Question 1

(a) *Solution:*

$$\nabla f(\beta) = \begin{pmatrix} \sum_{i=1}^{n} \left( y_i - \frac{\exp(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik})}{1+\exp(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik})} \right) \\ \sum_{i=1}^{n} \left( x_{i1} y_i - \frac{x_{i1}\exp(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik})}{1+\exp(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik})} \right) - 2\lambda\beta_1 \\ \vdots \\ \sum_{i=1}^{n} \left( x_{ik} y_i - \frac{x_{ik}\exp(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik})}{1+\exp(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik})} \right) - 2\lambda\beta_k \end{pmatrix}$$

*Supporting Work:*

$$f(\beta) = -\sum_{i=1}^{n} \log\left(1+\exp(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik})\right) + \sum_{i=1}^{n} y_i\left(\beta_0+\beta_1 x_{i}+\dots+\beta_k x_{ik}\right) - \lambda \sum_{i=1}^{k} \beta_i^2$$

$$\frac{\partial f}{\partial \beta_0} = -\sum_{i=1}^{n} \left(1 \cdot \exp(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik}) \cdot \frac{1}{1+\exp(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik})}\right) + \sum_{i=1}^{n} y_i \cancel{- 2\lambda\beta_0}$$

$$\boxed{\frac{\partial f}{\partial \beta_0} = \sum_{i=1}^{n} \left( y_i - \frac{\exp(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik})}{1+\exp(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik})} \right) \cancel{2\lambda\beta_0}}$$

$$\frac{\partial f}{\partial \beta_k} = -\sum_{i=1}^{n} \left(x_{ik} \cdot \exp(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik}) \cdot \frac{1}{1+\exp(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik})}\right) + \sum_{i=1}^{n} x_{ik} y_i - 2\lambda\beta_0$$

$$\boxed{\frac{\partial f}{\partial \beta_k} = \sum_{i=1}^{n} \left( x_{ik} y_i - \frac{x_{ik}\exp(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik})}{1+\exp(\beta_0+\beta_1 x_{i1}+\dots+\beta_k x_{ik})} \right) - 2\lambda\beta_k}$$

(b) *R Code:*

```
#clear environment
rm(list=ls())

#read data and set up data sets
ridge_data <- read.csv("logit_ridge.csv", header = FALSE)
colnames(ridge_data) <-
c("y","x1","x2","x3","x4","x5","x6","x7","x8","x9","x10","x11","x12","x13","x14","x15","x
16","x17","x18","x19","x20")
ridge_data$y <- as.numeric(ridge_data$y)
ridge_test <- ridge_data[1:10,]
ridge_train <- ridge_data[11:nrow(ridge_data),]

#initialize gradient ascent elements
#note – had to set eps to 0.0005 bc it took too long to run at 0.0001 and the hw did not
specify an error tolerance, only an alpha
alpha <- 10^-4
b_last <- matrix(ncol=21,nrow=1,data=0)
colnames(b_last) <-
c(“b0”,”b1”,”b2”,”b3”,”b4”,”b5”,”b6”,”b7”,”b8”,”b9”,”b10”,”b11”,”b12”,”b13”,”b14”,”b
15”,”b16”,”b17”,”b18”,”b19”,”b20”)
b <- b_last
b_history <- b_last
err <- 100
eps <- 0.0005
lambda <- 1
grad <- matrix(data=0,ncol=1,nrow=21)
y_train <- as.matrix(as.numeric(ridge_train[,1]))
x_train <- as.matrix(ridge_train[,2:21])

#calculation of grad descent
while(err > eps) {
	e <-
	exp(b_last[,1]+b_last[,2]*x_train[,1]+b_last[,3]*x_train[,2]+b_last[,4]*x_train[,3]
	+b_last[,5]*x_train[,4]+b_last[,6]*x_train[,5]+b_last[,7]*x_train[,6]+b_last[,8]*x_
	train[,7]+b_last[,9]*x_train[,8]+b_last[,10]*x_train[,9]+b_last[,11]*x_train[,10]+b
	_last[,12]*x_train[,11]+b_last[,13]*x_train[,12]+b_last[,14]*x_train[,13]+b_last[,
	15]*x_train[,14]+b_last[,16]*x_train[,15]+b_last[,17]*x_train[,16]+b_last[,18]*x_
	train[,17]+b_last[,19]*x_train[,18]+b_last[,20]*x_train[,19]+b_last[,21]*x_train[,
	20])

	grad[1,] <- sum(y_train-e/(1+e))

	for(j in 2:21){
```

```
                grad[j,] <- sum(x_train[j-1]*(y_train-e)/(1+e))-2*lambda*b_last[,j]
        }
        b = b_last + alpha*t(grad)
        err = norm(b – b_last, type = "2")
        b_last <- b
        b_history <- rbind(b_history, b_last)
}

b
```

*Output:*

```
> b
           b0        b1         b2         b3        b4         b5        b6
[1,] 17.30242 -1.606982 -0.2795531 -0.8849855 -1.44509 -1.460163 0.357651
            b7        b8         b9         b10        b11        b12       b13
[1,] -2.367504 -2.66819 -1.585938 0.01360522 0.3468647 -2.999609 -1.81987
             b14          b15         b16        b17        b18       b19         b20
[1,] -0.06484308 4.139413e-05 -0.7166331 -2.180653 -1.741077 -1.27292 -0.4380196
```

*Solution:*

Max likelihood estimates of b1 and b2 are -1.607 and -0.280 respectively.

(c) *R Code:*

```
#calculate prediction error
y_test <- as.matrix(as.numeric(ridge_test[,1]))
x_test <- as.matrix(ridge_test[,2:21])
y_prob <- matrix(ncol=1,nrow=10,data=0)
for(I in 1:10){
   y_prob[i] <- exp(b[,1] + x_test[I,]%*%b[,2:21])/(1+exp(b[,1] + x_test[I,]%*%b[,2:21]))
}
pred_err <- (y_test-y_prob)^2
mean(pred_err)
```

*Output:*

```
> mean(pred_err)
[1] 0.4835571
```

(not great I know but I could not improve my algorithm no matter how much debugging I did, switching between matrix notation and non-matrix, changing lambdas and tolerances… best I could do…)

Question 2

(a) $$Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$$

$$p(y; X, \beta) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} \varepsilon_i^2\right)$$

(b) $$p(\beta) = \prod_{j=1}^{p} \frac{1}{2\tau} e^{-|\beta_j|/\tau}$$

$$p(\beta | X, Y) \propto \mathcal{L}(Y | X, \beta)\, p(\beta | X) = \mathcal{L}(Y | X, \beta)\, p(\beta)$$

$$\mathcal{L}(Y | X, \beta)\, p(\beta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} \varepsilon_i^2\right)\left[\frac{1}{2\tau}\exp\left(\frac{-|\beta|}{\tau}\right)\right]$$

(c) $$\mathcal{L}(Y | X, \beta)\, p(\beta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \left(\frac{1}{2\tau}\right) \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} \varepsilon_i^2 - \frac{|\beta|}{\tau}\right)$$

$$\underset{\beta}{\text{maximize}} \left\{ \log\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \left(\frac{1}{2\tau}\right)\right] - \left(\frac{1}{2\sigma^2}\sum_{i=1}^{n} \varepsilon_i^2 + \frac{|\beta|}{b}\right)\right\}$$

$$\downarrow$$

$$\underset{\beta}{\text{minimize}} \left(\frac{1}{2\sigma^2}\sum_{i=1}^{n} \varepsilon_i^2 + \frac{|\beta|}{\tau}\right) = \min_{\beta} \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} \varepsilon_i^2 + \frac{2\sigma^2}{\tau}\sum_{j=1}^{p} |\beta_j|\right)$$

$$= \min_{\beta} \left(\sum_{i=1}^{n} \varepsilon_i^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right)$$

$$= \min_{\beta} \left(RSS + \lambda \sum_{j=1}^{p} |\beta_j|\right) \leftarrow \text{Lasso}$$