**The PDF of R2 for House Prices** [See problem 14.53 on p.475 of the text.]

The following table shows the assessed values and selling prices of 8 houses, constituting a random sample of all the houses sold recently in a metropolitan area.

Table of Assessed Value (x) and sell price (y)

x = [70.3 102 62.5 74.8 57.9 81.6 110.4 88];

y = [114.4 169.3 106.2 125 99.8 132.1 174.2 143.5];

(a) Obtain a linear model to predict house price from assessed value, based on this data

(b) Use simulations to estimate the pdf of R2, assuming that the statistics associated with (a) are the true values of the parameters.

% PROGRAM NAME: houseprices.m

% From problem 14.53 on p.475 of Miller & Miller

% DATA:

x = [70.3 102 62.5 74.8 57.9 81.6 110.4 88];

y = [114.4 169.3 106.2 125 99.8 132.1 174.2 143.5];

XY = [x' y'];

% Yhat = mX + b model parameter estimates

CXY = cov(XY);

m = CXY(1,2)/CXY(1,1);

MXY = mean(XY);

b = MXY(2) - m\*MXY(1);

yhat = m\*x + b; yhat = yhat';

% Correlation coefficient & coefficient of determination

r = m\*(CXY(1,1)/CXY(2,2))^.5;

R2 = r^2;

figure(1)

plot(x,y,'\*',x,yhat,'r')

xlabel('Assessed Value ($1000)')

ylabel('Sell Price ($1000)')

title('Scatter Plot of Assessed vs Sell Prices')

grid

pause

%===============================================

% Use simulations to otain pdf of R2

mx = MXY(1); my = MXY(2);

sigx = CXY(1,1)^.5; sigy = CXY(2,2)^.5; covxy = CXY(1,2);

sigw = (sigy^2 - (covxy/sigx)^2 )^.5;

n = 8; nsim = 5000;

X = mx + sigx\*randn(n,nsim);

W = sigw\*randn(n,nsim);

Y = m\*X + b + W;

sX = std(X); sY = std(Y); mX = mean(X); mY = mean(Y);

cXY = zeros(1,nsim);

for k = 1:nsim

cc = cov(X(:,k),Y(:,k));

cXY(k) = cc(1,2);

end

mvec = cXY./sX.^2;

rvec = mvec.\*(sX./sY);

R2vec = rvec.^2;

% Obtain histogram-based pdf of R2hat

db = 0.005; bvec = .9+db/2 : db : 1-db/2;

h = hist(R2vec,bvec);

fR2 = (nsim\*db)^-1 \* h;

figure(2)

bar(bvec,fR2)

title('Histogram-based pdf Estimate for R2 Based on 5000 Sims')

grid

From: <http://en.wikipedia.org/wiki/Coefficient_of_determination>

SS_\text{tot}=\sum_i (y_i-\bar{y})^2,A data set has values *yi*, each of which has an associated modelled value *fi* (also sometimes referred to as *ŷi*). Here, the values *yi* are called the observed values and the modelled values *fi* are sometimes called the predicted values. The "variability" of the data set is measured through different [sums of squares](http://en.wikipedia.org/wiki/Sum_of_squares):

the [total sum of squares](http://en.wikipedia.org/wiki/Total_sum_of_squares) (proportional to the sample variance);

SS_\text{reg}=\sum_i (f_i -\bar{y})^2,

the regression sum of squares, also called the [explained sum of squares](http://en.wikipedia.org/wiki/Explained_sum_of_squares).

SS_\text{err}=\sum_i (y_i - f_i)^2\, , the sum of squares of residuals, also called the [residual sum of squares](http://en.wikipedia.org/wiki/Residual_sum_of_squares).

R^2 \equiv 1 - {SS_{\rm err}\over SS_{\rm tot}}.\,The notations *SSR* and *SSE* should be avoided, since in some texts their meaning is reversed to **R**esidual sum of squares and **E**xplained sum of squares, respectively. The most general definition of the coefficient of determination is

In the above discussion, the four formulas include only numbers. In this problem we will address the random variables that the numbers result from. We will couch investigation in the framework of the simple *linear regression* problem: For the 2-D random variable , consider the prediction model:

. (1.1)

Denote the prediction error as: .

(5pts) To find the expressions for the model parameters *m* and *b*, we will adhere to the following requirements:

(R1):  (that is: ). In words, the model is required to be *unbiased.*

(R2):  (that is: ). In words, the model error is required to be uncorrelated with *X*

Use the relevant material in the Chapter 5 notes to carry out the steps leading to the expressions:

 and . (1.2)

**(b)(5pts)** Recall that in class we defined the operation  as , where *V* is any random variable and where  are the associated data collection random variables. For each of the four formulas given above, modify them so that they entail random variables. [**Note:** In (a) we used the random variable . And so, a number corresponds to the symbol . We chose the former because it more clearly emphasizes what actually is. ]