Introduction to Simple Linear Regression

Yang Feng

http://www.stat.columbia.edu/~yangfeng

About me

- Faculty in the Department of Statistics, Columbia University
- Operations Research & Financial Engineering PhD, 2010, Princeton University
- Taught the course over 5 times

My research interests.

- High-dimensional Statistical Learning
- Network models
- Nonparametric and semiparametric methods
- Bioinformatics

See my website for more details:
`http://www.stat.columbia.edu/~yangfeng/`

Course Description

Theory and practice of regression analysis, Simple and multiple regression, including testing, estimation, and confidence procedures, modeling, regression diagnostics and plots, polynomial regression, collinearity and confounding, model selection, geometry of least squares. Extensive use of the computer to analyze data.

Course website: `http://www.stat.columbia.edu/~yangfeng/LinReg`
Required Text: Applied Linear Regression Models (4th Ed.)
Authors: Kutner, Nachtsheim, Neter

Course Outline

First half of the course is single variable linear regression.

- Least squares
- Maximum likelihood, normal model
- Tests / inferences
- ANOVA
- Diagnostics
- Remedial Measures

Course Outline (Continued)

Second half of the course is 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
- Other topics (If time permits)
  - Principle Component Analysis
  - Generalized Linear Models
  - Introduction to Bayesian Inference

Requirements

- Calculus
  - Derivatives, gradients, convexity
- Linear algebra
  - Matrix notation, inversion, eigenvectors, eigenvalues, rank
- Probability and Statistics
  - Random variable
  - Expectation, variance
  - Estimation
  - Bias/Variance
  - Basic probability distributions
  - Hypothesis Testing
  - Confidence Interval

Software

R will be used throughout the course and it is required in all homework. An **R** tutorial session will be given (Bring your laptop with R and RStudio installed!).

Reasons for **R**:

- Completely free software. Can be downloaded from `http://cran.r-project.org/`
- RStudio is an easy to use IDE. `https://www.rstudio.com/`
- Available on various systems, PC, MAC, Linux, and even

Software

R will be used throughout the course and it is required in all homework. An **R** tutorial session will be given (Bring your laptop with R and RStudio installed!).

Reasons for **R**:

- Completely free software. Can be downloaded from `http://cran.r-project.org/`
- RStudio is an easy to use IDE. `https://www.rstudio.com/`
- Available on various systems, PC, MAC, Linux, and even iPhone and iPad!
- Advanced yet easy to use.

An Introduction to **R**: `http://cran.r-project.org/doc/manuals/R-intro.pdf`

R for Data Science: `http://r4ds.had.co.nz/`

Why regression?

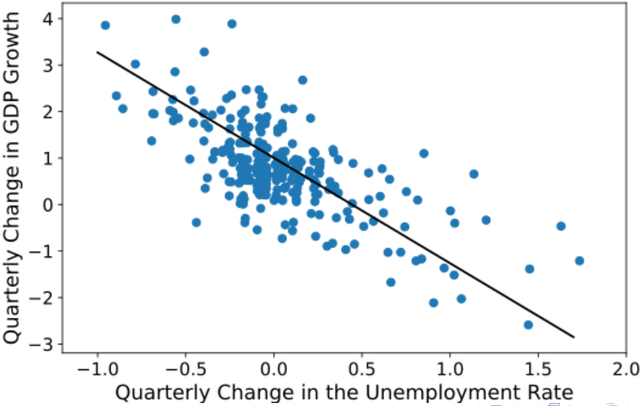
- Want to model a functional relationship between an “predictor variable” (input, independent variable, etc.) and a “response variable” (output, dependent variable, etc.)
  - Examples?
- But real world is noisy, no $f = ma$
  - Observation noise
  - Process noise
- Two distinct goals
  - (Estimation) Understanding the relationship between predictor variables and response variables
  - (Prediction) Predicting the future response given the new observed predictors.

History

- Sir Francis Galton, 19^{th} century
  - Studied the relation between heights of parents and children and noted that the children "regressed" to the population mean
- "Regression" stuck as the term to describe statistical relations between variables

Example Applications

Okun's law: The "gap version" states that for every 1% increase in the unemployment rate, a country's GDP will be roughly an additional 2% lower than its potential GDP.



Others

- Epidemiology
  - Relating lifespan to obesity or smoking habits etc.
- Science and engineering
  - Relating physical inputs to physical outputs in complex systems.
- Education and income
  - Relating the income to the length of education.

Aims for the course

- Given something you would like to predict (the response) and the available explanatory variables.
    - What kind of model should you use?
    - Which variables should you include?
    - Which transformations of variables and interaction terms should you use?

Aims for the course

- Given something you would like to predict (the response) and the available explanatory variables.
  - What kind of model should you use?
  - Which variables should you include?
  - Which transformations of variables and interaction terms should you use?
- Given a model and some data
  - How do you fit the model to the data?
  - How do you express confidence in the values of the model parameters?
  - How do you regularize the model to avoid over-fitting and other related issues?

Data for Regression Analysis

- Observational Data
  Example: relation between age of employee (X) and number of days of illness last year (Y)
  Cannot be controlled!

Data for Regression Analysis

- Observational Data
  Example: relation between age of employee (X) and number of days of illness last year (Y)
  Cannot be controlled!
- Experimental Data
  Example: an insurance company wishes to study the relation between productivity of its analysts in processing claims (Y) and length of training X.
  - **Treatment**: the length of training
  - **Experimental Units**: the analysts included in the study.

Data for Regression Analysis

- Observational Data
  Example: relation between age of employee (X) and number of days of illness last year (Y)
  Cannot be controlled!
- Experimental Data
  Example: an insurance company wishes to study the relation between productivity of its analysts in processing claims (Y) and length of training X.
  - **Treatment**: the length of training
  - **Experimental Units**: the analysts included in the study.
- Completely Randomized Design: Most basic type of statistical design
  Example: same example, but every experimental unit has an equal chance to receive any one of the treatments.

Questions?

Good time to ask now.

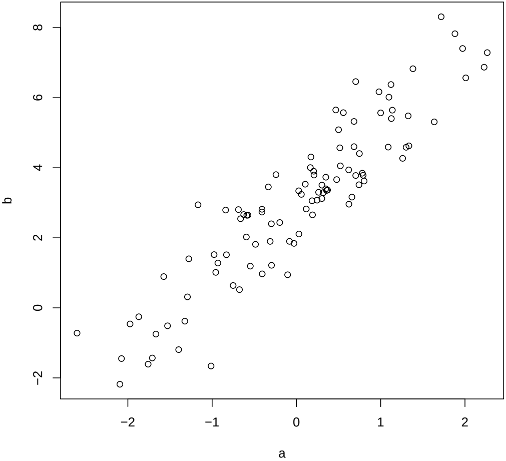
Simple Linear Regression

- Want to find parameters for a function of the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Distribution of error random variable not specified

Quick Example - Scatter Plot



Formal Statement of Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i value of the response variable in the i^{th} trial
- β_0 and β_1 are parameters
- X_i is a known constant, the value of the predictor variable in the i^{th} trial
- ϵ_i is a random error term with mean $\mathbb{E}(\epsilon_i) = 0$ and variance $\mathrm{Var}(\epsilon_i) = \sigma^2$
- ϵ_i and ϵ_j are uncorrelated
- $i = 1, \ldots, n$

Properties

- The response Y_i is the sum of two components
  - Constant term $\beta_0 + \beta_1 X_i$
  - Random term ϵ_i
- The expected response is

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

Expectation Review

- Suppose X has probability density function $f(x)$.
- Definition

$$\mathbb{E}(X) = \int xf(x)dx.$$

- Linearity property

$$\begin{aligned} \mathbb{E}(aX) &= a\,\mathbb{E}(X) \\ \mathbb{E}(aX + bY) &= a\,\mathbb{E}(X) + b\,\mathbb{E}(Y) \end{aligned}$$

- Obvious from definition

Example Expectation Derivation

Suppose p.d.f of X is

$$f(x) = 2x, 0 \leq x \leq 1$$

Expectation

$$\begin{aligned}
\mathbb{E}(X) &= \int_0^1 xf(x)dx \\
&= \int_0^1 2x^2 dx \\
&= \frac{2x^3}{3}|_0^1 \\
&= \frac{2}{3}
\end{aligned}$$

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

$$\mathbb{E}(XY) = \int_{XY} xyf(x, y)dxdy.$$

Expectation of a product of random variables

What if X and Y are independent? If X and Y are independent with density functions f_1 and f_2 respectively then

$$\begin{aligned}
\mathbb{E}(XY) &= \int_{XY} xyf_1(x)f_2(y)dxdy \\
&= \int_X \int_Y xyf_1(x)f_2(Y)dxdy \\
&= \int_X xf_1(x)[\int_Y yf_2(y)dy]dx \\
&= \int_X xf_1(x)\,\mathbb{E}(Y)dX \\
&= \mathbb{E}(X)\,\mathbb{E}(Y)
\end{aligned}$$

Regression Function

- The response Y_i comes from a probability distribution with mean

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

- This means the regression function is

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

Since the regression function relates the means of the probability distributions of Y for a given X to the level of X

Error Terms

- The response Y_i in the i^{th} trial exceeds or falls short of the value of the regression function by the error term amount ϵ_i
- The error terms ϵ_i are assumed to have constant variance σ^2

Response Variance

Responses Y_i have the same constant variance

$$\begin{aligned}\text{Var}(Y_i) &= \text{Var}(\beta_0 + \beta_1 X_i + \epsilon_i)\\ &= \text{Var}(\epsilon_i)\\ &= \sigma^2\end{aligned}$$

Variance (2^{nd} central moment) Review

- Continuous distribution

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \int (x - \mathbb{E}(X))^2 f(x)dx$$

- Discrete distribution

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \sum_i (X_i - \mathbb{E}(X))^2 P(X = X_i)$$

- Alternative Form for Variance:

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

Example Variance Derivation

$$f(x) = 2x, 0 \leq x \leq 1$$

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

Variance Properties

$$
\begin{aligned}
\mathrm{Var}(aX) &= a^2\,\mathrm{Var}(X) \\
\mathrm{Var}(aX + bY) &= a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) \text{ if } X \perp Y \\
\mathrm{Var}(a + cX) &= c^2\,\mathrm{Var}(X) \text{ if } a \text{ and } c \text{ are both constants}
\end{aligned}
$$

More generally

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

Covariance

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

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}((X - \mu)(Y - \nu)) \\ &= \mathbb{E}(XY) - \mu\nu. \end{aligned}$$

If $X \perp Y$,

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

Then

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

Formal Statement of Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i value of the response variable in the i^{th} trial
- β_0 and β_1 are parameters
- X_i is a known constant, the value of the predictor variable in the i^{th} trial
- ϵ_i is a random error term with mean $\mathbb{E}(\epsilon_i) = 0$ and variance $\mathrm{Var}(\epsilon_i) = \sigma^2$
- $i = 1, \ldots, n$

Example

An experimenter gave three subjects a very difficult task. Data on the age of the subject (X) and on the number of attempts to accomplish the task before giving up (Y) follow:

Table:

Subject i	1	2	3
Age X_i	20	55	30
Number of Attempts Y_i	5	12	10

Least Squares Linear Regression

- Goal: make Y_i and $b_0 + b_1 X_i$ close for all i.

Least Squares Linear Regression

- Goal: make Y_i and $b_0 + b_1 X_i$ close for all i.
- Proposal 1: minimize $\sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]$.
- Proposal 2: minimize $\sum_{i=1}^{n} |Y_i - (b_0 + b_1 X_i)|$.

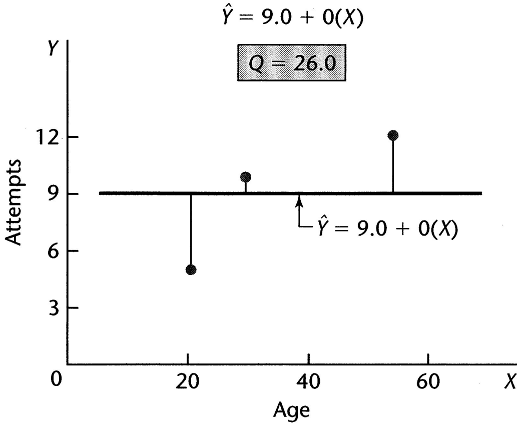
Least Squares Linear Regression

- Goal: make Y_i and $b_0 + b_1 X_i$ close for all i.
- Proposal 1: minimize $\sum_{i=1}^{n}[Y_i - (b_0 + b_1 X_i)]$.
- Proposal 2: minimize $\sum_{i=1}^{n}|Y_i - (b_0 + b_1 X_i)|$.
- Final Proposal: minimize

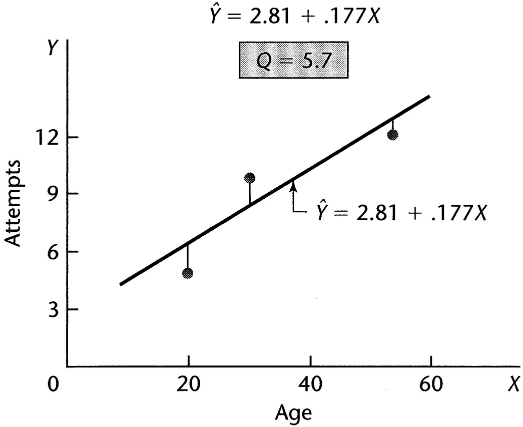
$$Q(b_0, b_1) = \sum_{i=1}^{n}[Y_i - (b_0 + b_1 X_i)]^2$$

- Choose b_0 and b_1 as estimators for β_0 and β_1.
- b_0 and b_1 will minimize the criterion Q for the given sample observations $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$.

Guess #1



Guess #2


$\hat{Y} = 2.81 + .177X$
$Q = 5.7$
Y
Attempts
12
9
6
3
0
20
40
60
X
Age
$\hat{Y} = 2.81 + .177X$

Function maximization

- Important technique to remember!
  - Take derivative
  - Set result equal to zero and solve
  - Test second derivative at that point
- Question: does this always give you the maximum?
- Going further: multiple variables, convex optimization

Function Maximization

Find the value of x that maximize the function

$$f(x) = -x^2 + \log(x), x > 0$$

Function Maximization

Find the value of x that maximize the function

$$f(x) = -x^2 + \log(x), x > 0$$

1. Take derivative

$$\begin{aligned} \frac{d}{dx}(-x^2 + \log(x)) &= 0 && (1) \\ -2x + \frac{1}{x} &= 0 && (2) \\ 2x^2 &= 1 && (3) \\ x^2 &= \frac{1}{2} && (4) \\ x &= \frac{\sqrt{2}}{2} && (5) \end{aligned}$$

2. Check second order derivative

$$\begin{aligned}
\frac{d^2}{dx^2}(-x^2+\log(x)) &= \frac{d}{dx}(-2x+\frac{1}{x}) && (6)\\
&= -2-\frac{1}{x^2} && (7)\\
&< 0 && (8)
\end{aligned}$$

Then we have found the maximum. $x^* = \frac{\sqrt{2}}{2}$, $f(x^*) = -\frac{1}{2}[1+\log(2)]$.

Least Squares Minimization

- Function to minimize w.r.t. b_0 and b_1
- b_0 and b_1 are called point estimators of β_0 and β_1 respectively

$$Q(b_0, b_1) = \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

- Find partial derivatives and set both equal to zero

$$\begin{aligned} \frac{\partial Q}{\partial b_0} &= 0 \\ \frac{\partial Q}{\partial b_1} &= 0 \end{aligned}$$

Normal Equations

- The result of this maximization step are called the normal equations. b_0 and b_1 are called point estimators of β_0 and β_1 respectively.

$$\begin{aligned} \sum Y_i &= nb_0 + b_1 \sum X_i && (9) \\ \sum X_i Y_i &= b_0 \sum X_i + b_1 \sum X_i^2 && (10) \end{aligned}$$

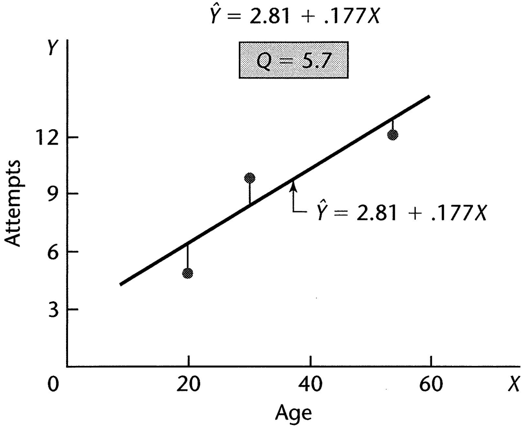
- This is a system of two equations and two unknowns.

Solution to Normal Equations

After a lot of algebra one arrives at

$$
\begin{aligned}
b_1 &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \\
b_0 &= \bar{Y} - b_1\bar{X} \\
\bar{X} &= \frac{\sum X_i}{n} \\
\bar{Y} &= \frac{\sum Y_i}{n}
\end{aligned}
$$

Least Square Fit



Properties of Solution

- The i^{th} residual is defined to be

$$e_i = Y_i - \hat{Y}_i$$

- The sum of the residuals is zero from (9).

$$\begin{aligned}\sum_i e_i &= \sum (Y_i - b_0 - b_1 X_i) \\ &= \sum Y_i - nb_0 - b_1 \sum X_i \\ &= 0\end{aligned}$$

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

$$\sum_i \hat{Y}_i = \sum_i Y_i$$

Properties of Solution

The sum of the weighted residuals is zero when the residual in the i^{th} trial is weighted by the level of the predictor variable in the i^{th} trial from (10).

$$\begin{aligned}
\sum_i X_i e_i &= \sum (X_i(Y_i - b_0 - b_1 X_i)) \\
&= \sum_i X_i Y_i - b_0 \sum X_i - b_1 \sum X_i^2 \\
&= 0
\end{aligned}$$

Properties of Solution

The sum of the weighted residuals is zero when the residual in the i^{th} trial is weighted by the fitted value of the response variable in the i^{th} trial

$$\begin{aligned}\sum_i \hat{Y}_i e_i &= \sum_i (b_0 + b_1 X_i) e_i \\ &= b_0 \sum_i e_i + b_1 \sum_i X_i e_i \\ &= 0\end{aligned}$$

Properties of Solution

The regression line always goes through the point

$$\bar{X}, \bar{Y}$$

Using the alternative format of linear regression model:

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \epsilon_i$$

The least squares estimator b_1 for β_1 remains the same as before. The least squares estimator for $\beta_0^* = \beta_0 + \beta_1\bar{X}$ becomes

$$b_0^* = b_0 + b_1\bar{X} = (\bar{Y} - b_1\bar{X}) + b_1\bar{X} = \bar{Y}$$

Hence the estimated regression function is

$$\hat{Y} = \bar{Y} + b_1(X - \bar{X})$$

Apparently, the regression line always goes through the point $(\bar{X}, \bar{Y})$.

Estimating Error Term Variance σ^2

- Review estimation in non-regression setting.
- Show estimation results for regression setting.

Estimation Review

- An estimator is a rule that tells how to calculate the value of an estimate based on the measurements contained in a sample
- i.e. the sample mean

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Point Estimators and Bias

- Point estimator

$$\hat{\theta} = f(\{Y_1, \ldots, Y_n\})$$

- Unknown quantity / parameter

$$\theta$$

- Definition: Bias of estimator

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

Example

- Samples

$$Y_i \sim N(\theta, \sigma^2)$$

- Estimate the population mean

$$\hat{\theta} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Sampling Distribution of the Estimator

- First moment

$$\begin{aligned}\mathbb{E}(\hat{\theta}) &= \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} Y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i) = \frac{n\theta}{n} = \theta\end{aligned}$$

- This is an example of an unbiased estimator

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta = 0$$

Variance of Estimator

- Definition: Variance of estimator

$$\mathsf{Var}(\hat{\theta}) = \mathbb{E}([\hat{\theta} - \mathbb{E}(\hat{\theta})]^2)$$

- Remember:

$$\begin{aligned} \mathsf{Var}(cY) &= c^2\,\mathsf{Var}(Y) \\ \mathsf{Var}(\sum_{i=1}^{n} Y_i) &= \sum_{i=1}^{n} \mathsf{Var}(Y_i) \end{aligned}$$

Only if the Y_i are independent with finite variance

Example Estimator Variance

$$\begin{aligned}\mathrm{Var}(\hat{\theta}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i)\\ &= \frac{n\sigma^2}{n^2}\\ &= \frac{\sigma^2}{n}\end{aligned}$$

Bias Variance Trade-off

- The mean squared error of an estimator

$$MSE(\hat{\theta}) = \mathbb{E}([\hat{\theta} - \theta]^2)$$

- Can be re-expressed

$$MSE(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + [\mathrm{Bias}(\hat{\theta})]^2$$

MSE $=$ Var $+$ Bias^2

Proof

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

Trade-off

- Think of variance as confidence and bias as correctness.
  - Intuitions (largely) apply
- Sometimes choosing a biased estimator can result in an overall lower MSE if it has much lower variance than the unbiased one.

s^2 estimator for σ^2 for Single population

Sum of Squares:

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

Sample Variance Estimator:

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

- s^2 is an unbiased estimator of σ^2.
- The sum of squares SSE has $n-1$ "degrees of freedom" associated with it, one degree of freedom is lost by using $\bar{Y}$ as an estimate of the unknown population mean μ.

Estimating Error Term Variance σ^2 for Regression Model

- Regression model
- Variance of each observation Y_i is σ^2 (the same as for the error term ϵ_i)
- Each Y_i comes from a different probability distribution with different means that depend on the level X_i
- The deviation of an observation Y_i must be calculated around its own estimated mean.

s^2 estimator for σ^2

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

- MSE is an unbiased estimator of σ^2

$$\mathbb{E}(MSE) = \sigma^2$$

- The sum of squares SSE has $n-2$ “degrees of freedom” associated with it.
- Cochran’s theorem (later in the course) tells us where degree’s of freedom come from and how to calculate them.

Normal Error Regression Model

- No matter how the error terms ϵ_i are distributed, the least squares method provides unbiased point estimators of β_0 and β_1
  - that also have minimum variance among all unbiased linear estimators
- To set up interval estimates and make tests we need to specify the distribution of the ϵ_i
- We will assume that the ϵ_i are normally distributed.

Normal Error Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i value of the response variable in the i^{th} trial
- β_0 and β_1 are parameters
- X_i is a known constant, the value of the predictor variable in the i^{th} trial
- $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
  note this is different, now we know the distribution
- $i = 1, \ldots, n$

Notational Convention

- When you see $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
- It is read as ϵ_i is identically and independently distributed according to a normal distribution with mean 0 and variance σ^2

Maximum Likelihood Principle

The method of maximum likelihood chooses as estimates those values of the parameters that are most consistent with the sample data.

Likelihood Function

If

$$Y_i \sim F(\Theta), i = 1 \dots n$$

then the likelihood function is

$$\mathcal{L}(\{Y_i\}_{i=1}^n, \Theta) = \prod_{i=1}^n f(Y_i; \Theta)$$

Maximum Likelihood Estimation

- The likelihood function can be maximized w.r.t. the parameter(s) Θ, doing this one can arrive at estimators for parameters as well.

$$\mathcal{L}(\{Y_i\}_{i=1}^n, \Theta) = \prod_{i=1}^n f(Y_i; \Theta)$$

- To do this, find solutions to (analytically or by following gradient)

$$\frac{d\mathcal{L}(\{Y_i\}_{i=1}^n, \Theta)}{d\Theta} = 0$$

Important Trick

(almost) Never maximize the likelihood function, maximize the log likelihood function instead.

$$\begin{aligned}\log(\mathcal{L}(\{Y_i\}_{i=1}^n, \Theta)) &= \log(\prod_{i=1}^n f(Y_i; \Theta)) \\ &= \sum_{i=1}^n \log(f(Y_i; \Theta))\end{aligned}$$

Usually the log of the likelihood is easier to work with mathematically.

ML Normal Regression

Likelihood function

$$
\begin{aligned}
\mathcal{L}(\{Y_i\}_{i=1}^n; \beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2}.
\end{aligned}
$$

We can maximize (how?) w.r.t. to the parameters $\beta_0, \beta_1, \sigma^2$, and get...

Maximum Likelihood Estimator(s)

- β_0
  b_0 same as in least squares case
- β_1
  b_1 same as in least squares case
- σ^2

$$\hat{\sigma}^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n}$$

- Note that ML estimator is biased as s^2 is unbiased and

$$s^2 = MSE = \frac{n}{n-2}\hat{\sigma}^2$$

Comments

- Least squares minimizes the squared error between the prediction and the true output
- The normal distribution is fully characterized by its first two central moments (mean and variance)