# 2022 August Qualifying Exam

Day 2

1. Scoliosis is an abnormal lateral curvature of the spine. Information on scoliosis is widely available across sites on the Internet including video sharing site YouTube, which has over 1 billion visitors per month. To study the popularity versus accuracy of YouTube videos for scoliosis, data from **50 independent** YouTube videos were collected. The primary outcome is the number of views on YouTube videos. The following variables were considered in this analysis.

   - Video ID: YouTube video ID number
   - Scol: Scoliosis-specific quality score. A higher score indicates better quality.
   - Age: The age of the videos in days.
   - Views: The number of views on the video

   Read the data with the code:

```
youtube <- read.csv("https://people.stat.sc.edu/gregorkb/youtube2.csv",
                    header = T,
                    stringsAsFactors = F)
```

   The first 5 observations of the data are:

| Video.ID | Scol | Age | Views |
|---|---|---|---|
| 188ptlrq1Qo | 3 | 1840 | 120125 |
| miWPVMmn-zQ | 14 | 2232 | 53547 |
| 9TWtrCmzaOw | 2 | 2842 | 382825 |
| BO8mtChRosg | 4 | 1822 | 8032 |
| ftkK0qIgjN0 | 7 | 1949 | 18385 |

   (a) Perform exploratory analysis on the variable **Views** and describe the characteristics of the variable **Views**. Describe the relationship between the variable **Views**, **Age** and **Scol**.

   (b) Let Y be 1 if a video has more than 20,000 views and 0 otherwise. Write down a regression model to model the probability of having more than 20,000 views based on the covariates: Age and scoliosis-specific quality score (Scol). Implement the described model, report the regression coefficient estimates with corresponding 95% confidence intervals. Interpret the regression coefficient estimates.

(c) Now consider a linear regression model with **Views** as the response variable (**Y**) and **Age**, **Scol** as the covariates. Write down the linear regression model and describe the model assumptions.

(d) Comment on the appropriateness of the model described in (c) for the data. Explain your reasons.

(e) i. Write down a Poisson regression model with **Views** as the response variable (**Y**) and **Age**, **Scol** as the covariates.

ii. Describe the model assumptions,

iii. implement the Poisson regression model, and

iv. interpret the regression coefficients.

(f) i. Comment on whether the model described in (e) is appropriate for the data. Why?

ii. Describe alternative approaches if better models could be considered (no need to perform the analyses, just describe).

iii. Describe an approach to compare goodness of fit between various models (no need to perform the analyses, just describe).

2. In a study of cardiovascular disease, researchers would like to examine the association between BMI (body mass index) and cholesterol level between 110 same-sex twin pairs. The twins are **randomly** ordered within a pair. The following variables are considered in this analysis.

- ID: ID number of the twin pair
- bmi1: Body mass index (BMI) of the first individual in the twin pair
- bmi2: BMI of the second individual in the twin pair
- age: age of the twin pair
- gender: gender of the twin pair
- dchol: difference of blood cholesterol level within the twin pair

Read the data with the code:

```
twindat <- read.table("https://people.stat.sc.edu/gregorkb/twin.txt",
                      header = T,
                      stringsAsFactors = F,
                      sep = "\t")
```

The first 5 observations of the data are:

| ID | bmi1 | bmi2 | age | gender | dchol |
|---|---|---|---|---|---|
| 1 | 24.91 | 24.38 | 48.08 | male | 0.99 |
| 2 | 26.67 | 22.27 | 60.74 | male | 3.51 |
| 3 | 29.96 | 25.34 | 50.89 | male | 3.08 |
| 4 | 23.74 | 22.86 | 48.63 | female | -0.05 |
| 5 | 26.17 | 28.58 | 51.96 | female | -1.72 |

(a) Write down a linear regression model for testing whether the differnce of BMI (Y) is associated with the difference of blood cholesterol level (X) within a twin pair. Explain whether the intercept term should be included in the model? State your hypothesis based on the model.

(b) Derive the least squares estimator for the regression coefficient and show that the estimator is unbiased. Derive the variance of the least squares estimator.

(c) Let the residuals be $e_i = Y_i - \widehat{Y}_i$; what is the covariance matrix of $\mathbf{e} = (e_1, e_2, ...e_n)'$? Are the residuals uncorrelated? Show your steps.

(d) Now consider the following multiple linear regression model:

$$E(Y_{ij}) = \beta_0 + \beta_1 \times gender_i + \beta_2 \times age_i,$$

where $Y_{ij}$ is the BMI for twin pair $i, i = 1, 2, 3, ..., 110$, individual $j, j = 1, 2$. Test the hypothesis: $\beta_1 = 0$ using the provided dataset and report the 95% confidence interval for $\beta_1$.

(e) Comment on the assumptions considered in (d) in light of the data.

(f) Describe two alternative approaches to address the issues mentioned in (e). (No need to perform the analyses, just describe)

3. Let $X \sim \text{Poisson}(\lambda)$ and let $Y = \mathbf{1}(X \geq m)$ for some $m \geq 1$. That is, $Y = 1$ if $X \geq m$ and $Y = 0$ otherwise. Consider making inference about $\lambda$ based on $n$ independent realizations $Y_1, \ldots, Y_n$ of $Y$, where $m \geq 1$ is known. Such data would arise if a researcher, after observing a Poisson count $X$, only recorded whether the count met or exceeded the threshold $m$.

(a) Write down the pmf of $Y$.

(b) Give the log-likelihood function for $\lambda$ based on $Y_1, \ldots, Y_n$.

(c) Describe how to obtain the value of the maximum likelihood estimator for $\lambda$ based on $Y_1, \ldots, Y_n$.

(d) Give the Fisher information $I_n(\lambda)$.

(e) Give the form of a Wald-type $(1 - \alpha) \times 100\%$ confidence interval for $\lambda$ based on $Y_1, \ldots, Y_n$.

(f) *Use R and submit all code*: For $n = 100$, suppose you observe $\sum_{i=1}^{100} Y_i = 14$ with $m = 4$.

i. Give the value of the maximum likelihood estimator for $\lambda$ (rounded to two decimal places).

ii. Give the bounds of your Wald-type confidence interval for $\lambda$ with $\alpha = 0.05$.