STA302/1001: Methods of Data Analysis

Instructor: Fang Yao

fyao@utstat.toronto.edu

http://www.utstat.utoronto.ca/fyao

Sidney Smith 6027B

Chapter 1: Introduction and Review

An Important Question

Why do we study statistics?

Because we are lazy and stupid!

Course Focus: Applied Regression Analysis

- What is “Regression Analysis”?
- What is “Applied”? Compared to old days...
- Regression Analysis: possibly the most widely studied and applicable statistical tool
- there are many types
- and you will learn the simplest, but important: linear regression analysis
- before we proceed, two more general questions:
- What is probability? What is statistics?

What is Probability?

- “forward logic” or deductive logic
- know the cause and predict the result
- you know the random behavior of something, and you calculate the chance of some other thing happens
- e.g., you toss a fair coin 3 times and calculate the probability of getting 2 heads
- (something to think about: does a fair coin or die exist?)

What is Statistics?

- “backward logic” or inductive logic
- observe the result and induce the cause
- you observe something and try to guess “what happened before” or “what was the truth”
- e.g., you tossed a coin 10 times and obtained 9 heads - is the coin fair?
- this is statistical inference
- parameter estimation, confidence intervals, hypothesis tests, prediction
- we need statistical models

Statistical Models

- "All models are wrong, some are useful" - George Box
- we don't know the truth so we approximate it
- a good example: Newton's Laws of Motion
- we know they are incorrect, but also extremely useful
- for some problems, the "truth" is not that easy to approximate
- e.g., problems in social science
- so we are stupid ...

Statistical Models (cont...)

- for some other problems, it is not worth our time or money or energy to find the “truth”
- can you think of any examples?
- so we are also lazy ...
- statistical modeling: we use randomness in our approximations
- we use randomness to handle things that
  1. we don’t know about
  2. we don’t want to spend too much time or money or energy on
  3. or both

Regression Analysis

- regression analysis: one kind of statistical modeling
- simplest case: $Y = a + bX + \epsilon$
- e.g., Y: income, X: years of education
- “how the changes in X impact on Y?”, or
  “given X, I want to predict Y”
- questions:
  1. how to estimate a and b?
  2. how to construct confidence intervals?
  3. good model? or should we use $Y = a + bX + cX^2 + \epsilon$?
  4. and many more ...

In This Course

- you will learn a set of new techniques
- some for estimation, some for constructing confidence intervals, some for other things
- you will need to know when to use which, and use it well
- intermediate goal: want to know “what happened”, make prediction, etc
- ultimate goal: ?
- read Chapter 1 of the textbook: data examples

Review of Basic (Appendix A.2 of Text)

- Let $u_1, \ldots, u_n$ be n random variables.
- Let $a_0, \ldots, a_n$ be $n+1$ constants.
- $\mathrm{E}(u_i)$: expectation of u_i
- interpretation: If we observe u_i many times, then $\mathrm{E}(u_i)$ is the average.
- Rules:

$$\begin{aligned} \mathrm{E}(a_0 + a_1 u_1) &= a_0 + a_1 \mathrm{E}(u_1) \\ \mathrm{E}(a_0 + a_1 u_1 + \cdots + a_n u_n) &= a_0 + a_1 \mathrm{E}(u_1) + \cdots + a_n \mathrm{E}(u_n) \end{aligned}$$

Review of Basic (cont...)

- $\mathrm{Var}(u_i) = \mathrm{E}(u_i - E(u_i))^2$: variance of u_i
- For independent random variables:

$$\mathrm{Var}(a_0 + a_1 u_1 + \cdots + a_n u_n) = a_1^2 \mathrm{Var}(u_1) + \cdots + a_n^2 \mathrm{Var}(u_n)$$

- $\mathrm{Cov}(u_i, u_j)$: covariance of u_i and u_j
- describes the way two random variables vary jointly
- $$\begin{aligned}\mathrm{Cov}(u_i, u_j) &= \mathrm{E}[(u_i - \mathrm{E}(u_i))(u_j - \mathrm{E}(u_j))] \\ &= \mathrm{Cov}(u_j, u_i)\end{aligned}$$
- $\mathrm{Cov}(u_i, u_i) = \mathrm{Var}(u_i)$

Review of Basic (cont...)

- rules:

$$
\begin{aligned}
\mathrm{Cov}(a_0 + a_1 u_1, a_3 + a_2 u_2) &= a_1 a_2 \mathrm{Cov}(u_1, u_2) \\
\mathrm{Var}(a_0 + a_1 u_1 + \cdots + a_n u_n) &= a_1^2 \mathrm{Var}(u_1) + \cdots + a_n^2 \mathrm{Var}(u_n) \\
&\quad + 2 \sum_{i<j} a_i a_j \mathrm{Cov}(u_i, u_j)
\end{aligned}
$$

- correlation coefficient:

$$
\rho(u_i, u_j) = \frac{\mathrm{Cov}(u_i, u_j)}{\sqrt{\mathrm{Var}(u_i)\mathrm{Var}(u_j)}}
$$

- Can you re-express the above in vector/matrix forms?

Review of Basic (cont...)

- conditional means and conditional variances
- what we have seen so far are *unconditional means* and *unconditional variances*
- conditional mean (same as conditional expectation): $\mathrm{E}(Y|X = x)$: the expectation of Y when the value of X is fixed at $X = x$
- similar for conditional variance: $\mathrm{Var}(Y|X = x)$: the variance of Y when X is fixed at $X = x$
- more on background: normal, Student's-t, confidence intervals and hypothesis testing.