3rd IIA-Penn State Astrostatistics School

19–27 July, 2008

Vainu Bappu Observatory, Kavalur

Correlation and Regression

Rajeeva Karandikar

Chennai Mathematical Institute

Adapted from notes prepared by Rajeeva Karandikar/Rahul Roy

Mean and variance

Recall the Expectation or the mean of a random variable X is defined as

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

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 μ_X and μ_Y, then the covariance of X and Y is defined by

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

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

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

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

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

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

Random vectors, mean vectors and covariance matrix

Let $Y_1, ..., Y_n$ be random variables. Then

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

is a p-dimesional random vector .

The mean vector $\mu = E\mathbf{Y}$ and covariance matrix $\Sigma = \text{Cov}(\mathbf{Y})$ are defined by

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \cdots & \Sigma_{1n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n1} & \cdots & \Sigma_{nn} \end{pmatrix}$$

where

$$\mu_i = EY_i,$$

$$\Sigma_{ij} = \begin{cases} \operatorname{Var}(Y_i) & \text{for } i = j \\ \operatorname{Cov}(Y_i, Y_j) & \text{for } i \neq j \end{cases}$$

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

$$E\left(\mathbf{AY}+\mathbf{b}\right)=\mathbf{A}E\mathbf{Y}+\mathbf{b},$$

$$\mathrm{Cov}\left(\mathbf{AY}+\mathbf{b}\right)=\mathbf{A}\mathrm{Cov}\left(\mathbf{Y}\right)\mathbf{A}^{\mathrm{T}}.$$

which is the basic result used in regression.

Note that the covariance matrix is a symmetric matrix. Further,

$$0 \leq \operatorname{Var}\left(\mathbf{a}^{\mathrm{T}}\mathbf{Y}\right) = \mathbf{a}^{\mathrm{T}}\operatorname{Cov}\left(\mathbf{Y}\right)\mathbf{a}$$

which implies that the covariance matrix is non-negative definite, which means there is a matrix $\mathbf{B}$ such that

$$\mathbf{B}\mathbf{B}^{\mathrm{T}} = \operatorname{Cov}\left(\mathbf{Y}\right)$$

Such a matrix $\mathbf{B}$ is called a square root of the covariance matrix. Actually there are several such matrices. One of the most useful and easy to find in computer software is the Cholesky square root which is a triangular matrix.

Multivariate normal distribution

We say that an n-dimensional random vector $\mathbf{Y}$ has a multivariate normal distribution with mean vector μ and covariance matrix Σ and write

$$\mathbf{Y} \sim N_n(\mu, \Sigma)$$

if $\mathbf{Y}$ has joint probability density function (pdf)

$$f(\mathbf{y}) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{y}-\mu)' \Sigma^{-1} (\mathbf{y}-\mu)\}, \quad \forall \mathbf{y}$$

Note that this function has the two worst things in matrices, the determinant and the inverse of a matrix.

Properties

Let $\mathbf{Y} \sim N_n(\mu, \Sigma)$ and $\mathbf{A}$ a $m \times n$ matrix, $\mathbf{b}$ a $m \times 1$ vector, then

$$\mathbf{AY} + \mathbf{b} \sim N_m\left(\mathbf{A}\mu + \mathbf{b}, \mathbf{A}\Sigma\mathbf{A}'\right)$$

$$Y_i \sim N_1\left(\mu_i, \Sigma_{ii}\right).$$

If U and V are random variables with joint normal distribution, then $\mathrm{Cov}(U, V) = 0$ implies U and V are independent.

Linear regression

Let $f : \mathbb{R}^n \to \mathbb{R}$ and $y = f(\mathbf{x})$, for $\mathbf{x} \in \mathbb{R}^n$.

The vector $\mathbf{x}$ is the vector of predictors and the scalar y is the response.

In simple linear regression we have one predictor and one response; in multiple linear regression we have several predictors and one response.

In this tutorial we will look at multiple linear regression with simple linear regression as a special case.

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.

$Y =$ Distance (in Megaparsecs) from earth

$X =$ 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`

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.

How do we do this?

Let Y_i be the response for the i^{th} data point and let $\mathbf{x}_i$ be the p-dimensional (row vector) of the predictors for the ith data point, $i = 1, \cdots, n$. (In the Hubble data set, x_i is scalar.)

We assume that

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

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

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

Note that σ^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 $\text{rank}(\mathbf{X}) = p$.

The normal model can be stated more compactly as

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

or as

$$\mathbf{Y} \sim N_n\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right)$$

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

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

Therefore the likelihood for this model is

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

Estimation of β

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

The ordinary least square (OLS) estimator of β 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 β is

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

To see this note that

$$\nabla q(\beta) = 2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta) = 2\left(\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\beta\right)$$

setting this equal to 0 we get the above formula for $\widehat{\beta}$.

For an algebraic derivation note that

$$\begin{aligned} q(\beta) &= \left\|\mathbf{Y} - \mathbf{X}\widehat{\beta}\right\|^2 + \left\|\mathbf{X}\widehat{\beta} - \mathbf{X}\beta\right\|^2 \\ &\geq \left\|\mathbf{Y} - \mathbf{X}\widehat{\beta}\right\|^2 = q\left(\widehat{\beta}\right) \end{aligned}$$

Note that

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

Therefore

$$\widehat{\beta} \sim N_p\left(\beta, \sigma^2\mathbf{M}\right)$$

Properties of the OLS estimator

1. (Gauss-Markov) For the non-normal model the OLS estimator is the best linear unbised estimator (BLUE), i.e., it has smaller variance than any othe 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. $\widehat{\beta} \to \beta$

The unbiased estimator of σ^2

In regression we typically estimate σ^2 by

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

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

$$\frac{(n-p)\,\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p} \text{ independently of } \widehat{\beta}$$

Properties of $\widehat{\sigma}^2$

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

The maximum likelihood estimator (MLE)

Looking at the likelihood above, we see that the OLS estimator maximizes the exponent so that $\widehat{\beta}$ is the MLE of β. To find the MLE of σ^2 differentiate $\log\left(L_Y\left(\widehat{\beta},\sigma^2\right)\right)$ with respect to σ, getting

$$\widehat{\sigma}^2_{MLE} = \frac{n-p}{n}\widehat{\sigma}^2$$

Note that if

$$p/n = q$$

then

$$E\widehat{\sigma}^2_{MLE} = (1-q)\,\sigma^2,\ \widehat{\sigma}^2_{MLE} \to (1-q)\,\sigma^2$$

so the MLE is not unbiased and is not consistent unless $p/n \to 0$.

Interval estimators and tests

We first discuss infererence about β_i the ith component of β. Note that $\widehat{\beta}_i$ the ith component of the OLS estimator is the estimator of β_i. Further

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

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

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

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

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

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

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

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

$$\widehat{\delta} = \mathbf{a}'\widehat{\beta} \sim N_1\left(\delta, \sigma^2\mathbf{a}'\mathbf{M}\mathbf{a}\right)$$

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

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

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 δ is

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

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

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

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
$\mu_0 = \mathbf{x}_0\beta = EY_0$.
$\widehat{\mu}_0 = \mathbf{x}_0\widehat{\beta}$ is the obvious estimator of μ_0 and

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

and therefore a confidence interval for μ_0 is

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

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

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

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

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

The derivation of this interval is based on the fact that

$$\text{Var}\left(Y_0 - \widehat{\mu}_0\right) = \sigma^2 + \sigma^2_{\widehat{\mu}_0}$$

The hat matrix

The hat matrix $\mathbf{H}$ is defined as

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}$$

$\mathbf{H}$ is a symmetric idempotent matrix, i.e

$$\mathbf{H}' = \mathbf{H}, \qquad \mathbf{H}^2 = \mathbf{H}$$

Let $\mu = \mathbf{X}\beta$, $\widehat{\mu} = \mathbf{X}\widehat{\beta}$. Then

$$\widehat{\mu} = \mathbf{HY}$$

which is why $\mathbf{H}$ is called the hat matrix. Now let

$$\mathbf{H}^{\perp} = \mathbf{I} - \mathbf{H}$$

then $\mathbf{H}^{\perp}$ is also a symmetric idempotent matrix which is orthogonal to $\mathbf{H}$, i.e

$$\mathbf{H}'\mathbf{H}^{\perp} = \mathbf{0}$$

Then

$$(n-p)\,\widehat{\sigma}^2 = \left\|\mathbf{H}^{\perp}\mathbf{Y}\right\|$$

Note that

$$\mathbf{Y} = \mathbf{H}\mathbf{Y} + \mathbf{H}^{\perp}\mathbf{Y}$$

We think of think of $\mathbf{HY}$ as having information about the signal μ and $\mathbf{H}^{\perp}\mathbf{Y}$ as having information about the noise $Y - \mu$. For the rest of this talk, we shall use $\mathbf{H}$ for this matrices

R^2, adjusted R^2 and predictive R^2

Let

$$T^2 = \sum \left(Y_i - \overline{Y}\right)^2, \; S^2 = \left\| \mathbf{Y} - \mathbf{X}\widehat{\beta} \right\|^2$$

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.

Note that

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

which suggests that this might be improved by substituting unbiased estimator for the MLE's getting adjusted R^2

$$R_a^2 = \frac{\frac{T^2}{n-1} - \frac{S^2}{n-p}}{\frac{T^2}{n-1}} = 1 - \frac{n-1}{n-p}\left(I - R^2\right)$$

Both R^2 and adjusted R^2 suffer from the fact that the fit is being evaluated with the same data used to compute it and the therefore the fit looks better than it is. A better procedure is based on cross-validation.

Suppose we delete the ith observation and compute $\widehat{\beta}_{-i}$ the OLS estimator of β without the ith observation.

We do this for all i. We also compute $\overline{Y}_{-i}$

$$\overline{Y}_{-i} = \sum_{j \neq i} Y_j / (n-1)$$

the sample mean of theY_i without the ith one.

Then let

$$T_p^2 = \sum \left(Y_i - \overline{Y}_{-i}\right)^2 = \frac{nT^2}{n-1}$$

$$S_p^2 = \sum \left(Y_i - \mathbf{x}_i\widehat{\beta}_{-i}\right)^2 = \sum \left(\frac{Y_i - \mathbf{x}_i\widehat{\beta}}{1 - H_{ii}}\right)^2$$

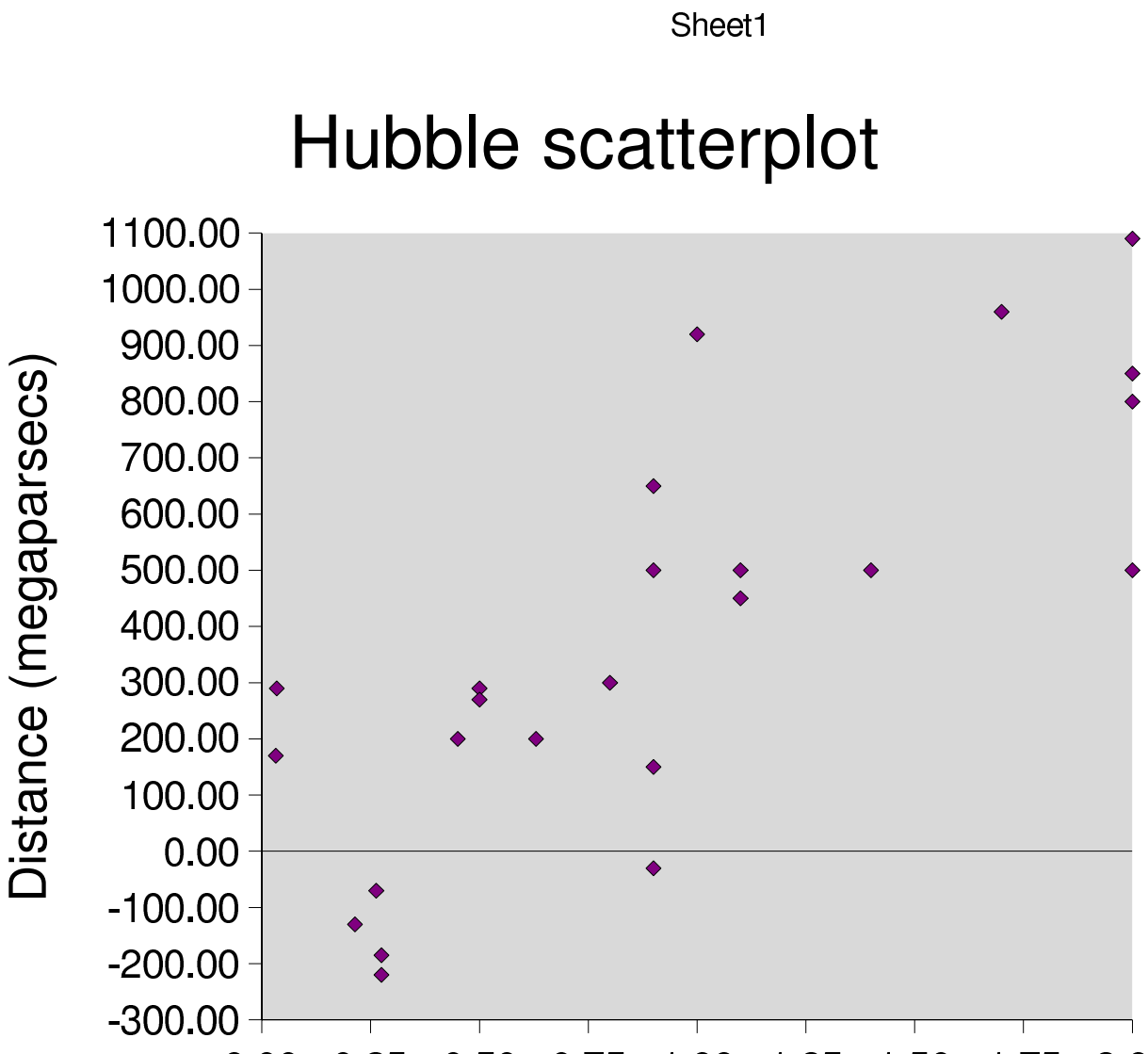
(where H_{ii} is the ith diagonal of the hat matrix).

Then predictive R^2 is defined as

$$R_p^2 = \frac{T_p^2 - S_p^2}{T_p^2}$$

Predictive R^2 computes the fit to the ith observation without using that observation and is therefore a better measure of the fit of the model that R^2 or adjusted R^2.

Back to Hubble's data



The ML Method for Linear Regression Analysis

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

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

Statistical model:

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, \ldots, n$$

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

Parameters: α, β, σ^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, $Y_1, \ldots, Y_n$

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

$$= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right]$$

Use partial derivatives to maximize L over all α, β 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

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