LINEAR REGRESSION MODELS W4315

HOMEWORK 1 QUESTIONS

October 26, 2011

Professor: 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 $b_1 = \frac{\sum_i((X_i - \bar{X})(Y_i - \bar{Y}))}{\sum_i (X_i - \bar{X})^2}$.

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. $\epsilon_i \sim_{iid} N(0, \sigma^2)$).

3. (20 points) Let $X_1, X_2, X_3, \ldots, X_{100}$ be iid normal random variables with mean μ and variance σ^2. Assume that we want to estimate the mean μ from observations. Consider two estimators, X_1 and $\bar{X} = \frac{1}{100}\sum_{i=1}^{100} X_i$:
a. Show these two estimators are both unbiased. Also derive the distribution of each estimator. Which estimator do you think is better? Why?
b. With $\mu = 0$ and $\sigma^2 = 100$, write matlab code to generate $X_1, X_2, X_3, \ldots, X_{100} \sim \mathcal{N}(\mu, \sigma^2)$ using only looping constructs and the builtin function *randn*. Using only looping constructs, write a function with inferface

$$\bar{\mathrm{x}} = \mathrm{mean}(\mathbf{x})$$

for the estimator $\bar{X}$ where $\mathbf{x}$ is a vector (of unknown length) consisting of the observed values of $X_1, X_2, X_3, \ldots, X_{100}$. Generate 100 different datasets, denote the output of the function above as $\hat{\mu}_i$, and plot the density histogram of the set $\{\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \ldots, \hat{\mu}_{100}\}$ using the function *bar*. Using *hold on* and *plot* overlay the probability density of $\bar{X}$ on the plot (see matlab function *normpdf*).

4. (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.

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

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

a. Write a matlab function with interface

$$\{\mathrm{b}_0, \mathrm{b}_1\} = \mathrm{linregress}(\mathbf{x}, \mathbf{y})$$

and use it to obtain a regression function estimate.

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

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 for this problem on www.mhhe.com/KutnerALRM4e. Use MATLAB, do not use any other programming language. Only basic MATLAB operators are allowed, do not use any built-in functions to do the regression, i.e. the function *regress* cannot be used except, perhaps, to verify that your answer is correct before submitting your own implementation.