STAT 512: Statistical Inference **Autumn 2022**

Quiz Session 6: Multinomial Distribution and MLE for Simple Linear Regression

Yikun Zhang *November 16, 2022*

Problem 1 (Midterm problem in Autumn 2018, 2019)**.** *Suppose that a sample of size n is taken at random* with replacement *from the population of all UW students. Each student in the sample is recorded as either male (M) or female (F) and as either a Washington resident (R) or non-resident (N). The data are presented in a two-way contingency table as:*

$$\begin{array}{c} \\ M \\ F \end{array} \begin{array}{c} \begin{array}{cc} R & N \end{array} \\ \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} \end{array}.$$

That is, X_{11} is the number of students in the sample who are both M and R, X_{12} is the number of students in the sample who are both M and N, etc. Thus, $X_{11} + X_{12} + X_{21} + X_{22} = n$. Let

$$\begin{array}{c} \\ M \\ F \end{array} \begin{array}{c} \begin{array}{cc} R & N \end{array} \\ \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \end{array}$$

denote the corresponding population proportions, that is, p_{11} is the proportion of students in the UW population who are both M and R, p_{12} is the proportion of students in the population who are both M and N, etc. Hence, $p_{11} + p_{12} + p_{21} + p_{22} = 1$.

- *(a) What is the distribution of $(X_{11}, X_{12}, X_{21}, X_{22})$?*
- *(b) What is the conditional distribution of $(X_{11}, X_{12})|X_{11} + X_{12}$? What is the conditional correlation Corr$(X_{11}, X_{12})|X_{11} + X_{12}$?*
- *(c) Find Corr$(X_{11}, X_{12})|X_{11} + X_{12} + X_{21}$.*
- *(d) What is the conditional distribution of $(X_{11}, X_{12})|X_{11} + X_{21}$? What is the conditional correlation Corr$(X_{11}, X_{12})|X_{11} + X_{21}$?*

Solution. (a) Based on the "sampling with replacement" setting, we know that

$$(X_{11}, X_{12}, X_{21}, X_{22}) \sim \text{Multinomial}_4(n; p_{11}, p_{12}, p_{21}, p_{22}).$$

(b) According to the calculations and results in Section 7.2 of Lecture 7 notes (see also Chapter 7 in Perlman 2020a), we know that the conditional distribution of $(X_{11}, X_{12})|X_{11} + X_{12}$ is

$$(X_{11}, X_{12})|X_{11} + X_{12} \sim \text{Multinomial}_2\left(X_{11} + X_{12}; \frac{p_{11}}{p_{11} + p_{12}}, \frac{p_{12}}{p_{11} + p_{12}}\right).$$

Given $X_{11} + X_{12}$, X_{11} and X_{12} is negatively linear correlated, so $\text{Corr}(X_{11}, X_{12})|X_{11} + X_{12} = -1$.

(c) Recall from Section 7.2 of Lecture 7 notes that for any multinomial random vector $(X_1, ..., X_k) \sim \text{Multinomial}_4(n; p_1, ..., p_k)$, the covariance between any two components X_i, X_j with $1 \leq i \neq j \leq k$ can be computed as:

$$\text{Cov}(X_i, X_j) = \frac{1}{2}\left[\text{Var}(X_i + X_j) - \text{Var}(X_i) - \text{Var}(X_j)\right]$$

$$= \frac{1}{2}\left[n(p_i+p_j)(1-p_i-p_j) - np_i(1-p_i) - np_j(1-p_j)\right]$$
$$= -np_ip_j.$$

Hence, given that

$$(X_{11}, X_{12}, X_{21})|X_{11}+X_{12}+X_{21} \sim \text{Multinomial}_3\left(X_{11}+X_{12}+X_{21}; \frac{p_{11}}{p_{11}+p_{12}+p_{21}}, \frac{p_{12}}{p_{11}+p_{12}+p_{21}}, \frac{p_{21}}{p_{11}+p_{12}+p_{21}}\right),$$

we can obtain from the above results as:

$$\text{Cov}(X_{11}, X_{12})|X_{11}+X_{12}+X_{21} = -(X_{11}+X_{12}+X_{21})\left(\frac{p_{11}}{p_{11}+p_{12}+p_{21}}\right)\left(\frac{p_{12}}{p_{11}+p_{12}+p_{21}}\right).$$

Additionally,

$$\text{Var}(X_{11}|X_{11}+X_{12}+X_{21}) = (X_{11}+X_{12}+X_{21})\left(\frac{p_{11}}{p_{11}+p_{12}+p_{21}}\right)\left(\frac{p_{12}+p_{21}}{p_{11}+p_{12}+p_{21}}\right)$$

and

$$\text{Var}(X_{12}|X_{11}+X_{12}+X_{21}) = (X_{11}+X_{12}+X_{21})\left(\frac{p_{12}}{p_{11}+p_{12}+p_{21}}\right)\left(\frac{p_{11}+p_{21}}{p_{11}+p_{12}+p_{21}}\right).$$

Therefore,

$$\begin{aligned}\text{Corr}(X_{11}, X_{12})|X_{11}+X_{12}+X_{21} &= \frac{\text{Cov}(X_{11}, X_{12})|X_{11}+X_{12}+X_{21}}{\sqrt{\text{Var}(X_{11}|X_{11}+X_{12}+X_{21})\cdot\text{Var}(X_{12}|X_{11}+X_{12}+X_{21})}} \\ &= -\sqrt{\frac{p_{11}p_{12}}{(p_{12}+p_{21})(p_{11}+p_{21})}}.\end{aligned}$$

(d) Notice that the conditional distribution of $(X_{11}, X_{12})|X_{11}+X_{21}$ is identical to the conditional distribution of $(X_{11}, X_{12})|(X_{11}+X_{21}, X_{12}+X_{22})$. By the results in Section 7.2 of Lecture notes, we know that X_{11} and X_{12} is conditionally independent given $(X_{11}+X_{21}, X_{12}+X_{22})$. Thus, the conditional distribution of $(X_{11}, X_{12})|X_{11}+X_{21}$ is the product of two independent binomial distributions as:

$$(X_{11}, X_{12})|X_{11}+X_{21} \sim \text{Binomial}\left(X_{11}+X_{12}, \frac{p_{11}}{p_{11}+p_{21}}\right) \otimes \text{Binomial}\left(X_{12}+X_{22}, \frac{p_{12}}{p_{12}+p_{22}}\right),$$

where $\otimes$ stands for the product of two independent distributions. Finally, the conditional correlation $\text{Corr}(X_{11}, X_{12})|X_{11}+X_{21}$ is zero. □

Problem 2. *Assume that we want to estimate θ from some data $\boldsymbol{X} = (X_1, \dots, X_n)$, where $X_i \sim P_\theta$ are independent and identically distributed. An estimator $\hat{\theta} = T(\boldsymbol{X})$ has been constructed and we quantify its performance via the squared loss function $L(T(\boldsymbol{X}), \theta) = (T(\boldsymbol{X}) - \theta)^2$, where T is some deterministic function. The risk function $R(T(\boldsymbol{X}), \theta)$ is defined as the expected value of the loss function as (see also Section 7.1 in Hogg et al. 2012):*

$$R(T(\boldsymbol{X}), \theta) = \mathbb{E}\left[L(T(\boldsymbol{X}), \theta)\right] = \mathbb{E}\left[(T(\boldsymbol{X}) - \theta)^2\right]$$

Show that $R(T(\boldsymbol{X}), \theta) = Bias^2(T(\boldsymbol{X})) + Var(T(\boldsymbol{X}))$ where $Bias(T(\boldsymbol{X})) = \mathbb{E}[T(\boldsymbol{X})] - \theta$.

Proof. By direct calculations,

$$(T(\boldsymbol{X}) - \theta)^2 = (T(\boldsymbol{X}) - \mathbb{E}[T(\boldsymbol{X})] + \mathbb{E}[T(\boldsymbol{X})] - \theta)^2$$

$$= (T(\boldsymbol{X}) - \mathbb{E}[T(\boldsymbol{X})])^2 + (\mathbb{E}[T(\boldsymbol{X})] - \theta)^2 + 2\left(T(\boldsymbol{X}) - \mathbb{E}[T(\boldsymbol{X})]\right)\left(\mathbb{E}[T(\boldsymbol{X})] - \theta\right).$$

Now, taking the expectation yields that

$$\begin{aligned}
R\left(T(\boldsymbol{X}), \theta\right) &= \mathbb{E}\left[(T(\boldsymbol{X}) - \theta)^2\right] \\
&= \mathbb{E}\left[(T(\boldsymbol{X}) - \mathbb{E}[T(\boldsymbol{X})])^2\right] + \mathbb{E}\left[(\mathbb{E}[T(\boldsymbol{X})] - \theta)^2\right] + \mathbb{E}\left[2\left(T(\boldsymbol{X}) - \mathbb{E}[T(\boldsymbol{X})]\right)\left(\mathbb{E}[T(\boldsymbol{X})] - \theta\right)\right] \\
&= \mathrm{Var}(T(\boldsymbol{X})) + (\mathbb{E}[T(\boldsymbol{X})] - \theta)^2 + 2\left(\mathbb{E}[T(\boldsymbol{X})] - \theta\right) \cdot \underbrace{\mathbb{E}\left[T(\boldsymbol{X}) - \mathbb{E}[T(\boldsymbol{X})]\right]}_{=0} \\
&= \mathrm{Var}(T(\boldsymbol{X})) + \left(\mathbb{E}\left[T(\boldsymbol{X})\right] - \theta\right)^2 \\
&= \mathrm{Var}(T(\boldsymbol{X})) + \mathrm{Bias}^2(T(\boldsymbol{X}))
\end{aligned}$$

The result follows. □

Problem 3 (MLE of simple linear regression; Exercises 7.19–7.21 in Casella and Berger 2002)**.** *Suppose that the random variables $Y_1, ..., Y_n$ satisfy*

$$Y_i = \beta x_i + \epsilon_i, \quad i = 1, \dots, n$$

where $x_1, \dots, x_n$ are fixed constants and $\epsilon_1, \dots, \epsilon_n \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$, $\sigma^2 > 0$ is unknown.

(a) Find the maximum likelihood estimators (MLEs) of β and σ^2. Show that the MLE $\hat{\beta}_{MLE}$ of β is unbiased. What is its variance?

(b) Show that $\sum_{i=1}^n Y_i / \sum_{i=1}^n x_i$ is also an unbiased estimator of β. What is its variance? Show that it is larger than the variance in (a).

(c) Show that $\frac{1}{n}\sum_{i=1}^n (Y_i/x_i)$ is also an unbiased estimator of β. What is its variance? Compare it to the variances in (a) and (b).

Solution. (a) The log-likelihood is given by

$$\begin{aligned}
\log L(\beta, \sigma^2 | \boldsymbol{Y}) &= \sum_{i=1}^n \log p(Y_i | \beta, \sigma^2) \\
&= \sum_{i=1}^n \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y_i - \beta x_i)^2\right] \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta x_i)^2 \\
&= -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\lambda) - \frac{\lambda}{2}\sum_{i=1}^n (Y_i - \beta x_i)^2,
\end{aligned}$$

where $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ and we take $\lambda = \frac{1}{\sigma^2}$. Taking the partial derivatives with respect to β and λ (or equivalently, σ^2) yields that

$$\begin{aligned}
\frac{\partial}{\partial \beta} \log L(\beta, \sigma^2 | \boldsymbol{Y}) &= \frac{1}{\sigma^2}\sum_{i=1}^n x_i (Y_i - \beta x_i), \\
\frac{\partial}{\partial \lambda} \log L(\beta, \sigma^2 | \boldsymbol{Y}) &= \frac{n}{2\lambda} - \frac{1}{2}\sum_{i=1}^n (Y_i - \beta x_i)^2.
\end{aligned}$$

Given that $\frac{\partial^2}{\partial \beta^2} \log L(\beta, \sigma^2|\boldsymbol{Y}) = -\frac{1}{\sigma^2}\sum_{i=1}^n x_i^2 < 0$, the log-likelihood $\log L(\beta, \sigma^2|\boldsymbol{Y})$ is strictly concave with respect to β for any fixed $\sigma^2 > 0$. Hence, the solution $\beta^* = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$ to the equation $\frac{\partial}{\partial \beta} \log L(\beta, \sigma^2|\boldsymbol{Y}) = 0$ did maximize the log-likelihood $\frac{\partial}{\partial \beta} \log L(\beta, \sigma^2|\boldsymbol{Y})$ for any fixed $\sigma^2 > 0$. The partial maximum is

$$\max_{\beta} \log L(\beta, \sigma^2|\boldsymbol{Y}) = -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\lambda) - \frac{\lambda}{2}\sum_{i=1}^n (Y_i - \beta^* x_i)^2,$$

and $\frac{\partial^2}{\partial \lambda^2} \log L(\beta, \sigma^2|\boldsymbol{Y}) = -\frac{n}{2\lambda^2} < 0$. It implies that $\max_\beta \log L(\beta, \sigma^2|\boldsymbol{Y}) = \log L\left(\beta^*, \frac{1}{\lambda}|\boldsymbol{Y}\right)$ is strictly concave with respect to λ and has its unique maximum at $\hat{\lambda} = \frac{n}{\sum_{i=1}^n (Y_i - \beta^* x_i)^2}$ by solving $\frac{\partial}{\partial \lambda} \log L\left(\beta^*, \frac{1}{\lambda}|\boldsymbol{Y}\right) = 0$. Therefore, $(\beta^*, \hat{\lambda})$ jointly maximizes the log-likelihood $\log L(\beta, \sigma^2|\boldsymbol{Y})$, so it leads to the MLEs as:

$$\hat{\beta}_{MLE} = \beta^* = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \quad \text{and} \quad \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (Y_i - \beta^* x_i)^2 = \frac{\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n Y_i^2\right) - \left(\sum_{i=1}^n x_i Y_i\right)^2}{n\left(\sum_{i=1}^n x_i^2\right)}.$$

Finally, the expectation and variance of $\hat{\beta}_{MLE}$ are given by

$$\begin{aligned}
\mathbb{E}\left(\hat{\beta}_{\text{MLE}}\right) &= \mathbb{E}\left(\frac{\sum_{i=1} Y_i x_i}{\sum_{i=1}^n x_i^2}\right) \\
&= \frac{1}{\sum_{i=1}^n x_i^2}\sum_{i=1}^n x_i \mathbb{E}(Y_i) \\
&= \frac{1}{\sum_{i=1}^n x_i^2}\sum_{i=1}^n \beta x_i^2 \\
&= \beta,
\end{aligned}$$

showing that $\hat{\beta}_{MLE}$ is an unbiased estimator of β, and

$$\begin{aligned}
\text{Var}\left(\hat{\beta}_{\text{MLE}}\right) &= \text{Var}\left(\frac{\sum_{i=1} Y_i x_i}{\sum_{i=1}^n x_i^2}\right) \\
&= \frac{1}{\left(\sum_{i=1}^n x_i^2\right)^2}\sum_{i=1}^n x_i^2 \text{Var}(Y_i) \\
&= \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.
\end{aligned}$$

Notes: Indeed, the distribution of $\hat{\beta}_{MLE}$ is $N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n x_i^2}\right)$.

(b) By direct calculation, we have that

$$\begin{aligned}
\mathbb{E}\left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}\right) &= \frac{1}{\sum_{i=1}^n x_i}\sum_{i=1}^n \mathbb{E}(Y_i) \\
&= \frac{1}{\sum_{i=1}^n x_i}\sum_{i=1}^n \beta x_i \\
&= \beta
\end{aligned}$$

showing that $\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}$ is also an unbiased estimator of β, and

$$\text{Var}\left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}\right) = \frac{1}{\left(\sum_{i=1}^n x_i\right)^2}\sum_{i=1}^n \text{Var}(Y_i)$$

$$= \frac{\sigma^2}{\left(\sum_{i=1}^n x_i\right)^2/n}$$

The denominator here is $\frac{\left(\sum_{i=1}^n x_i\right)^2}{n} = n(\overline{x}_n)^2$ where $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$. Given that

$$0 \le \sum_{i=1}^n (x_i - \overline{x}_n)^2 = \sum_{i=1}^n x_i^2 - n(\overline{x}_n)^2,$$

it implies that

$$\sum_{i=1}^n x_i^2 \ge n(\overline{x}_n)^2$$

and thus, $\text{Var}(\hat{\beta}_{\text{MLE}}) \le \text{Var}\left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}\right)$. Here, the equality holds only when $x_1 = \cdots = x_n$.

(c) By direct calculations, we have that

$$\begin{aligned}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \frac{Y_i}{x_i}\right) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\frac{Y_i}{x_i}\right)\\ &= \frac{1}{n}\sum_{i=1}^n \frac{\beta x_i}{x_i}\\ &= \beta\end{aligned}$$

showing that $\frac{1}{n}\sum_{i=1}^n \frac{Y_i}{x_i}$ is also an unbiased estimator of β, and

$$\begin{aligned}\text{Var}\left(\frac{1}{n}\sum_{i=1}^n \frac{Y_i}{x_i}\right) &= \frac{1}{n^2}\sum_{i=1}^n \text{Var}\left(\frac{Y_i}{x_i}\right)\\ &= \frac{1}{n^2}\sum_{i=1}^n \frac{1}{x_i^2}\text{Var}(Y_i)\\ &= \frac{\sigma^2}{n^2}\sum_{i=1}^n \frac{1}{x_i^2}\end{aligned}$$

By the Cauchy-Schwarz inequality, we know that

$$\left(\sum_{i=1}^n \frac{1}{x_i^2}\right)\left(\sum_{i=1}^n x_i^2\right) \ge \left(\sum_{i=1}^n x_i \cdot \frac{1}{x_i}\right)^2 = n^2,$$

and thus,

$$\frac{\sigma^2}{n^2}\left(\sum_{i=1}^n \frac{1}{x_i^2}\right) \ge \frac{\sigma^2}{\sum_{i=1}^n x_i^2},$$

i.e., $\text{Var}\left(\frac{1}{n}\sum_{i=1}^n \frac{Y_i}{x_i}\right) \ge \text{Var}(\hat{\beta}_{MLE})$.

Notice, however, that the variances of estimators in (b) and (c) are *not comparable*. That is, we cannot conclude whether $\frac{n\sigma^2}{\left(\sum_{i=1}^n x_i\right)^2}$ or $\frac{\sigma^2}{n^2}\sum_{i=1}^n \frac{1}{x_i^2}$ is bigger without further information on $x_i, i = 1, ..., n$. Consider the following two cases:

- *Case 1*: Take $x_i = \frac{(-1)^i}{\sqrt{n}}$ for $i = 1, ..., n$ so that $|\sum_{i=1}^n x_i| \le \frac{1}{\sqrt{n}} \to 0$ as $n \to \infty$. It implies that

$$\text{Var}\left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}\right) = \frac{n\sigma^2}{\left(\sum_{i=1}^n x_i\right)^2} \ge n^3\sigma^2 \to \infty \quad \text{as } n \to \infty,$$

while

$$\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{x_i}\right) = \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\frac{1}{x_i^2} = \sigma^2 < \infty.$$

In this case,

$$\text{Var}\left(\frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i}\right) = \frac{n\sigma^2}{\left(\sum_{i=1}^{n} x_i\right)^2} > \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\frac{1}{x_i^2} = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{x_i}\right).$$

- *Case 2*: When $x_i, i = 1, ..., n$ are all positive and not identical, we can apply the Jensen's inequality to the convex function $f(u) = \frac{1}{u^2}$ for $u > 0$ to obtain that

$$\frac{1}{\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} \leq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{x_i^2}\right),$$

which in turn shows that $\frac{n\sigma^2}{\left(\sum_{i=1}^{n} x_i\right)^2} \leq \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\frac{1}{x_i^2}$. Thus, in this case, $\text{Var}\left(\frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i}\right) \leq \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{x_i}\right)$ and the equality does not hold when $x_i, i = 1, ..., n$ are not the same.

Notes: It is WRONG to apply the Jensen's equality to the function $f(u) = \frac{1}{u^2}$ without restricting to $\mathbb{R}^+$, because $f(u)$ is not convex in $\mathbb{R}$; see Figure 1 below. □

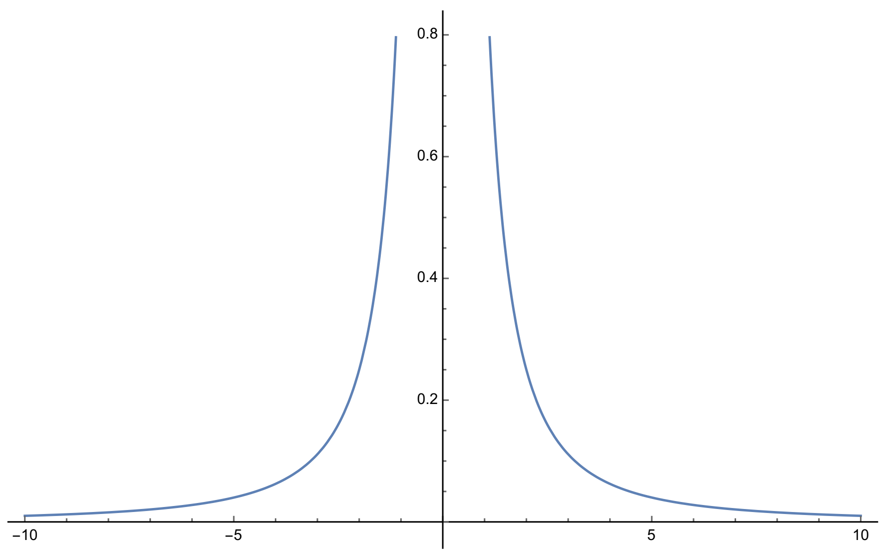

Figure 1: Plot of $\frac{1}{x^2}$ on $x \in [-10, 10]$.

Remark 1. *There are two different arguments about why $\hat{\beta}_{MLE}$ must attain the minimum variance among all the unbiased estimators:*

- *On the one hand, by maximizing the log-likelihood under Gaussian and homoscedastic assumptions on the errors $\epsilon_i, i = 1, ..., n$, $\hat{\beta}_{MLE}$ coincides with the ordinary least square solution. According to Gauss-Markov theorem*[1]*, $\hat{\beta}_{MLE}$ is the best linear unbiased estimator (BLUP) of β under the mean square error criterion (see Problem 2).*

- *On the other hand, $\hat{\beta}_{MLE}$ is a function of the two-dimensional complete and sufficient statistic for (β, σ^2) as:*

$$\left(\sum_{i=1}^{n} Y_i^2, \sum_{i=1}^{n} x_i Y_i\right),$$

[1] See https://en.wikipedia.org/wiki/Gauss-Markov_theorem.

so it is also the uniformly minimum-variance unbiased estimator (UMVUE); see Section 12.3 in Perlman [2020b].

A statistic $T(\boldsymbol{X})$ *is called* sufficient *for* θ *if the conditional distribution of the sample* $\boldsymbol{X}$ *given the value of* $T(\boldsymbol{X})$ *does not depend on* θ*; see Definition 6.2.1 in Casella and Berger [2002]. Also,* $T(\boldsymbol{X})$ *is called* complete *if* $\mathbb{E}_\theta g(T) = 0$ *for all* θ *implies* $\mathbb{P}_\theta\left(g(T) = 0\right) = 1$ *for all* θ*, where* $\mathbb{E}_\theta$ *and* $\mathbb{P}_\theta$ *are taken with respect to the distribution of* $T(\boldsymbol{X})$*; see Definition 6.2.21 in Casella and Berger [2002]. These concepts will be discussed in detail during STAT 513.*

References

G. Casella and R. Berger. *Statistical Inference*. Duxbury advanced series. Thomson Learning, 2nd ed. edition, 2002.

R. Hogg, J. McKean, and A. Craig. *Introduction to Mathematical Statistics*. Pearson Education, 7th edition, 2012.

M. Perlman. Probability and Mathematical Statistics I (STAT 512 Lecture Notes), 2020a. URL `https://sites.stat.washington.edu/people/mdperlma/STAT%20512%20MDP%20Notes.pdf`.

M. Perlman. Probability and Mathematical Statistics II (STAT 513 Lecture Notes), 2020b. URL `https://sites.stat.washington.edu/people/mdperlma/STAT%20513%20MDP%20Notes.pdf`.