data 102

hw2.

1. → racial disparity of error rate. (COMPAS)

→ more recidivate: P(black & high risk) > P(white & high risk).

→ positive/error rate & calibration by group.

→ classifier $\hat{Y}$ → 1 → high risk
→ 0 → low risk & Y = true outcome (reality).

→ A = race & classifier base on score.

a.

1).

⇒ $P(\hat{Y}=1 \mid (Y=0, A=\text{black})) > P(\hat{Y}=1 \mid (Y=0, A=\text{white}))$.

2).

⇒ $P(\hat{Y}=0 \mid (Y=1, A=\text{white})) > P(\hat{Y}=0 \mid (Y=1, A=\text{black}))$.

b.

⇒ Yes. ProPublica's statement in 1.(a) violate the three criteria of fairness.

It violate the second one: equalizing the error rate.

⇒ the equalization of error rate require $P(\hat{Y}=1 \mid (Y=0, A=\text{black})) == P(\hat{Y}=1 \mid (Y=, A=\text{white}))$ and $P(\hat{Y}=0 \mid (Y=1, A=\text{white})) == P(\hat{Y}=0 \mid (Y=1, A=\text{black}))$, since this is clearly not the case in our 1.a., thus ProPublica violate equalizing the error rate fairness criteria.

c.

⇒ $P(Y=1 \mid (\hat{Y}=1, A=\text{black})) \approx P(Y=1 \mid (\hat{Y}=1, A=\text{white}))$.

⇒ and $P(Y=1 \mid (\hat{Y}=0, A=\text{black})) \approx P(Y=1 \mid (\hat{Y}=0, A=\text{white}))$.

d.

⇒ Northpointe's claim satisfy the last criteria=
equalizing the column-wise rate.

e. → P_a = prop of group a that recidivate.

→ $P_a = P(Y=1 \mid A=a)$.

→ TPR_a & FPR_a = true/false positive rate.

→ PPV_a = positive predictive value ⇒ $PPV_a = P(Y=1 \mid \hat{Y}=1, A=a)$.

→ NPV_a = negative predictive value

→ prove: $PPV_a = (TPR_a \cdot P_a) / (TPR \cdot P_a + FPR_a * (1-P_a))$ & $a \in \{\text{black}, \text{white}\}$.

→ bayes rule & PPV_a → $P(\hat{Y}=1, Y=1 \mid A=a)$
→ $P(\hat{Y}=1 \mid A=a)$.

⇒ in the back page.

$\Rightarrow PPV_a = P(Y=1 \mid \hat{Y}=1, A=a)$

$= \frac{P((\hat{Y}=1, A=a) \mid Y=1) * P(Y=1 \mid A=a)}{P(\hat{Y}=1 \mid A=a)}$ ← bayes's rule

$= \frac{P(\hat{Y}=1 \mid Y=1, A=a) * P(Y=1 \mid A=a)}{P(\hat{Y}=1 \mid A=a)}$ ← $TPR_a = P(\hat{Y}=1 \mid Y=1, A=a)$

← $P_a = P(Y=1 \mid A=a)$.

$= \frac{TPR_a * P_a}{P(\hat{Y}=1 \mid A=a)}$

← $TNR_a = P(\hat{Y}=0 \mid Y=0, A=a)$

← $FPR_a = P(\hat{Y}=1 \mid Y=0, A=a)$.

$\Rightarrow P(\hat{Y}=1 \mid A=a) = \frac{\#TP_a + \#FP_a}{\#A=a}$

$= \frac{\#TPR_a * (\#Y=1 \mid A=a) + \#FPR_a * (\#Y=0 \mid A=a)}{\#A=a}$

$= \#TPR_a * \frac{\#Y=1 \mid A=a}{\#A=a} + \#FPR_a * \frac{\#Y=0 \mid A=a}{\#A=a}$.

$= \#TPR_a * P_a + \#FPR_a (1 - P_a)$

$\Rightarrow$ thus, $PPV_a = \frac{TPR_a * P_a}{TPR_a * P_a + FPR_a * (1 - P_a)}$.

f. → recidivism not independant of race

→ $P(Y=1 \mid A=black) \neq P(Y=1 \mid A=white)$.

→ non-zero fine & false positive rate.

→ equalizing two error rate ∝ can't have same positive prediction value.

⇒ equalizing error rate

→ $P(\hat{Y}=1 \mid Y=0, A=black) == P(\hat{Y}=1 \mid Y=0, A=white) = FPR = 1 - TNR$.

→ $P(\hat{Y}=0 \mid Y=1, A=black) == P(\hat{Y}=0 \mid Y=1, A=white) = FNR = 1 - TPR$

⇒ for $PPV_a == PPV_b$

→ $\frac{TPR_a * P_a}{P(\hat{Y}=1 \mid A=a)} = \frac{TPR_b * P_b}{P(\hat{Y}=1 \mid A=b)}$

→ $\frac{TPR_a * P_a}{TPR_a * P_a + FPR_a * (1 - P_a)} = \frac{TPR_b * P_b}{TPR_b * P_b + FPR_b * (1 - P_b)}$

→ $\frac{(1 - FNR_a) * P_a}{(1 - FNR_a) * P_a + FPR_a * (1 - P_a)} = \frac{(1 - FNR_b) * P_b}{(1 - FNR_b) * P_a + FPR_b * (1 - P_b)}$.

⇒ since $1 - FNR_a == 1 - FNR_b$ and $FPR_a == FPR_b$

⇒ P_a have to be equal to P_b for the equation to work

⇒ however, we know recidivism is not independent of race.

⇒ $P_a = P(Y=1 \mid A=black) \neq P_b = P(Y=1 \mid A=white)$

⇒ thus, $PPV_a \neq PPV_b$ if the condition hold true.

⇒ thus, if equalizing error rates hold, then the two group can't have same PPV.

stats 102

hw 2.

2. X = application & Y = invitation. (1 → invite, 0 → not.)

→ A = religious minority group.

→ $\hat{Y}$ = better by recruiter.

a.

⇒ equalize positive rate = $P(\hat{Y}=1 \mid A=1) = P(\hat{Y}=1 \mid A=0)$.

⇒ $P(\hat{Y}=1 \mid A=1) = 100/500 = 1/5 = 3/15$

⇒ $P(\hat{Y}=1 \mid A=0) = 1200/(3000+1500) = 4/15$

⇒ $P(\hat{Y}=1 \mid A=0) \neq P(\hat{Y}=1 \mid A=1)$

⇒ thus, this classifier does not satisfied equalizing positive rate criterion.

⇒ equalize error rate =

$P(\hat{Y}=1 \mid Y=0, A=0) = P(\hat{Y}=1 \mid Y=0, A=1)$

$P(\hat{Y}=0 \mid Y=1, A=0) = P(\hat{Y}=0 \mid Y=1, A=1)$

⇒ $P(\hat{Y}=1 \mid Y=0, A=1) = 40/400 = 1/10$

⇒ $P(\hat{Y}=1 \mid Y=0, A=0) = 300/3000 = 1/10$

$\Rightarrow P(\hat{Y}=0 \mid Y=1, A=1) = 40/100 = 2/5$

$\Rightarrow P(\hat{Y}=0 \mid Y=1, A=0) = 600/1500 = 2/5$

$\Rightarrow P(\hat{Y}=1 \mid Y=0, A=0) = = P(\hat{Y}=1 \mid Y=0, A=1)$

$P(\hat{Y}=0 \mid Y=1, A=0) = = P(\hat{Y}=0 \mid Y=1, A=1)$

$\Rightarrow$ thus, this satisfied equalizing error rate criteria.

b. threshold = $\hat{Y} = \mathbb{1}\{R(X) > t\}$

→ if not satisfied & group depend threshold.

$\Rightarrow$ suppose the original threshold is t_0 & t_1,

and it dose not have equalizing criteria

$\Rightarrow$ let the new threshold be t_0' & t_1'

$\Rightarrow$ from a we know that we have non-equalized positive rate

$\Rightarrow P(\hat{Y}=1 \mid A=1) < P(\hat{Y}=1 \mid A=0)$

$\Rightarrow$ so we need to increase the $\hat{Y}=1$ of group $A=1$ and decrease $\hat{Y}=1$ for $A=0$.

$\Rightarrow$ thus, t_1' will be a decrease for t_1,

t_0' will be an increase for t_0

c. → historical data.

1).

⇒ If the religious group are getting "free in", the reality if

$Y=1 \mid A=1$ is going to be large

⇒ so the chance a religious person getting an invite is already high

⇒ as we even lower more t_i' threshold,

more religious person are being considered interview by the company.

⇒ thus, the company is going to interview even more

religious people.

2).

⇒ the religious group have "better" application score than non-religous group,

so the prediction of # $\hat{Y}=1$ is going to be high

⇒ since we lower t_i' even more, we are getting even more $\hat{Y}=1$

for the religious group.

⇒ thus, as the result, the positive rate for for religious group will go higher

the company is likely to recruit more people from the

religious group

3. → X = input & $Y \in \{0, 1\}$ = output

→ score func $R(x)$ & decision rule

→ $\mathbb{P}(Y=1 \mid R(x)=r) = r$ & $r \in [0, 1]$.

→ $\mathbb{P}(Y=1 \mid R(x)=r, A=a) = r$ = calibrated by group.

→ estimate & plot (calibrate) score func. = calibration plot.

a. → dataset heart.csv

→ column 1 = score (logistic $\in [0, 1]$).

→ colum 2 = reality of whether heart disease.

→ column 3 = binary label A.

→ calibration_plot()

→ input = probability array, array of score, reality.

→ return = $\mathbb{P}(\hat{Y}=1 \mid R(x) \in [r[i], r[i+1]])$.

→ generate prob = r & output rate = p.

b. → generate prob against rate.

→ perf func = y=x line

c. → calibration plot to specific group A = a.

→ plot ↗ male = A=1

→ female = A=0.

4. → linear regression & gauss-markov theorm

→ $y^{(i)} = \langle \beta^*, x^{(i)} \rangle + \epsilon^{(i)}$ & $\epsilon^{(i)}$ independent dist.

→ OLS: $\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} x^{(i)} x^{(i)T}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} x^{(i)} y^{(i)}\right)$.

→ if independent $\epsilon^{(i)}$ & $\mathbb{E}[\epsilon^{(i)} \mid x^{(i)}] = 0$

→ for $n \to \infty$ = $\hat{\beta} = \beta^*$

→ $\mathbb{E}[\hat{\beta}] = \beta^*$ & unbaised $\hat{\beta}$.

→ if $\text{var}[\epsilon^{(i)}]$ same for all i

→ $\text{var}(\hat{\beta})$ = lowest unbaised estimator.

→ OLS = best linear unbaised estimator (best mean lowest variance).

a. → whether $\hat{\beta}$ is unbaised estimator of β^*

→ independent $z^{(i)}$

1).

⇒ $y^{(i)} = \langle \beta^*, x^{(i)} \rangle + \sin(x^{(i)}) \cdot z^{(i)}$ & $z^{(i)} \sim N(0,1)$.

⇒ $\mathbb{E}[\epsilon^{(i)} \mid x^{(i)}] = 0$ & $\epsilon^{(i)}$ independent from each other

⇒ $\mathbb{E}[\sin(x^{(i)}) \cdot z^{(i)} \mid x^{(i)}]$

$= \sin(x^{(i)}) * \mathbb{E}(z^{(i)})$ ← since $x^{(i)}$ is given, we can treat $\sin(x^{(i)})$ as a constant

⇒ since $z^{(i)}$ is $N(0,1)$ dist

$z^{(i)}$ = standard norm distribution $\begin{cases} \text{mean} = 0 \\ \text{variance} \end{cases}$

$\Rightarrow \sin(x^{(i)}) * \mathbb{E}(z^{(i)}) = \sin(x^{(i)}) * 0$

$= 0$

$\Rightarrow \mathbb{E}\left[\sin(x^{(i)}) \cdot z^{(i)} \mid x^{(i)}\right] = 0$

→ statisfied $\mathbb{E}\left[\epsilon^{(i)} \mid x^{(i)}\right] = 0$

⇒ since $x^{(i)}$ independent variable → $\sin(x^{(i)})$ independent variable

⇒ $z^{(i)}$ independent from each other and x

⇒ $\sin(x^{(i)}) \cdot z^{(i)}$ = independent from each other

→ statisfied $\epsilon^{(i)}$ independent from each other

⇒ thus, $\hat{\beta}$ is an unbaised estimate of β^*.

2)

$\Rightarrow \mathbb{E}\left[z^{(i)^2} \mid x^{(i)}\right] = \mathbb{E}\left[z^{(i)^2}\right]$ ← $z^{(i)}$ independent from $x^{(i)}$.

$\Rightarrow z^{(i)} = Nu(0,1)$ dist

$\Rightarrow$ let $z^{(i)^2} = V \rightarrow f_V(v) = \frac{1}{\sqrt{2\pi}} V^{-\frac{1}{2}} e^{-\frac{1}{2}v}$

$= \text{gamma}(\frac{1}{2}, \frac{1}{2})$ dist

$= \text{chi-square}(1)$ dist

$\Rightarrow \mathbb{E}\left[\text{gamma}(\frac{1}{2}, \frac{1}{2}) \text{ dist}\right] = \frac{1}{2} / \frac{1}{2}$

$= 1$

$\Rightarrow \mathbb{E}[z^{(i)^2} | X^{(i)}] = 1$

$\rightarrow$ not satisfied $\mathbb{E}[\epsilon^{(i)} | X^{(i)}] = 0$

$\Rightarrow$ thus, $\hat{\beta}$ is not an unbaised estimator of β^*.

3).

$\Rightarrow$ for the gauss-markov theorm, we assume $y^{(i)} = \beta^* X^{(i)} + \Sigma^{(i)}$

$\Rightarrow$ thus $y^{(i)}$ must be a linear transformation of $X^{(i)}$.

$\Rightarrow \cos(\langle \beta^*, X^{(i)} \rangle)$ indicate not linear relationship

$\Rightarrow y^{(i)} = \cos(\langle \beta^*, x^{(i)} \rangle) + z^{(i)}$ indicate that $y^{(i)}$ is not a linear transformation of $x^{(i)}$

$\Rightarrow$ the gauss-markov theorm don't hold true anymore

$\Rightarrow$ thus, $\hat{\beta}$ is not an unbaised estimate of β^*.

4).

$\Rightarrow \mathbb{E}[z^{(i)} | X^{(i)}] = \mathbb{E}[z^{(i)}]$

$= 0$ $\leftarrow$ mean $(N(0, |X^{(i)}|)) = 0$

$\rightarrow$ satisfied $\mathbb{E}[\epsilon^{(i)} | X^{(i)}] = 0$.

$\Rightarrow$ however $\text{var}[\epsilon^{(i)}] \neq \text{var}[\epsilon^{(i')}]$

$\Rightarrow$ thus, $\hat{\beta}$ is an unbaised estimate of β^*

but it's not the best linear unbaised estimator of β^*.

5). $\rightarrow y^{(i)} = \langle \beta^*, x^{(i)} + z^{(i)} \mathbb{1} \rangle$

$\rightarrow z^{(i)} \nearrow^{1\ (p=1/2)}_{\rightarrow -1\ (p=1/2)}$ & $\mathbb{1}$ = all-ones vector

$\rightarrow x^{(i)} + z^{(i)} \mathbb{1} = x^{(i)} + z^{(i)}$ element-wise.

$\Rightarrow y^{(i)} = \langle \beta^*, x^{(i)} + z^{(i)} \mathbb{1} \rangle$

$= [x^{(i)} + z^{(i)} \mathbb{1}] \cdot \beta^*$ $\leftarrow$ a dot product.

$= [x^{(i)}] \cdot \beta^* + [z^{(i)} \mathbb{1}] \cdot \beta^*$

$= \langle \beta^*, x^{(i)} \rangle + \langle \beta^*, z^{(i)} \mathbb{1} \rangle$.

$\Rightarrow \varepsilon^{(i)} = \langle \beta^*, z^{(i)} \mathbb{1} \rangle$

$= \beta^* \cdot \sum_{i=1}^{d} z^{(i)}$

$\Rightarrow \mathbb{E}[\varepsilon^{(i)} | x^{(i)}] = \mathbb{E}[\beta^* \cdot \sum_{i=1}^{d} z^{(i)} | x^{(i)}]$

$= \beta^* \cdot \mathbb{E}[\sum_{i=1}^{d} z^{(i)}]$

$\Rightarrow$ since $z^{(i)} \nearrow^{1\ (p=\frac{1}{2})}_{\rightarrow -1\ (p=\frac{1}{2})}$

$\Rightarrow \mathbb{E}[\sum_{i=1}^{d} z^{(i)}] = \frac{1}{2} * \sum_{i=1}^{d}(1) + \frac{1}{2} * \sum_{i=1}^{d}(-1)$

$= \frac{1}{2} * d * 1 + \frac{1}{2} * d * (-1)$

$= 0$

$\Rightarrow \beta^* \cdot \mathbb{E}\left[\sum_{i=1}^{d} z^{(i)}\right] = \beta^* \cdot 0$

$= 0$

$\Rightarrow \mathbb{E}\left[\langle \beta^*, z^{(i)} \mathbb{1} \rangle \mid x^{(i)}\right] = 0$

→ satisfied $\mathbb{E}\left[\varepsilon^{(i)} \mid x^{(i)}\right] = 0$

$\Rightarrow z^{(i)} \mathbb{1}$ independent of each other → $\langle \beta^*, z^{(i)} \mathbb{1} \rangle$ independent

→ satisfied $\varepsilon^{(i)}$ independent of each other

$\Rightarrow$ thus, $\hat{\beta}$ is an unbaised estimator of β^*.

b. → intercept only model = $x^{(i)} = 1$ & $d = 1$

→ $y^{(i)} = \beta^* + \epsilon^{(i)}$

$\Rightarrow \hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^{n} x^{(i)} x^{(i)T}\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x^{(i)} y^{(i)}\right)$

$= \left(\frac{1}{n} \sum_{i=1}^{n} (1)\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} y^{(i)}\right)$

$= 1 * \left(\frac{1}{n} \sum_{i=1}^{n} y^{(i)}\right)$

$= \mathbb{E}(Y)$.

$\Rightarrow$ thus, the OLS $\hat{\beta} = \mathbb{E}(Y)$

= mean of our target variable (value).

Untitled

February 21, 2020

[131]:
```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
import scipy.stats
%matplotlib inline
```

[132]:
```
dataframe = pd.read_csv("heart.csv",header = None)
```

[133]:
```
def calibration_plot(probabilities, realities, label):
    recurr = len(probabilities) - 1
    #label = np.repeat(label,recurr + 1)
    holder = []
    for i in np.arange(recurr):
        total_score = 0.0
        true_positive = 0.0
        for a in np.arange(len(realities)):
            score = realities[a]
            if score > probabilities[i] and score < probabilities[i + 1]:
                total_score = total_score + 1
                if label[a] == 1:
                    true_positive = true_positive + 1
        #print(total_score)
        #print(true_positive)
        if total_score == 0 or true_positive == 0:
            rate = 0
        else:
            rate = true_positive / total_score
        #print(rate)
        holder = np.append(holder,rate)
    return holder
```

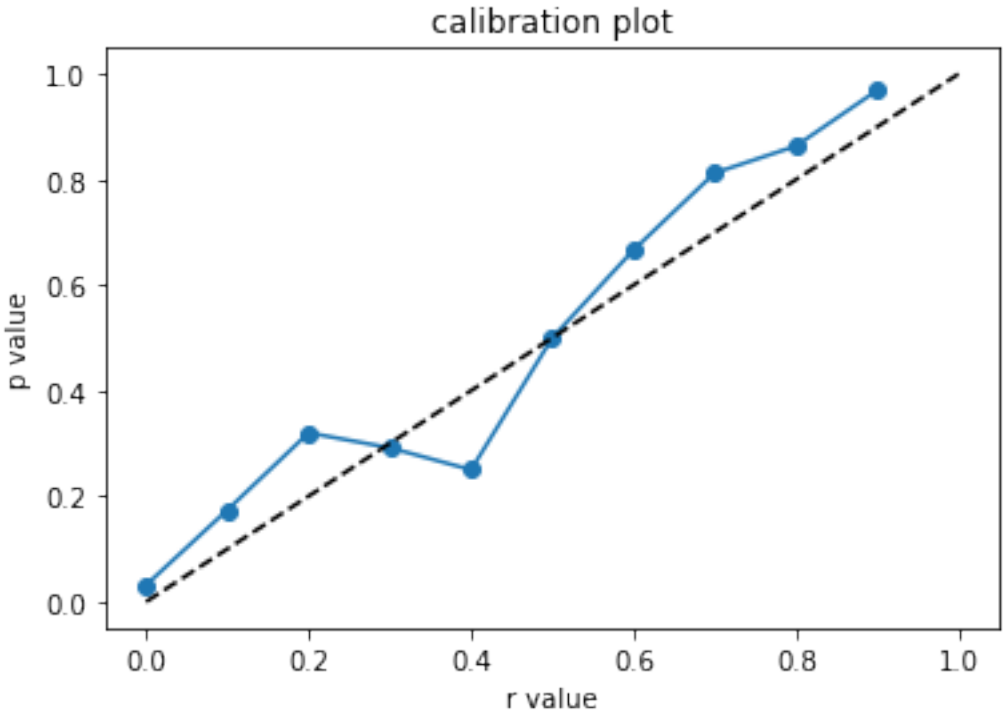
[134]:
```
#calibration_plot(np.arange(0,1.1,1),dataframe['0.94249713'],datafram[1])
first_column = np.asarray(dataframe[0])
second_column = np.asarray(dataframe[1])
outcome = calibration_plot(np.arange(0,1.1,0.1),first_column,second_column)
outcome
```

```
#dataframe
```

[134]:
```
array([0.03174603, 0.17391304, 0.32      , 0.29166667, 0.25      ,
       0.5       , 0.66666667, 0.8125    , 0.86363636, 0.96923077])
```

[135]:
```
plt.plot(np.arange(0,1,0.1),outcome,'-o')
plt.plot([0,1], [0, 1], "--k")
plt.xlabel('r value')
plt.ylabel('p value')
plt.title('calibration plot')
```

[135]:
```
Text(0.5, 1.0, 'calibration plot')
```

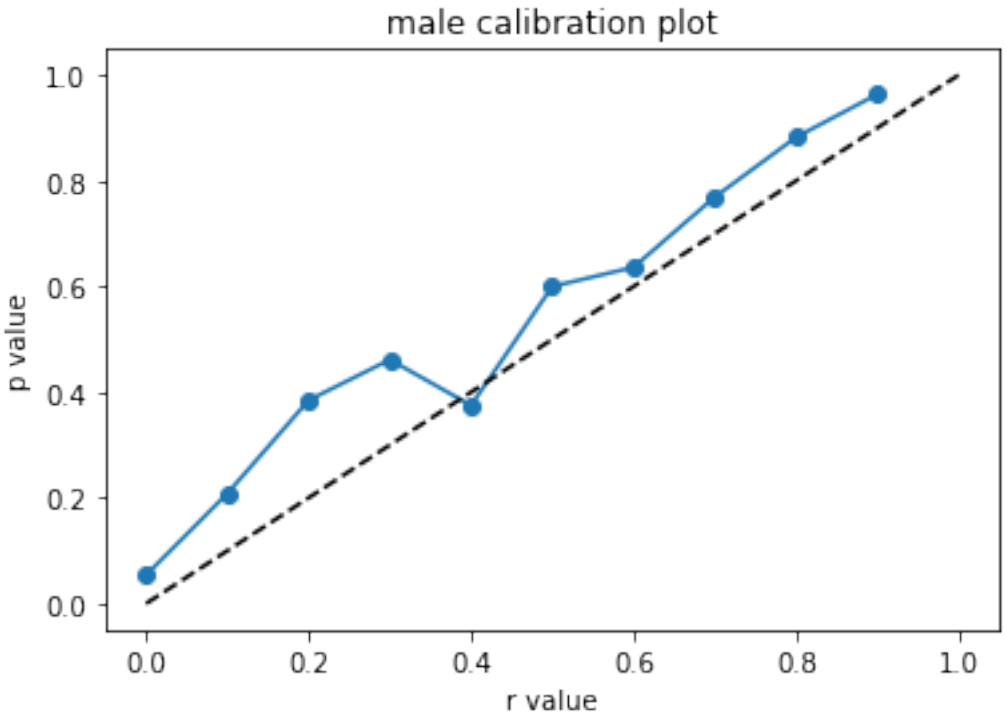


Write up: 1. Yes, the score function is close to calibration. From the plot we can see that the calibration line is roughly close to the perfect calibration line. So although it's not fully calibrated, the score function is close to calibration. 2.set:(0.29166667,0.3) For an ideal calibration line, we are expected the r value to be exactly equal to the p value. So the outcome of the re function is exactly the same as the positive rate.The x is the positive rate we are expected to get(if fully calibrated) and the y is the positive rate we actually get from the score funciton. So this difference in x and y tells us that the score function is only roughtly but not fully calibrated. And those differences in x and y exist because the score fucntion isn't perfectly calibrated.

[136]:
```
male = dataframe[dataframe[2]==1]
m_first_column = np.asarray(male[0])
m_second_column = np.asarray(male[1])
#male
```

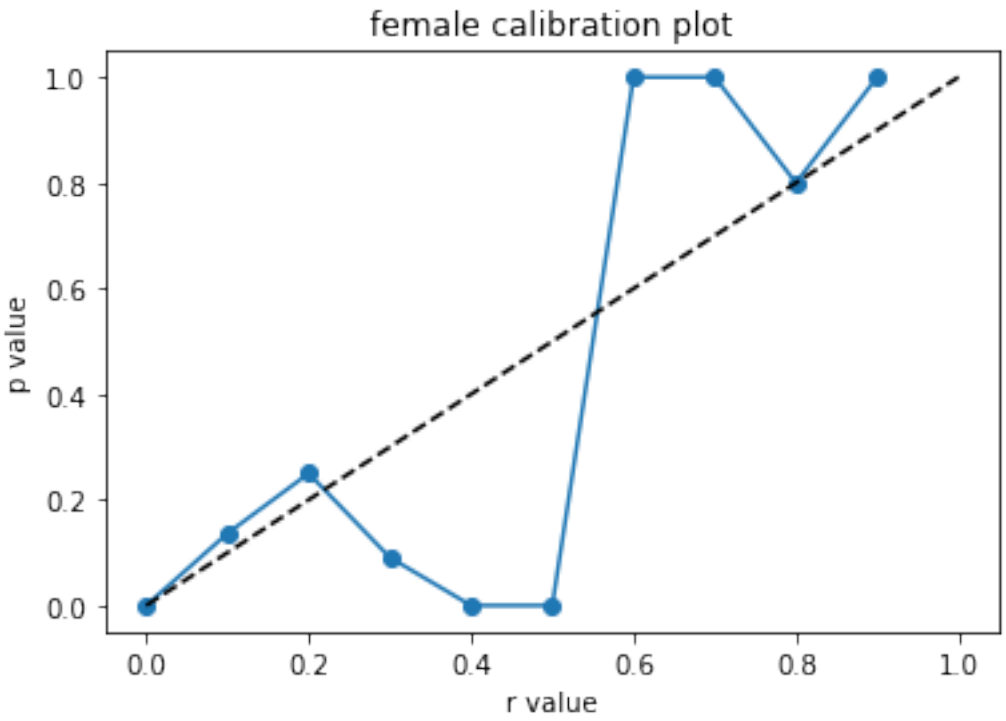
```
m_outcome = calibration_plot(np.arange(0,1.1,0.
 ↪1),m_first_column,m_second_column)
plt.plot(np.arange(0,1,0.1),m_outcome,'-o')
plt.plot([0,1], [0, 1], "--k")
plt.xlabel('r value')
plt.ylabel('p value')
plt.title('male calibration plot')
```

[136]: `Text(0.5, 1.0, 'male calibration plot')`



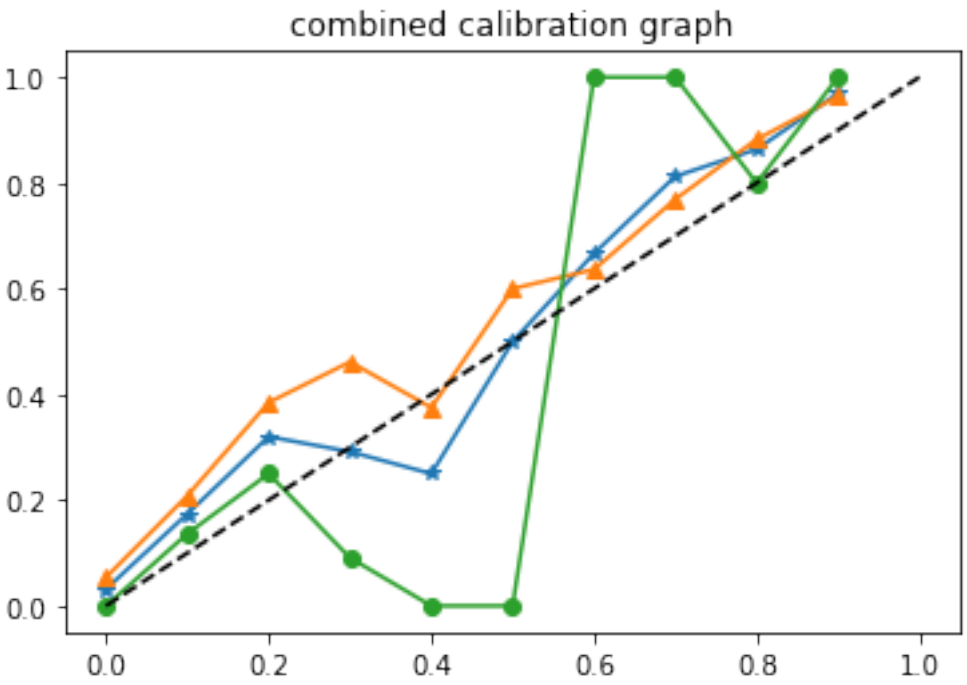
[141]:
```
female = dataframe[dataframe[2]==0]
f_first_column = np.asarray(female[0])
f_second_column = np.asarray(female[1])
#male
f_outcome = calibration_plot(np.arange(0,1.1,0.
 ↪1),f_first_column,f_second_column)
plt.plot(np.arange(0,1,0.1),f_outcome,'-o')
#plt.plot(np.arange(0,1,0.1),m_outcome,'-o')
plt.plot([0,1], [0, 1], "--k")
plt.xlabel('r value')
plt.ylabel('p value')
plt.title('female calibration plot')
```

[141]: `Text(0.5, 1.0, 'female calibration plot')`



[148]:
```
plt.plot(np.arange(0,1,0.1),outcome,'-*',np.arange(0,1,0.1),m_outcome,'-^',np.
↪arange(0,1,0.1),f_outcome,'-o')
plt.plot([0,1], [0, 1], "--k")
plt.title("combined calibration graph")
```

[148]:
```
Text(0.5, 1.0, 'combined calibration graph')
```



Write up: As we can see from the graph above, the distinction between the calibration by group are very distinct. The male group is way more calibrated compare to female and the female group have very bad clibration from the score function. Compare to the first graph(the not gender mix one), the male groups have more calibrated score function while the female groups have really bad calibratred score function. So for male groups, the score function is closer to calibration function. And the male group is a lot more calibrated compare to the female group.

[ ]: