HW9

Load the *Pima* diabtes dataset. Remove missing values from the predictors glucose, diastolic, & bmi. The predictor pregnant does have zero's, but in this context a zero does make sense (no children); age & diabetes do not contain zero's.

```
## [1] 768   9
```

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

```
## [1] 724   9
```

```
##     pregnant         glucose         diastolic        triceps
##  Min.   : 0.000   Min.   : 44.00   Min.   : 24.0   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.75   1st Qu.: 64.0   1st Qu.: 0.00
##  Median : 3.000   Median :117.00   Median : 72.0   Median :24.00
##  Mean   : 3.866   Mean   :121.88   Mean   : 72.4   Mean   :21.44
##  3rd Qu.: 6.000   3rd Qu.:142.00   3rd Qu.: 80.0   3rd Qu.:33.00
##  Max.   :17.000   Max.   :199.00   Max.   :122.0   Max.   :99.00
##     insulin           bmi           diabetes           age
##  Min.   :  0.00   Min.   :18.20   Min.   :0.0780   Min.   :21.00
##  1st Qu.:  0.00   1st Qu.:27.50   1st Qu.:0.2450   1st Qu.:24.00
##  Median : 48.00   Median :32.40   Median :0.3790   Median :29.00
##  Mean   : 84.49   Mean   :32.47   Mean   :0.4748   Mean   :33.35
##  3rd Qu.:130.50   3rd Qu.:36.60   3rd Qu.:0.6275   3rd Qu.:41.00
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00
##       test
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3439
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

Fit a *binomial regression* model with the result of the diabetes *test* as a response and *pregnant*, *glucose*, *diastolic*, *bmi*, *diabetes* and *age* as predictors.

```
##
## Call:
## glm(formula = test ~ pregnant + glucose + diastolic + bmi + diabetes +
##     age, family = binomial(link = logit), data = pima)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8062  -0.7229  -0.4049   0.7173   2.3959
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.962146   0.820892 -10.918  < 2e-16 ***
## pregnant     0.117863   0.033418   3.527  0.00042 ***
## glucose      0.035194   0.003605   9.763  < 2e-16 ***
## diastolic   -0.008916   0.008618  -1.035  0.30084
## bmi          0.090926   0.015740   5.777 7.61e-09 ***
## diabetes     0.960515   0.306415   3.135  0.00172 **
## age          0.016944   0.009834   1.723  0.08489 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 931.94  on 723  degrees of freedom
## Residual deviance: 672.86  on 717  degrees of freedom
## AIC: 686.86
##
## Number of Fisher Scoring iterations: 5
```

1. Refering to slides 1 of week 12, it is **not possible** to use to the deviance to test the goodness of fit given that the response (test) is a binary one (0,1).

2. The ratio of the odds of testing positive for a woman with a BMI at the first quartile compared to a woman with BMI at the third quartile, with all other predictors held constant is **0.4371729**. A unit increase in x1 (BMI) with all other predictors held fixed leads to an increase in B4 (bmi coeff) in log-odd; equivalently odds being multiplied by exp(B4). Hence, the ratio would be exp(B4* x1=27.5) / exp(B4* x1=36.60), where B4 is equal to the coefficient of from the model (0.090926).

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.20   27.50   32.40   32.47   36.60   67.10
```

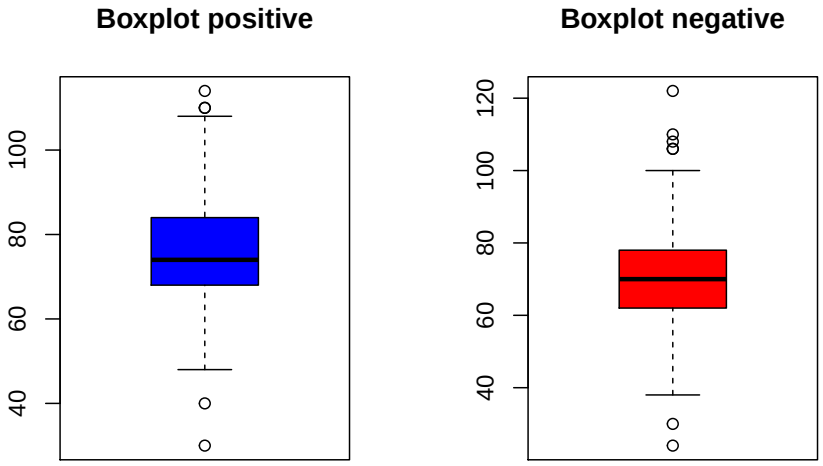
```
## [1] 0.4371729
```

3. Women who test positive for diabetes have **higher** diastolic blood pressure. This can be first checked informally through a numerical/graphical summary, and as well more rigorously through a two-sample t-test. The conclusion suggests that there is indeed a statistical significant difference of distolic pressure between diabeteic and non-diabetic. The diastolic blood pressure is **NOT significant** as shown in the summary of the binomial regression model. Although the results might seem contradictory, they are not. The key is to recognize that in the in the two-sample t-test we are only dealing with one predictor (disatolic); whereas in the binomial regression model we are dealing 6 different predictors. Often times,

one predictor becomes insignifancant, if its effect is already represented by another predictor. Hence, in our case although we see that positive cases have higher diastolic pressure, this relationship could be already be taken into account by one of the 5 remaining predictors in the model.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.00   68.00   74.00   75.25   84.00  114.00
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   24.00   62.00   70.00   70.91   78.00  122.00
```




```
##
##  Welch Two Sample t-test
##
## data:  pos and neg
## t = 5.9701, df = 3576.1, p-value = 2.602e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.561798 12.979177
## sample estimates:
## mean of x mean of y
##  47.60171  37.83123
```

4. The probability of testing positive for a 30-year old woman who has been pregnant once, has glucose measurement of 100, diastolic blood pressure 70, BMI 25, and diabetes pedigree measurement of 0.6 is **0.0697082**.

```
test= data.frame(test=1,age=30, pregnant=1,glucose=100, diastolic=70, bmi=25,diabetes=0.6)
ilogit(predict(logit,test))
```

```
##         1
## 0.0697082
```