STANFORD UNIVERSITY

STATS 245 Final Project

Does Homelessness Make People Less Physically Competent?

Name: Anyu Zhu
ID: 006402340
Date: August 2019

Summary

This report analyzes whether homelessness makes people less physically competent, based on certain data set. The relationship between homelessness and physical component score is evaluated using linear regression. Potential con-founders are tested to improve the accuracy of the model. Regression method and propensity scores are adopted during the process and significance of the factors included have been detailedly tested.

0.1 Relationship Between Factors

0.1.1 The Relationship Between Exposure(homeless) and Outcome(pcs)

We regress the exposure "homeless" onto outcome "pcs", and get the first regression model:

$$pcs = -2.064 * homeless + 49.001$$

The p-value is 0.04216, which indicates that the relationship between homelessness and physical competence is significant. There is a negative correlation between "pcs" and "homeless". Homeless people tend to have lower physical component score. The adjusted R-squared of the model is 0.006926, which is quite small. We need to move on to explore the con-founders and adjust the model.

0.1.2 Selection of Con-founders

Apart from the exposure and outcome, there are in total fourteen factors where we should select the potential con-founders to improve the model. We adpoupt the standard that: if the changes in β is larger than 10% after including the factor into our regression model, we see the factor as a possible potential con-founder and move on to furtehr selection process.

Variable	β_1 - β_1'	Ratio	larger than 10%
age	-0.4124	0.1998	Yes
anysubstance	-1.2021	0.5824	Yes
a15a	0.2192	-0.1062	Yes
a15b	-1.4643	0.7094	Yes
cesd	-0.5392	0.2613	Yes
d1	-0.3616	0.1752	Yes
daysanysub	-1.2119	0.5871	Yes
daysdrink	-1.2053	0.5840	Yes
drugrisk	-0.1139	0.0552	No
female	0.3534	-0.1712	Yes
g1b	-0.3922	0.1900	Yes
i1	-0.9443	0.4575	Yes
mcs	-0.1540	0.0746	No
pss_fr	-0.2309	0.1119	Yes

By comparing the ratio of changes in β_1, we can select out the possible potential con-founders: if the ratio is larger than 10%, then the factor can be selected as a potential con-founder. We then move on to test whether the factors can improve the model by comparing the p-value and adjusted R-squared of the regression model after adding the factor. The following part only displays the potential con-founders which improves the regression model (significant p-value and larger adjusted R-square), the evaluation process of the rest factors will be included in appendix.

- age: the regression model after including the age factor: The p-value is reduced to 1.412e-06 and adjusted R-squared increased to 0.0539.

$$pcs = -1.6515 * homeless - 0.31074 * age + 59.88949$$

- a15b: the regression model after considering number of nights on streets: the p-value is reduced to 2.039e-08 and adjusted R-squared increased to 0.0763.

$$pcs = -0.3056 * homeless - 0.2986 * age - 0.0652 * a15b + 59.4655$$

- cesd: after adding depression scale to the improved model above: the p-value is decreased to 3.888e-15 and adjusted R-squared increased to 0.1437.

$$pcs = -0.1990 * homeless - 0.3019 * age - 0.0461 * a15b - 0.2304 * cesd + 66.9135$$

- d1: the times hospitalized for medical problems is also counted as a con-founder, the p-value is decreased to less than 2.2e-16 and adjusted R-squared is 0.1787 now.

$$pcs = -0.0817 * homeless - 0.2598 * age - 0.0423 * a15b - 0.2172 * cesd - 0.3396 * d1 + 65.9327$$

- female: the regression model after considering the gender: The p-value is less than 2.2e-16 and adjusted R-squared increased to 0.1865.

$$pcs = -0.3484 * homeless - 0.2531 * age - 0.0420 * a15b - 0.2011 * cesd - 0.3359 * d1 - 2.5465 * female + 65.8722$$

- g1b: The next con-founder which can improve the regression model is thoughts of suicide, the p-value is less than 2.2e-16 and adjusted R-squared increased to 0.1875.

$$pcs = -0.1586 * homeless - 0.2590 * age - 0.0440 * a15b - 0.1855 * cesd - 0.3325 * d1 - 2.3870 * female - 1.3638 * g1b + 65.8371$$

- i1: The next con-founder we need to consider is the daily consumption of drinks: p-value is significantly small and adjusted R-squared increased to 0.1879.

$$pcs = -0.0301 * homeless - 0.2472 * age - 0.0405 * a15b - 0.1806 * cesd - 0.3158 * d1 - 2.5434 * female - 1.2646 * g1b - 0.0282 * i1 + 65.6235$$

In all, after all evaluation process, we selected seven con-founders in total: age, number of nights on street (a15b), depression scale (cesd), times hospitalized for medical situation (d1), gender (female), thoughts of suicide (g1b), and daily drinks consumption (i1).

0.1.3 Relationship Between Con-founders and Exposure

We then evaluate the relationship between con-founders and exposure. Since exposure: homeless has binary values, for binary factors (gender, thoughts of suicide), we compare the number of homeless people in different categories; for factors which have continuous value, we adopt logistic regression and compare the coefficient of the factor in the regression model. The following are the results:

Binary con-founders:

Factors	male	female	χ^2 and p-value	not suicide	suicide	χ^2 and p-value
not homeless	177	67	129.6	189	55	100.63
homeless	169	40	$<$ 2.2e-16	137	72	$<$ 2.2e-16

From the table, we can see that: The chi-square test shows the significance in differences between frequencies. Although male takes up a much larger proportion of sample size, the proportion of homeless among male is also larger than that of female. Number of people who had thoughts of suicide is smaller in sample, among these people who had negative thoughts, homeless people takes a larger proportion. We then draw to the conclusion that negative thoughts and homelessness has positive correlation, and gender also has a significant effect on exposure.

Continuous Con-founders:

Con-founder	age	cesd	d1	i1
Coefficient	0.0224	0.0141	0.0253	0.0271
p-value	0.0685	0.0647	0.1921	1.71e-6

Based on the coefficients generated from logistic regression, we can see that these four con-founders may potentially cause homelessness. From the p-values, we can see that except for "times hospitalized for health problems" (d1) is much less significant than others, number of drinks is a important driving factor of homelessness, and cesd, age also has effect. In further process, we will consider whether d1 is an appropirate con-founder.

* A special factor "a15b" fails to run the logistic regression, we count the population of homeless people in different range of "nights on street": 53.86% of sample (244 people) have 0 nights on street and is not homeless, the rest are all homeless and spend different numbers of nights on street. Thus, "a15b" has a positive correlation with homelessness.

0.1.4 Relationship Between Con-founders and Outcome

The following is the evaluation of the relationship between con-founders and outcome: physical component score. Since physical competence is of continuous values, we calculate the correlation between pcs and con-founders (except gender and thoughts of suicide, which are binary con-founders).

Continuous con-founders:

cor	age	a15b	cesd	d1	i1
pcs	-0.2287	-0.1915	-0.2927	-0.2576	-0.1963

From the correlation coefficients, the above con-founders have negative correlation with physical component score: older people tend to have lower physical competence; increasing number of nights on street cause slightly decrease in physical competence; higher depression scale does harm to physical competence; number of hospitalized indicates weaker physical competence, and drinking decrease the physical competence score.

Binary con-founders: We calculate the average physical component score within each category.

mean	female	male	not suicide	suicide
pcs	45.0164	48.9862	49.0408	45.5014

By the average values, we can draw to the conclusion that male tend to have higher physical competence; People who never experienced negative thoughts related to suicide have higher physical component scores.

0.2 Casual Inference and Improved Model

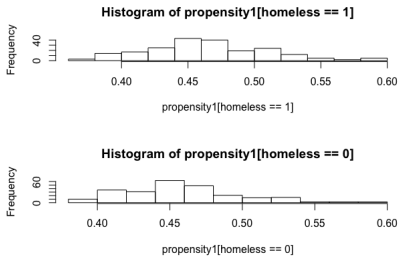
After basic selection process of potential con-founders, our current model is:

$$\begin{aligned} pcs = &- 0.0301 * homeless - 0.2472 * age - 0.0405 * a15b - 0.1806 * cesd - 0.3158 * d1 \\ &- 2.5434 * female - 1.2646 * g1b - 0.0282 * i1 + 65.6235 \end{aligned}$$

To further increase the accuracy of the model, we use propensity scores to test the validity of all the factors included in the model. Since the number of con-founders is quite large, we use propensity score measure the sample and the con-founders.

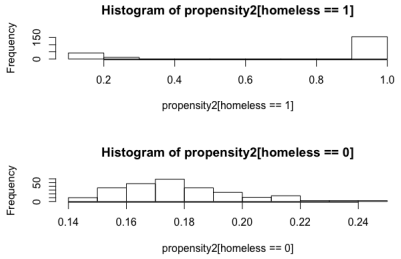
- Test for age:

  From the histogram generated from the propensity scores, we can see there is reasonable amount of overlapping between homeless and non-homeless group, so we keep age as an effective con-founder.



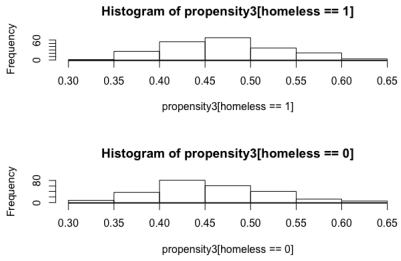
- Test for a15b:

  From the histogram generated from the propensity scores, the amount of overlapping areas is small, indicating a lack of good balance. Thus we remove "a15b" from con-founders. Possible reason: number of nights on streets is so hlghly related to the factor "homeless" itself.



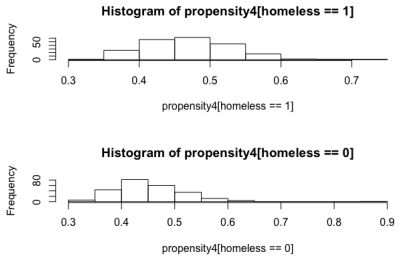
- Test for cesd:

  From the histogram of propensity scores of age and cesd, we can see a significant amount of overlap between homeless and non-homeless, indicating that cesd also has similar effect on both homeless and non-homeless. So we keep it as a con-founder.



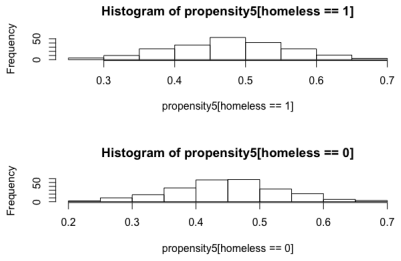
- Test for d1:

  From the histogram of propensity scores including d1, we can see that the area of overlapping decreased significantly. Thus we remove d1 from confounders. Possible reason: times of hospitalized for medical problems strongly indicates the level of pcs, and is highly related with other factors.



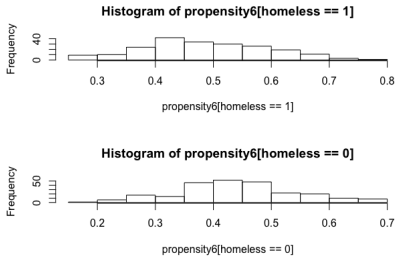
- Test for gender:

  From the histogram of the propensity scores of model including gender, we can see the propensity scores for both homeless group and non-homeless group overlap well. This indicates that gender affect both groups at similar level. So we keep it as an effective con-founder.



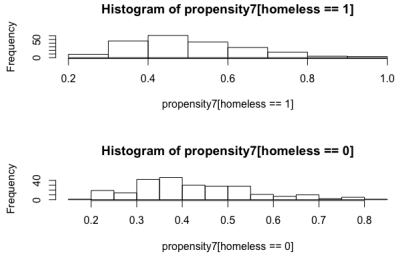
- Test for g1b:

  From the histogram of the propensity scores of model including "g1b", we can see that the area of overlapping is significant, indicating the similarity of two sample groups. Thus we include g1b as a significant con-founder.



- Test for i1:

  From the histogram of propensity scores of model including "i1", we can observe a significant overlap of two groups. Data won't extrapolate outside the range when we adjust Thus we keep i1 as an effective con-founder.



After going through all the con-founders and going through the histograms, we include the propensity score as a continuous covariate in the regression model. Our regression model looks like the following:

$$pcs = -0.8608 * homeless - 14.8605 * propensity + 55.3018$$

The p-value is 3.559e-05, which is far less than 0.05, indicating the significance of the model. From the histogram of propensity scores above, we can see that there is a good matching between homeless group and

non-homeless group. The coefficient of homeless in this model is negative, which means being homeless has negative effect on physical competence.

0.3 Conclusion

After all the analysis process, to answer the question: Does homelessness makes people less physical competent? We can draw to the conclusion that with a negative coefficient of the factor "homeless", being homeless has negative effect on people's physical competence. However, we cannot judge their relationship by simply conducting regression between homelessness and people's physical component scores. Some other factors also plays important roles: age; depression scale; gender; thoughts of getting suicide; number of daily drinks. Through doing regression, calculating correlations, and comparing propensity scores, we find these factors has significant on both the exposure(homeless) and the outcome(physical competence), they all exert similar effect on both the homeless group and non-homeless groups. From our final model with propensity, we can see that apart from the factor "homeless" itself, these factors in total actually decrease people's physical competence significantly.

0.4 Appendix

0.4.1 Code

```
health<-read.csv("health.csv")
attach(health)
model1<-lm(pcs~homeless)
summary(model1) # -2.064

c1<-lm(pcs~homeless + age)
summary(c1) # -1.65156
(-2.064 + 1.65156)/(-2.064) # 0.1998256

c2<-lm(pcs~homeless + anysubstatus)
summary(c2) # -0.8619
(-2.064 + 0.8619)/(-2.064) # 0.5824128

c3<-lm(pcs~homeless + a15a)
summary(c3) # -2.283210
(-2.064 + 2.283210)/(-2.064) # -0.1062064

c4<-lm(pcs~homeless + a15b)
summary(c4) # -0.59970
(-2.064 + 0.59970)/(-2.064) # 0.7094477

c5<-lm(pcs~homeless + cesd)
summary(c5) # -1.52478
(-2.064 + 1.52478)/(-2.064) # 0.26125

c6<-lm(pcs~homeless + d1)
summary(c6) # -1.7024
(-2.064 + 1.7024)/(-2.064) # 0.1751938

c7<-lm(pcs~homeless + daysanysub)
summary(c7) # -0.852082
(-2.064 + 0.852082)/(-2.064) # 0.5871696

c8<-lm(pcs~homeless + daysdrink)
summary(c8) # -0.8586766
(-2.064 + 0.8586766)/(-2.064) # 0.5839745

c9<-lm(pcs~homeless + drugrisk)
summary(c9) # -1.9501
(-2.064 + 1.9501)/(-2.064) # 0.05518411
```

```
c9<-lm(pcs~homeless + drugrisk)
summary(c9) # -1.9501
(-2.064 + 1.9501)/(-2.064) # 0.05518411

c10<-lm(pcs~homeless + female)
summary(c10) # -2.4174
(-2.064 + 2.4174)/(-2.064) # -0.1712209

c11<-lm(pcs~homeless + g1b)
summary(c11) # -1.6718
(-2.064 + 1.6718)/(-2.064) # 0.1900194

c12<-lm(pcs~homeless + i1)
summary(c12) # -1.11969
(-2.064 + 1.11969)/(-2.064) # 0.4575145

c13<-lm(pcs~homeless + mcs)
summary(c13) # -1.91003
(-2.064 + 1.91003)/(-2.064) # 0.07459787
```

```r
c14<-lm(pcs~homeless + pss_fr)
summary(c14) # -1.8331
(-2.064 + 1.8331)/(-2.064) # 0.1118702

model2<-lm(pcs~homeless + age)
summary(model2)

# delete anysubstatus
delete1<-lm(pcs~homeless + age + anysubstatus)
summary(delete1)

# delete a15a
delete2<-lm(pcs~homeless + age + a15a)
summary(delete2)

model3<-lm(pcs~homeless + age + a15b)
summary(model3)

model5<- lm(pcs~homeless + age + a15b + cesd)
summary(model5)
model6<- lm(pcs~homeless + age + a15b + cesd + d1)
summary(model6)

delete3<-lm(pcs~homeless + age + a15b + cesd + d1 + daysanysub)
summary(delete3)

delete4<-lm(pcs~homeless + age + a15b + cesd + d1 + daysdrink)
summary(delete4)

model7<-lm(pcs~homeless + age + a15b + cesd + d1 + female)
summary(model7)

model8<-lm(pcs~homeless + age + a15b + cesd + d1 + female + g1b)
summary(model8)

model9<-lm(pcs~homeless + age + a15b + cesd + d1 + female + g1b + i1)
summary(model9)

delete5<-lm(pcs~homeless + age + a15b + cesd + d1 + female + g1b + i1 + pss_fr)
summary(delete5)
# relationship btw confounders and exposure
xtabs(~ homeless + female, data = health)
xtabs(~ homeless + g1b, data = health)

logi1<-glm(homeless~age, data = health, family = "binomial")
summary(logi1)

logi3<-glm(homeless~cesd, data = health, family = "binomial")
summary(logi3)

logi4<-glm(homeless~d1, data = health, family = "binomial")
summary(logi4)

logi5<-glm(homeless~i1, data = health, family = "binomial")
summary(logi5)
obs<-c(177,67,169,40)
obs2<-c(189, 55, 137, 72)
exp<-c(0.25, 0.25, 0.25,0.25)
chisq.test(obs, p = exp)
chisq.test(obs2, p = exp)
```

```r
# con-founder and outcome
cor(age,pcs)
cor(a15b,pcs)
cor(cesd,pcs)
cor(d1,pcs)
cor(i1,pcs)

mean(pcs[female == 0])  # male
mean(pcs[female == 1])  # female
mean(pcs[g1b == 0])    # not suicide
mean(pcs[g1b == 1])   # suicide

glm1<-glm(homeless ~ age, family = binomial, data = health)
propensity1 = glm1$fitted
summary(propensity1)
par(mfrow = c(2,1))
hist(propensity1[homeless == 1])
hist(propensity1[homeless == 0])
```

```r
# drop a15b
glm2<-glm(homeless ~ age + a15b, family = binomial, data = health)
propensity2 = glm2$fitted
summary(propensity2)
par(mfrow = c(2,1))
hist(propensity2[homeless == 1])
hist(propensity2[homeless == 0])

glm3<-glm(homeless ~ age + cesd, family = binomial, data = health)
propensity3 = glm3$fitted
summary(propensity3)
par(mfrow = c(2,1))
hist(propensity3[homeless == 1])
hist(propensity3[homeless == 0])
```

```r
# drop d1
glm4<-glm(homeless ~ age + cesd + d1, family = binomial, data = health)
propensity4 = glm4$fitted
summary(propensity4)
par(mfrow = c(2,1))
hist(propensity4[homeless == 1])
hist(propensity4[homeless == 0])

glm5<-glm(homeless ~ age + cesd + female, family = binomial, data = health)
propensity5 = glm5$fitted
summary(propensity5)
par(mfrow = c(2,1))
hist(propensity5[homeless == 1])
hist(propensity5[homeless == 0])

glm6<-glm(homeless ~ age + cesd + female + g1b, family = binomial, data = health)
propensity6 = glm6$fitted
summary(propensity6)
par(mfrow = c(2,1))
hist(propensity6[homeless == 1])
hist(propensity6[homeless == 0])
```

```r
glm7<-glm(homeless ~ age + cesd + female + g1b + i1, family = binomial, data = health)
propensity7 = glm7$fitted
summary(propensity7)
par(mfrow = c(2,1))
hist(propensity7[homeless == 1])
hist(propensity7[homeless == 0])

lm=lm(pcs~homeless+propensity7, data = health)
summary(lm)
```

0.4.2 Additional selection process of con-founders

- After including "anysubstatus": p-value: 0.003393; Adjusted R-squared: 0.04309. p-value increased while adjusted R-squared decreased compared with the model including only age. So we do not take it as a con-founder.
- After including "a15a": p-value: 4.654e-06; Adjusted R-squared: 0.05303. p-value increased while adjusted R-squared decreased compared with that only includes age. So we do not consider it as a con-founder.
- After including "daysanysub": p-value: 8.36e-11; Adjusted R-squared: 0.1979. p-value increased and adjusted R-squared decreased compared with the model after including "d1". So we drop this factor.
- After including "pss_fr": Adjusted R-squared: 0.187, a little bit smaller than that of the model after including "i1", which means less accuracy of the model. So we drop "pss_fr".