CaravanHealth

Amy Jung

2022-04-26

```
library(data.table)

library(sandwich)
library(lmtest)

library(AER)

library(ggplot2)
library(patchwork)

library(stargazer)
library(cregg)
```

Load Data

```
##        patient_id date_of_birth patient_age gender zip_code date_of_delivery
##     1:     101511    1992-02-24          24      F    37027       2017-01-23
##     2:     101143    1992-08-22          25      F    37013       2017-09-16
##     3:     193330    1983-02-22          34      F    37250       2017-10-07
##     4:     142808    1978-02-25          38      F    37027       2017-01-12
##     5:     142808    1970-09-11          47      F    37250       2017-12-13
##    ---
## 64149:     102809    1989-07-22          27   <NA>    37122       2017-02-02
## 64150:     136397    1981-09-28          35      F    37013       2017-08-22
## 64151:     190534    1976-06-25          40      F    37250       2017-04-21
## 64152:     137050    1969-05-11          47      F    37250       2017-03-06
## 64153:     157354    1998-08-29          18      F    37250       2017-08-21
##        icd_10_diagnosis_code length_of_stay total_claim_cost
##     1:                   O80              5         15590.55
##     2:                   O82              7         37930.46
##     3:                   O82              7         24965.33
##     4:                   O80              4          8669.38
##     5:                   O82              8         21954.30
##    ---
## 64149:                   O80              4         11357.88
## 64150:                   O82              8         22995.74
## 64151:                   O69              6         18503.37
## 64152:                   O80              7         11688.88
## 64153:                   O69             11         31829.80
```

Regression

Model 1: only length_of_stay

The intuition is that length_of_stay would be the best indicator of total_claim_cost (more time at hospital = greater cost).

Null hypothesis: no correlation between length_of_stay and total_claim_cost.

```
# model with only length_of_stay

mod_1 = lm(total_claim_cost ~
                 length_of_stay, data=d)

mod_1$cluster_se <- sqrt(diag(vcovCL(mod_1)))

stargazer(
  mod_1,
  se = list(mod_1$cluster_se),
  type = 'text'
)
```

```
##
## ================================================
##                         Dependent variable:
##                     ----------------------------
##                           total_claim_cost
## ------------------------------------------------
## length_of_stay              4,228.162***
##                               (47.564)
##
## Constant                   -4,302.775***
##                              (231.021)
##
## ------------------------------------------------
## Observations                   64,153
## R2                             0.134
## Adjusted R2                    0.134
## Residual Std. Error   25,907.470 (df = 64151)
## F Statistic         9,895.957*** (df = 1; 64151)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Interpretation: The correlation between length_of_stay and total_claim_cost is a statistically significant. An increase in 1 day in the hospital leads to an increase in $4,228.16 in claim costs, according to model 1.

Model 2: length_of_stay + icd_10_diagnosis_code + zip_code + patient_age

The intuition is that icd_10_diagnosis_code, zip_code, and patient_age would also have an impact on total_claim_costs – with ICD codes used for billing purposes, zip codes relating to SDH (social determinants of health), and patient age maybe correlating with more procedures needed for older mothers.

```r
# model with other variables

mod_2 = lm(total_claim_cost ~
                 length_of_stay +
                 icd_10_diagnosis_code +
                 as.factor(zip_code) +
                 patient_age,
               data=d)

mod_2$cluster_se <- sqrt(diag(vcovCL(mod_2)))

stargazer(
  mod_2,
  se = list(mod_2$cluster_se),
  type = 'text'
)
```

```
##
## ======================================================
##                               Dependent variable:
##                           -----------------------------
##                                total_claim_cost
## ------------------------------------------------------
## length_of_stay                   2,452.569***
##                                    (53.035)
##
## icd_10_diagnosis_codeO67         12,233.250***
##                                    (711.954)
##
## icd_10_diagnosis_codeO69         8,857.200***
##                                    (726.090)
##
## icd_10_diagnosis_codeO80         -9,072.846***
##                                    (453.496)
##
## icd_10_diagnosis_codeO82         7,714.229***
##                                    (582.724)
##
## as.factor(zip_code)37013            685.245
##                                   (2,968.024)
##
## as.factor(zip_code)37027            446.126
##                                   (2,971.945)
##
## as.factor(zip_code)37122            831.668
##                                   (2,976.875)
##
## as.factor(zip_code)37167            713.181
##                                   (2,974.235)
##
## as.factor(zip_code)37250            615.119
##                                   (2,968.470)
##
```

```
## patient_age                           -18.152**
##                                        (9.161)
##
## Constant                             7,644.931**
##                                      (3,010.982)
##
## -----------------------------------------------------
## Observations                           64,153
## R2                                      0.206
## Adjusted R2                             0.206
## Residual Std. Error         24,806.080 (df = 64141)
## F Statistic             1,511.571*** (df = 11; 64141)
## =====================================================
## Note:                     *p<0.1; **p<0.05; ***p<0.01
```

Model 2 Constant is when: ICD code O03, length of stay 0, patient age 0, zip code 35896.

Interpretation: The correlation of length_of_stay, icd_10_diagnosis_code, patient_age to total_claim_cost is statistically significant. The correlation between zip_code and total_claim_cost is not significant.

Interesting insights according to Model 2: - with the other variables factored in (compared to Model 1), the coefficient for length_of_stay decreases by 60% (from 4,228.162 to 2,452.569), which indicates omitted variable bias in Model 1 - ICD code O80 has a negative significant coefficient (-9,072.846) - an increase in patient age decreases total_claim_cost (-18.152)