

**Summer School in Statistics for Astronomers V**  
**June 1 - June 6, 2009**

**Regression**

Mosuk Chow  
Statistics Department  
Penn State University.

Adapted from notes prepared by RL Karandikar

# Mean and variance

Recall we had defined the **Expectation** or the **mean** of a random variable  $X$  as

$$\mu = E(X) = \begin{cases} \sum_i x_i P(X = x_i) & \text{for discrete } X \\ \int_{-\infty}^{\infty} xf(x)dx & \text{for continuous } X \text{ with density } f(x) \end{cases}$$

and the **variance** of a random variable  $X$  as

$$\sigma^2 = \text{Var}(X) = E(X^2) - \mu^2 = \begin{cases} \sum_i x_i^2 P(X = x_i) - \mu^2 \\ \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2 \end{cases}$$

If  $X$  and  $Y$  are random variables with means  $\mu_X$  and  $\mu_Y$ , then the **covariance** of  $X$  and  $Y$  is defined by

$$\text{Cov}(X, Y) = E(X - \mu_X)(Y - \mu_Y) = E(XY) - \mu_X\mu_Y$$

The **correlation coefficient**  $\rho(X, Y)$  of  $X$  and  $Y$  is defined by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}$$

## Properties of mean and variance

$$E(aX + b) = aEX + b,$$

$$\text{Var}(aX + b) = a^2\text{Var}(X)$$

$$E(aX + bY + c) = aEX + bEY + c$$

$$\text{Var}(aX + bY + c) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y)$$

It may be shown that for any  $n \times n$  matrix  $\mathbf{A}$  and  $n \times 1$  vector  $\mathbf{b}$

$$E(\mathbf{AY} + \mathbf{b}) = \mathbf{AEY} + \mathbf{b},$$

$$\text{Cov}(\mathbf{AY} + \mathbf{b}) = \mathbf{ACov}(\mathbf{Y})\mathbf{A}^T.$$

which is the basic result used in regression.

## Hubble's data (1929)

In 1929 Edwin Hubble investigated the relationship between distance and radial velocity of extra-galactic nebulae (celestial objects). It was hoped that some knowledge of this relationship might give clues as to the way the universe was formed and what may happen later. His findings revolutionized astronomy and are the source of much research today. Given here is the data which Hubble used for 24 nebulae.

$X$  = Distance (in Megaparsecs) from earth

$Y$  = The recession velocity (in km/sec)

$X$	$Y$	$X$	$Y$	$X$	$Y$	$X$	$Y$
.032	170	.034	290	.214	-130	.263	-70
.275	-185	.275	-220	.45	200	.5	290
.5	270	.63	200	.8	300	.9	-30
.9	650	.9	150	.9	500	1.0	920
1.1	450	1.1	500	1.4	500	1.7	960
2.0	500	2.0	850	2.0	800	2.0	1090

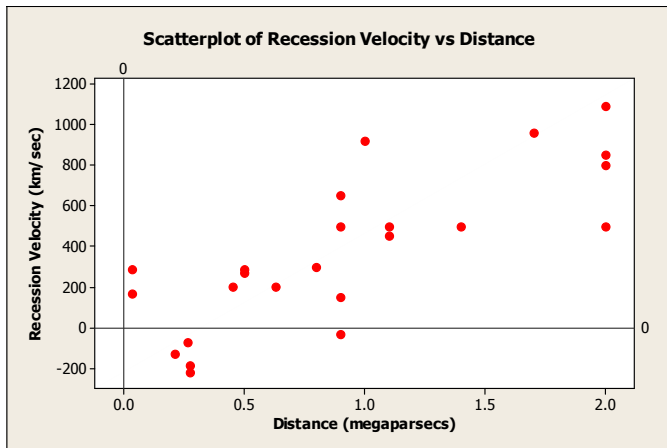
[lib.stat.cmu.edu/DASL/Datafiles/Hubble.html](http://lib.stat.cmu.edu/DASL/Datafiles/Hubble.html)

From this data-set Hubble obtained the relation

$$\text{Recession Velocity} = H_0 \times \text{Distance}$$

where  $H_0$  is Hubble's constant thought to be about 75 km/sec/Mpc.

# Back to Hubble's data



# The ML Method for Linear Regression Analysis

Scatterplot data:  $(x_1, y_1), \dots, (x_n, y_n)$

Basic assumption: The  $x_i$ 's are non-random measurements; the  $y_i$  are observations on  $Y$ , a random variable

Statistical model:

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \dots, n$$

Errors  $\epsilon_1, \dots, \epsilon_n$ : a random sample from  $N(0, \sigma^2)$

Parameters:  $\alpha, \beta, \sigma^2$

$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$ : The  $Y_i$ 's are independent

The  $Y_i$  are not identically distributed, because they have differing means

The likelihood function is the joint density function of the observed data,  $Y_1, \dots, Y_n$

$$\begin{aligned} L(\alpha, \beta, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(Y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right] \\ &= (2\pi\sigma^2)^{-n/2} \exp \left[ -\frac{\sum_{i=1}^n (Y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right] \end{aligned}$$

Use partial derivatives to maximize  $L$  over all  $\alpha, \beta$  and  $\sigma^2 > 0$  (Wise advice: Maximize  $\ln L$ )

The ML estimators are:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

and

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

Using this on Hubble's data we get

$$\hat{\beta} = 454.16, \quad \hat{\alpha} = -40.78$$

where the intercept term is not significant. If we use regression through origin, the result is

$$\hat{\beta} = 423.94$$

The result from the historical data set obtained by Hubble is very different from the current estimate of the Hubble's constant due to various issues of measurement errors and censoring, etc. Those issues will be discussed in a subsequent lecture.

## How do we obtain the least squares estimates?

Let  $Y_i$  be the response for the  $i^{\text{th}}$  data point and let  $\mathbf{x}_i$  be the  $p$ -dimensional (row vector) of the predictors for the  $i$ th data point,  $i = 1, \dots, n$ . There are  $p-1$  predictors.

We assume that

$$Y_i = \mathbf{x}_i\beta + e_i,$$

where  $\beta$ , an unknown parameter, is a  $p \times 1$  column vector, and

$$e_i \sim N_1(0, \sigma^2), \text{ and the } e_i \text{ are independent.}$$

Note that  $\sigma^2$  is another parameter for this model.

We further assume that the predictors are linearly independent. Thus we could have the second predictor be the square of the first predictor, the third one the cube of the first one, etc, so this model includes polynomial regression.

We often write this model in matrices. Let

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}, \mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

so that  $\mathbf{Y}$  and  $\mathbf{e}$  are  $n \times 1$  and  $\mathbf{X}$  is  $n \times p$ . The assumed linear independence of the predictors implies that the columns of  $\mathbf{X}$  are linearly independent and hence  $\text{rank}(\mathbf{X}) = p$ .

$$\mathbf{x}_i = (1, x_{i,1}, \dots, x_{i,p-1}).$$

The normal model can be stated more compactly as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}, \quad \mathbf{e} \sim N_n(0, \sigma^2 \mathbf{I})$$

Note that:

$$\beta = (\beta_0, \beta_1, \dots, \beta_{p-1})^T.$$

The input matrix  $\mathbf{X}$  is of dimension  $n \times (p)$ :

$$\begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p-1} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p-1} \end{pmatrix}$$

Therefore, using the formula for the multivariate normal density function, we see that the joint density of  $\mathbf{Y}$  is

$$\begin{aligned} f_{\beta, \sigma^2}(\mathbf{y}) &= (2\pi)^{-n/2} |\sigma^2 \mathbf{I}|^{-1/2} \exp\left\{-\frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)^T (\sigma^2 \mathbf{I})^{-1} (\mathbf{y} - \mathbf{X}\beta)\right\} \\ &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\beta\|^2\right\} \end{aligned}$$

Therefore the likelihood for this model is

$$L_{\mathbf{Y}}(\beta, \sigma^2) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{X}\beta\|^2\right\}$$

## Estimation of $\beta$

First, we note that the assumption on the  $\mathbf{X}$  matrix implies that  $\mathbf{X}^T \mathbf{X}$  is invertible.

The ordinary least square (OLS) estimator of  $\beta$  is found by minimizing

$$q(\beta) = \sum (Y_i - \mathbf{x}_i \beta)^2 = \|\mathbf{Y} - \mathbf{X}\beta\|^2$$

The formula for the OLS estimator of  $\beta$  is

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Note that

$$\begin{aligned} E\hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E\mathbf{Y} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta = \beta \\ \text{Cov}(\hat{\beta}) &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2 \mathbf{I} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} = \sigma^2 \mathbf{M} \end{aligned}$$

Therefore

$$\hat{\beta} \sim N_p(\beta, \sigma^2 \mathbf{M})$$

## Properties of the OLS estimator

1. (Gauss-Markov) For the non-normal model the OLS estimator is the best linear unbiased estimator (BLUE), i.e., it has smaller variance than any other linear unbiased estimator.
2. For the normal model, the OLS is the best unbiased estimator i.e., has smaller variance than any other unbiased estimator
3. Typically, the OLS estimator is consistent, i.e.  $\hat{\beta} \rightarrow \beta$

## The unbiased estimator of $\sigma^2$

In regression we typically estimate  $\sigma^2$  by

$$\hat{\sigma}^2 = \left\| \mathbf{Y} - \mathbf{X}\hat{\beta} \right\|^2 / (n - p)$$

which is called the unbiased estimator of  $\sigma^2$ . we first state the distribution of  $\hat{\sigma}^2$ .

$$\frac{(n - p) \hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p}^2 \text{ independently of } \hat{\beta}$$

## Properties of $\hat{\sigma}^2$

1. For the general model  $\hat{\sigma}^2$  is unbiased.
2. For the normal model  $\hat{\sigma}^2$  is the best unbiased estimator.
3.  $\hat{\sigma}^2$  is consistent.

## Interval estimators and tests

We first discuss inference about  $\beta_i$  the  $i$ th component of  $\beta$ . Note that  $\hat{\beta}_i$  the  $i$ th component of the OLS estimator is the estimator of  $\beta_i$ .

Further

$$\text{Var}(\hat{\beta}_i) = \sigma^2 M_{ii}$$

which implies that the standard error of  $\hat{\beta}_i$  is

$$\hat{\sigma}_{\hat{\beta}_i} = \hat{\sigma} \sqrt{M_{ii}}$$

Therefore we see that a  $1 - \alpha$  confidence interval for  $\beta_i$  is

$$\beta_i \in (\hat{\beta}_i - t_{n-p}^{\alpha/2} \hat{\sigma}_{\hat{\beta}_i}, \hat{\beta}_i + t_{n-p}^{\alpha/2} \hat{\sigma}_{\hat{\beta}_i}).$$

To test the null hypothesis  $\beta_i = c$  against one and two-sided alternatives we use the t-statistic

$$t = \frac{\hat{\beta}_i - c}{\hat{\sigma}_{\hat{\beta}_i}} \sim t_{n-p}.$$

Now consider inference for  $\delta = \mathbf{a}^T \beta$ , let

$$\widehat{\delta} = \mathbf{a}^T \widehat{\beta} \sim N_1 (\delta, \sigma^2 \mathbf{a}' \mathbf{M} \mathbf{a})$$

therefore we see that  $\widehat{\delta}$  is the estimator of  $\delta$ , and

$$\text{Var} (\widehat{\delta}) = \sigma^2 \mathbf{a}' \mathbf{M} \mathbf{a}$$

so that the standard error of  $\widehat{\delta}$  is

$$\widehat{\sigma}_{\widehat{\delta}} = \widehat{\sigma} \sqrt{\mathbf{a}' \mathbf{M} \mathbf{a}}$$

Therefore the confidence interval for  $\delta$  is

$$\delta \in (\hat{\delta} - t_{n-p}^{\alpha/2} \hat{\sigma}_{\hat{\delta}}, \hat{\delta} + t_{n-p}^{\alpha/2} \hat{\sigma}_{\hat{\delta}})$$

and the test statistic for testing  $\delta = c$  is given by

$$\frac{\hat{\delta} - c}{\hat{\sigma}_{\hat{\delta}}} \sim t_{n-p} \text{ under the null hypothesis}$$

There are tests and confidence regions for vector generalizations of these procedures.

Let  $\mathbf{x}_0$  be a row vector of predictors for a new response  $Y_0$ . Let

$$\mu_0 = \mathbf{x}_0\beta = EY_0.$$

$\hat{\mu}_0 = \mathbf{x}_0\hat{\beta}$  is the obvious estimator of  $\mu_0$  and

$$\text{Var}(\hat{\mu}_0) = \sigma^2 \mathbf{x}_0 \mathbf{M} \mathbf{x}'_0 \Rightarrow \hat{\sigma}_{\hat{\mu}_0} = \hat{\sigma} \sqrt{\mathbf{x}_0 \mathbf{M} \mathbf{x}'_0}$$

and therefore a confidence interval for  $\mu_0$  is

$$\mu_0 \in (\hat{\mu}_0 - t_{n-p}^{\alpha/2} \hat{\sigma}_{\hat{\mu}_0}, \hat{\mu}_0 + t_{n-p}^{\alpha/2} \hat{\sigma}_{\hat{\mu}_0})$$

A  $1 - \alpha$  prediction interval for  $Y_0$  is an interval such that

$$P(a(\mathbf{Y}) \leq Y_0 \leq b(\mathbf{Y})) = 1 - \alpha$$

A  $1 - \alpha$  prediction interval for  $Y_0$  is

$$Y_0 \in (\hat{\mu}_0 - t_{n-p}^{\alpha/2} \sqrt{\hat{\sigma}^2 + \hat{\sigma}_{\hat{\mu}_0}^2}, \hat{\mu}_0 + t_{n-p}^{\alpha/2} \sqrt{\hat{\sigma}^2 + \hat{\sigma}_{\hat{\mu}_0}^2})$$

The derivation of this interval is based on the fact that

$$\text{Var}(Y_0 - \hat{\mu}_0) = \sigma^2 + \sigma_{\hat{\mu}_0}^2$$

Let

$$T^2 = \sum (Y_i - \bar{Y})^2, \quad S^2 = \left\| \mathbf{Y} - \mathbf{X}\hat{\beta} \right\|^2$$

be the numerators of the variance estimators for the regression model and the intercept only model. We think of these as measuring the "variation" under these two models. Then the coefficient of determination  $R^2$  is defined by

$$R^2 = \frac{T^2 - S^2}{T^2}$$

Note that

$$0 \leq R^2 \leq 1$$

Note that  $T^2 - S^2$  is the amount of variation in the intercept only model which has been explained by including the extra predictors of the regression model and  $R^2$  is the proportion of the variation left in the intercept only model which has been explained by including the additional predictors.