Midterm CSC121

(TRUE)
a) Yes, this is because the regression line fits well

(FALSE)
b) No, this is because the regression line is not near $y=x$, instead it is around $y=20x$, thus, sample B is around 20x of sample A.

c) FALSE

d) FALSE

e) TRUE

f) TRUE

Problem 2

a) $x_{ij} \sim \text{Poisson}(s_i, \beta_j \tilde{I}_i)$

Likelihood = $\prod_{i=1}^{6}$ (PMF of poisson distribution)

$$= \prod_{i=1}^{6} \left(\frac{e^{-s_i \beta_j \tilde{I}_i} \times (s_i \beta_j \tilde{I}_i)^{x_{ij}}}{x_{ij}!} \right)$$

MLE : Take log of the value above

$$\log(\mathcal{L}_n) = \cancel{\log} \sum_{i=1}^{6} -s_i \beta_j \tilde{I}_i + x_{ij} \log(s_i \beta_j \tilde{I}) - \log(x_{ij}!)$$

taking derivative w.r.t β_j

$$\frac{\partial(\log \mathcal{L}_n)}{\partial \cancel{\log} \beta_j} = \sum_{i=1}^{6} -s_i \tilde{I}_j + \frac{x_{ij}}{\beta_j}$$

derivative = 0

$$\sum_{i=1}^{6} -s_i \tilde{I}_j = \sum_{i=1}^{6} \frac{x_{ij}}{\beta_j}$$

$$\beta_j = \frac{\sum_{i=1}^{6} x_{ij}}{\tilde{I}_j \sum_{i=1}^{6} - s_i}$$

double derivative = $\frac{-x_{ij}}{\beta_j}$, which is always going to be negative as all variables are positive, thus ~~minimizing~~ maximizing likelihood.

b) This accounts for the avg of the raw counts from sample 1-6. S_i and $\hat{S}_i$ are normalization factors ~~c~~. This would give a count for B_j, divided by 6 to ensure the # samples are considered

c) $$X_{ij} = \text{Poisson}(S_i B_j \tilde{I}_{ij})$$

$$\text{Likelihood} = \prod_{i=1}^{6} (\text{PMF of poisson distribution})$$

$$= \prod_{i=1}^{6} \left(\frac{e^{-S_i B_j \tilde{I}_{ij}} \times (S_i B_j \tilde{I}_{ij})^{X_{ij}}}{X_{ij}!} \right)$$

MLE: Take log of the value above

$$\log(\mathcal{L}_n) = \sum_{i=1}^{6} - S_i B_j \tilde{I}_{ij} + X_{ij} \log(S_i B_j \tilde{I}) - \log(X_{ij}!)$$

taking derivative w.r.t B_j

$$\frac{\partial(\log \mathcal{L}_n)}{\partial B_j} = \sum_{i=1}^{6} - S_i \tilde{I}_{ij} + \frac{X_{ij}}{B_j}$$

derivative $= 0$

$$\sum_{i=1}^{6} - S_i \tilde{I}_{ij} = \sum_{i=1}^{6} \frac{X_{ij}}{B_j}$$

$$B_j = \frac{\sum_{i=1}^{6} X_{ij}}{\sum_{i=1}^{6} (-S_i \tilde{I}_{ij})}$$ → Terms need to be in the bracket.

double derivative is also negative similar to a)

d) Now the bottom term does a summation over the lengths. Thus, it now calculates the ~~the~~ denominator, where for each sample, it is ~~s~~ multiplied by the length. Thus, it now accounts for the ~~its~~ ~~length~~ effective length of each sample.

$$\sum_{i=1}^{6} -S_i \tilde{I}_{ij} \neq \left(\sum_{i=1}^{6} -S_i \right) \tilde{I}_j$$

Problem 3

$N = 100$

a) 100 reads $\quad$ ~~30~~ $N_B = 30 \quad N_A = 70$

$$P(AB \mid O_1, O_2, O_3, \dots O_{100}) = \frac{P(O_1, O_2, O_3 \dots O_{100} \mid P(AB))\, P(AB}{P(O_1, O_2, O_3 \dots O_{100})}$$

$O_i = A \quad E_i = 0$

$$P(\underbrace{O_i = A \mid E_i = 0, c_i = A, G = AB}_{1})\,\underbrace{P(E_i = 0)}_{1-\gamma}\,\underbrace{P(c_i = A \mid G = AB)}_{0.5}$$

$$P(\underbrace{O_i = A \mid E_i = 0, c_i = B, G = AB}_{0})\,\underbrace{P(E_i = 0)}_{1-\gamma}\,\underbrace{P(c_i = B \mid G = AB)}_{0.5}$$

$O_i = B \quad E_i = 0$

$$P(\underbrace{O_i = B \mid E_i = 0, c_i = A, G = AB}_{0})\,\underbrace{P(E_i = 0)}_{1-\gamma}\,\underbrace{P(c_i = A \mid G = AB)}_{0.5}$$

$$P(\underbrace{O_i = B \mid E_i = 0, c_i = B, G = AB}_{1})\,\underbrace{P(E_i = 0)}_{1-\gamma}\,\underbrace{P(c_i = B \mid G = AB)}_{0.5}$$

$O_i = A \quad E_i = 1$

$$P(\underbrace{O_i = A \mid E_i = 1, c_i = A, G = AB}_{0})\,\underbrace{P(E_i = 1)}_{\gamma}\,\underbrace{P(c_i = A \mid G = AB)}_{0.5}$$

$$P(\underbrace{O_i = A \mid E_i = 1, c_i = B, G = AB}_{1})\,\underbrace{P(E_i = 1)}_{\gamma}\,\underbrace{P(c_i = B \mid G = AB)}_{0.5}$$

$O_i = B \quad E_i = 1$

$$P(\underbrace{O_i = B \mid E_i = 1, c_i = A, G = AB}_{1})\,\underbrace{P(E_i = 1)}_{\gamma}\,\underbrace{P(c_i = A \mid G = AB)}_{0.5}$$

$$P(\underbrace{O_i = B \mid E_i = 1, c_i = B, G = AB}_{0})\,\underbrace{P(E_i = 1)}_{\gamma}\,\underbrace{P(c_i = B \mid G = AB)}_{0.5}$$

Total = 1

$P(O_1, O_2 \ldots O_{100} | AB) = \left(\frac{1}{2}\right)^N$

$P(O_1, O_2 \ldots O_{100} | AA) = \underline{(1-\gamma)^{N_A} (\gamma)^{N-N_A}}$

$\hookrightarrow \prod_{i:O_i=A} P(O_i = A | AA) \prod_{i:O_i=B} P(O_i = B | AA)$

Similarly

$P(O_1, O_2 \ldots O_{100} | BB) = (\gamma)^{N_A} (1-\gamma)^{N-N_A}$

$$\boxed{\text{Total Probs} = \frac{\left(\frac{1}{2}\right)^{100} \cdot \frac{1}{4}}{\frac{1}{4} \cdot \left(\frac{1}{2}\right)^{100} + \frac{(1-\gamma)^{30} \gamma^{(70)}}{2} + \frac{(1-\gamma)^{70} \gamma^{30}}{4}}}$$

b) $P(C_i = A | G = AB) = S$

$P(C_i = B | G = AB) = 1 - S$

$P(O_i = B | C_i = A, G = AB) = \gamma$

$P(O_i = B | C_i = B, G = AB) = 1 - \gamma$

$P(O_i = B | G = AB) = P(O_i = B | G = AB)$

$= S \times \gamma + (1-S)(1-\gamma)$

$= 1 - S - \gamma + 2\gamma S$

Expected Count = $100 \times (1 - S - \gamma + 2\gamma S)$

rounded up/down to nearest read.

c) $P(AB \mid O_1 \ldots\ldots O_N) \geq \alpha$

$P(AB \mid O_1 O_2 \ldots O_N)$

$$= \frac{\left(\frac{1}{2}\right)^N \overbrace{P(AB)}^{1/4}}{\frac{\gamma^{N_A}(1-\gamma)^{N_B}}{4} + \left(\frac{1}{2}\right)^N \cdot \frac{1}{4} + \frac{\gamma^{N_B}(1-\gamma)^{N_A}}{2}} \geq \alpha$$

$N_A = N_B = N/2$

$$\frac{\cancel{4} \times \frac{1}{\cancel{4}}\left(\frac{1}{2}\right)^N}{\cancel{4} \times \frac{\gamma^{N/2}(1-\gamma)^{N/2}}{\cancel{4}} + \left(\frac{1}{2}\right)^N \cdot \frac{1 \times \cancel{4}}{\cancel{4}} + \frac{\gamma^{N/2}(1-\gamma)^{N/2}}{\cancel{2}} \times \cancel{4}2} \geq \alpha$$

$\cancel{\frac{1}{4}\left(\frac{1}{2}\right)^N}$ $(0.5)^N \geq 3\gamma^{N/2}(1-\gamma)^{N/2}\alpha + \left(\frac{1}{2}\right)^N \alpha$

$(0.5)^N (1-\alpha) \geq 3\gamma^{N/2}(1-\gamma)^{N/2}\alpha$

$\underline{Z}\,(0.5)^N - 3(\gamma^{N/2})(1-\gamma)^{N/2} - 1 \geq 0$

solving for inequality, we get

$\alpha\alpha_N$