Stat 111: Homework 6

Daniel de Castro

March 25, 2022

1. (a) The PDF of a Poisson random variable Y_1 is

$$P(Y_1 = y_1|\lambda) = e^{-\lambda}\frac{\lambda^{y_1}}{y_1!}$$

To rewrite this in NEF form, we let

$$\theta = \log\lambda$$
$$\psi(\theta) = e^{\theta}$$
$$h(y) = \frac{1}{y_1!}$$

Thus, in the case of the conditional Poisson variable,

$$\psi(\theta x) = e^{\theta x}$$

(b)

$$f_{Y_j|X_j=x_j}(y_j|\theta) = e^{\theta x_j y_j - \psi(\theta x_j)}h(x_j, y_j)$$

Thus,

$$f_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}|\theta) = \prod_{j=1}^{n} e^{\theta x_j y_j - \psi(\theta x_j)}h(x_j, y_j)$$

$$f_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}|\theta) = e^{\theta\sum_{j=1}^{n} x_j y_j - \sum_{j=1}^{n}\psi(\theta x_j)}\prod_{j=1}^{n} h(x_j, y_j)$$

To get the likelihood, we remove the constant $h(x_j, y_j)$ factors:

$$L(\theta; \boldsymbol{y}, \boldsymbol{x}) = e^{\theta\sum_{j=1}^{n} x_j y_j - \sum_{j=1}^{n}\psi(\theta x_j)}$$

(c) The sufficient statistics for θ are $\sum_{j=1}^{n} x_j y_j$.

(d) The log-likelihood is

$$l(\theta; \boldsymbol{x}, \boldsymbol{y}) = \theta\sum_{j=1}^{n} x_j y_j - \sum_{j=1}^{n}\psi(\theta x_j)$$

The score is

$$s(\theta; \boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{n} x_j y_j - \sum_{j=1}^{n} x_j\frac{\partial\psi(\theta x_j)}{\partial\theta}$$

(e) The derivative of the score with respect to θ is

$$s'(\theta; \boldsymbol{x}, \boldsymbol{y}) = -\sum_{j=1}^{n} x_j^2\frac{\partial^2\psi(\theta x_j)}{\partial\theta^2}$$

We can now find the Fisher information:

$$\mathcal{I}_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\theta) = E[-s'(\theta; \boldsymbol{x}, \boldsymbol{y})]$$

$$\mathcal{I}_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\theta) = E\left[\sum_{j=1}^{n} x_j^2 \frac{\partial^2 \psi(\theta x_j)}{\partial \theta^2}\right]$$

Because by the definition of the NEF $\psi(\theta x_j)$ does not contain any Y_j terms, we have

$$\mathcal{I}_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\theta) = \sum_{j=1}^{n} x_j^2 \frac{\partial^2 \psi(\theta x_j)}{\partial \theta^2}$$

(f) A convex function will have a positive second derivative; a concave function will have a negative second derivative. The second derivative of the log-likelihood function is given by

$$s'(\theta; \boldsymbol{x}, \boldsymbol{y}) = -\sum_{j=1}^{n} x_j^2 \frac{\partial^2 \psi(\theta x_j)}{\partial \theta^2}$$

Note that because we are working with an NEF, $\frac{\partial^2 \psi(\theta x_j)}{\partial \theta^2} = \text{Var}(Y_1 | \boldsymbol{X} = \boldsymbol{x})$. By definition, variance cannot be negative. Thus, $s'(\theta; \boldsymbol{x}, \boldsymbol{y}) \geq 0$, and so $\frac{\partial}{\partial \theta} l(\theta; \boldsymbol{x}, \boldsymbol{y}) \leq 0$. The log-likelihood is concave.

(g) I would advise Joey to compute the MLE of θ using an iterative method, such as the Newton-Raphson method. Because the log-likelihood function is concave, this algorithm will converge with a good initial guess, and Joey could make a good initial guess by inspection of the graph of the log-likelihood function.

2. (a) This is a linear regression, because it is linear in the parameters.

(b) This is not a linear regression, because the second term includes an expression of the form e^{θ}, which is not linear with respect to the parameter θ used.

(c) This is not a linear regression, because the first term is not linear with respect to the parameter θ_0, specifically because of the use of the square root function.

3. (a)

$$\hat{\theta}_n = \frac{1}{\sum_{j=1}^{n} x_j^2} \sum_{j=1}^{n} x_j y_j = \frac{1}{P_n} \sum_{j=1}^{n} x_j y_j$$

$$\hat{\theta}_n = \frac{1}{P_n} \left(\sum_{j=1}^{n-1} x_j y_j + x_n y_n \right) = \frac{\hat{\theta}_{n-1} P_{n-1} + x_n y_n}{P_n}$$

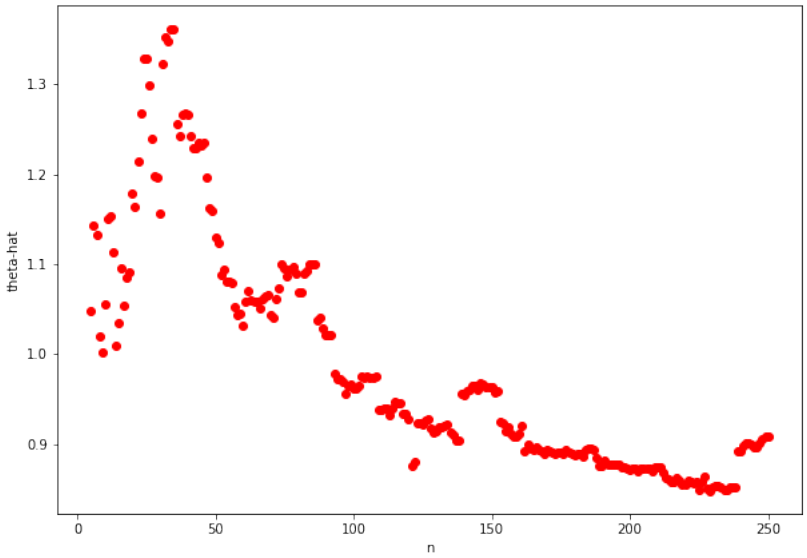
$$\hat{\theta}_n = \frac{\hat{\theta}_{n-1} P_{n-1} + x_n y_n + \hat{\theta}_{n-1} x_n^2 - \hat{\theta}_{n-1} x_n^2}{P_n}$$

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \frac{x_n y_n - \hat{\theta}_{n-1} x_n^2}{P_n}$$

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \frac{x_n v_n}{P_n}$$

(b) Below are the code and plot to answer this question.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("predictRecursive.csv")

theta = np.zeros(len(df.index))
theta[0] = df.iloc[0,1]/df.iloc[0,0]

P_n = df.iloc[0,0]**2
for i in range(1, len(theta)):
    P_n += df.iloc[i,0]**2
    theta[i] = theta[i-1] + df.iloc[i,0] * (df.iloc[i,1] - df.iloc[i,0] *
    ↪  theta[i-1]) / P_n

n = np.arange(5,len(theta) + 1)

plt.figure(figsize=[10,7])
plt.xlabel("n")
plt.ylabel("theta-hat")
plt.plot(n, theta[4:], 'ro')
plt.savefig("3b.png")
plt.show()
```



(c) First, we make a numerical argument. As we progress through the sequence, the variability declines, because compared to $\hat{\theta}_{n-1}$, $\frac{x_n v_n}{P_n}$ will be progressively smaller, especially since P_n will grow larger fairly quickly as n increases.

Now, we make a theoretical argument. First, consider that

$$\hat{\theta}_n = \frac{n^{-1}\sum_{j=1}^{n} X_j Y_j}{n^{-1}\sum_{j=1}^{n} X_j^2}$$

By the Law of Large Numbers,

$$n^{-1}\sum_{j=1}^{n} X_j Y_j \xrightarrow{p} E[XY]$$

and

$$n^{-1}\sum_{j=1}^{n} X_j^2 \xrightarrow{p} E[X^2]$$

Thus, by Slutsky's Theorem,

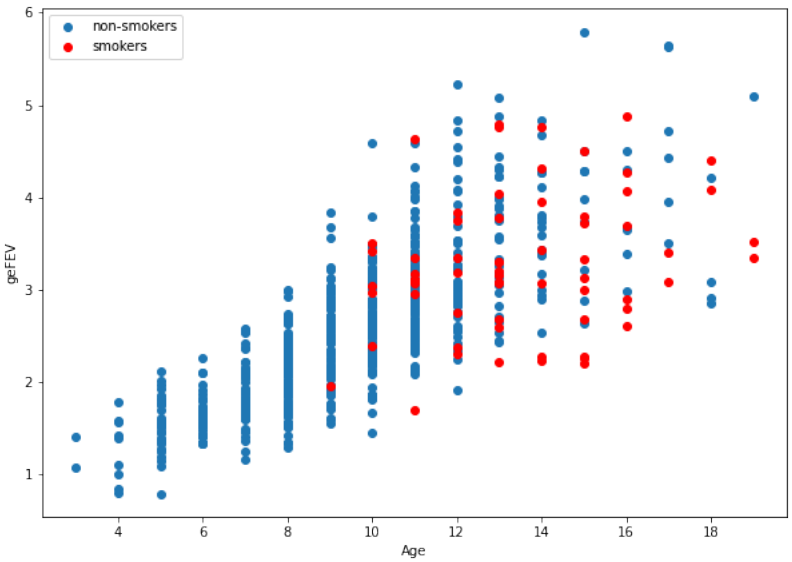
$$\frac{n^{-1}\sum_{j=1}^{n} X_j Y_j}{n^{-1}\sum_{j=1}^{n} X_j^2} \xrightarrow{p} \frac{E[XY]}{E[X^2]}$$

$$\hat{\theta}_n \xrightarrow{p} \frac{E[XY]}{E[X^2]}$$

By the definition of convergence in probability, as $n \to \infty$, $P\left(\hat{\theta}_n - \frac{E[XY]}{E[X^2]}\right) \to 0$. Thus, it follows that as n increases, the variability of $\hat{\theta}_n$ decreases.

4. (a) Below are the code and plot to answer this question.

```
fev = pd.read_csv("lung.csv")

plt.figure(figsize=[10,7])
plt.plot(fev[fev.smoke == 0].age,fev[fev.smoke == 0].geFEV, "o",
↪  label="non-smokers")
plt.plot(fev[fev.smoke == 1].age,fev[fev.smoke == 1].geFEV, "ro",
↪  label="smokers")
plt.legend()
plt.xlabel("Age")
plt.ylabel("geFEV")
plt.savefig("4a.png")
plt.show()
```



(b) i. For the non-smokers, I used the following code to find the sample quantiles.

```
np.quantile(fev[fev.smoke == 0].geFEV, [0.05,0.5,0.95])
```
The 0.05 sample quantile was 1.4246, the 0.5 sample quantile was 2.465, and the 0.95 sample quantile was 4.2292.
For the smokers, I used a similar line of code:
```
np.quantile(fev[fev.smoke == 1].geFEV, [0.05,0.5,0.95])
```
The 0.05 sample quantile was 2.22, the 0.5 sample quantile was 3.169, and the 0.95 sample quantile was 4.7322.

ii. It would appear that the reason that descriptive statistics reveal that the FEV is higher for smokers than non-smokers is because the sample of non-smokers includes many children under the age of ten (a total of 308), whose lung volumes are probably much smaller, while the sample of smokers includes only one individual less than ten years of age. The higher number of younger children in the sample of non-smokers likely drags the sample quantiles down lower than they would be if we excluded these young children from the comparison between the sample of smokers and non-smokers.

(c) i. For the rest of this problem, set X represent age and let Y represent FEV. The following code calculates the quantity $\hat{\beta}_{Y\sim X} = \hat{\theta}$ for the smokers.

```
fev_nm_s = fev[fev.smoke == 1].geFEV - np.mean(fev[fev.smoke == 1].geFEV)
age_nm_s = fev[fev.smoke == 1].age - np.mean(fev[fev.smoke == 1].age)
n_s = fev_nm_s.size
top = 0
bottom = 0
for i in range(n_s):
    top += fev_nm_s.iloc[i] * age_nm_s.iloc[i]
    bottom += age_nm_s.iloc[i] ** 2
theta_s = top / bottom
theta_s
```

For the smokers, we found that

$$\hat{\theta} \approx 0.0799$$

where Y represents FEV and X represents age. To find a 95% confidence interval for $\beta_{Y\sim X} = \theta$, we can use the result from pages 174 and 175 of the Stat 111 book:

$$\hat{\theta} = \frac{n^{-1}\sum_{j=1}^{n} X_j Y_j}{n^{-1}\sum_{j=1}^{n} X_j^2}$$

If we let $U_j = Y_j - \theta X_j$, then

$$\hat{\theta} = \theta + \frac{n^{-1}\sum_{j=1}^{n} X_j U_j}{n^{-1}\sum_{j=1}^{n} X_j^2}$$

By the Central Limit Theorem,

$$n^{1/2}\left(n^{-1}\sum_{j=1}^{n} X_j U_j\right) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(XU))$$

By the Law of Large Numbers,

$$n^{-1}\sum_{j=1}^{n} X_j^2 \xrightarrow{p} E[X^2]$$

So by Slutsky's Theorem,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathrm{Var}(XU)}{E[X^2]^2}\right)$$

We can use the method of moments to estimate the variance with the following expression:

$$\text{AsyVar} = \frac{n^{-1}\sum_{j=1}^{n} X_j^2 (Y_j - \hat{\theta} X_j)^2}{\left(n^{-1}\sum_{j=1}^{n} X_j^2\right)^2}$$

The 95% confidence interval is thus given by

$$C(\boldsymbol{X}, \boldsymbol{Y}) = \left[\hat{\theta} - 1.96\frac{\sqrt{\text{AsyVar}}}{\sqrt{n}}, \hat{\theta} + 1.96\frac{\sqrt{\text{AsyVar}}}{\sqrt{n}}\right]$$

We use the following code to calculate this confidence interval:

```
num = 0
for i in range(n_s):
    num += age_nm_s.iloc[i] ** 2 * (fev_nm_s.iloc[i] - theta_s *
    ↪  age_nm_s.iloc[i]) ** 2
AsyVar_s = (num / n_s) / (bottom / n_s) ** 2
(theta_s - 1.96 * np.sqrt(AsyVar_s) / np.sqrt(n_s), theta_s + 1.96 *
↪  np.sqrt(AsyVar_s) / np.sqrt(n_s))
```

The confidence interval given by the above code was

$$C(\boldsymbol{X}, \boldsymbol{Y}). = [0.0157, 0.144]$$

ii. The following code calculates the quantity $\hat{\beta}_{Y\sim X} = \hat{\theta}$ for the non-smokers.

```
fev_nm_c = fev[fev.smoke == 0].geFEV - np.mean(fev[fev.smoke == 0].geFEV)
age_nm_c = fev[fev.smoke == 0].age - np.mean(fev[fev.smoke == 0].age)
n_c = fev_nm_c.size
top = 0
bottom = 0
for i in range(n_c):
    top += fev_nm_c.iloc[i] * age_nm_c.iloc[i]
    bottom += age_nm_c.iloc[i] ** 2
theta_c = top / bottom
theta_c
```

For the non-smokers, we found that

$$\hat{\theta} \approx 0.243$$

where Y represents FEV and X represents age. To find a 95% confidence interval for $\beta_{Y\sim X} = \theta$, we can use the result from pages 174 and 175 of the Stat 111 book as before:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\text{Var}(XU)}{E[X^2]^2}\right)$$

We again estimate the variance of this distribution using the following expression:

$$\text{AsyVar} = \frac{n^{-1}\sum_{j=1}^{n} X_j^2(Y_j - \hat{\theta}X_j)^2}{\left(n^{-1}\sum_{j=1}^{n} X_j^2\right)^2}$$

This gives the 95% confidence interval

$$C(\boldsymbol{X}, \boldsymbol{Y}) = \left[\hat{\theta} - 1.96\frac{\sqrt{\text{AsyVar}}}{\sqrt{n}}, \hat{\theta} + 1.96\frac{\sqrt{\text{AsyVar}}}{\sqrt{n}}\right]$$

We used similar code to calculate this result:

```
num = 0
for i in range(n_c):
    num += age_nm_c.iloc[i] ** 2 * (fev_nm_c.iloc[i] - theta_c *
    ↪  age_nm_c.iloc[i]) ** 2
AsyVar_c = (num / n_c) / (bottom / n_c) ** 2
(theta_c - 1.96 * np.sqrt(AsyVar_c) / np.sqrt(n_c), theta_c + 1.96 *
↪  np.sqrt(AsyVar_c) / np.sqrt(n_c))
```

The confidence interval given by the code was

$$C(\boldsymbol{X}, \boldsymbol{Y}). = [0.223, 0.262]$$

iii. From parts (i) and (ii) we have

$$\sqrt{n}(\hat{\theta}_s - \theta_s) \xrightarrow{d} \mathcal{N}\left(0, \frac{\text{Var}(X_s U_s)}{E[X_s^2]^2}\right)$$

and

$$\sqrt{n}(\hat{\theta}_c - \theta_c) \xrightarrow{d} \mathcal{N}\left(0, \frac{\text{Var}(X_c U_c)}{E[X_c^2]^2}\right)$$

where θ_s corresponds to the smokers and θ_c corresponds to the non-smokers. Thus, we have

$$\hat{\theta}_s \xrightarrow{d} \mathcal{N}\left(\theta_s, \frac{\text{Var}(X_s U_s)}{n_s E[X_s^2]^2}\right)$$

and

$$\hat{\theta}_c \xrightarrow{d} \mathcal{N}\left(\theta_c, \frac{\text{Var}(X_c U_c)}{n_n E[X_c^2]^2}\right)$$

Assuming that the individuals in the smoker and non-smoker populations are independent,

$$\hat{\theta}_s - \hat{\theta}_c \xrightarrow{d} \mathcal{N}\left(\theta_s - \theta_c, \frac{\text{Var}(X_s U_s)}{n_s E[X_s^2]^2} + \frac{\text{Var}(X_c U_c)}{n_c E[X_c^2]^2}\right)$$

Similarly to what we did in parts (i) and (ii), we can estimate the variance of the above distribution using the following method of moments estimator:

$$\text{AsyVar} = \frac{n_s^{-1} \sum_{j=1_s}^{n_s} X_j^2 (Y_j - \hat{\theta}_s X_j)^2}{\left(n_s^{-1} \sum_{j=1_s}^{n_s} X_j^2\right)^2} + \frac{n_c^{-1} \sum_{j=1_c}^{n_c} X_j^2 (Y_j - \hat{\theta}_c X_j)^2}{\left(n_c^{-1} \sum_{j=1_c}^{n_c} X_j^2\right)^2}$$

Note that the two terms of this AsyVar have already been calculated in parts (i) and (ii). This yields the following confidence interval:

$$C(\boldsymbol{X_s}, \boldsymbol{X_c}, \boldsymbol{Y_s}, \boldsymbol{Y_c}) = \left[\left(\hat{\theta}_s - \hat{\theta}_c\right) - 1.96\sqrt{\text{AsyVar}}, \left(\hat{\theta}_s - \hat{\theta}_c\right) + 1.96\sqrt{\text{AsyVar}}\right]$$

We calculate this interval using the following code:

```
AsyVar = (1/n_s) * AsyVar_s + (1/n_c) * AsyVar_c
((theta_s - theta_n) - 1.96 * np.sqrt(AsyVar), (theta_s - theta_n) + 1.96 *
↪  np.sqrt(AsyVar))
```

This code returned the interval $[-0.230, -0.0957]$. The calculated difference $\hat{\theta}_s - \hat{\theta}_c$ was -0.163.