---
title: "36-402 Homework 10"
author: "Jacky Liu"
date: "2/12/21"
output: pdf_document

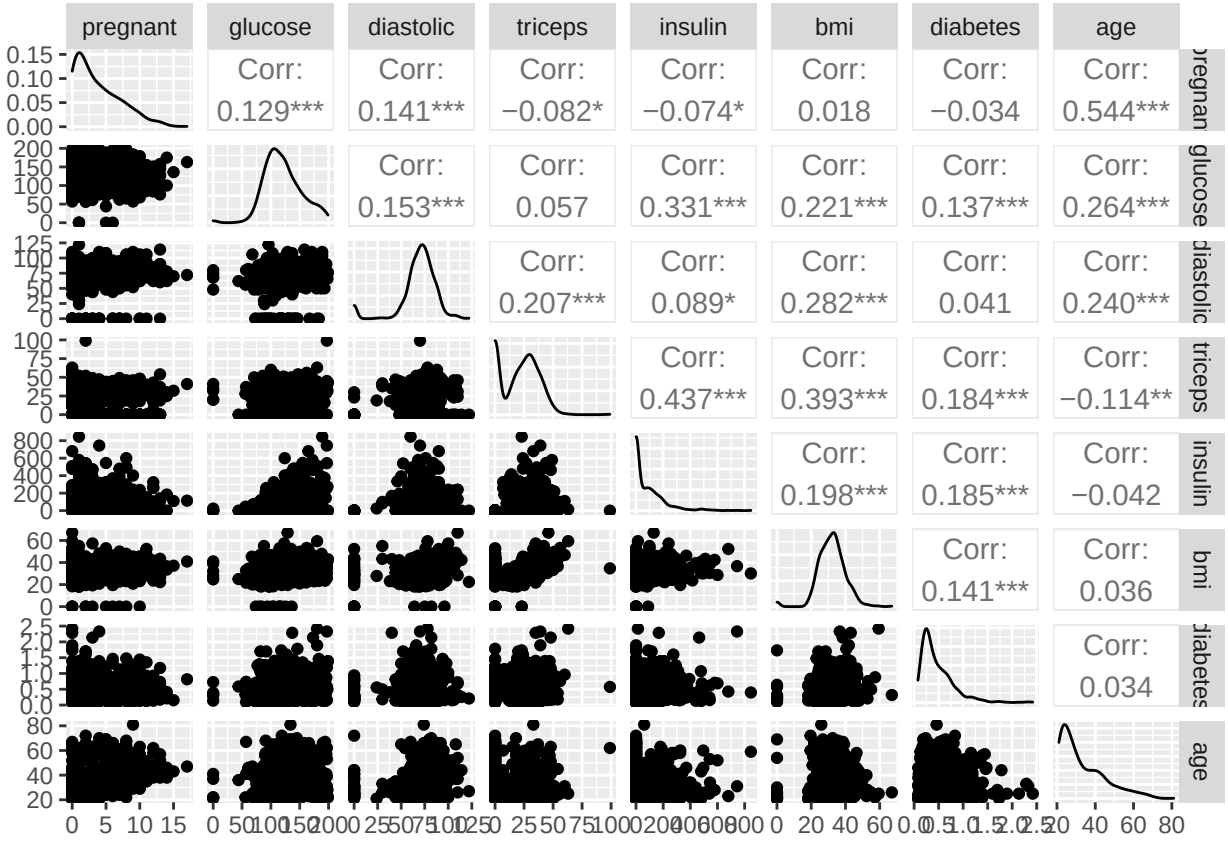
```
pima = read.csv("pima.csv")
```

Problem 1

Problem 1 (a)

Plots:

```
library(tidyverse)
library(GGally)
ggpairs(pima, columns=1:8)
```



Summaries:

```
means = vector(length = 8)
SDs = vector(length = 8)
mins = vector(length = 8)
for (i in 1:8){
```

```
  means[i] = mean(pima[,i])
  SDs[i] = sd(pima[,i])
  mins[i] = min(pima[,i])
}
means
```

```
## [1]   3.8450521 120.8945312  69.1054688  20.5364583  79.7994792  31.9925781
## [7]   0.4718763  33.2408854
```

```
SDs
```

```
## [1]   3.3695781  31.9726182  19.3558072  15.9522176 115.2440024   7.8841603
## [7]   0.3313286  11.7602315
```

```
mins
```

```
## [1]  0.000  0.000  0.000  0.000  0.000  0.000  0.078 21.000
```

The above values show the means, standard deviations, and minimums of each variable in the same order as in the GGPairs plots above.

For variables such as bmi, diastolic, and triceps, the value of 0 usually doesn't make sense. Therefore, it is most likely a filler value for NA.

```
# Replace 0 values with NA
for(i in 1:length(pima$glucose)) {
  if(pima$glucose[i] == 0){
    pima$glucose[i] = NA
  }
  if(pima$diastolic[i] == 0){
    pima$diastolic[i] = NA
  }
  if(pima$triceps[i] == 0){
    pima$triceps[i] = NA
  }
  if(pima$insulin[i] == 0){
    pima$insulin[i] = NA
  }
  if(pima$bmi[i] == 0){
    pima$bmi[i] = NA
  }
}
```

```
# create new data frame with NA entries omitted
newpima = na.omit(pima)
length(newpima[,1])
```

```
## [1] 392
```

Problem 1 (b)

```r
model1 = glm(test ~ ., family = binomial, data=newpima)
summary(model1)
```

```
##
## Call:
## glm(formula = test ~ ., family = binomial, data = newpima)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7823  -0.6603  -0.3642   0.6409   2.5612
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.004e+01  1.218e+00  -8.246  < 2e-16 ***
## pregnant     8.216e-02  5.543e-02   1.482  0.13825
## glucose      3.827e-02  5.768e-03   6.635 3.24e-11 ***
## diastolic   -1.420e-03  1.183e-02  -0.120  0.90446
## triceps      1.122e-02  1.708e-02   0.657  0.51128
## insulin     -8.253e-04  1.306e-03  -0.632  0.52757
## bmi          7.054e-02  2.734e-02   2.580  0.00989 **
## diabetes     1.141e+00  4.274e-01   2.669  0.00760 **
## age          3.395e-02  1.838e-02   1.847  0.06474 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.02  on 383  degrees of freedom
## AIC: 362.02
##
## Number of Fisher Scoring iterations: 5
```

It seems that diastolic, triceps, and insulin do not have statistically significant association and may not be contributing to the fit.

Problem 1 (c)

```r
model2 = glm(test ~ 1, data = newpima)
anova(model2, model1, test="Rao")
```

```
## Analysis of Deviance Table
##
## Model 1: test ~ 1
## Model 2: test ~ pregnant + glucose + diastolic + triceps + insulin + bmi +
##     diabetes + age
##   Resid. Df Resid. Dev Df Deviance    Rao  Pr(>Chi)
## 1       391      86.89
## 2       383     344.02  8  -257.13 30.044 0.0002077 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is small which suggests that model 1 is a better fir to the data than model 2. This means Model 1 is a significant improvement on Model 2.

Problem 1 (d)

Do women with signs of diabetes have higher 2-hour serum insulin values?

```
NOdiabetes = newpima[which(newpima$test == 0),]
YESdiabetes = newpima[which(newpima$test == 1),]
t.test(NOdiabetes$insulin, YESdiabetes$insulin)
```

```
##
##  Welch Two Sample t-test
##
## data:  NOdiabetes$insulin and YESdiabetes$insulin
## t = -5.7337, df = 207.88, p-value = 3.429e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -102.11966  -49.86272
## sample estimates:
## mean of x mean of y
##  130.8550  206.8462
```

Based on our statistical test above, the p-value is very small, which means women with signs of diabetes most likely have higher 20hour serum insulin values.

The insulin coefficient is in model 1 is 0.52757, which makes it not significant.

These answers are not contradictory because model 1 is not simply only between insulin and diabetes test; there are other factors in the model as well. If you fit the same model except with everything taken out except insulin and test, we can see that it is significant in this case (p=0.005653).

Problem 1 (e)

```
model3 <- step(model1, direction="backward", trace=0)
anova(model3, model1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: test ~ pregnant + glucose + bmi + diabetes + age
## Model 2: test ~ pregnant + glucose + diastolic + triceps + insulin + bmi +
##     diabetes + age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       386     344.89
## 2       383     344.02  3   0.8639   0.8341
```

The p-value is large, which means that we fail to reject the null hypothesis. This suggests that model 3 is a better fit to the data.

Problem 1 (g)

```
preds = predict(model3, data.frame(pregnant=3, glucose=103, diastolic=70, tricep=29.2, insulin=160, bmi=
```

The probability of this Pima woman testing positive based on Model 3 is 0.1593196.

```
c(preds$fit - qnorm(0.1)*preds$se.fit, preds$fit + qnorm(0.1)*preds$se.fit)
```

```
##         1         1
## 0.1931330 0.1255062
```

The 90% confidence interval is (0.1255062, 0.1931330).

Problem 1 (h)

```
pred2 = predict(model3, data.frame(pregnant=3, glucose=103, diastolic=70, tricep=29.2, insulin=160, bmi=
pred3 = predict(model3, data.frame(pregnant=3, glucose=103, diastolic=70, tricep=29.2, insulin=160, bmi=
pred2
```

```
## $fit
##       1
## -1.6633
##
## $se.fit
## [1] 0.1969939
##
## $residual.scale
## [1] 1
```

```
pred3
```

```
## $fit
##         1
## -2.066119
##
## $se.fit
## [1] 0.239052
##
## $residual.scale
## [1] 1
```

The log-odds are different by 0.402819.

```
ci2 = c(pred2$fit - qnorm(0.1)*pred2$se.fit, pred2$fit + qnorm(0.1)*pred2$se.fit)
ci3 = c(pred3$fit - qnorm(0.1)*pred3$se.fit, pred3$fit + qnorm(0.1)*pred3$se.fit)
ci2-ci3
```

```
##         1         1
## 0.3489199 0.4567191
```

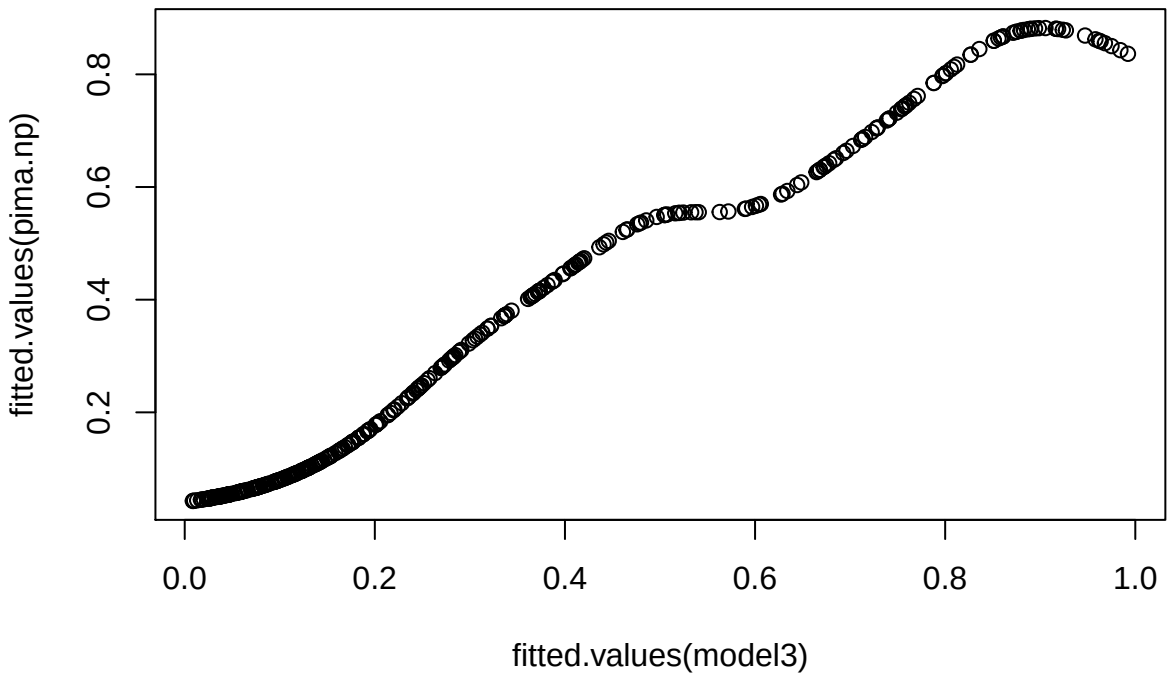
The 90% confidence interval for the difference is (0.3489199, 0.4567191).

Problem 1 (i)

```
library(np)
```

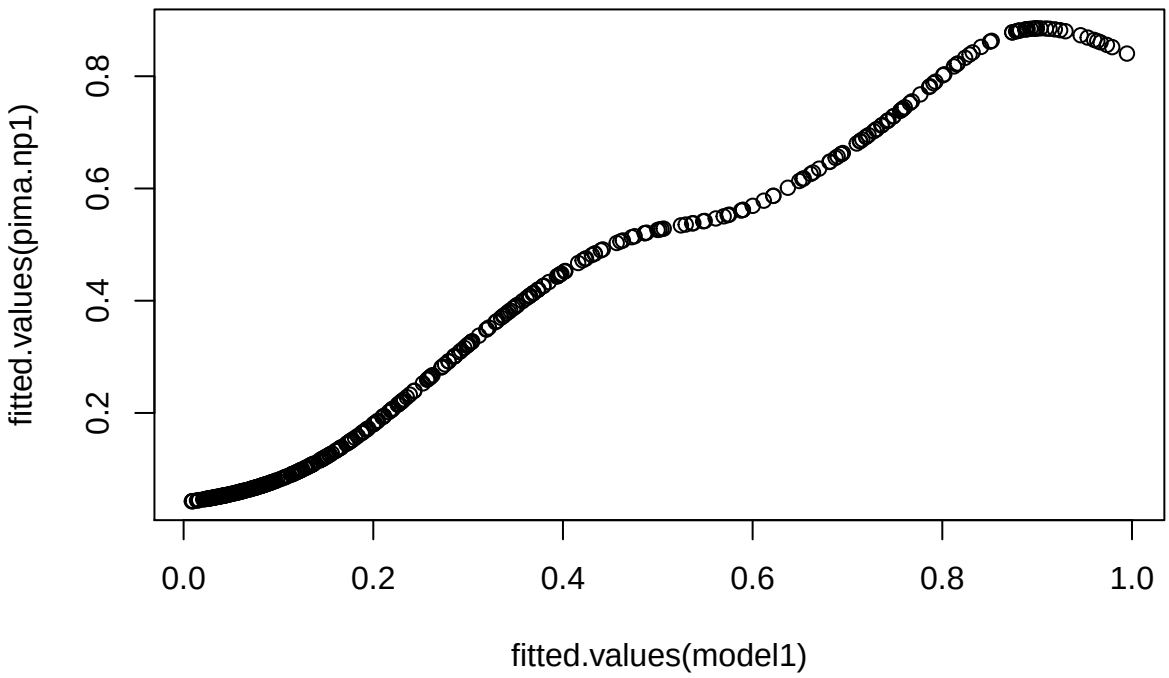
```
## Nonparametric Kernel Methods for Mixed Datatypes (version 0.60-10)
## [vignette("np_faq",package="np") provides answers to frequently asked questions]
## [vignette("np",package="np") an overview]
## [vignette("entropy_np",package="np") an overview of entropy-based methods]
```

```
pima.np = npreg(newpima$test ~ fitted.values(model3), bws=0.075)
plot(fitted.values(model3), fitted.values(pima.np))
```



The kernel regression's fitted values follow the y=x line fairly closely. This suggests that model 3 is fairly well calibrated.

```
pima.np1 = npreg(newpima$test ~ fitted.values(model1), bws=0.075)
plot(fitted.values(model1), fitted.values(pima.np1))
```



Model 1 also seems fairly well calibrated. None of these models appear to be noticeably better calibrated than the other.

Problem 2

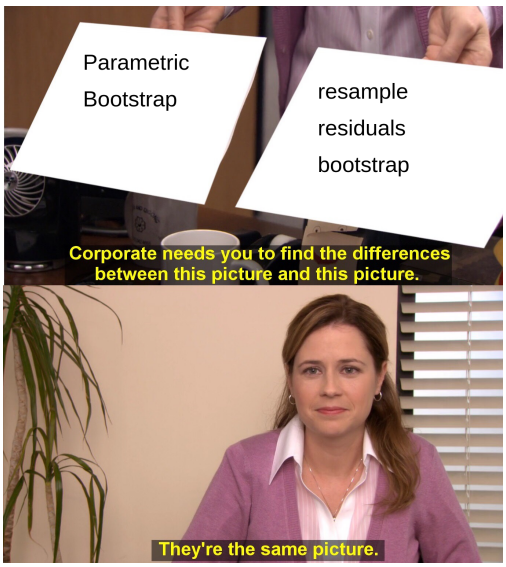


Figure 1: low effort meme (source: original)