Propensity Scoring Exercise in R

Question 7: The following data have been taken from nurses rounding in a facility. The time they spent with patients has been recorded. In addition, several characteristics of the patients have also been recorded and standardized. Do any of the nurses have a significant impact on overall satisfaction in the unit?

Steps to solve this problem:

1. Regression of satisfaction on Nurse 1 & Dx

First we will run ordinary logistic regression in R using the code:

model=lm(Satisfaction ~ Nurse1 + MI + CHF + Diabetes +Injuries + LungCancer +Age + UnderStaff + Pain , data = sat)
summary(model)

Outcome: satisfaction
Treatment: Nurse1
Confounders: MI, CHF, Diabetes, Injuries, Lung Cancer, Age, Understaff, pain

```
Call:
lm(formula = Satisfaction ~ Nurse1 + MI + CHF + Diabetes + Injuries +
    LungCancer + Age + UnderStaff + Pain, data = sat)

Residuals:
    Min      1Q  Median      3Q     Max
-47.380  -0.632   0.340   1.324   5.422

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.973219   0.232486 309.581  < 2e-16 ***
Nurse1       0.019915   0.007252   2.746  0.00607 **
MI           0.686428   0.121338   5.657 1.72e-08 ***
CHF          0.769107   0.114293   6.729 2.11e-11 ***
Diabetes     0.742842   0.113402   6.551 6.94e-11 ***
Injuries     1.636057   0.119442  13.698  < 2e-16 ***
LungCancer   2.641309   0.123826  21.331  < 2e-16 ***
Age          0.023165   0.002184  10.608  < 2e-16 ***
UnderStaff   2.196382   0.123285  17.816  < 2e-16 ***
Pain         1.614099   0.122912  13.132  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.826 on 2489 degrees of freedom
Multiple R-squared:  0.3634,	Adjusted R-squared:  0.3611
F-statistic: 157.8 on 9 and 2489 DF,  p-value: < 2.2e-16
```

Looking at the regression results we can say that nurse1, MI, CHF, Diabetes, Injuries, lung cancer, age, understaff, pain significantly predict satisfaction.

2. Propensity of Nurse 1 participating on Dx, age (Confounding)

Run ordinary logistic regression of Nurse1 on confounding using the code below:

propensity=lm(Nurse1 ~ MI + CHF + Diabetes +Injuries + LungCancer +Age + UnderStaff + Pain , data = sat)
summary(propensity)

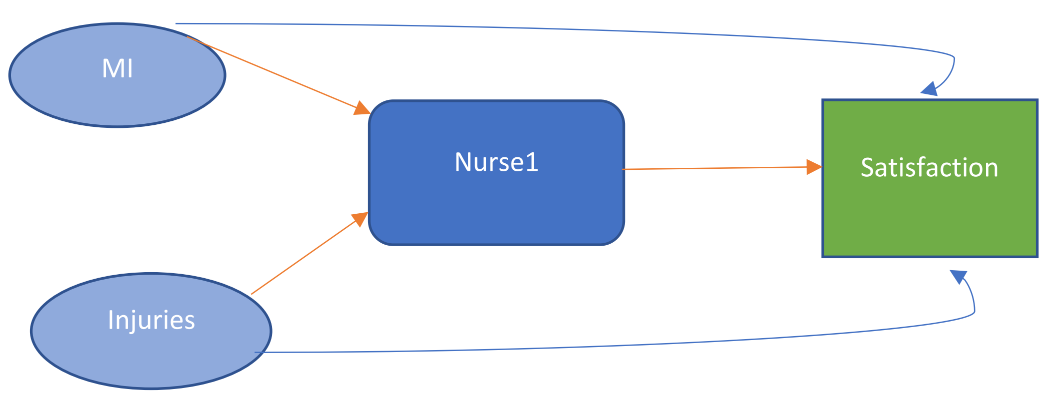
Outcome: Nurse 1 (treatment)

Confounders: MI, CHF, Diabetes, Injuries, Lung Cancer, Age, Understaff, pain

```
> propensity=lm(Nurse1 ~ MI + CHF + Diabetes +Injuries + LungCancer +Age + UnderStaff + Pain , data = sat)
> summary(propensity)

Call:
lm(formula = Nurse1 ~ MI + CHF + Diabetes + Injuries + LungCancer +
    Age + UnderStaff + Pain, data = sat)

Residuals:
     Min       1Q   Median       3Q      Max
-20.1361  -5.5597  -0.4757   5.5214  20.7098

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.449141   0.510701  38.083   <2e-16 ***
MI          -5.742573   0.314932 -18.234   <2e-16 ***
CHF          0.162256   0.315815   0.514    0.607
Diabetes    -0.001855   0.313369  -0.006    0.995
Injuries    -4.439621   0.317843 -13.968   <2e-16 ***
LungCancer   0.505592   0.342026   1.478    0.139
Age         -0.001564   0.006034  -0.259    0.796
UnderStaff  -0.324898   0.340617  -0.954    0.340
Pain         0.353133   0.339574   1.040    0.298
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.811 on 2490 degrees of freedom
Multiple R-squared:  0.1938,    Adjusted R-squared:  0.1912
F-statistic: 74.82 on 8 and 2490 DF,  p-value: < 2.2e-16
```

The regression results suggest that:

1. MI and Injuries predict Nurse1 time spend on patients.
2. Nurse 1 is spending less time with MI and Injured patients.
3. MI and Injuries and are affecting Nurse 1 time.
4. Nurse1 is affecting satisfaction. (from Step one regression)
5. Nurse 1 is highly correlated with MI and Injuries patients



Run Predict command. It will predict the number of minutes nurse 1 is spends with all patients using the code:

```
predicted=predict(propensity,sat)
```

If the data is binarized we skip this step, convert the equation prob=(predicted time -Min)/(Max-Min) in R :

```
prob=((predicted-min(predicted))/(max(predicted)-min(predicted)))
```

To use logistic regression, we need to binarized the treatment Nurse1:

```
#create a vector containing the original Nurse1 data
nurse1binary=sat$Nurse1

##we will binarize the original nurse1
a=which(nurse1binary>mean(nurse1binary))
b=which (nurse1binary <=mean(nurse1binary))
nurse1binary[a]=1
nurse1binary[b]=0
nurse1binary

##replace Nurse1 with the binarized
sat$Nurse1=nurse1binary
```

Run Logistic regression using the code below: this will give us the predicted probability

```
propensity=glm(Nurse1 ~ MI + CHF + Diabetes +Injuries + LungCancer +Age + UnderStaff + Pain ,
data = sat, family=binomial)
summary(propensity)
```

```
> sat$Nurse1=nurse1binary
> propensity=glm(Nurse1 ~ MI + CHF + Diabetes +Injuries + LungCancer +Age + UnderStaff + Pain , data = sat, family=binomial)
> summary(propensity)

Call:
glm(formula = Nurse1 ~ MI + CHF + Diabetes + Injuries + LungCancer +
    Age + UnderStaff + Pain, family = binomial, data = sat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8632  -0.9641  -0.5852   0.9885   1.9277

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.2157864  0.1497811   8.117 4.77e-16 ***
MI          -1.8270590  0.0915405 -19.959  < 2e-16 ***
CHF          0.0327068  0.0918459   0.356    0.722
Diabetes     0.0893806  0.0911241   0.981    0.327
Injuries    -0.9966703  0.0924085 -10.785  < 2e-16 ***
LungCancer   0.1166890  0.0997476   1.170    0.242
Age         -0.0002021  0.0017548  -0.115    0.908
UnderStaff  -0.0641306  0.0989606  -0.648    0.517
Pain         0.0981476  0.0986736   0.995    0.320
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3464.3  on 2498  degrees of freedom
Residual deviance: 2868.3  on 2490  degrees of freedom
AIC: 2886.3

Number of Fisher Scoring iterations: 4
```

3. Satisfaction & weighted regression using regression using propensity scores score (weight)

W=(T/p) +((1-T)/1-p)

It is the same as :

if Treatment =1 then 1/p, if Treatment =0 then 1/1-p

The code in R:

```
##Step 3 Weighting the regression using propensity score.
weights=predicted
sat=cbind(sat,weights)
sat$weights=(sat$Nurse1/sat$weights) + ((1-sat$Nurse1)/(1-sat$weights))
fix(sat)
```

Run logistic Regression with the weighted variable:

```
propensity2=glm(Nurse1 ~ MI + CHF + Diabetes +Injuries + LungCancer +Age + UnderStaff + Pain
, data = sat, family=binomial,weights=weights)
summary(propensity2)
```

```
> propensity2=glm(Nurse1 ~ MI + CHF + Diabetes +Injuries + LungCancer +Age + UnderStaff + Pain , data = sat, family=binomial, data=sat, we
ights=weights)
Error in glm(Nurse1 ~ MI + CHF + Diabetes + Injuries + LungCancer + Age +  :
  formal argument "data" matched by multiple actual arguments
> summary(propensity2)
Error in summary(propensity2) : object 'propensity2' not found
> propensity2=glm(Nurse1 ~ MI + CHF + Diabetes +Injuries + LungCancer +Age + UnderStaff + Pain , data = sat, family=binomial,weights=weigh
ts)
Warning message:
In eval(family$initialize) : non-integer #successes in a binomial glm!
> summary(propensity2)

Call:
glm(formula = Nurse1 ~ MI + CHF + Diabetes + Injuries + LungCancer +
    Age + UnderStaff + Pain, family = binomial, data = sat, weights = weights)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.773  -1.482  -1.271   1.513   2.985

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.0095550  0.0924015   0.103    0.918
MI          -0.0001549  0.0570091  -0.003    0.998
CHF         -0.0216572  0.0571342  -0.379    0.705
Diabetes     0.0039358  0.0567706   0.069    0.945
Injuries    -0.0054910  0.0576550  -0.095    0.924
LungCancer  -0.0067766  0.0617617  -0.110    0.913
Age          0.0001325  0.0010979   0.121    0.904
UnderStaff  -0.0180397  0.0616432  -0.293    0.770
Pain        -0.0237987  0.0615784  -0.386    0.699

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6927.4  on 2498  degrees of freedom
Residual deviance: 6927.0  on 2490  degrees of freedom
AIC: 6881.7

Number of Fisher Scoring iterations: 4
```

If we looked at the logistic regression after the propensity score with the treatment: Nurse 1 and covariates we can see that the effect of MI and injuries was removed.

If we run the original ordinary regression we can see that the regression is significant and the effect of MI and injuries on Nurse 1 is removed.

```
> model2=lm(Satisfaction ~ Nurse1 + MI + CHF + Diabetes +Injuries + LungCancer +Age + UnderStaff + Pain , data = sat,weights==weights )
> summary(model2)

Call:
lm(formula = Satisfaction ~ Nurse1 + MI + CHF + Diabetes + Injuries +
    LungCancer + Age + UnderStaff + Pain, data = sat, subset = weights ==
    weights)

Residuals:
    Min      1Q  Median      3Q     Max
-47.492  -0.621   0.342   1.301   5.416

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 72.143184   0.209424 344.485  < 2e-16 ***
Nurse1       0.283818   0.128377   2.211   0.0271 *
MI           0.687301   0.125376   5.482 4.63e-08 ***
CHF          0.770540   0.114350   6.738 1.98e-11 ***
Diabetes     0.737876   0.113484   6.502 9.54e-11 ***
Injuries     1.604818   0.117952  13.606  < 2e-16 ***
LungCancer   2.645039   0.123871  21.353  < 2e-16 ***
Age          0.023143   0.002185  10.593  < 2e-16 ***
UnderStaff   2.193501   0.123338  17.784  < 2e-16 ***
Pain         1.615703   0.122975  13.138  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.828 on 2489 degrees of freedom
Multiple R-squared:  0.3627,	Adjusted R-squared:  0.3604
F-statistic: 157.4 on 9 and 2489 DF,  p-value: < 2.2e-16
```

```
1         80.0      0   14.0   29.0     0     1         0        0          1  65.0          1     0
> predicted[1:5]
        1         2         3         4         5 
0.7837918 0.1786602 0.3422069 0.5667819 0.1938573 
> sat[2,]
# A tibble: 1 x 12
  Satisfaction Nurse1 Nurse2 Nurse3    MI   CHF Diabetes Injuries LungCancer   Age UnderStaff  Pain
         <dbl>  <dbl>  <dbl>  <dbl> <int> <int>    <int>    <int>      <int> <dbl>      <int> <int>
1         79.8      0   6.94   5.43     1     1        0        1          1  13.7          1     0
> sat[1,]
# A tibble: 1 x 12
  Satisfaction Nurse1 Nurse2 Nurse3    MI   CHF Diabetes Injuries LungCancer   Age UnderStaff  Pain
         <dbl>  <dbl>  <dbl>  <dbl> <int> <int>    <int>    <int>      <int> <dbl>      <int> <int>
1         80.0      0   14.0   29.0     0     1        0        0          1  65.0          1     0
> 
```

.By looking at the predicted 5 patients we see that patient 1 has a probability of 0.7837
And patients 2 is 0.17866 . if we looked at the original data for patient 1 and 2 then we can see that patient 2 has MI and Injuries and Patient 1 doesn't have MI or injuries. from the result we expected that the nurse will spend more time with patient 1 and not because from the regression before PS , MI and injuries were statistically significant and predict lower probability of nurse time.