**ECE 5403: Machine Learning** **(Due: 03/23/2020)**

# Submission Assignment #1

*Instructor:* Chun-Shu Wei *Name:* Student name, *Student Id:* Student Id

**Course Policy**: Read all the instructions below carefully before you start working on the assignment, and before you make a submission. For this assignment, please hand in the following two things: a pdf file and a ipynb file.

- PDF file: contains both your results and explanations. Please name this pdf file as **HW1_StudentID_Name.pdf** and **remember to type your Student ID and Name in pdf (e.g. HW1_9400000_chunshuwei)**.
- Ipynb file: write the comment to explain your code. Please name this ipynb file as **HW1_StudentID_Name.ipynb**
- Please name your assignment as **HW1_StudentID_Name.zip**. The archive file contains source code(ipynb file) and report (pdf file).
- Implementation will be graded by completeness, algorithm correctness, model description, and discussion.
- PLAGIARISM IS STRICTLY PROHIBITED.
- Please submit your assignment as ONE single zip file on the E3 system. Paper submission is not allowed. Inserting clear scanned image of handwritten derivations is accepted. Denote date and time on the first page.
- Submission deadline: **2020.3.23 11:59:59 PM**.

**Deriving the Ordinary Least Squares Estimates**
Let $\{(x_i, y_i)|i = 1, ..., n\}$ be a random sample of size n from the population. Given the following simple linear regression:

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{0.1}$$

where $u_i$ is the error term for observation $i$ since it contains all factors affecting $y_i$ other than $x_i$. Assume that **u is uncorrelated with x** in the population. Therefore, the expected value of u is zero, and the covariance between x and u is zero:

$$E[u] = 0 \tag{0.2}$$

and

$$Cov(x, u) = E[xu] = 0 \tag{0.3}$$

In terms of the observable variables x, y ,and the unknown parameters $\beta_0$ and $\beta_1$, equations (0.2) and (0.3) can be expressed as

$$E[y - \beta_0 - \beta_1 x] = 0 \tag{0.4}$$

and

$$E[x(y - \beta_0 - \beta_1 x)] = 0 \tag{0.5}$$

respectively. Equations (0.4) and (0.5) imply two restrictions on the joint probability distribution of (x,y) in the population. Since there are two unknown parameters to estimate, we could use equations (0.4) and (0.5) to estimate $\beta_0$ and $\beta_1$. Given a sample of data, we choose estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ to solve the sample counterparts of equations (0.4) and (0.5):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{0.6}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{0.7}$$

This is an example of the method of moments approach to estimation. Since

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{0.8}$$

As $\bar{y}$ is also the sample average of the $y_i$ and likewise for x. This equation allows us to write $\hat{\beta}_0$ in terms of $\hat{\beta}_1$, y, and x:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{0.9}$$

Therefore, once we have the slope estimate $\hat{\beta}_1$, it is straightforward to obtain the intercept estimate $\hat{\beta}_1$, given y and x. Dropping the $\frac{1}{n}$ in (0.7) (since it does not affect the solution) and plugging (0.9) into (0.7) yields

$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i) = 0 \tag{0.10}$$

which, upon rearrangement, gives

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}) \tag{0.11}$$

From basic properties of the summation operator

$$\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

and

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

. Therefore, provided that

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$$

, the estimated slope is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{0.12}$$

The estimates given in (0.9) and (0.12) are called the **ordinary least squares (OLS)** estimates of $\beta_0$ and $\beta_1$. To justify this name, for any $\hat{\beta}_0$ and $\hat{\beta}_1$ define a fitted value for y when $x = xi$ as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{0.13}$$

This is the value we predict for y when $x = x_i$ for the given intercept and slope. There is a fitted value for each observation in the sample. The **residual** for observation $i$ is the difference between the actual $y_i$ and its fitted value:

$$\hat{u}_i = y_i - \hat{y} = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \tag{0.14}$$

Again, there are n such residuals. Notice that residuals are not the same as the errors in (0.1). The fitted values and residuals are indicated in Figure 1. Now, suppose we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to make the sum of squared residuals,

$$\sum_{i=1}^{n} \hat{u_i}^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{0.15}$$

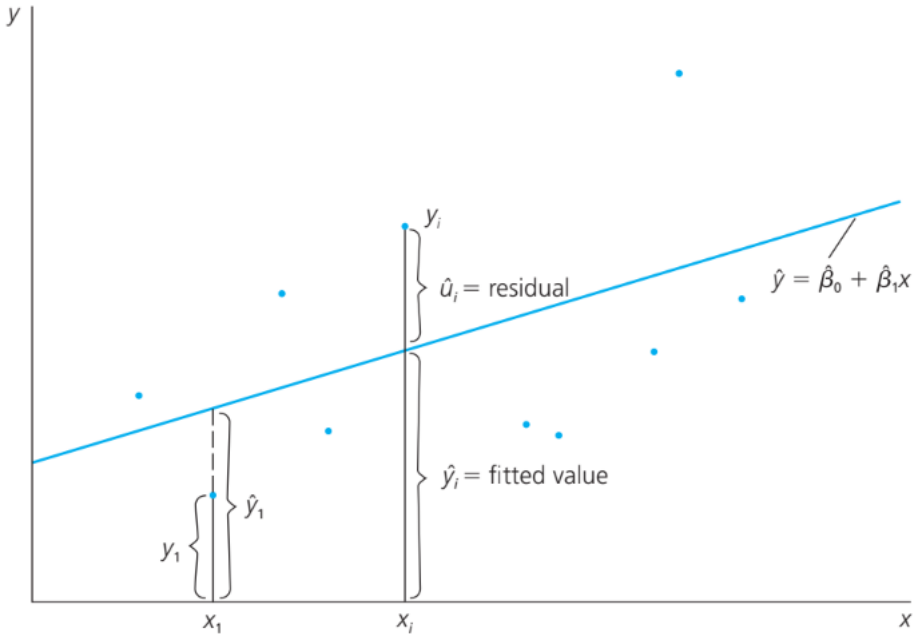


Figure 1: Fitted values and residuals

as small as possible. The appendix to this chapter shows that the conditions necessary for $(\hat{\beta}_0, \hat{\beta}_1)$ to minimize (0.15) are given exactly by equations (0.6) and (0.7), without $\frac{1}{n}$. Equations (0.6) and (0.7) are often called the **first order conditions** for the OLS estimates, a term that comes from optimization using calculus. From our previous calculations, we know that the solutions to the OLS first order conditions are given by (0.9) and (0.12). The name "ordinary least squares" comes from the fact that these estimates minimize the sum of squared residuals. With OLS, we will be able to derive unbiasedness, consistency, and other important statistical properties relatively easily. Plus, as the motivation in equations (0.4) and (0.5) suggests, and as we will see in Section 2-5, OLS is suited for estimating the parameters appearing in the conditional mean function (0.16).

$$E[y|x] = \beta_0 + \beta_1 x \tag{0.16}$$

Once we have determined the OLS intercept and slope estimates, we form the OLS regression line:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{0.17}$$

where it is understood that $\hat{\beta}_0$ and $\hat{\beta}_1$ have been obtained using equations (0.9) and (0.12). The notation $\hat{y}$, emphasizes that the predicted values from equation (0.17) are estimates. The intercept, $\hat{\beta}_0$, is the predicted value of $y$ when $x = 0$, although in some cases it will not make sense to set $x = 0$. In those situations, $\hat{\beta}_0$ is not, in itself, very interesting. When using (0.17) to compute predicted values of $y$ for various values of $x$, we must account for the intercept in the calculations. Equation (0.17) is also called the sample regression function (SRF) because it is the estimated version of the population regression function (0.17).

**Unbiasedness of OLS** We begin by establishing the unbiasedness of OLS under a simple set of assumptions. For future reference, it is useful to number these assumptions using the prefix "SLR" for simple linear regression. The first assumption defines the population model.

**Theorem 0.1** *Assumption of Ordinary Least Squares: Gauss–Markov assumptions*

- *Linear in Parameters (Correct specification): In the population model, the dependent variable, $y$, is related to the independent variable, $x$, and the error (or disturbance), $u$, as*

$$y = \beta_0 + \beta_1 x + u \tag{0.18}$$

  *where $\beta_0$ and $\beta_1$ are the population intercept and slope parameters, respectively.*

- *Random Sampling: We have a random sample of size $n$, $\{(x_i, y_i)|i = 1, .., n\}$, following the population model (0.18).*

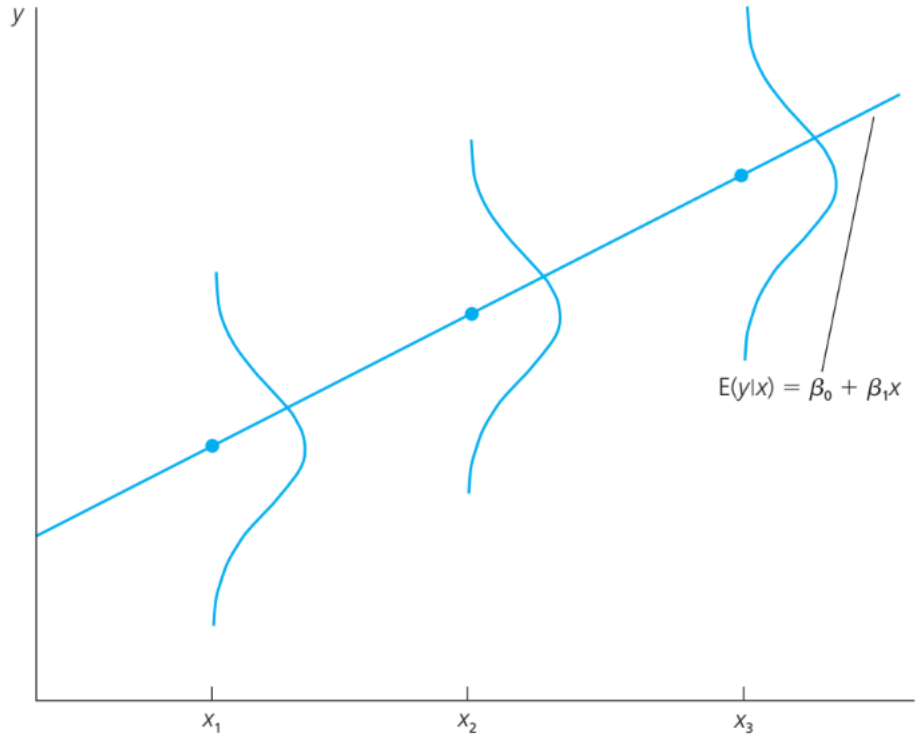


Figure 2: $E[y|x]$: population regression function

- *Sample Variation in the Explanatory Variable: The sample outcomes on x, namely,* $\{(x_i, y_i)|i = 1, .., n\}$*, are not all the same value.*
- ***Zero Conditional Mean*** *(Strict exogeneity): The error u has an expected value of zero given any value of the explanatory variable. In other words,*

$$E[u|x] = 0 \tag{0.19}$$

- *Homoskedasticity: The error u has the same variance given any value of the explanatory variable. In other words,*

$$Var[u|x] = \sigma^2 \tag{0.20}$$

*Now, we are ready to show that the OLS estimators are unbiased. Now, we are ready to show that the OLS estimators are unbiased. To this end, we use the fact that*

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i \tag{0.21}$$

*to write the OLS slope estimator in equation (0.12) as*

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{0.22}$$

*Because we are now interested in the behavior of* $\hat{\beta}_1$ *across all possible samples,* $\hat{\beta}_1$ *is properly viewed as a random variable. We can write* $\hat{\beta}_1$ *in terms of the population coefficient and errors by substituting the right-hand side of (0.18) into (0.21). We have*

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{0.23}$$

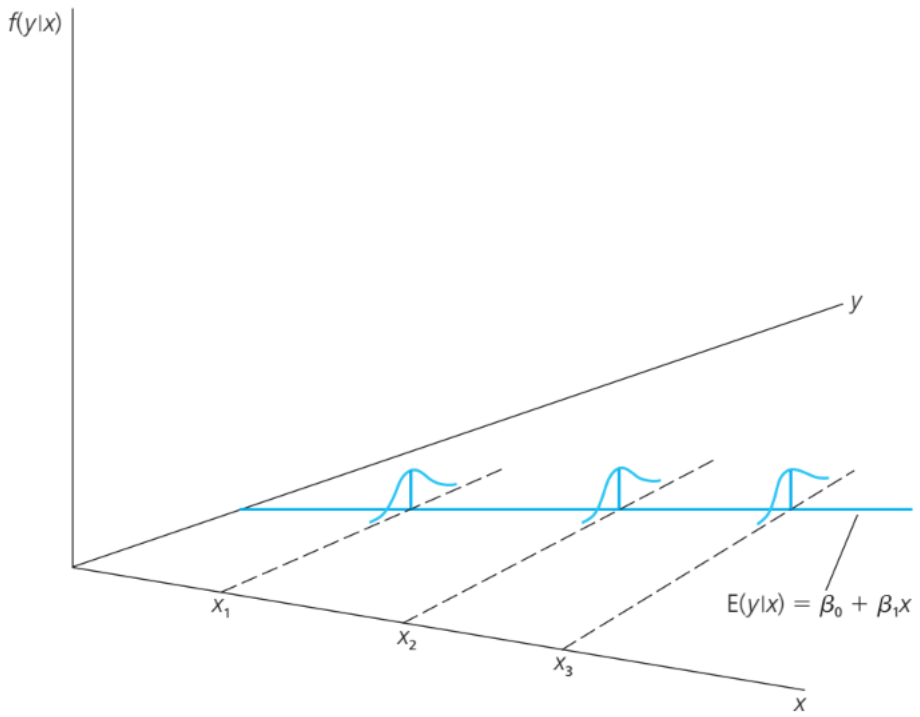


Figure 3: The simple regression model under homoskedasticity.

*Using the algebra of the summation operator, write the numerator of* $\hat{\beta}_1$ *as*

$$\sum_{i=1}^{n}(x_i-\bar{x})\beta_0+\sum_{i=1}^{n}(x_i-\bar{x})\beta_1 x_i+\sum_{i=1}^{n}(x_i-\bar{x})u_i=\beta_0\sum_{i=1}^{n}(x_i-\bar{x})+\beta_1\sum_{i=1}^{n}(x_i-\bar{x})x_i+\sum_{i=1}^{n}(x_i-\bar{x})u_i \quad (0.24)$$

*As* $\sum_{i=1}^{n}(x_i-\bar{x})=0$ *and* $\sum_{i=1}^{n}(x_i-\bar{x})x_i=\sum_{i=1}^{n}(x_i-\bar{x})^2$ *Therefore, we can write the numerator of* $\hat{\beta}_1$ *as* $\beta_1\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(x_i-\bar{x})u_i$*. Putting this over the denominator gives*

$$\hat{\beta}_1=\beta_1+\frac{\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\beta_1+\frac{\sum_{i=1}^{n}d_i u_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (0.25)$$

*where* $d_i=x_i-\bar{x}$*.*

**Problem 1. Unbiasedness of OLS** *(10+10=20 points)*

*Suppose that Population model is:*

$$y=\beta_0+\beta_1 x+u$$

*and OLS regression is :*

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$$

*. Using Assumptions 1 through 4 in Theorem 0.1 to show that:*

***(a)***

$$E[\hat{\beta}_0]=\beta_0$$

***(b)***

$$E[\hat{\beta}_1]=\beta_1$$

**Problem 2: Python code Exercise** *(5+8+7+10=30 points)*

*Data: MEAP93.csv*
*Notice: Please hand in with your code and results. Do not use a ready-made regression function in python. (e.g. sklearn.linear_model.LinearRegression, statmodels)*
*Let math10 denote the percentage of tenth graders at a high school receiving a passing score on a standardized mathematics exam. Suppose we wish to estimate the effect of the federally funded school lunch program on student performance. If anything, we expect the lunch program to have a positive ceteris paribus effect on performance: all other factors being equal, if a student who couldn't afford to eat regular meals becomes eligible for the school lunch program, his or her performance should improve. Let lnchprg denote the percentage of students who are eligible for the lunch program. Then, a simple regression model would be*

$$math10 = \beta_0 + \beta_1 lnchprg + u \tag{0.26}$$

*where u contains school and student characteristics that affect overall school performance. Using the data in MEAP93.csv on 408 Michigan high schools for the 1992–1993 school year:*
***(a)*** *Please estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ in (0.27).*

$$\hat{math10} = \hat{\beta}_0 + \hat{\beta}_1 lnchprg \tag{0.27}$$

***(b)*** *By Equation (0.27), How will an increment of 10 percent in the number of student eligible for the lunch program affect the percentage of students passing the math exam?*

***(c)*** *Try to explain the result in (b). If it is unreasonable, please list some possible reasons.*
***(d)*** *Estimate the model parameters*

$$math10 = \beta_0 + \beta_1 log(expend) + \beta_2 lnchprg + u \tag{0.28}$$

*and report the results in the usual form, including the sample size and R-squared. What sign of the slope coefficient do you expect? What is its meaning?*

**Problem 3: Essay Question** *(10+10=20 points)*

***(a)*** *Please describe the relationship between data, information, knowledge, and wisdom, and give an example in the daily life. (limit 500 words)*
***(b)*** *Please compare supervised learning and unsupervised learning and give an example for each. (limit 500 words)*

**Problem 4: Paper Review** *(10+20=30 points)*

*Assigned Paper: Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105)*
*Please review the paper and answer the following questions.*
***(a)*** *Please summarize the main technical contribution of this paper. (limit 300 words)*

***(b)*** *Please analyze this work according to the machine learning procedures. Provide short and precise description of what was done in each of the following steps of machine learning. (limit 500 words in sum)*

- *Defining problem*
- *Gathering data*
- *Data preparation*
- *Model development*
- *Training*
- *Evaluation*
- *Parameter tuning*