LINEAR REGRESSION MODELS W4315

HOMEWORK 1 ANSWERS

January 26, 2010

Instructor: Frank Wood

1. (20 points) Let $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ be a linear regression model with distribution of error terms unspecified (but with mean $E(\epsilon) = 0$ and variance $V(\epsilon_i) = \sigma^2$ (σ^2 finite) known). Show that $s^2 = MSE = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}$ is an unbiased estimator for σ^2. $\hat{Y}_i = b_0 + b_1 X_i$ where $b_0 = \bar{Y} - b_1 \bar{X}$ and $b_1 = \frac{\sum_i((X_i - \bar{X})(Y_i - \bar{Y}))}{\sum_i (X_i - \bar{X})^2}$

Answer:

First, let's denote the followings:

$$\hat{e}_i = y_i - \hat{y}_i$$
$$SXX = \sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$SYY = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
$$SXY = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Now we set out to prove the following equation which accomplishes essentially the final result:

$$Var\hat{e}_i = E\hat{e}_i^{\,2} = (\tfrac{n-2}{n} + \tfrac{1}{SXX}(\tfrac{1}{n}\sum_{j=1}^{n} x_j^2 + x_i^2 - 2(x_i - \bar{x})^2 - 2x_i\bar{x}))\sigma^2$$

To prove the above, realize that:

$$\begin{aligned}
Var(\hat{e}_i) &= Var(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
&= Var((y_i - \beta_0 - \beta_1 x_i) - (\hat{\beta}_0 - \beta_0) - x_i(\hat{\beta}_1 - \beta_1)) \\
&= Var(y_i) + Var(\hat{\beta}_0) + x_i^2 Var(\hat{\beta}_1) - 2Cov(y_i, \hat{\beta}_0) - 2x_i Cov(y_i, \hat{\beta}_1) + 2x_i Cov(\hat{\beta}_0, \hat{\beta}_1)
\end{aligned}$$

The last equation holds because the covariance between any random variable and a constant is zero, and all the y_i's are independent entailing that the $Cov(y_i, y_j) = 0, i \neq j$

$Var(y_i) = \sigma^2$

Notice that(some algebras needed here, and the following tricks are crucial in reducing the amount of calculation):
$\sum (x_i - \bar{x}) = 0$
$\beta_1 = \frac{\sum x_i - \bar{x} y_i}{SXX}$
So now we have:

$$\begin{aligned}
Var(\beta_1) &= Var(\frac{SXY}{SXX}) \\
&= Var(\frac{\sum (x_i - \bar{x})y_i}{SXX}) \\
&= \frac{1}{SXX^2} \sum x_i - \bar{x}^2 Var(y_i) \\
&= \frac{\sigma^2}{SXX}
\end{aligned}$$

And:

$$\begin{aligned}
Var(\beta_0) &= Var(\bar{y} - \hat{\beta}_1 \bar{x}) \\
&= Var(\sum (\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{SXX})y_i) \\
&= \sum (\frac{1}{n} - \frac{x_i - \bar{x}}{SXX}\bar{x})^2 \sigma^2 \\
&= \sum [\frac{1}{n^2} + \frac{SXX * \bar{x}^2}{SXX^2} - \frac{2}{n} \frac{\bar{x}(x_i - \bar{x})}{XSS}] \sigma^2 \\
&= [\frac{1}{n} + \frac{n\bar{x}^2}{SXX}] \sigma^2 \\
&= \frac{\sum {x_i}^2}{n * SXX} \sigma^2
\end{aligned}$$

For the other terms in the decomposition of $Var(\hat{e_i})$, we have:

$$\begin{aligned}
Cov(y_i, \hat{\beta}_1) &= Cov(y_i, \frac{sum x_i - \bar{x} y_i}{SXX}) \\
&= \frac{x_i - \bar{x}}{SXX} Var(y_i) \\
&= \frac{x_i - \bar{x}}{SXX} \sigma^2
\end{aligned}$$

and:

$$
\begin{aligned}
Cov(y_i, \hat{\beta}_0) &= Cov(y_i, \bar{y} - \hat{\beta}_1 \bar{x}) \\
&= Cov(y_i, \frac{\sum y_i}{n} - \sum (x_i - \bar{x}) y_i SXX \bar{x}) \\
&= \frac{\sigma^2}{n} + \bar{x} \frac{x_i - \bar{x}}{SXX} \sigma^2
\end{aligned}
$$

At last, we have:

$$
\begin{aligned}
Cov(\hat{\beta}_0, \hat{\beta}_1) &= Cov(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1) \\
&= Cov(\frac{\sum y_i}{n} - \sum \frac{(x_i - \bar{x})\bar{x}}{SXX} y_i, \sum \frac{(x_i - \bar{x}) y_i}{SXX}) \\
&= \sum_{i=i}^{n} (\frac{1}{n} - \frac{x_i - \bar{x}}{SXX} \bar{x}) \frac{x_i - \bar{x}}{SXX} \sigma^2 \\
&= -\frac{\bar{x}}{SXX} \sigma^2
\end{aligned}
$$

Then plug in all the parts back to the decomposition of $Var(\hat{e_i})$, we have:

$Var(\hat{e_i}) = (\frac{n-1}{n} + \frac{1}{SXX}(\frac{1}{n} \sum_{j=1}^{n} {x_j}^2 + x_i^2 - 2(x_i - \bar{x})^2 - 2x_i \bar{x}))\sigma^2$

Thus,

$$
\begin{aligned}
E\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} E\hat{e_i}^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} [\frac{n-2}{n} + \frac{1}{SXX}(\frac{1}{n} \sum_{j=1}^{n} x_j^2 + x_i^2 - 2(x_i - \bar{x})^2 - 2x_i \bar{x})]\sigma^2 \\
&= [\frac{n-2}{n} + \frac{1}{nS_{xx}} \{\sum_{j=1}^{n} x_j^2 + \sum_{i=1}^{n} x_i^2 - 2SXX - 2\frac{1}{n}(\sum_{i=1}^{n} x_i)^2\}]\sigma^2 \\
&= (\frac{n-2}{n} + 0)\sigma^2 \\
&= \frac{n-2}{n} \sigma^2
\end{aligned}
$$

where the third equation holds because: $\sum x_i \bar{x} = \frac{1}{n}(\sum x_i)^2$

and the second to last equation holds since $\sum x_i^2 - \frac{1}{n}(\sum x_i)^2 = SXX$

From the above equation, the result flows.

2. (20 points) Derive the maximimum likelihood estimators $\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\sigma}^2$ for parameters β_0, β_1, and σ^2 for the normal linear regression model (i.e. $\epsilon_i \sim_{iid} N(0, \sigma^2)$).

Answer:

To figure the MLE of the parameters, we need to first write down the likelihood function of the data, so under normal assumption, we have the log-likelihood function as follows:

$logL(\beta_0, \beta_1, \sigma^2|x, y) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log\sigma^2 - \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$.

For any fixed value of σ^2, $logL$ is maximized as a function of β_0 and β_1, that minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \tag{1}$$

But to minimize this function is just to principle behind LSE, so it's apparent that the MLE of β_0 and β_1 are the same as their LSE's. Now, substituting in the log-likelihood, to find the MLE of σ^2 we need to maximize

$$-\frac{n}{2}log(2\pi) - \frac{n}{2}log\sigma^2 - \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{2\sigma^2}$$

This maximization problem is nothing but MLE of σ^2 in ordinary normal sampling problems, which is easily given as

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

If you are not familiar with the MLE in normal sampling setting, you can take derivative with respect to σ^2 (N.B. not σ), and then set the derivative to be zero. The solution of the equation is just the MLE of σ^2.

3. (10 points) File "unif.txt"(you can find in on professor Wood's website) contains 200 numbers randomly generated from uniform distribution U(2,5). Read these numbers into MATLAB(using command say, "textread" or simply copy paste from the .txt file) and do the followings:

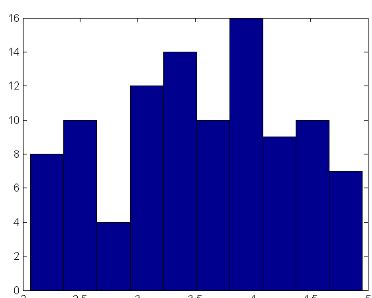
a. Take these 200 number as a population. Use command 'randsample' of MATLAB, draw 100 numbers out of this population randomly with replacement and plot one histogram of these 100 numbers.

b. Use command 'rand' in MATLAB, draw 100 samples directly from U(2,5) and plot another histogram of these numbers.
c. Compare the two histograms, what can you say about the difference between the distribution of samples from the population and from the uniform distribution itself? For the convenience of comparison, you may want to overlay two histograms onto one graph and to see if any apparent difference.

Answer:

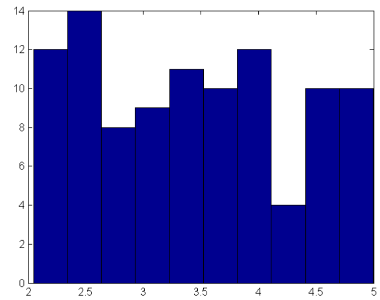
a. The code is as follows:
$data = textread('unif.txt');$
$data = data(1 : end - 1);$
$sample_a = randsample(data, 100, true);$
$hist(sample_a);$
Comments on the code:
1. In MATLAB, add a semicolon at the end of the command whose results you don't want to display in the command window. It will save some time when you run codes which are intense computationally.
2. When using command "textread", keep an eye on the data size read into MATLAB. Usually there will be a zero attached at the end as a symbol to end the file, be sure to eliminate it, otherwise your inference may be influenced.
3. Command "randsample" draws sample w/o replacement from a designated population.
4. Use help system in MATLAB whenever you have doubts about any command. It goes like "help COMMAND" b. The command is

Figure 1:



$sample_b = 2 + 3. * rand(1, 100);$
$hist(sample_b);$
Comments:
"rand(M,N)" gives an M*N matrix whose elements are drawn from U(0,1). To draw from

Figure 2:



U(a,b), simply apply the transformation a+(b-a).*[sequence of U(0,1)];

c. From the two graphs above, we can see that drawing samples from a certain population and drawing samples from a certain distribution which generates that population are de facto the same thing. The histograms look similar. The reason that the histograms are not absolutely "flat" is due to the sample size.

4. (10 points) File "normal.txt"(professor's website) contain 200 randomly generated numbers from Normal distribution N(-1,2^2). Like in problem 3, do the followings:

a. Take these 200 number as a population. Use command 'randsample' of MATLAB, draw 100 numbers out of this population randomly with replacement and plot a histogram of these 100 numbers.

b. Use command 'randn' in MATLAB, draw 100 samples directly from N(-1,2^2) and plot a histogram of these numbers.

c. Compare the two histograms, can you get the similar conclusion as that of problem 3?

Answer:

Combine the 2 histograms together to compare. The codes are as followings: $data = textread('normal.txt');$

$data = data(1 : end - 1)$

$sample_a = randsample(data, 100, true);$

$figure(21);$

$subplot(121);$

$hist(sample_a);\ title('samplesfromthepopulation');$

$sample_b = -1 + 2. * randn(1, 100);$

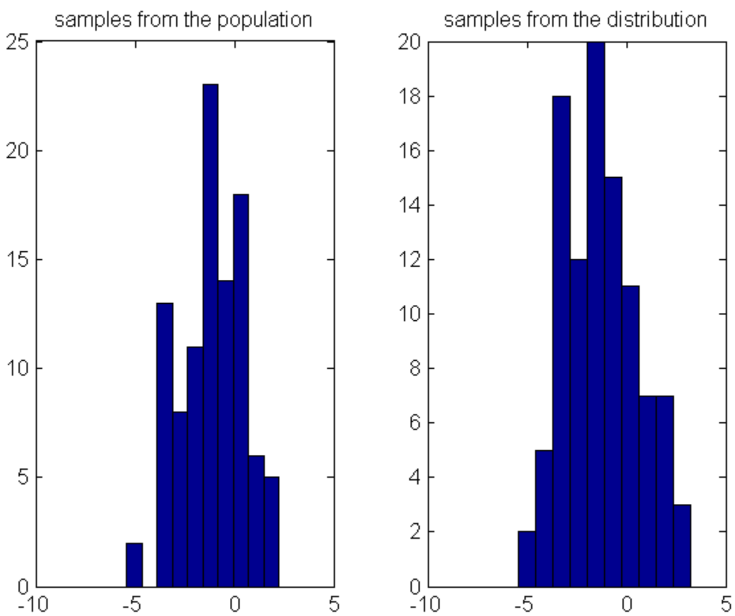
$subplot(122);$

$hist(sample_b);\ title('samplesfromthedistribution');$

And the graph is:

From the graph, we can get similar conclusion as that of problem 3. From these 2 questions,

Figure 3:



we should keep in mind that drawing samples from a distribution(e.g. uniform, normal) is in essence the same thing as drawing samples from a population which is generated by the corresponding distribution.

5. (40 points) Copier maintenance.[1] The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair services on this copier. The data below have been collected from 45 recent calls on users to perform routine preventive number of minutes spent by the service person. Assume that first-order regression model($Y_i = b_0 + b_1 X_i + \epsilon_i$) is appropriate.

i:	**1**	**2**	**3**	...	**43**	**44**	**45**
X_i	2	4	3	...	2	4	5
Y_i	20	60	46	...	27	61	77

a. Obtain estimated regression function.

b. Plot the estimated regression function and the data. How well does the estimated regression function fit the data?

[1] This is problem 1.20 in "Applied Linear Regression Models(4th edition)" by Kutner etc.)

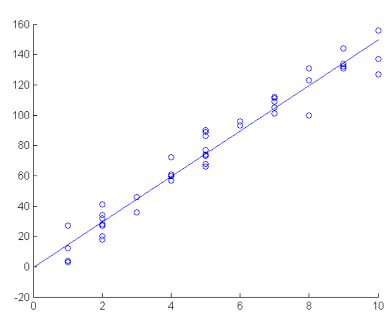
c. interpret b_0 in your estimated regression function. Does b_0 provide any relevant information here? Explain.
d. Obtain a point estimate of the mean service time when $X = 5$ copiers are serviced.

Notice: You can get data of this problem on www.mhhe.com/KutnerALRM4e. Use MATLAB to finish this problem, do not use any other program language. Only basic MATLAB commands are allowed, do not use any built-in command to run regression automatically, e.g. 'regress' in MATLAB is forbidden.

Answer:

a. The code is as follows:
$data = textread('5.txt');$
$X = data(:, 2);$
$Y = data(:, 1);$
$avg\_x = mean(X);$
$avg\_y = mean(Y);$
$SXX = sum((X - avg\_x).^2);$
$SXY = sum((X - avg\_x). * (Y - avg\_y));$
$b\_1 = SXY/SXX;$
$b\_0 = avg\_y - b\_1 * avg\_x;$
So the regression line is: $\quad Y_i = -0.5802 + 15.0352X_i + \epsilon_i$
b. The code and the graph are as followings: $clf;$
$figure(2);$
$scatter(X, Y);$
$holdon;$
$plot([010], [b\_0, b\_0 + b\_1 * 10]);$
From the graph, we can see the fit pretty well. There are too many points drifting far away

Figure 4:



from the fitted line and the overall trend is reflected by the line.
c. b_0 means the expected total number of minutes spent by the service person if the number of copiers serviced $X_i = 0$. It doesn't provide much relevant information, since this case doesn't make much sense. $X = 0$ means the person calling doesn't have any copier then there is no point he's making the call.
d. Plug $X = 5$ into the regression equation and after calculation we have $Y = 74.5961$.