Regularization and logistic regression

Question 1

Suppose we fit “Lasso Regression” to a data set, which has 100 features (X1,X2…X100). Now, we rescale one of these feature by multiplying with 10 (say that feature is X1), and then refit Lasso regression with the same regularization parameter.

Now, which of the following option will be correct?
A. It is more likely for X1 to be excluded from the model
B. It is more likely for X1 to be included in the model
C. Can’t say
D. None of these

Solution: B

Big feature values =⇒ smaller coefficients =⇒ less lasso penalty =⇒ more likely to have be kept

Question 2

Suppose you have fitted a multiple regression model on a dataset. Now, you are using Ridge regression with tuning parameter lambda to reduce its complexity. Choose the options below which describes relationship of bias and variance with lambda.

A. In case of very large lambda; bias is low, variance is low
B. In case of very large lambda; bias is low, variance is high
C. In case of very large lambda; bias is high, variance is low
D. In case of very large lambda; bias is high, variance is high

Solution: C
If lambda is very large it means model is less complex. Also remember adding regularization will cause bias in the solution, but tradeoff is smaller variance. So in this case bias is high and variance in low.

Question 3

Write a function to realize gradient descent in R. Understand how learning rate affects convergence.

```
num.iterations <- 1000

# Download South African heart disease data
loan <- read.csv("loan.csv", stringsAsFactors = FALSE)
loan$annual_inc <- log(loan$annual_inc)
x <- loan[,c("annual_inc", "dti")]
y <- loan$loan_status_binary

# Function to standardize input values
zscore <- function(x, mean.val=NA) {
     if(is.matrix(x)) return(apply(x, 2, zscore, mean.val=mean.val))
     if(is.data.frame(x)) return(data.frame(apply(x, 2, zscore,
mean.val=mean.val)))
     if(is.na(mean.val)) mean.val <- mean(x)
     sd.val <- sd(x)
     if(all(sd.val == 0)) return(x) # if all the values are the same
     (x - mean.val) / sd.val
}

# Standardize the features or use scale function directly
x.scaled <- zscore(x)

# Gradient descent function
grad <- function(x, y, theta) {
  gradient <- (1 / nrow(y)) * (t(x) %*% (1/(1 + exp(-x %*% t(theta)))
- y))
  return(t(gradient))
}

gradient.descent <- function(x, y, alpha=0.1, num.iterations=500,
threshold=1e-5, output.path=FALSE) {

    # Add x_0 = 1 as the first column
    m <- if(is.vector(x)) length(x) else nrow(x)
    if(is.vector(x) || (!all(x[,1] == 1))) x <- cbind(rep(1, m), x)
    if(is.vector(y)) y <- matrix(y)
    x <- apply(x, 2, as.numeric)

    num.features <- ncol(x)

    # Initialize the parameters
    theta <- matrix(rep(0, num.features), nrow=1)
```

```
    # Look at the values over each iteration
    theta.path <- theta
    for (i in 1:num.iterations) {
      theta <- theta - alpha * grad(x, y, theta)
      if(all(is.na(theta))) break
      theta.path <- rbind(theta.path, theta)
      if(i > 2) if(all(abs(theta - theta.path[i-1,]) < threshold))
break
    }

    if(output.path) return(theta.path) else
return(theta.path[nrow(theta.path),])
}

unscaled.theta <- gradient.descent(x=x, y=y,
num.iterations=num.iterations, output.path=TRUE)
scaled.theta <- gradient.descent(x=x.scaled, y=y,
num.iterations=num.iterations, output.path=TRUE)

summary(glm( loan_status_binary ~ annual_inc + dti, family =
binomial, data=loan))
library(ggplot2)
qplot(1:(nrow(scaled.theta)), scaled.theta[,1], geom=c("line"),
xlab="iteration", ylab="theta_1")
qplot(1:(nrow(scaled.theta)), scaled.theta[,2], geom=c("line"),
xlab="iteration", ylab="theta_2")

# Look at output for various different alpha values
vary.alpha <- lapply(c(1e-12, 1e-9, 1e-7, 1e-3, 0.1, 0.9),
function(alpha) gradient.descent(x=x.scaled, y=y, alpha=alpha,
num.iterations=num.iterations, output.path=TRUE))

par(mfrow = c(2, 3))
for (j in 1:6) {
     plot(vary.alpha[[j]][,2], ylab="area (alpha=1e-9)",
xlab="iteration", type="l")
}
```

Question 4

A five year follow-up study on 600 disease free subjects was carried out to assess the effect of whether having exposure E or not (of smoking for example) on the development (or not) of a certain disease. The variables AGE (continuous) and obesity status (boolean), which were determined at the start of the follow-up and were to be considered as control variables in analyzing the data.

(1) State the logit form of a logistic regression model that assesses the effect of the 0/1 exposure variable E controlling for the confounding effects of AGE and OBS and the interaction effects of AGE with E and OBS with E.
(2) Given above model you have, give a formula for the odds ratio for the exposure-disease relationship that controls for the confounding and interactive effects of AGE and OBS.
(3) Now use the formula from above to write an expression for the estimated odds ratio for the exposure-disease relationship when AGE=40 and OBS=1.

Predictor	Value of Predictor for Person who is Exposed	Value of Predictor for Person who is Not Exposed
E	1	0
AGE	AGE_1	AGE_0
OBS	OBS_1	OBS_0
AGEE	AGE_1	0
OBSE	OBS_1	0

Solution:
(1) $\text{logit}[\pi] = \beta_0 + \beta_1 *E + \beta_2 *AGE + \beta_3 *OBS + \beta_4 *AGE \cdot E + \beta_5 *OBS \cdot E$
π = Probability of disease,
$AGE \cdot E$ = AGE * E. This is a created variable that is the interaction of AGE with E
$OBS \cdot E$ = OBS * E Similarly, this is the interaction of OBS with E.

(2) OR = exp { logit[π for exposed person] - logit[π for NON exposed person] }
= exp { [$\beta_0 + \beta_1 *E + \beta_2 *AGE_1 + \beta_3 *OBS_1 + \beta_4 *AGE_1.E + \beta_5 *OBS_1.E$] - [$\beta_0 + \beta_2 * AGE_0 + \beta_3 *OBS_0$] } = exp { $\beta_1 + \beta_2 *(AGE_1 - AGE_0) + \beta_3 *(OBS_1 - OBS_0) + \beta_4 *AGE_1.E + \beta_5 * OBS_1.E$ }
(3) OR = exp { $\beta_1 + (40)\beta_4 + \beta_5$ }

Question 5

Build the best logistic regression model to predict loan will be default (delay) or not. Add regularization to control for multicollinearity.