STATISTICS 3A03

Applied Regression Analysis with SAS

Angelo J. Canty

Office : Hamilton Hall 209

Phone : (905) 525-9140 extn 27079

E-mail : cantya@mcmaster.ca

SAS Labs :

L1	Friday	11:30	in BSB 249
L2	Tuesday	11:30	in BSB 244
L3	Thursday	8:30	in BSB 249

Office Hours :

Monday	10:30-11:30
Tuesday	11:30-12:30
Thursday	10:30-11:30

or at other times by e-mail appointment.

Website : All lecture notes will be available on the website and any announcements about the course will be made there.

http://www.math.mcmaster.ca/canty/teaching/stat3a03

Required Text : *Regression Analysis By Example* (5th Edition) Samprit Chatterjee & Ali S. Hadi. Wiley (2012)

Optional (recommended) Text : *Applied Linear Regression* (4th Edition) Sanford Weisberg. Wiley (2014)

Distributions and Moments

* The distribution of a continuous random variable, Y, is usually described by a Probability Density Function, $f_Y(y)$.

* Two important characteristics of a distribution are the mean $\mathsf{E}(Y)$ and the variance $\mathsf{Var}(Y)$.

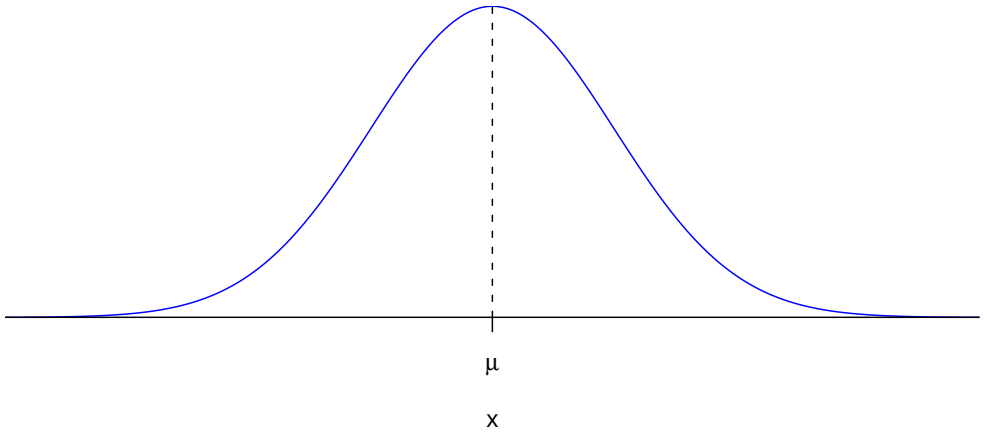
* These describe the central location and amount of spread in the distribution of the random variable Y.

* Two important rules for calculating moments are
  1. $\mathsf{E}(aY+b) = a\mathsf{E}(Y)+b$ for any constants $a, b \in \mathbb{R}$.
  2. $\mathsf{Var}(aY+b) = a^2\mathsf{Var}(Y)$ for any constants $a, b \in \mathbb{R}$.

The Normal Distribution

* The Normal or Gaussian distribution is commonly used to model naturally occurring phenomena.
* It is actually a family of distributions characterized by two parameters, the mean μ and the variance σ^2.
* The density function is given by

$$f(y;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \qquad -\infty < x < \infty$$



Random Vectors

* Two random variables X and Y have a joint distribution.

* In the continuous case there is a joint density function $f(x, y)$.

* The marginal pdf for one variable alone can be found by integrating over the other.

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

Conditional Distributions

* We are often interested in the distribution of Y given the additional information that $X = x$ for some value x.

* This gives rise to the conditional distribution of Y given $X = x$.

* The conditional density is

$$f_{Y|X}(y \mid x) \;=\; \frac{f(x,y)}{f_X(x)}$$

* In this course we will be particularly interested in modelling the conditional expected value

$$\mathrm{E}(Y \mid X = x) \;=\; \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x)\, dy$$

* This will be a function of the value of x used.

Random Samples

* In statistics we deal with random samples.

* A set of random variables $Y_1, \ldots, Y_n$ is a random sample if
  1. All of the random variables have the same marginal distribution $f_Y(y)$.
  2. The set of random variables is independent

$$f(y_1, \ldots, y_n) \;=\; f_Y(y_1) f_Y(y_2) \cdots f_Y(y_n)$$

* A random sample is also called an independent and identically distributed (*iid*) set of random variables.

Moments of Linear Combinations

* Many important estimators are linear combinations of independent random variables.

* The moments of these estimators can often be found using the following theorem.

Theorem 1

If $Y_1, \ldots, Y_n$ are independent random variables and $a_1, \ldots, a_n$ are real constants then

1. $$\mathsf{E}\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i \mathsf{E}(Y_i)$$

2. $$\mathsf{Var}\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i^2 \mathsf{Var}(Y_i).$$

Statistical Inference

* In statistical inference we have a random sample $Y_1, \ldots, Y_n$ with common density $f(y; \theta)$.

* θ is a parameter of the distribution whose value is unknown.

* We wish to use the sample to make inferences about θ.

* The three types of inference we make are
  1. Estimation of θ.
  2. Confidence Intervals for θ.
  3. Hypothesis Testing that θ takes on a certain value.

Inference for Normal Distributions

* The sample mean

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

is the usual estimator for μ.

* The sample variance

$$S^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

is the usual estimator for σ^2.

* These are both unbiased estimators in that

$$\mathrm{E}(\overline{Y}) \;=\; \mu \qquad \text{and} \qquad \mathrm{E}(S^2) \;=\; \sigma^2$$

Inference for Normal Distributions (ctd)

* The expectations are taken relative to the distributions of $\overline{Y}$ and S^2 when considering repeated sampling from the normal distribution.

* These are referred to as the Sampling Distributions of the estimators.

* If $Y_1, \ldots, Y_n$ are *iid* $\mathsf{N}(\mu, \sigma^2)$ then the sampling distribution of $\overline{Y}$ is

$$\overline{Y} \sim \mathsf{N}(\mu, \sigma^2/n).$$

* When σ^2 is known we can use

$$Z = \frac{\sqrt{n}(\overline{Y} - \mu)}{\sigma} \sim \mathsf{N}(0, 1)$$

to make inference.

Inference for Normal Distributions (ctd)

* Usually (and always in this course) σ^2 is unknown.

* For that reason we use

$$T = \frac{\sqrt{n}(\overline{Y} - \mu)}{S} \sim t_{n-1}$$

  for inference.

* t_{n-1} denote's the Student's t distribution with $n-1$ degrees of freedom.

* Symmetric distribution about 0 but with heavier tails than the standard normal to account for the variability in S as an estimator of σ.

Inference for Normal Distributions (ctd)

* A confidence interval is a random interval which contains the true value of the parameter θ for a large percentage (usually 95%) of samples.

* A confidence interval for the mean is

$$\overline{y} \ \pm \ t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

* In hypothesis testing we specify a null value for the parameter (often, but not always, 0) and see if the data supports this value.

* Evidence against the null hypothesis is usually based on a p-value.

* If the p-value is small we say that there is evidence against the null hypothesis.

Introduction to Regression

* The aim of regression is to model the dependence of one variable Y on a set of variables $X_1, \ldots, X_p$.

* Y is called the dependent variable or the response variable.

* $X_1, \ldots, X_p$ are called the independent variables or covariates.

* In this course Y will be a continuous or quantitative variable but the covariates may be continuous or discrete.

The General Regression Model

* A general model for predicting Y given the covariates $X_1, \ldots, X_p$ would be

$$Y = f(X_1, \ldots, X_p) + \varepsilon$$

* The term ε is usually called the Random Error and explains the variability of the random variable Y about $f(X_1, \ldots, X_p)$.

* If we specify that $\mathrm{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$ and that ε is independent of the covariates then we see that this implies

$$\begin{aligned} \mathrm{E}(Y \mid X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) &= f(x_1, x_2, \ldots, x_p) \\ \mathrm{Var}(Y \mid X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) &= \sigma^2 \end{aligned}$$

* In most modelling situations the form of f will be determined by the analyst and it will typically depend on a set of unknown parameters $\beta_0, \beta_1, \ldots, \beta_p$.

The Linear Regression Model

* In this course we will deal with the situation where the parameters enter the function f in a linear way.

* This results in the linear model

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

* Furthermore we will generally assume that the random error ε is normally distributed with mean 0 and unknown variance σ^2 and is independent of the covariates.

* The aim of regression is then to make inferences about the $p + 2$ unknown parameters in this model.

Example: Fish age and Length

* The growth pattern of game fish, such as the smallmouth bass, is of interest to agencies who manage stocks in inland lakes.

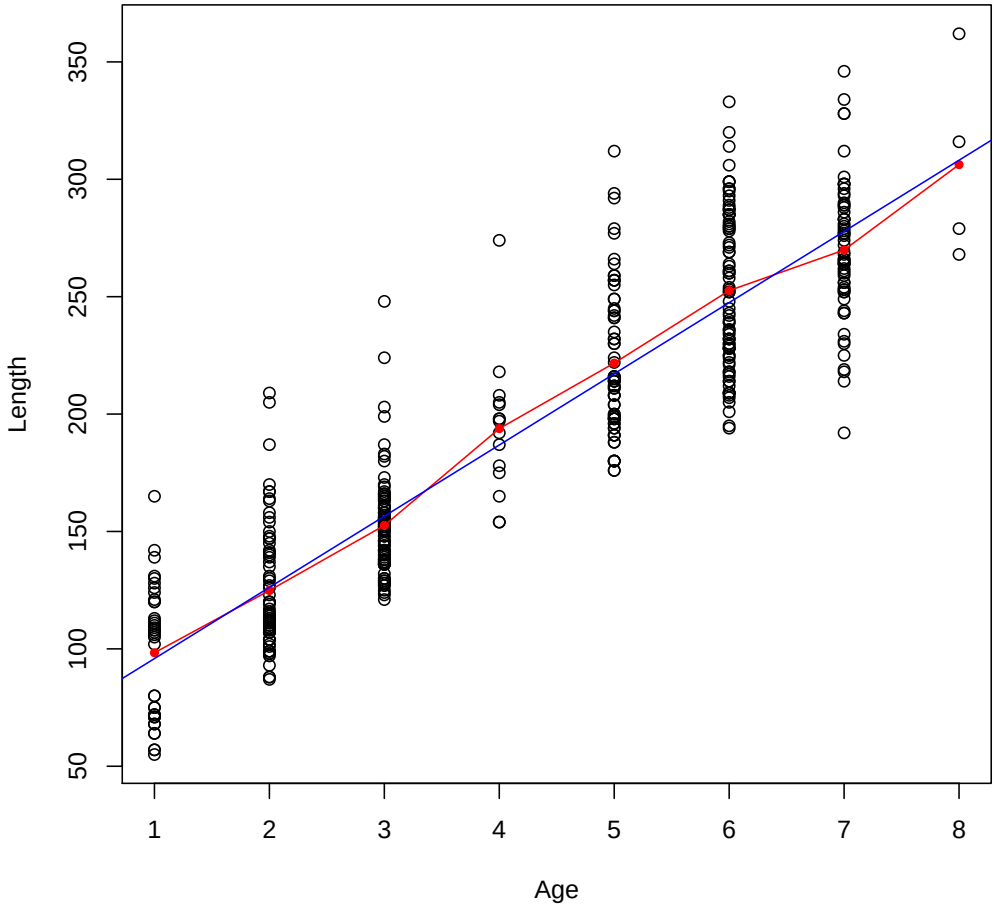
* In this example 439 smallmouth bass, of at least one year old, were caught in West Bearskin Lake in Minnesota.

* For each fish, their age was measured using annular rings on their scales. Their length was also measured.

* We wish to fit the model

$$\text{Length} = \beta_0 + \beta_1 \text{Age} + \varepsilon$$

Example: Fish age and Length



Example: Fish age and Length

* The red dots are the mean lengths examining only fish of a given age.

* The blue line is the fitted regression line

$$\text{Length} = 65.53 + 30.32 \times \text{Age}$$

* The blue line seems to follow the mean ages very well.

* Individual fish, however, vary about this line quite a bit.

* The fitted line tells us that, on average, fish grow 30.32mm per year after 1 year of age.

Example: Mother and Daughter Heights

* How does a mother's height affect her daughter's height?

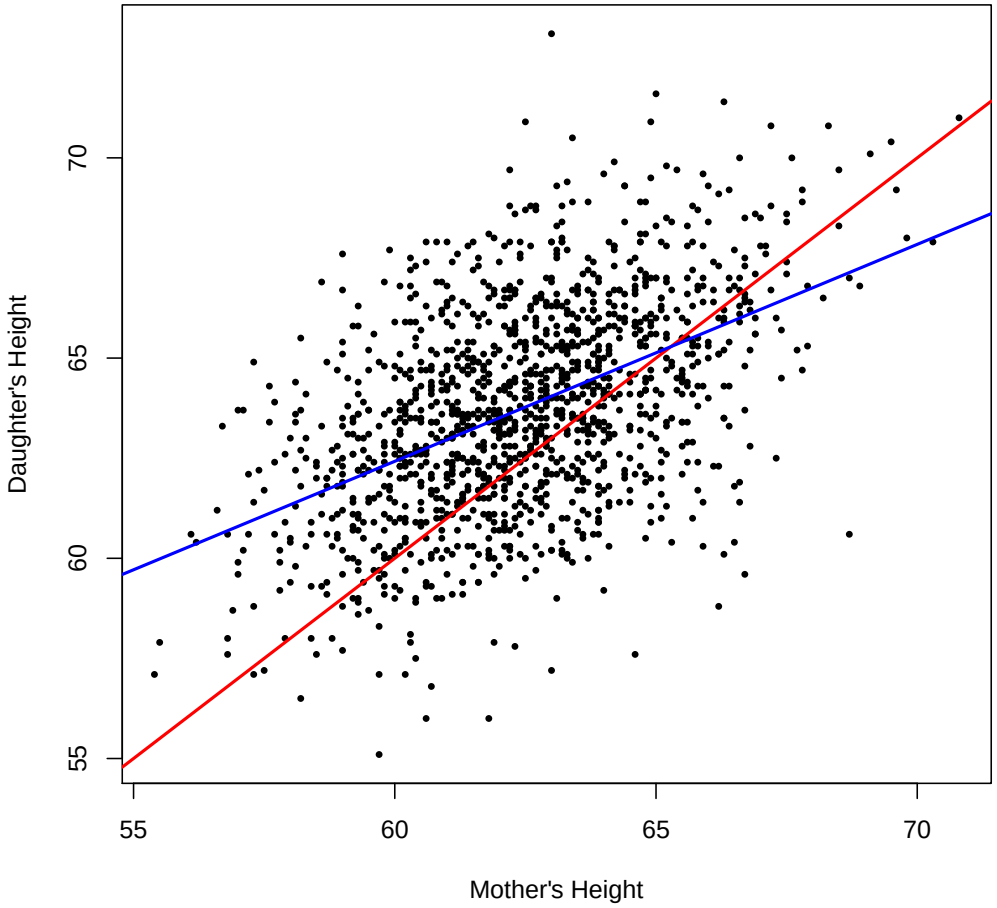
* We would expect tall mothers to have tall daughters and short mothers to have short daughters.

* Do daughters tend to be the same height as their mother, on average?

* To answer this question, the famous statistician Karl Pearson recorded the heights of 1375 British mothers and their adult daughters in the period 1893–1898.

* The data were published in 1903 in one of the first publications to look at heredity of physical traits using real data.

Example: Mother and Daughter Heights



Example: Mother and Daughter Heights

* The red line is the line

$$\text{Daughter Height} = \text{Mother Height}$$

* The blue line is the fitted line

$$\text{Daughter Height} = 29.92 + 0.54 \times \text{Mother Height}$$

* The fitted line shows that short mothers have shorter than average daughters but they tend to be taller than their mothers.

* Conversely, tall mothers have taller than average daughters but they are shorter than their mothers.

* This phenomenon is known as Regression to the Mean.

The Regression Process

1. The researcher must clearly define the question(s) of interest in the study.

2. The response variable Y must be decided on, based on the question of interest.

3. A set of potentially relevant covariates, which can be measured, needs to be defined.

4. Data is collected.

Data Collection

* In some cases, called designed experiments the data can be collected in a controlled setting which will hold constant variables which are not of interest.

* Controlled experiments also facilitate setting certain values of the covariates which are of interest to us.

* These types of experiments are common in areas such as industrial process control, animal models for medical research etc.

* In many studies, however, the data are collected by choosing a random sample of n individuals and observing Y and $X_1, \ldots, X_p$ for those individuals.

Data Collection

* We generally assume that each individual selected is independent of all others.

* In that case we have a random sample of data which can be organized as

Subject	Y	X_1	X_2	$\cdots$	X_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

The Regression Process (continued)

5. Model Specification.

   - What is the form of the model?
   - What assumptions will we make?

6. Decide on a method for fitting the specified model.

7. Fit the model - typically using software such as SAS.

8. Examine the fitted model for violations of assumptions.

9. Conduct hypothesis testing for questions of interest.

10. Report the results from statistical inference.