Practice 3 Solutions

In this practice you will go through a case study using logistic regression model for binary response. You can refer to `prac3.R` script file for the R commands to be used.

The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The data may be found in the the dataset `pima` in `faraway` package.

1. *Create a factor version of the* ***`test`*** *results and use this to produce an interleaved histogram to show how the distribution of* ***`insulin`*** *differs between those testing positive and negative. Do you notice anything unbelievable about the plot?*

   Variable `insulin` has large number of 0 values for both negative and positive cases, with 236 0's for negatives and 138 0's for positives. It is possible a positive case may have 0 insulin (e.g. for a type I diabetes patient). It is not possible for a negative case to have 0 insulin level. This suggests that 0 here is a code for missing value.

```
library(faraway); data(pima, package="faraway")
help(pima); dim(pima); head(pima)   #768 rows by 9 columns
summary(pima)
```

```
    pregnant         glucose        diastolic         triceps         insulin
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.0
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00   Median : 30.5
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54   Mean   : 79.8
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.0
      bmi           diabetes           age             test            test.f
 Min.   : 0.00   Min.   :0.0780   Min.   :21.00   Min.   :0.000   negative:500
 1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00   1st Qu.:0.000   positive:268
 Median :32.00   Median :0.3725   Median :29.00   Median :0.000
 Mean   :31.99   Mean   :0.4719   Mean   :33.24   Mean   :0.349
 3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00   3rd Qu.:1.000
 Max.   :67.10   Max.   :2.4200   Max.   :81.00   Max.   :1.000
```

```
pima$test.f <- factor(pima$test)
levels(pima$test.f) <- c("negative","positive"); pima[1,]
```

```
par(mfrow=c(1,2)); plot(insulin ~ test.f, pima)
plot(jitter(test,0.1) ~ jitter(insulin), pima, xlab="insulin", ylab="signs of diabetes", pch="*")
library(ggplot2)
ggplot(pima, aes(x=insulin, color=test.f)) + geom_histogram(position="dodge", binwidth=30)
```

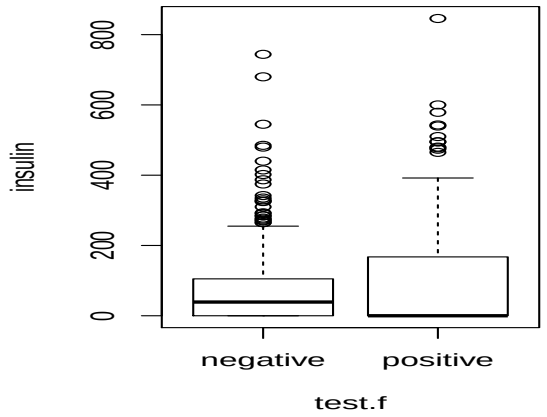
```
ggplot(pima, aes(x=insulin, color=test.f))+geom_histogram(position="dodge", binwidth=10, aes(y=..density..))
```

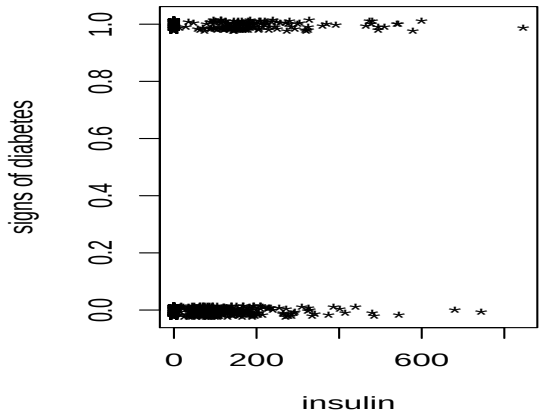
```
summary(pima$insulin)
```

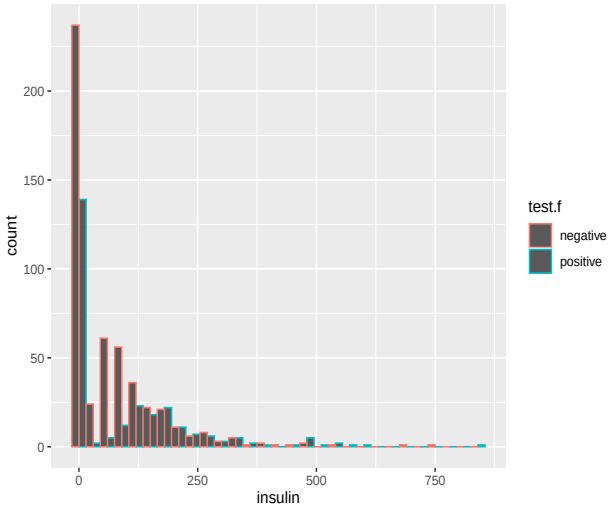
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     0.0    30.5    79.8   127.2   846.0
```

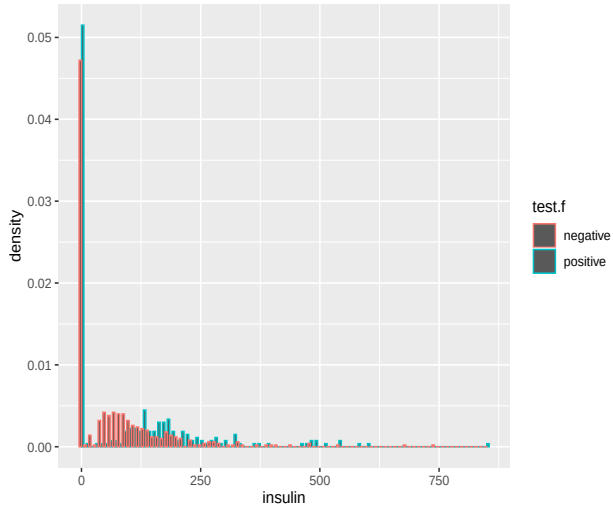
```
summary(pima$test.f[pima$insulin==0])
```

```
negative positive
     236      138
```









2. *Replace the zero values of* ***insulin*** *with the missing value code* ***NA****. Recreate the interleaved histogram plot and comment on the distribution.*

   The **insulin** distribution for positives is more spread-out than that for negatives. The **insulin** levels for both groups are mixed up, but more likely to be smaller for the negatives.

```
pima$insulinN <- pima$insulin
pima$insulinN[pima$insulin==0]<-NA
summary(pima$insulinN)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  14.00   76.25  125.00  155.55  190.00  846.00     374
```

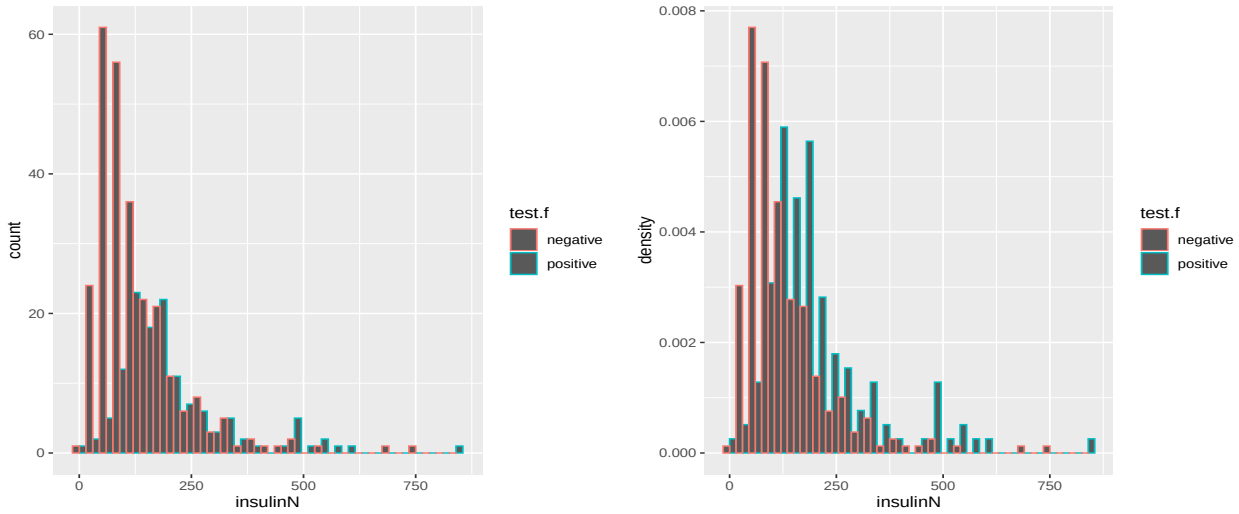
```
summary(pima$insulinN[pima$test.f=="negative"])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   15.0    66.0   102.5   130.3   161.2   744.0     236
```

```
summary(pima$insulinN[pima$test.f=="positive"])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   14.0   127.5   169.5   206.8   239.2   846.0     138
```

```
ggplot(pima, aes(x=insulinN, color=test.f))+geom_histogram(position="dodge", binwidth=30)
ggplot(pima, aes(x=insulinN, color=test.f))+geom_histogram(position="dodge", binwidth=30, aes(y=..density..))
```



3. *Replace the incredible zeroes in other variables with the missing value code. Fit a model with the result of the diabetes test as the response and all the other variables as predictors. How many observations were used in the model fitting? Why is this less than the number of observations in the data frame?*

   336 observations were used. The other $768 - 336 = 432$ individuals contain NAs, thus are excluded from analysis be default in R.

```
pima$pregnantN=pima$pregnant; pima$pregnantN[pima$pregnant==0]<-NA
summary(pima$pregnantN); table(pima$pregnant)

pima$glucoseN=pima$glucose; pima$glucoseN[pima$glucose==0.0]<-NA; summary(pima$glucoseN)

pima$diastolicN=pima$diastolic; pima$diastolicN[pima$diastolic==0.0]<-NA; summary(pima$diastolicN)

pima$tricepsN=pima$triceps; pima$tricepsN[pima$triceps==0.0]<-NA; summary(pima$tricepsN)

pima$bmiN=pima$bmi; pima$bmiN[pima$bmi==0.0]<-NA; summary(pima$bmiN); summary(pima)

lmodNA <- glm(test ~ pregnantN+glucoseN+diastolicN+tricepsN+insulinN+bmiN+diabetes+age, family = binomial, pima)
summary(lmodNA)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.083e+01  1.423e+00  -7.610 2.73e-14 ***
pregnantN    7.364e-02  5.973e-02   1.233   0.2176
glucoseN     3.616e-02  6.249e-03   5.785 7.23e-09 ***
diastolicN   5.993e-03  1.320e-02   0.454   0.6497
tricepsN     1.110e-02  1.869e-02   0.594   0.5527
insulinN     3.231e-05  1.445e-03   0.022   0.9822
bmiN         7.615e-02  3.174e-02   2.399   0.0164 *
diabetes     1.097e+00  4.777e-01   2.297   0.0216 *
age          4.075e-02  1.919e-02   2.123   0.0337 *
---
    Null deviance: 426.34  on 335  degrees of freedom
Residual deviance: 288.92  on 327  degrees of freedom
  (432 observations deleted due to missingness)
AIC: 306.92
```

4. *Refit the model but now without the* ***insulin*** *and* ***triceps*** *predictors. How many observations were used in fitting this model? Devise a test to compare this model with that in the previous question.*

   Not including `insulin` and `triceps` into the model, the model is fitted using 625 observations. So it can not be compared with the model in 3. because the number of observations used in 3. is 336. These two models can only be compared of each other based on the same data. We make this possible by using data `pimaN` which removes all cases containing NAs. The results can be seen in comparing `lmodNA1` with `lmodNA2` in `R`, with p-value of 0.8386. Thus there is no significant difference between the two models in terms of adequacy of fit.

```
lmodNAA <- glm(test ~ pregnantN+glucoseN+diastolicN+bmiN+diabetes+age, family = binomial, pima)
summary(lmodNAA)


Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.354750   0.915697 -10.216  < 2e-16 ***
pregnantN    0.130695   0.037880   3.450  0.00056 ***
glucoseN     0.035337   0.003900   9.061  < 2e-16 ***
diastolicN  -0.008673   0.009422  -0.920  0.35734
bmiN         0.098547   0.017768   5.546 2.92e-08 ***
diabetes     1.020669   0.336136   3.036  0.00239 **
age          0.016642   0.010553   1.577  0.11478
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 807.12  on 624  degrees of freedom
Residual deviance: 577.80  on 618  degrees of freedom
  (143 observations deleted due to missingness)
AIC: 591.8


pimaN <- na.omit(pima)
lmodNA1 <- glm(test ~ pregnantN+glucoseN+diastolicN+tricepsN+insulinN+bmiN+diabetes+age, family = binomial, pimaN)
lmodNA2 <- glm(test ~ pregnantN+glucoseN+diastolicN+bmiN+diabetes+age, family = binomial, pimaN)
anova(lmodNA2, lmodNA1, test="Chi")


Analysis of Deviance Table

Model 1: test ~ pregnantN + glucoseN + diastolicN + bmiN + diabetes +
    age
Model 2: test ~ pregnantN + glucoseN + diastolicN + tricepsN + insulinN +
    bmiN + diabetes + age
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       329     289.27
2       327     288.92  2  0.35199   0.8386
```

5. *Use AIC to select a model. You will need to take account of the missing values. Which predictors are selected? How many cases are used in your selected model?*

   336 cases are used in the selected model. Refer to `summary(lmodNAr)` for detail.

```
lmodNAr <- step(lmodNA1, trace=0)
summary(lmodNAr)


Call:
glm(formula = test ~ glucoseN + bmiN + diabetes + age, family = binomial, data = pimaN)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8112  -0.6673  -0.3433   0.6128   2.6207
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.810466   1.253806  -8.622  < 2e-16 ***
glucoseN      0.036394   0.005495   6.624 3.51e-11 ***
bmiN          0.089165   0.024301   3.669 0.000243 ***
diabetes      1.055880   0.465979   2.266 0.023455 *
age           0.059405   0.014515   4.093 4.26e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 426.34  on 335  degrees of freedom
Residual deviance: 291.12  on 331  degrees of freedom
AIC: 301.12

Number of Fisher Scoring iterations: 5
```

6. *Create a variable that indicates whether the case contains a missing value. Use this variable as a predictor of the test result. Is missingness associated with the test result? Refit the selected model, but now using as much of the data as reasonable. Explain why it is appropriate to do this.*

   The missingness is not significantly associated with the test result, with the p-values of 0.34, 0.3397 and 0.3402 from the `R` output all > 0.05. Finally, the selected model is fitted using 752 cases, 16 less than the 768 available cases. The 16 deleted cases have NAs in either `glucose` or `bmi`.

```
pima$misInd<-apply(pima,1, anyNA); xtabs(~test.f+misInd, pima)
```

```
          misInd
test.f     FALSE TRUE
  negative   225  275
  positive   111  157
```

```
summary(glm(test.f~misInd, family=binomial, pima))$coef
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.7066     0.1160  -6.092 1.12e-09 ***
misIndTRUE    0.1460     0.1532   0.954     0.34
```

```
anova(glm(test.f~misInd, family=binomial, pima), test="Chi")
```

```
Analysis of Deviance Table
Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                     767     993.48
misInd  1  0.91167       766     992.57   0.3397
```

```
chisq.test(pima$test.f, pima$misInd, correct=F)
```

```
        Pearson's Chi-squared test

data:  pima$test.f and pima$misInd
X-squared = 0.90974, df = 1, p-value = 0.3402
```

```
lmodNArs <- glm(test ~ glucoseN + bmiN + diabetes + age, family = binomial, data = pima)
summary(lmodNArs)
```

```
Call: glm(formula = test ~ glucoseN + bmiN + diabetes + age, family = binomial, data = pima)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7389  -0.7362  -0.4103   0.7239   2.4344

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.302177   0.728380 -12.771  < 2e-16 ***
glucoseN     0.035281   0.003517  10.030  < 2e-16 ***
bmiN         0.086372   0.014448   5.978 2.25e-09 ***
diabetes     0.866221   0.298356   2.903 0.003692 **
age          0.028764   0.007852   3.663 0.000249 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 974.75  on 751  degrees of freedom
Residual deviance: 716.30  on 747  degrees of freedom
  (16 observations deleted due to missingness)
AIC: 726.3

Number of Fisher Scoring iterations: 5
```

7. *Using the last fitted model of the previous question, what is the difference in the log-odds of testing positive for diabetes for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant? Then calculate the associated odds ratio value, and give a 95% confidence interval for this odds ratio.*

```
summary(pima$bmiN)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  18.20   27.50   32.30   32.46   36.60   67.10      11
```

   - $Q_1(\texttt{bmiN}) = 27.5$ and $Q_3(\texttt{bmiN}) = 36.6$
   - Estimated log-odds difference $= \hat{\beta}_2 \times (Q_3 - Q_1) = 0.086372 \times (36.6 - 27.50) = 0.7859852$
   - Estimated odds ratio (OR) $= e^{0.086372 \times (36.6 - 27.50)} = 2.194568$
   - Approximate 95% CI for the log-odds difference is

$$0.7859852 \pm 1.96 \times 0.014448 \times (36.6 - 27.5) = (0.5282907, 1.0436797)$$

   - Approx 95% C.I. for the OR $= (e^{0.5282907}, e^{1.0436797}) = (1.696031, 2.839647)$.

8. *Do women who test positive have higher diastolic blood pressures? Is the diastolic blood pressure significant in the logistic regression model? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.*

   Diastolic values tend to be higher for those positives. But the interleaved histograms of the `diastolicN` between those testing positive and negative do not seem to be significantly different. However, both the two-sample t test and the Wilcoxon rank-sum test suggest the positive cases have significantly higher diastolic blood pressures (with p-values of 0.03576 and 3.779×10^{-5} respectively).

   On the other hand, `diastolicN` is not found to be significant to the odds of positive test vs. negative test based on the aforementioned logistic models. The means a given

difference between the diastolic pressures of two women does not lead to a significant value of odds ratio of positive test vs. negative test between the two women. Therefore, although the two answers appear to be contradictory, they are actually not.

```
summary(pima$diastolicN[pima$test.f=="negative"])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  24.00   62.00   70.00   70.88   78.00  122.00      19
```

```
summary(pima$diastolicN[pima$test.f=="positive"])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  30.00   68.00   74.50   75.32   84.00  114.00      16
```

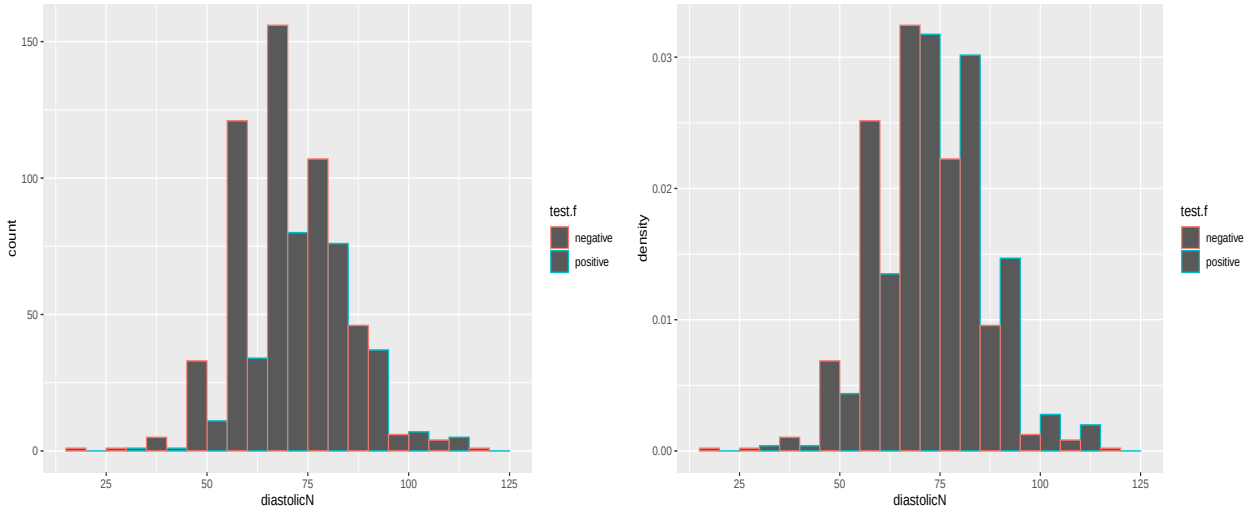
```
t.test(diastolic~test.f, alternative="less",data=pima, var.equal=T)
```

```
        Two Sample t-test
data:  diastolic by test.f
t = -1.8047, df = 766, p-value = 0.03576
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.2309752
sample estimates:
mean in group negative mean in group positive
              68.18400               70.82463
```

```
wilcox.test(diastolic~test.f, alternative="less",data=pima)
```

```
        Wilcoxon rank sum test with continuity correction
data:  diastolic by test.f
W = 55414, p-value = 3.779e-05
alternative hypothesis: true location shift is less than 0
```

```
ggplot(pima, aes(x=diastolicN, color=test.f)) + geom_histogram(position="dodge", binwidth=10)
ggplot(pima, aes(x=diastolicN, color=test.f)) + geom_histogram(position="dodge", binwidth=10, aes(y=..density..))
```



There is always some freedom in deciding which specifications of the method in use, in what order to apply them, and how to interpret the results. So there may not be one clear right answer and good analysts may come up with different models.

It is always a good idea to record your data analysis results and turn them into a technical or research report.