Homework 1
STAT W4315: Linear Regression Models
Simple Linear Regression Model

DUE: Thursday, February 4, 12:00 noon

(1) Please sign your home work with your name and UNI number.

(2) Homework must be submitted into the Statistics Homework Boxes room 904 on the 9th floor of SSW building.

(3) Homework is due Thursday, February 4, 12:00 noon.

(4) No late homework, under any circumstances, will be accepted.

(5) At the end of semester, one of your lowest homework scores will be dropped before the final grade is calculated.

During the lecture using the least squares method for the Simple Linear Regression Model we have derived the following **normal equations**:

$$\sum_{i=1}^{N} Y_i = Nb_0 + b_1 \sum_{i=1}^{N} X_i$$

$$\sum_{i=1}^{N} X_i Y_i = b_0 \sum_{i=1}^{N} X_i + b_1 \sum_{i=1}^{N} X_i^2,$$

where b_0 and b_1 are the least squares estimators of β_0 and β_1, respectively.

We have also stated the following properties of the corresponding fitted regression line $\widehat{Y}_i = b_0 + b_1 X_i$, and model residuals $e_i = Y_i - \widehat{Y}_i$, for $i = 1, \ldots, N$:

(1) The sum of the residuals is zero:

$$\sum_{i=1}^{N} e_i = 0,$$

(2) The sum of the square residuals $\sum_{i=1}^{N} e_i^2$ is minimized, i.e., for all $a_0 \in \mathbb{R}$ and $a_1 \in \mathbb{R}$,

$$\sum_{i=1}^{N} e_i^2 \leq \sum_{i=1}^{N} \left(Y_i - a_0 - a_1 X_i\right)^2 .$$

(3) The sum of the observed values Y_i equals the sum of the fitted values $\widehat{Y}_i$

$$\sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} \widehat{Y}_i.$$

(4) The sum of the residuals weighted by the predictors X_i is zero

$$\sum_{i=1}^{N} X_i e_i = 0.$$

(5) The sum of the residuals weighted by the fitted value of the response variables Y_i is zero

$$\sum_{i=1}^{N} \widehat{Y}_i e_i = 0.$$

(6) The regression line always goes through the point $\left(\bar{X}, \bar{Y}\right)$, where $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ and $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$.

Problem 1 (30 points)

Please derive mathematically the Properties (1)-(6) listed above using the assumptions of the Linear Regression Model and the Normal Equations. Derivation of each property is worth 5 points.

Problem 2 (24 points)

Geometric interpretation and intuition behind the properties (1)-(6) is independent of their mathematical derivation. Please in each sentence identify to which property it refers the closest and select one of the three options (A / B / C) which you find the most correct for this property. Correct answer to every sentence is worth 4 points (2 for the right property and 2 for the right choice A,B,C). You can use each property only once.

- Based on Property (   ), the regression line always goes through the "center of mass" point ($\left(\bar{X}, \bar{Y}\right)$ / $(0,0)$ / $\left(\bar{X} + \bar{Y}, \bar{X} - \bar{Y}\right)$).

- Based on Property (   ), the regression line minimizes the sum of the (horizontal / vertical / shortest) distances between the data (X_i, Y_i) and the regression line $\widehat{Y_i} = b_0 + b_1 X_i$.

- Based on Property (   ), the fitted regression line is (parallel / orthogonal / at 45 degree angle) with the model residuals.

- Based on Property (   ), in the language of descriptive statistics, the regression line is an unbiased estimator of the (regression coefficients / mean of the error term / mean of the predicted variable).

- Based on Property (   ), in the language of descriptive statistics, the residuals have (maximal / minimal / zero) mean.

- Based on Property (   ), the model residuals are (parallel / orthogonal / at 45 degree angle) with the independent variable.

Problem 3 - Problem 1.19 & 1.23 in the ALRM book (26 points)

The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a students grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). The results of the study follow. Assume that first-order regression model, i.e.,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where:

- Y_i is the value of the response variable in the ith trial.

- β_0 and β_1 are parameters.

- X_i is a known constant, namely, the value of the predictor variable in the ith trial.

- ε_i is a random error term with mean $\mathbb{E}[\varepsilon_i] = 0$ and variance σ^2; ε_i and ε_j are uncorrelated so that their covariance is zero (i.e., $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for all i, j; $i \neq j$, $i, j = 1, \ldots, n$;

is appropriate. (The data is in Data_HW1_Problem3.txt file on courseworks.)

(a)(3 points) Obtain the least squares estimates of β_0 and β_1, and state the estimated regression function.

(b)(3 points) Plot the estimated regression function and the data. Does the estimated regression function appear to fit the data well?

(c)(5 points) Obtain a point estimate of the mean freshman GPA for students with ACT test score $X = 30$.

(d)(5 points) What is the point estimate of the change in the mean response when the entrance test score increases by one point?

(e)(5 points) Obtain the residuals e_i. Do they sum to zero in accord with property 1 in Problem 1?

(f)(5 points) Estimate σ^2 and σ. In what units is σ expressed?

Problem 4 (20 points)

Consider the square of a standard Gaussian random variable with $\mu = 0$ and $\sigma = 1$. Show that the pdf of Z^2 is

$$f_{Z^2}(u) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-u/2} u^{-1/2} \mathbb{I}\{u > 0\}.$$

Hint: Since the transformation $Y = X^2$ is not monotone we cannot use

$$f_Y(y) = f_X(h(y)) \left| \frac{\mathrm{d}h}{\mathrm{d}y} \right|. \tag{1}$$

But consider the "half normal" density,

$$f_{|Z|}(u) = 2f(u)\,\mathbb{I}\{u \geq 0\}$$

where $f(u)$ is the pdf of a standard Gaussian random variable. This is the pdf of the absolute value of a Gaussian variable. The transformation $g(|u|) = u^2$ is $1 - 1$, so you can apply (1) to obtain the pdf of Z^2.