d}y.$$

- If X and Y are *independent* with density function f_X and f_Y, respectively, then $f(x, y) = f_X(x) f_Y(y)$. Hence,

$$\begin{aligned}
\mathbb{E}(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \; f(x, y) \, \mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \; f_X(x) f_Y(y) \, \mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty} y \; f_Y(y) \left\{ \int_{-\infty}^{\infty} x \; f_X(x) \, \mathrm{d}x \right\} \mathrm{d}y \\
&= \int_{-\infty}^{\infty} y \; f_Y(y) \, \mathbb{E}(X) \, \mathrm{d}y \\
&= \mathbb{E}(X) \int_{-\infty}^{\infty} y \; f_Y(y) \, \mathrm{d}y = \mathbb{E}(X) \, \mathbb{E}(Y).
\end{aligned}$$

Regression Function

Since,

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X} + \varepsilon, \text{ and } \mathbb{E}(\varepsilon) = \mathbf{0},$$

then

- the response Y_i comes from a probability distribution with mean

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 X_i,$$

- this means that the regression function provides the mean of $\mathbf{Y}$ for a given $\mathbf{X}$,

$$\mathbb{E}(\mathbf{Y}) = \beta_0 + \beta_1 \mathbf{X},$$

- the predicted variable Y_i differs from the value of the regression function by the error term amount ε_i

$$\varepsilon_i = Y_i - \mathbb{E}(Y_i).$$

Variance (Second Central Moment) Review

- Discrete distributions: Let X be a random variable with $\mathbb{P}\left(X = x_i\right)$, for $i = 1, \ldots, N$, if $\sum_{i=1}^{N} x_i^2 \mathbb{P}\left(X = x_i\right) < \infty$, then the variance of X is defined as

$$\mathrm{Var}\left(X\right) = \mathbb{E}\left(\left(X - \mathbb{E}\left(X\right)\right)^2\right) = \sum_{i=1}^{N} \left(x_i - \mathbb{E}\left(X\right)\right)^2 \mathbb{P}\left(X = x_i\right).$$

- Continuous distributions: Let X be a random variable with probability density function $f\left(x\right)$, if $\int x^2 f\left(x\right) \mathrm{d}x < \infty$, then the variance of X is defined as

$$\mathrm{Var}\left(X\right) = \mathbb{E}\left(\left(X - \mathbb{E}\left(X\right)\right)^2\right) = \int \left(x - \mathbb{E}\left(X\right)\right)^2 f(x)\mathrm{d}x.$$

- Note that

$$\begin{aligned}\mathrm{Var}\left(X\right) &= \mathbb{E}\left(\left(X - \mathbb{E}\left(X\right)\right)^2\right) \\ &= \mathbb{E}\left(X^2\right) - 2\mathbb{E}\left(X\right)\mathbb{E}\left(X\right) + \left(\mathbb{E}\left(X\right)\right)^2 \\ &= \mathbb{E}\left(X^2\right) - \left(\mathbb{E}\left(X\right)\right)^2.\end{aligned}$$

Example of Variance Derivation

Suppose p.d.f. of X is $f(x) = 2x, 0 \leq x \leq 1$, then

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}\left(X^2\right) - \left(\mathbb{E}(X)\right)^2 \\
&= \int_0^1 2xx^2 \mathrm{d}x - \left(\frac{2}{3}\right)^2 \\
&= \frac{2x^4}{4} \Big|_0^1 - \frac{4}{9} \\
&= \frac{1}{2} - \frac{4}{9} \\
&= \frac{1}{18}.
\end{aligned}$$

Example Variance of Normal Distribution

Suppose $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$, we have seen that $\mathbb{E}(X) = \mu$. Then

$$
\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}(X-\mu)^2 &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \mathrm{d}x \\
&\text{(setting } z = (x-\mu)/\sigma) \\
&= \sigma^2 \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} \mathrm{d}z}_{\text{variance of } \mathrm{N}(0,1)} \\
&= \sigma^2.
\end{aligned}
$$

Variance Properties

- $\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$,
- $\mathrm{Var}(aX + bY) = a^2 \mathrm{Var}(X) + b^2 \mathrm{Var}(Y)$, if X and Y are independent,
- $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$, if a and b are constant,
- More generally

$$\mathrm{Var}\left(\sum_i a_i X_i\right) = \sum_i \sum_j a_i a_j \mathrm{Cov}(X_i, X_j).$$

Covariance

The covariance between two real-valued random variables X an Y, with expected values $\mathbb{E}(X) = \mu$ and $\mathbb{E}(Y) = \nu$, is defined as

$$\operatorname{Cov}(X, Y) = \mathbb{E}((X - \mu)(Y - \nu))$$
$$\mathbb{E}(XY) - \mu\nu.$$

If X is independent of Y, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) = \mu\nu$. Hence,

$$\operatorname{Cov}(X, Y) = 0,$$

for independent random variables.

Response Variance

Since,

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X} + \boldsymbol{\varepsilon}, \text{ and } \mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0},$$

then

$$\begin{aligned} \text{Var}(Y_i) &= \text{Var}\Big(\underbrace{\beta_0 + \beta_1 X_i}_{\text{constant}} + \varepsilon_i\Big) \\ &= \text{Var}(\varepsilon_i) \\ &= \sigma^2. \end{aligned}$$

Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{for } i = 1, \dots, N,$$

where

- Y_i value of the dependent variable for $i = 1, \dots, N$,
- X_i value of the explanatory variable for $i = 1, \dots, N$,
- β_0 and β_1 are unknown parameters to be estimated, and
- ε_i is an *unobserved*, random error term with mean $\mathbb{E}(\varepsilon_i) = 0$ and variance $\mathrm{Var}(\varepsilon_i) = \sigma^2$,
- ε_i and ε_j are uncorrelated, for $i \neq j$, $i, j = 1, \dots, N$.

Properties of Simple Linear Regression Model

- The expected value of the predicted variable is

$$\mathbb{E}(Y_i) = \mathbb{E}(\beta_0 + \beta_1 X_i + \varepsilon_i) = \beta_0 + \beta_1 X_i + \mathbb{E}(\varepsilon_i) = \beta_0 + \beta_1 X_i,$$

since $\mathbb{E}(\varepsilon_i) = 0$.

- This means that the regression function provides the mean of $\mathbf{Y}$ for a given $\mathbf{X}$,

$$\mathbb{E}(\mathbf{Y}) = \beta_0 + \beta_1 \mathbf{X}.$$

- The variance of the predicted variable is given by

$$\mathrm{Var}(Y_i) = \mathrm{Var}\left(\underbrace{\beta_0 + \beta_1 X_i}_{\textbf{constant}} + \varepsilon_i\right) = \mathrm{Var}(\varepsilon_i) = \sigma^2.$$

- The predicted variable Y_i differs from the value of the regression function by the error term amount ε_i

$$\varepsilon_i = Y_i - \mathbb{E}(Y_i).$$

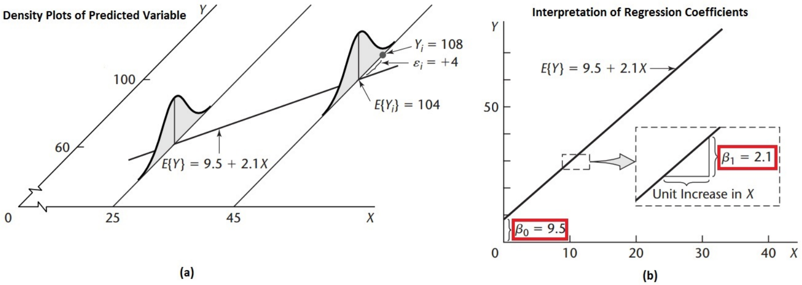
- The error terms are assumed to be uncorrelated, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$. So the predicted variables are uncorrelated, $\mathrm{Cov}(Y_i, Y_j) = 0$.

Meaning of Regression Parameters

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, N,$$

- β_0 and β_1 are called **regression coefficients**:
    - β_1 is the *slope* of the regression line. It indicates the change in the mean of the predicted variable Y per unit increase in X.
    - β_0 is the *intercept* of the regression line. When it is possible that $X = 0$, then β_0 gives the mean of Y at $X = 0$. If it is not possible that $X = 0$, then β_0 has no interpretation.
- We do not know the values of the regression coefficients, and we need to estimate them from the data.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, N,$$



- β_0 and β_1 are called **regression coefficients**:
  - β_1 is the slope of the regression line. It indicates the change in the mean of the predicted variable Y per unit increase in X.
  - β_0 is the intercept of the regression line. When it is possible that $X = 0$, then β_0 gives the mean of Y at $X = 0$. If it is not possible that $X = 0$, then β_0 has no interpretation.
- We do not know the values of the regression coefficients, and we need to estimate them from the data $(\mathbf{Y}, \mathbf{X})$.

Estimation of Regression Function

- Given the data (Y_i, X_i), for $i = 1, \ldots, N$, we want to find "good" estimators of the regression parameters β_0 and β_1.
- We could search for β_0 and β_1 which minimize:

(1) Least Absolute Deviations

$$Q_1(\beta_0, \beta_1) = \sum_{i=1}^{N} |Y_i - \beta_0 - \beta_1 X_i|$$

(2) Least Squares

$$Q_2(\beta_0, \beta_1) = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_i)^2$$

What is the difference between these two criteria?

Illustration of Least Squares Criterion Q_2

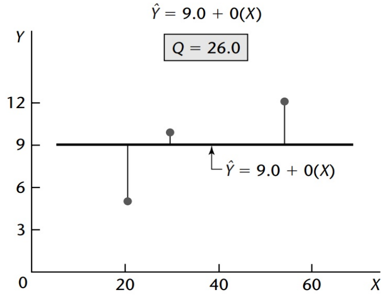


Figure 1.9 in ALRM book.

- The regression line in Figure (a) uses regression coefficients $\beta_0 = 9$ and $\beta_1 = 0$.
- Clearly the regression line in Figure (a) is not a good fit. It has very large deviations for two of the observations.
- The regression line in Figure (b) has much better fit, as indicated by the least squares criterion Q_2.

Least Squares Minimization

$$\{b_0, b_1\} = \arg\min_{\beta_0, \beta_1} Q_2(\beta_0, \beta_1),$$

where $Q_2(\beta_0, \beta_1) = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_i)^2$.

- Find partial derivatives and set both equal to zero:

$$\frac{\partial Q_2}{\partial \beta_0} = 0, \text{ and } \frac{\partial Q_2}{\partial \beta_1} = 0.$$

- In result we obtain the **normal equations**.

$$\sum_{i=1}^{N} Y_i = N b_0 + b_1 \sum_{i=1}^{N} X_i$$

$$\sum_{i=1}^{N} X_i Y_i = b_0 \sum_{i=1}^{N} X_i + b_1 \sum_{i=1}^{N} X_i^2.$$

- b_0 and b_1 are called the estimators of β_0 and β_1, respectively.

Deriving Normal Equations

$$\frac{\partial Q_2}{\partial \beta_0} = -2\sum_{i=1}^{N}(Y_i - \beta_0 - \beta_1 X_i) \qquad \frac{\partial Q_2}{\partial \beta_1} = -2\sum_{i=1}^{N}X_i(Y_i - \beta_0 - \beta_1 X_i)$$

$$\Downarrow \qquad\qquad \Downarrow$$

$$\frac{\partial Q_2}{\partial \beta_0} = 0 \qquad \frac{\partial Q_2}{\partial \beta_1} = 0$$

$$\Downarrow \qquad\qquad \Downarrow$$

$$\sum_{i=1}^{N}(Y_i - b_0 - b_1 X_i) = 0 \qquad \sum_{i=1}^{N}X_i(Y_i - b_0 - b_1 X_i) = 0$$

$$\Downarrow \qquad\qquad \Downarrow$$

$$\sum_{i=1}^{N}Y_i - Nb_0 - b_1\sum_{i=1}^{N}X_i = 0 \qquad \sum_{i=1}^{N}X_iY_i - b_0\sum_{i=1}^{N}X_i - b_1\sum_{i=1}^{N}X_i^2 = 0$$

$$\Downarrow \qquad\qquad \Downarrow$$

$$\sum_{i=1}^{N}Y_i = Nb_0 + b_1\sum_{i=1}^{N}X_i \qquad \sum_{i=1}^{N}X_iY_i = b_0\sum_{i=1}^{N}X_i + b_1\sum_{i=1}^{N}X_i^2$$

Using the second partial derivatives we can show that a minimum is obtained with the least squares estimators b_0 and b_1.

Solution to Normal Equations

The normal equations can be solved simultaneously for b_0 and b_1 to get

$$b_1 = \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2} \qquad \text{and} \qquad b_0 = \bar{Y} - b_1\bar{X}$$

where $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ and $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$.

Model Error Term vs. the Residuals

- The model error term is the difference between the observed value of the predicted variable Y_i and unknown regression line

$$\varepsilon_i = Y_i - \mathbb{E}(Y_i) = Y_i - \beta_0 - \beta_1 X_i.$$

- The residual is the difference between the observed value of the predicted variable Y_i and the corresponding fitted value $\widehat{Y}_i$

$$e_i = Y_i - \widehat{Y}_i = Y_i - b_0 - b_1 X_i.$$

- Model error is unknown/unobserved.
- The residual can be computed from the estimated model.

Properties of Fitted Regression Line

(1) The sum of the residuals is zero:

$$\sum_{i=1}^{N} e_i = 0.$$

(2) The sum of the square residuals $\sum_{i=1}^{N} e_i^2$ is minimized.

(3) The sum of the observed values Y_i equals the sum of the fitted values $\widehat{Y}_i$

$$\sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} \widehat{Y}_i.$$

(4) The sum of the residuals weighted by the predictors X_i is zero

$$\sum_{i=1}^{N} X_i e_i = 0.$$

(5) The sum of the residuals weighted by the fitted value of the response variables Y_i is zero

$$\sum_{i=1}^{N} \widehat{Y}_i e_i = 0.$$

(6) The regression line always goes through the point $(\bar{X}, \bar{Y})$, where $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ and $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$.

Estimation Review

- An estimator is a *random variable* which summarizes the rule for calculating an estimate of a given quantity based on observed sample.
- Point estimator $\widehat{\theta} = \phi(X_1, Y_1, \ldots, X_N, Y_N)$ of unknown quantity/paramter θ, e.g., the sample mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$.
- Interval estimator is a set of possible (or probable) values of an unknown quantity/paramter θ, e.g., confidence intervals.
- Definition: Bias of an estimator

$$\text{Bias}(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \theta.$$

Example: Sample mean vs. Population Mean

Let $Y_1, \ldots, Y_N$ be independent observations drawn from a population with unknown finite mean θ, then the sample mean $\widehat{\theta} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ is an unbiased estimator of θ:

$$
\begin{aligned}
\mathbb{E}\left(\widehat{\theta}\right) &= \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(Y_i\right) = \frac{N\theta}{N} = \theta.
\end{aligned}
$$

- Hence,

$$
\mathbb{E}\left(\widehat{\theta}\right) - \theta = 0.
$$

Variance of an Estimator

- Definition: Variance of an estimator $\text{Var}\left(\widehat{\theta}\right) = \mathbb{E}\left(\left(\widehat{\theta} - \mathbb{E}\left(\widehat{\theta}\right)\right)^2\right)$
- Example: sample mean

$$
\begin{aligned}
\text{Var}\left(\widehat{\theta}\right) &= \text{Var}\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) \\
&= \frac{1}{N^2}\sum_{i=1}^{N} \text{Var}\left(Y_i\right) \\
&= \frac{N\sigma^2}{N^2} \\
&= \frac{\sigma^2}{N}.
\end{aligned}
$$

Estimation of the Variance σ^2 of the Error Terms ε_i

- The variance σ^2 of the error terms ε_i needs to be estimated to obtain an indicator of the variability of the probability distributions of Y.
- Intuitively, inference regarding the regression function and the prediction of Y require an estimate of σ^2.
- Single Population:
  Let $Y_1, \ldots, Y_N$ be independent observations drawn from a population with unknown variance σ^2, then an unbiased estimator of σ^2 is given by

$$s^2 = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}{N-1}$$

- Regression Model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, for $i = 1, \ldots, N$.
  We need to compute the deviations of each observation Y_i around its own mean. Therefore, in regression model we use the *Sum of Square Errors* (SSE)

$$SSE = \sum_{i=1}^{N} \left(Y_i - \widehat{Y}_i\right)^2 .$$

  Now two degrees of freedom are lost because both β_0 and β_1 have to be estimated to obtain the estimates of $\widehat{Y}_i$. Hence, the appropriate Mean Square Error (MSE) or s^2, is

$$s^2 = MSE = \frac{SSE}{N-2} = \frac{\sum_{i=1}^{N} \left(Y_i - \widehat{Y}_i\right)^2}{N-2}$$

- It can be shown that MSE is an unbiased estimator of σ^2 and $\mathbb{E}\left(s^2\right) = \sigma^2$.

MSE & the Bias vs. Variance Trade-Off

- Definition The mean squared error (MSE) of an estimator $\widehat{\theta}$ is given by

$$\text{MSE}\left(\widehat{\theta}\right) = \mathbb{E}\left(\left(\widehat{\theta} - \theta\right)^2\right)$$

- Can be rewritten as

$$\text{MSE}\left(\widehat{\theta}\right) = \text{Var}\left(\widehat{\theta}\right) + \left(\text{Bias}\left(\widehat{\theta}\right)\right)^2$$

$$\begin{aligned}
& \boxed{MSE(\hat{\theta})} \\
= \quad & \mathbb{E}((\hat{\theta} - \theta)^2) \\
= \quad & \mathbb{E}(([\hat{\theta} - \underline{\mathbb{E}(\hat{\theta})}] + [\underline{\mathbb{E}(\hat{\theta})} - \theta])^2) \\
= \quad & \mathbb{E}([\hat{\theta} - \mathbb{E}(\hat{\theta})]^2) + 2\,\mathbb{E}([\mathbb{E}(\hat{\theta}) - \theta][\hat{\theta} - \mathbb{E}(\hat{\theta})]) + \mathbb{E}([\mathbb{E}(\hat{\theta}) - \theta]^2) \\
= \quad & \text{Var}(\hat{\theta}) + 2\,\mathbb{E}(\ \mathbb{E}(\hat{\theta})[\hat{\theta} - \mathbb{E}(\hat{\theta})] - \theta[\hat{\theta} - \mathbb{E}(\hat{\theta})]\ ) + (\text{Bias}(\hat{\theta}))^2 \\
= \quad & \text{Var}(\hat{\theta}) + 2(0 + 0) + (\text{Bias}(\hat{\theta}))^2 \\
\boxed{= \quad & \text{Var}(\hat{\theta}) + (\text{Bias}(\hat{\theta}))^2}
\end{aligned}$$

- Sometimes choosing a biased estimator can result in an overall lower MSE, if it has much lower variance than the unbiased one.

Gauss-Markov Thm: Least Squares Estimator is a BLUE

- BLUE = Best Linear Unbiased Estimator

Recall the Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, N,$$

where

- ε_i is an *unobserved*, random error term with mean $\mathbb{E}(\varepsilon_i) = 0$ and variance $\mathrm{Var}(\varepsilon_i) = \sigma^2$,
- ε_i and ε_j are uncorrelated, for $i \neq j$, $i, j = 1, \ldots, N$.

Gauss-Markov Theorem:

Theorem

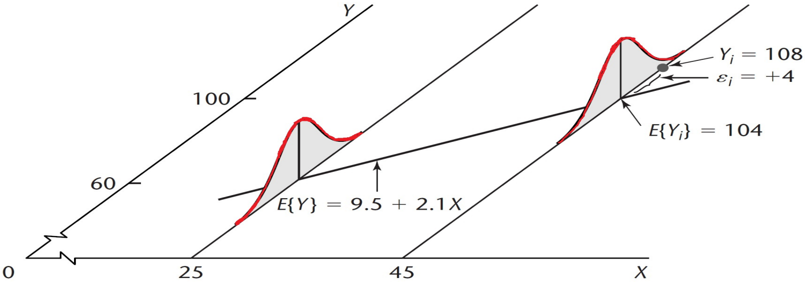
Under the conditions of Simple Linear Regression Model given above, the least squares estimators b_0 and b_1 are unbiased and have minimum variance among all unbiased linear estimators.

Distribution of the Error Term ε

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, N,$$

where

- ε_i is an *unobserved*, random error term with mean $\mathbb{E}(\varepsilon_i) = 0$ and variance $\mathrm{Var}(\varepsilon_i) = \sigma^2$,
- ε_i and ε_j are uncorrelated, for $i \neq j$, $i, j = 1, \ldots, N$.



Normal Error Regression Model

- Gauss-Markov theorem implies that no matter what is the form of the distribution of the error terms ε_i (and hence of Y_i), the least squares estimator is a BLUE among all unbiased linear estimators.
- However, to set up interval estimates, and make tests, we need to impose some assumption about the form of the distribution of the ε_i.
- The most standard assumption is that $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}\left(0, \sigma^2\right)$, i.e., the error terms ε_i are independent and identically distributed (i.i.d.) with the normal distribution with mean 0 and variance σ^2.

Normal Error Regression Model

In result, we get the Normal Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, N,$$

where

- Y_i is a known value of the dependent variable for $i = 1, \ldots, N$,
- X_i is a known value of the explanatory variable for $i = 1, \ldots, N$,
- β_0 and β_1 are unknown parameters to be estimated, and
- $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}\left(0, \sigma^2\right)$ is an unobserved error term for $i = 1, \ldots, N$.

Comments:

- The Normal Regression Model is the same as Simple Regression Model with unspecified error distribution, except that it assumes that the errors ε_i are normally distributed.
- Since the Normal Regression Model assumes that the errors are i.i.d., then they have to be also uncorrelated like in the Simple Regression Model.
- By Gauss-Markov Theorem, under the conditions of Normal Regression Model, the Least Squares Estimator is still a BLUE.
- The value of ε_i has no effect on the value of any other ε_j, nor on any other Y_j, for $j \neq i$.
- $Y_i \sim \mathrm{N}\left(\beta_0 + \beta_1 X_i, \sigma^2\right)$, and Y_i are independent (they are not i.i.d. because they have different means).

Normal Error Regression Model

- Least Squares Minimization

$$\{b_0, b_1\} = \arg\min_{\beta_0,\beta_1} Q_2(\beta_0, \beta_1),$$

where $Q_2(\beta_0, \beta_1) = \sum_{i=1}^{N}(Y_i - \beta_0 - \beta_1 X_i)^2$, and σ^2 is estimated, from the model residuals $e_i = Y_i - \widehat{Y}_i$, by

$$s^2 = MSE = \frac{SSE}{N-2} = \frac{\sum_{i=1}^{N}\left(Y_i - \widehat{Y}_i\right)^2}{N-2}$$

- Maximum Likelihood estimation
We know the distribution of ε_i, hence we also know the distribution of Y_i, $Y_i \sim \mathrm{N}\left(\beta_0 + \beta_1 X_i, \sigma^2\right)$, for $i = 1, \ldots, N$. Therefore, we can find parameters which maximize the log-likelihood function

$$\left\{\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2\right\} = \arg\max_{\beta_0,\beta_1,\sigma^2} \log L\left(\beta_0, \beta_1, \sigma^2\right).$$

Likelihood Function

- If $Y_i \overset{\text{i.i.d.}}{\sim} F(\theta)$, for $i = 1, \ldots, N$, where $\theta \in \Theta$, then the likelihood function is given by

$$L(\theta; Y_1, \ldots, Y_N) = \prod_{n=1}^{N} f_{Y_i}(Y_i; \theta).$$

  - It is the product of p.d.f.'s evaluated at the observations.
  - It is a function of the parameter vector θ.

- The Log-Likelihood Function is given by

$$\log L(\theta; Y_1, \ldots, Y_N) = \sum_{n=1}^{N} \log f_{Y_i}(Y_i; \theta).$$

Maximum Likelihood Estimator

- If you maximize $\log L(\theta; Y_1, \ldots, Y_N)$ with respect to parameters θ (and if a maximum exists), you get the maximum-likelihood estimator (MLE) of θ:

$$\widehat{\theta}_{\text{mle}} = \arg\max_{\theta \in \Theta} \log L(\theta; Y_1, \ldots, Y_N).$$

Comments:

- An MLE estimate is the same regardless of whether we maximize the likelihood or the log-likelihood function, since log is a strictly monotonically increasing function:

$$\widehat{\theta}_{\text{mle}} = \arg\max_{\theta \in \Theta} \log L(\theta; Y_1, \ldots, Y_N) = \arg\max_{\theta \in \Theta} L(\theta; Y_1, \ldots, Y_N).$$

- For many models, a maximum likelihood estimator can be found as an explicit function of the observed data $Y_1, \ldots, Y_N$.
- For many other models, however, no closed-form solution to the maximization problem is known or available, and an MLE has to be found numerically using optimization methods.

Maximum Likelihood Estimator for Normal Error Regression Model

- The joint density function for all the observations $Y_1, \ldots, Y_N$, by the independence property, is given by

$$f\left(y_1, \ldots, y_N \mid X_1, \ldots, X_N; \beta_0, \beta_1, \sigma^2\right) = \prod_{n=1}^{N} f_{Y_i}\left(y_i \mid X_i; \beta_0, \beta_1, \sigma^2\right).$$

- $\varepsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$ implies that $Y_i \sim \mathrm{N}\left(\beta_0 + \beta_1 X_i, \sigma^2\right)$, and

$$f_{Y_i}\left(y_i \mid X_i; \beta_0, \beta_1, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2}.$$

- Once we have the joint density function for all the observations, we can build the likelihood function.

Maximum Likelihood Estimator for Normal Error Regression Model

- In the Normal Regression Model the Log-Likelihood Function is given by

$$
\begin{aligned}
\log L\left(\theta;\mathbf{Y},\mathbf{X}\right) &= \sum_{n=1}^{N} \log f_{Y_i}\left(Y_i;\theta\right) \\
&= \sum_{n=1}^{N} \log\left\{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i-\beta_0-\beta_1 X_i}{\sigma}\right)^2}\right\} \\
&= -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(Y_i-\beta_0-\beta_1 X_i\right)^2 .
\end{aligned}
$$

- If you maximize it with respect to the parameters β_0, β_1, and σ^2, you get...

Maximum Likelihood Estimators for Normal Error Regression Model

- $\widehat{\beta}_0 = b_0$
- $\widehat{\beta}_1 = b_1$
- $\widehat{\sigma}^2 = \frac{\sum_{i=1}^{N} \left(Y_i - \widehat{Y}_i \right)^2}{N}$
- The ML estimator of σ^2 is biased, as s^2 is unbiased and $\widehat{\sigma}^2 = \frac{N-2}{N} s^2$.
- But $\lim_{N \to \infty} \frac{N-2}{N} = 1$, hence for large N the difference between s^2 and $\widehat{\sigma}^2$ is small.

Maximum Likelihood Estimators for Normal Error Regression Model

Comments:

- $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased.
- $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have minimum variance among all unbiased linear estimators.
- In addition, the maximum likelihood estimators, $\widehat{\beta}_0$ and $\widehat{\beta}_1$, and hence also least square estimators b_0 and b_1, for the normal error regression model are
    - consistent,
    - sufficient,
    - minimum variance unbiased, i.e., they have minimum variance in the class of all unbiased estimators (linear or otherwise).

Summary of Lecture 1 (Chapter 1 in ALRM book)

- Simple Linear Regression Model
- Normal Equations
- Bias vs. Variance Trade-off
- Gauss-Markov Theorem
- Normal Error Regression Model
- Maximum Likelihood Estimator

Next Lecture: ALRM Book Chap. 2

- Inference concerning β_1.
- Inference concerning β_0.
- Interval estimation of $\mathbb{E}\left(Y_h\right)$.
- Prediction of new observation.
- Confidence Bands for regression line.