# Generalized Linear Regression Homework 1

Brian Cervantes Alvarez
October 7, 2024

## Problem 1

Let the genotype frequencies be in equilibrium with probabilities:

| Haptoglobin Type | Expression | Count |
|---|---|---|
| Hpl-1 | $(1-\theta)^2$ | 10 |
| Hpl-2 | $2\theta(1-\theta)$ | 68 |
| Hpl-3 | $\theta^2$ | 112 |
| **Total** | | **190** |

### Part A

The likelihood function for the observed data can be written as,

$$L(\theta) = \binom{n}{n_1, n_2, n_3}(1-\theta)^{2n_1}[2\theta(1-\theta)]^{n_2}\theta^{2n_3}$$

Next, we can take the log-likelihood,

$$\ell(\theta) = n_1 \log[(1-\theta)^2] + n_2 \log[2\theta(1-\theta)] + n_3 \log[\theta^2]$$

Take derivative and solve for $\theta$,

$$\frac{d\ell(\theta)}{d\theta} = \frac{-(2n_1+n_2)}{1-\theta} + \frac{n_2+2n_3}{\theta} = 0,$$

$$\hat{\theta} = \frac{n_2+2n_3}{2(n_1+n_2+n_3)} = 0.7684211.$$

```
# Parameters
n1 <- 10; n2 <- 68; n3 <- 112; n_total <- 190
# MLE estimate
theta_mle <- (n2 + 2 * n3) / (2 * n_total)
theta_mle
```

```
[1] 0.7684211
```

# Problem 1

## Part B

The Fisher information $I(\theta)$ was solved algebraically below,

$$I(\theta) = \frac{n}{\theta(1-\theta)}$$

Thus, the asymptotic variance can be written and computed as follows,

$$\text{Var}(\hat{\theta}) = \frac{1}{I(\theta)} = \frac{\theta(1-\theta)}{n}$$

```
# Asymptotic variance
asymp_var <- theta_mle * (1 - theta_mle) / n_total
asymp_var
```

```
[1] 0.0009365797
```

Therefore, the asymptotic variance of $\hat{\theta}$ is $9.3657968 \times 10^{-4}$.

## Problem 1

### Part C

We can use the following Wald's test statistic,

$$Z = \frac{\hat{\theta} - 0.5}{\sqrt{\text{Var}(\hat{\theta})}}$$

```
# Test statistic
z_stat <- (theta_mle - 0.5) / sqrt(asymp_var)
z_stat
```

```
[1] 8.770901
```

Using the standard normal distribution as our reference distribution,

```
p_val <- 2 * pnorm(abs(z_stat), lower.tail = FALSE)
p_val
```

```
[1] 1.772422e-18
```

Hence, the Wald $Z$-statistic is 8.770901, with a p-value of $1.7724219 \times 10^{-18}$. Thus, we reject $H_0$.

## Problem 2

No, the least squares estimate would not be the same. Since regression on the mean responses of duplicates, $\bar{Y}_i$, removes **half of the variability** associated with each $Y_{i,1}$ and $Y_{i,2}$. Even though the estimates of $\beta_0$ and $\beta_1$ remain unbiased, they would have a smaller variance.

Given that the two observations for each $x_i$ are independent, the error variance can be estimated as,

$$\hat{\sigma}^2 = \frac{1}{2}\sum (Y_{i,1} - \bar{Y}_i)^2 + (Y_{i,2} - \bar{Y}_i)^2$$

Since each observation has half the variability, the estimate for the error variance remains unchanged.

### Illustration

```r
# Assume Y_i1 and Y_i2 are known
Y_i1 <- c(1, 2, 3); Y_i2 <- c(1.1, 1.9, 2.8); Y_bar <- (Y_i1 + Y_i2) / 2
# Error variance estimate
sigma_sq <- sum((Y_i1 - Y_bar)^2 + (Y_i2 - Y_bar)^2) / (2 * length(Y_i1))
sigma_sq
```

```
[1] 0.005
```

Thus, $\hat{\sigma}^2$ is 0.005 from this specific example

## Problem 3

Given a simple linear regression model,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

The likelihood function for the model is based on the normality assumption of $\epsilon_i$. The log-likelihood function is calculated to be,

$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Using this, we can directly derive the MLEs for $\beta_0$ and $\beta_1$ by taking their respective derivatives,

### Deriving $\hat{\beta}_1$

To find the MLE of $\beta_1$, we first differentiate the log-likelihood with respect to $\beta_1$ and set it equal to zero,

$$\frac{\partial \ell}{\partial \beta_1} = -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

Rearranging,

$$\sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$

Define $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, and subsitute,

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \beta_1 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Solving for $\beta_1$, we get its MLE,

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

## Deriving $\hat{\beta}_0$

Do the same steps as above, but for $\hat{\beta}_0$

$$\frac{\partial \ell}{\partial \beta_0} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

Rearranging,

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

Simplifying,

$$\bar{y} = \beta_0 + \beta_1 \bar{x}$$

We get,

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

Thus, we have derived the MLEs for both $\beta_0$ and $\beta_1$.

## Minimizing Var$(\hat{\beta}_1)$,

In order to minimize $\text{Var}(\hat{\beta}_1)$, we need to maximize $\sum(x_i - \bar{x})^2$. Our best choice for the $x_i$ values is to distribute them symmetrically and evenly across the interval $[-1, 1]$. We can achieve this by setting the $x_i$ values at the endpoints, $x_i \in \{-1, 1\}$, or placing them symmetrically within the bounds of this interval.