Q1

a.) $\hat{\beta}^R = (X^TX + \lambda I)^{-1} X^T y$

$(X^TX + \lambda I)^{-1} = (X^TX \cdot (I + \lambda (X^TX)^{-1}))^{-1} = (I + \lambda (X^TX)^{-1})^{-1} (X^TX)^{-1}$

$\Rightarrow \hat{\beta}^R = (I + \lambda (X^TX)^{-1})^{-1} (X^TX)^{-1} X^T y = (I + \lambda (X^TX)^{-1})^{-1} \hat{\beta}^{RSS}$

$E[\hat{\beta}^R] = E[(I + \lambda (X^TX)^{-1})^{-1} \hat{\beta}^{RSS}] = (I + \lambda (X^TX)^{-1})^{-1} \beta^{true}$

$E[\hat{\beta}^R] - \beta^{true} = \beta^{true} \left[(I + \lambda (X^TX)^{-1})^{-1} - 1 \right]$ • $\lambda \to \infty$ this quantity converges to $-\beta^{true}$

b.) Covariance matrix is defined by ; $C_{\beta OLS} = E[(\hat{\beta}^e - \beta^{true})(\hat{\beta}^e - \beta^{true})^T]$

$\hat{\beta}^{OLS} = (X^TX)^{-1} X^T y$ if $X^TX = I_p$ then $\hat{\beta}^{OLS} = X^T y$

$\hat{\beta}^{OLS} = X^T y = X^T (X \beta^{true} + \varepsilon_i) = \overbrace{(X^T X)}^{I_p} \beta^{true} + X^T \varepsilon_i = X^T \varepsilon_i + \beta^{true}$

$C_{\beta OLS} = E[(X^T\varepsilon_i)(X^T\varepsilon_i)^T] = E[(X^T\varepsilon_i)(\varepsilon_i^T X)] = E[\varepsilon_i \varepsilon_i^T] = Cov(\varepsilon_i, \varepsilon_i^T) = \sigma^2 I$

c.) $\hat{\beta}^R = (X^TX + \lambda I)^{-1} X^T y$ if $X^TX = I_p$ then $\hat{\beta}^R = X^T y \cdot \frac{1}{\lambda+1} = \frac{1}{\lambda+1} \left[\overbrace{X^T X}^{I_p} \beta^{true} + X^T \varepsilon_i \right]$

$\hat{\beta}^R - \beta^{true} = \frac{-\lambda}{\lambda+1} \beta^{true} + \frac{1}{\lambda+1} (X^T \varepsilon_i)$

$E\left[\left(\frac{-\lambda}{\lambda+1} \beta^{true} + \frac{1}{\lambda+1} (X^T\varepsilon_i) \right) \left(\frac{-\lambda}{\lambda+1} \beta^{true} + \frac{1}{\lambda+1} (X^T\varepsilon_i) \right)^T \right] = E\left[\left(\frac{-\lambda}{\lambda+1} \beta^{true} + \frac{1}{\lambda+1} (X^T\varepsilon_i) \right) \left(\varepsilon_i^T X \cdot \frac{1}{\lambda+1} + \beta^{true} \frac{-\lambda}{\lambda+1} \right) \right]$

$= \frac{\lambda^2}{(\lambda+1)^2} \beta^{2}_{true} + 0 + 0 + E\left[\frac{1}{(\lambda+1)^2} X^T \varepsilon_i \varepsilon_i^T X \right] = \frac{1}{(\lambda+1)^2} \underbrace{E[\varepsilon_i \varepsilon_i^T]}_{\sigma^2 I} + \frac{\lambda^2}{(\lambda+1)^2} \beta^2_{true}$

$C_{\beta R} = \frac{1}{(\lambda+1)^2} \left[\lambda^2 \beta^2_{true} + \sigma^2 \right] I$

d.) We expect variance of covariance of Ridge to be less than OLS. Trace is related with element at diagonal

$\sigma^2 \geq \frac{1}{(\lambda+1)^2} (\lambda^2 \beta^{2}_{true} + \sigma^2)$

Q2

a.) $L(\beta) = \prod_{y_i=1} \pi(x_i) \prod_{y_i=0} (1-\pi(x_i))$ where $\pi(x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$

$$I(\beta) = -\ln L(\beta) = -\sum_{y_i=1} \ln \pi(x_i) + \sum_{y_i=0} \ln(1-\pi(x_i)) = -\sum_{i=1}^{n} y_i \beta^T x_i - \ln(1+e^{\beta^T x_i})$$

$$\frac{\partial I(\beta)}{\partial \beta} = -\sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n} x_i \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}} = -x^T y + x^T \pi$$

$$\beta^{new} = \beta^{old} + \varphi(x^T(\pi - y))$$

b.)

c.) $$P(y=1 \mid X=x) = \frac{e^{-0.0130 - 1.5115x}}{1+e^{-0.013 - 1.5115x}}$$

$$P(y=0 \mid X=x) = \frac{1}{1+e^{-0.013 - 1.5115x}}$$

*(Stopped at iteration 27)

Q3

a.) After plotting error vs p plot, we cold see that after $p=5$, error converges to some fixed point. To make model simpler, the optimal p becomes 5

b.) As λ gets smaller, the error decreases. In the experiment I used λ values $\in [0, 100]$
In the experiment increasing λ caused error to be bigger therefore the optimal λ value is close to 0. we can select $\lambda = 1$

c.) Same as part b. Error gets lower if λ gets lower. Therefore we can select $\lambda = 1$

d.) In the Ridge, I fond $\beta = [0, 0.0238, 0.0256, -0.3717, -0.0247, 0.8421]$
for $p=5$ which is found to be optimal. (with $\lambda = 1$)

In the Lasso, I fond $\beta = [0, 0.0447, 0.0218, -0.4326, -0.0211, 0.8898]$
for $p=5$ which is found to be optimal (with $\lambda = 1$)

Q4

a.) $L(\beta) = \prod_{i=1}^{n} p(y_i | x_i, \beta) = \prod_{i=1}^{n} \frac{\mu^{y_i}}{y_i!} e^{-\mu}$

$-I(\beta) = \ln L(\beta) = -\sum_{i=1}^{n} (y_i \ln\mu - \ln y_i! - \mu) = -\sum_{i=1}^{n} (y_i \beta^T x_i - \ln y_i!) - e^{\beta^T x_i})$

$$\frac{\partial I(\beta)}{\partial \beta} = -\sum_{i=1}^{n} \left(y_i x_i^T - e^{\beta^T x_i} x_i^T\right) = -\left[X^T y - X^T \hat{\mu}_i\right] \text{ where } \mu_i = e^{\beta^T x_i}$$

$$\beta^{new} = \beta^{old} + \varphi(y_i - \mu_i) x_i$$

b.) MLE estimates that I found are

$\hat{\beta}_0 = 1.1889$

$\hat{\beta}_1 = 1.2584$