# EE708: Fundamentals of Data Science and Machine Intelligence

## Assignment 2

*Based on Module 3: Regression analysis and modeling*

1. When the training set is small, the contribution of variance to error may be more than that of bias, and in such a case, we may prefer a simple model even though we know it is too simple for the task. Can you give an example?
2. What is the effect of changing $\lambda$ on bias and variance?
$$E = \sum_t [r^t - g(x^t|w)]^2 + \lambda \sum_i w_i^2$$
3. On average, do people gain weight as they age? Based on a dataset of 250 samples, some summary statistics for both age ($x$) and weight ($y$) are:
$$\sum_{i=1}^{n} x_i = 11211.00 \qquad \sum_{i=1}^{n} y_i = 44520.80 \qquad \sum_{i=1}^{n} x_i y_i = 1996904.15$$
$$\sum_{i=1}^{n} x_i^2 = 543503.00 \qquad \sum_{i=1}^{n} y_i^2 = 8110405.02$$
Assume that the two variables are related according to the simple linear regression model.
   a. Calculate the least squares estimates of the slope and intercept.
   b. Use the equation of the fitted line to predict the weight that would be observed, on average, for a man who is 25 years old.
   c. Suppose that the observed weight of a 25-year-old man is 170 lbs. Find the residual for that observation.
   d. Was the prediction for the 25-year-old in part (c) an overestimate or underestimate? Explain briefly.
4. An article in Concrete Research presented 14 data samples on compressive strength $x$ and intrinsic permeability $y$ of various concrete mixes and cures. Summary quantities are:
$$\Sigma y_i = 572 \qquad \sum x_i = 43 \qquad \sum x_i y_i = 1697.8$$
$$\sum y_i^2 = 23{,}530 \qquad \sum x_i^2 = 157.42$$
Assume that the two variables are related according to the simple linear regression model.
   a. Calculate the least squares estimates of the slope and intercept. Estimate $\sigma^2$.
   b. Use the equation of the fitted line to predict what permeability would be observed when the compressive strength is $x = 4.3$.
   c. Give a point estimate of the mean permeability when compressive strength is $x = 3.7$.
   d. Suppose that the observed value of permeability at $x = 3.7$ is $y = 46.1$. Calculate the value of the corresponding residual.
5. A study was performed to investigate the shear strength of soil $y$ as it relates to depth in feet $x_1$ and % moisture content $x_2$. Ten observations were collected, and the following summary quantities were obtained:
$$\sum x_{i1} = 223 \qquad \sum x_{i1}^2 = 5{,}200.9 \qquad \sum x_{i1} y_i = 43{,}550.8$$
$$\sum x_{i2} = 553 \qquad \sum x_{i2}^2 = 31{,}729 \qquad \sum x_{i2} y_i \; 104{,}736.8$$
$$\sum y_i = 1{,}916 \qquad \sum y_i^2 = 371{,}595.6 \qquad \sum x_{i1} x_{i2} = 12{,}352$$
   a. Set up the least squares normal equations for the model
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$
   b. Estimate the parameters in the model in part (a).
   c. What is the predicted strength when $x_1 = 18$ feet and $x_2 = 43\%$?

6. A regression model is described between the percent body fat (%BF) measured by immersion and BMI from a study on 250 male subjects. The researchers also measured 13 physical characteristics of each man, including his age (yrs), height (in), and waist size (in). Write out the regression model of the percent of body fat with both height and waist as predictors with the given information:

$$(X'X)^{-1} = \begin{bmatrix} 2.9705 & -4.0042E-2 & -4.1679E-2 \\ -0.4004 & 6.0774E-4 & -7.3875E-5 \\ -0.00417 & -7.3875E-5 & 2.5766E-4 \end{bmatrix} \text{ and } (X'y) = \begin{bmatrix} 4757.9 \\ 334335.8 \\ 179706.7 \end{bmatrix}$$

7. Let us say we have two variables $x_1$ and $x_2$ and we want to make a quadratic fit using them, namely

$$f(x_1, x_2) = w_0 + w_1x_1 + w_2x_2 + w_3x_1x_2 + w_4x_1^2 + w_5x_2^2$$

Derive the least square estimates of $w_i, i = 0,1,\ldots,5$, given $N$ data samples.

**Programming Questions:**

8. Assume a linear model and add 0-mean Gaussian noise to generate 100 samples.
   a. Divide your sample into training and testing sets (80:20).
   b. Use linear regression for the training half. Compute the mean squared error (MSE) on the testing set.
   c. Plot the fitted model along with the data.
   d. Repeat the same for polynomials of degrees 2 and 3 as well.
9. Implement logistic regression using dataset A2_P2.csv. Write a code for gradient descent with learning rates of 0.01 and 0.05. For each learning rate:
   a. Plot variation of mean squared error for 20 iterations.
   b. Specify the final weight value.
10. Write a code to implement regression models using dataset A2_P3.csv. Divide the dataset into training and testing sets (80:20). Implement the following models using the training dataset and compute MSE on the test dataset:
    a. Linear regression.
    b. Linear regression with LASSO regularization $\left(\frac{\lambda}{2} = 1\right)$.
    c. Linear regression with ridge regularization $\left(\frac{\lambda}{2} = 0.1\right)$.

    Use bar plots to compare MSE and feature coefficients (weights) for the three methods.