Oral and Facial Study

Tianjian Xie, Amstrong Wang, Haochen Pan, Nuo Chen, Haochen Zhu

2022-10-28

Abstract

This study looks at the GCS and the ISS for Maine Portland medical center and how they associate with patients from rural/urban. Comparison based on the KS test of the urban and rural population showed no statistical evidence that the distribution of ISS nor GCS scores differ for patients from rural and urban locations. Regression analysis showed an association between whether a patient has a rural or urban origin and ISS after accounting for some factors. However, this result should be taken with caution since there is a potential confounding effect that individuals closer to Maine Medical Center are less likely to have a more severe injury, which would result in biased observation that the urban locals have a lower ISS. Since the data lacks information on the actual distance from the hospital, the result may be falsely attributing ISS differences to urban origin instead of distance. The regression analysis for GCS suggests no association between GCS and urban versus rural origin. However, the model checking results show substantial enough concerns with model fit. Overall, there does not seem to be evidence of a difference between each score and rural or urban patient origin, and that future modeling that includes confounding distance variables is recommended to make statements about the relationship between urban and rural populations and ISS, GCS, or days in the hospital.

Introduction

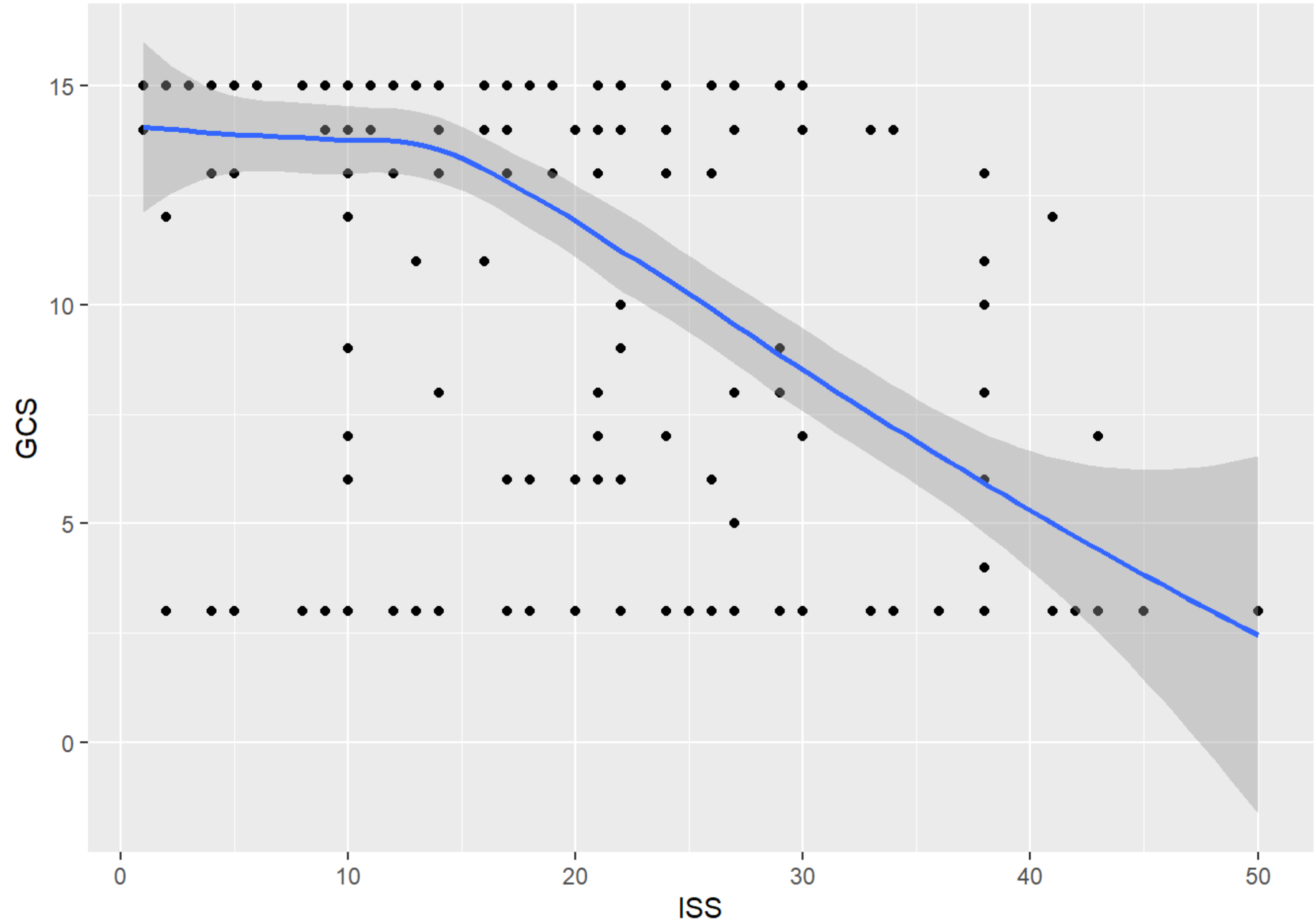
Our client is David Cartier from Boston University Henry M. Goldman School of Dental Medicine. The client's goal is to analyze the different types of facial fractures from Maine Portland medical center and compare injuries between rural and urban populations.The client asked us to compare two different injury assessment scores between rural and urban populations.These scores are the injury severity score (ISS) and Glasgow coma scale (GCS). The client also asked us to assess the relationship between injury mechanisms and these two scores, as well as explore factors affecting the frequency of hospital admissions.

Data Description

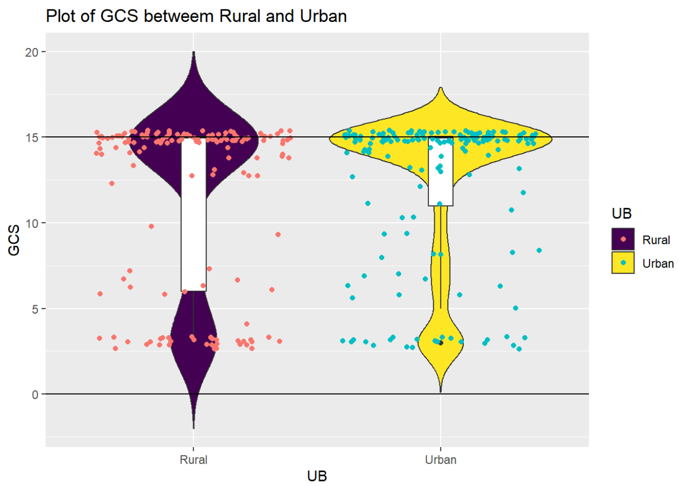
The data consists of de-identified medical records for 318 patients from Maine Medical Center. Patients are classified as coming from rural areas or urban areas based on their Zip code and 2010 census data.

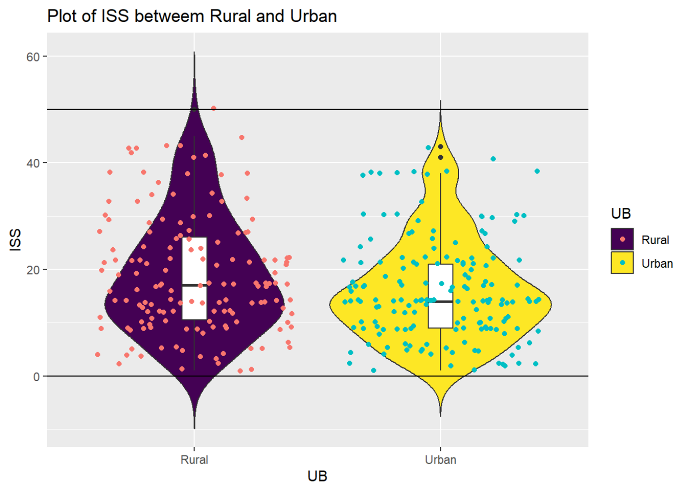
The Glasgow Coma Scale (GCS) is used to objectively describe the extent of impaired consciousness in all types of acute medical and trauma patients. The scale assesses patients according to three aspects of responsiveness:eye-opening, motor, and verbal responses. This scale is from 1 to 15 which 13-15 being Mild, 9-12 being Moderate, and 3-8 being Severe.

All injuries are assigned from an internationally recognized dictionary that describes over 2000 injuries. Multiple injuries are scored by adding together the squares of the three highest AIS scores. The ISS is the Injury severity score, which can range from 1 to 75, but we only have from 1 to 50.

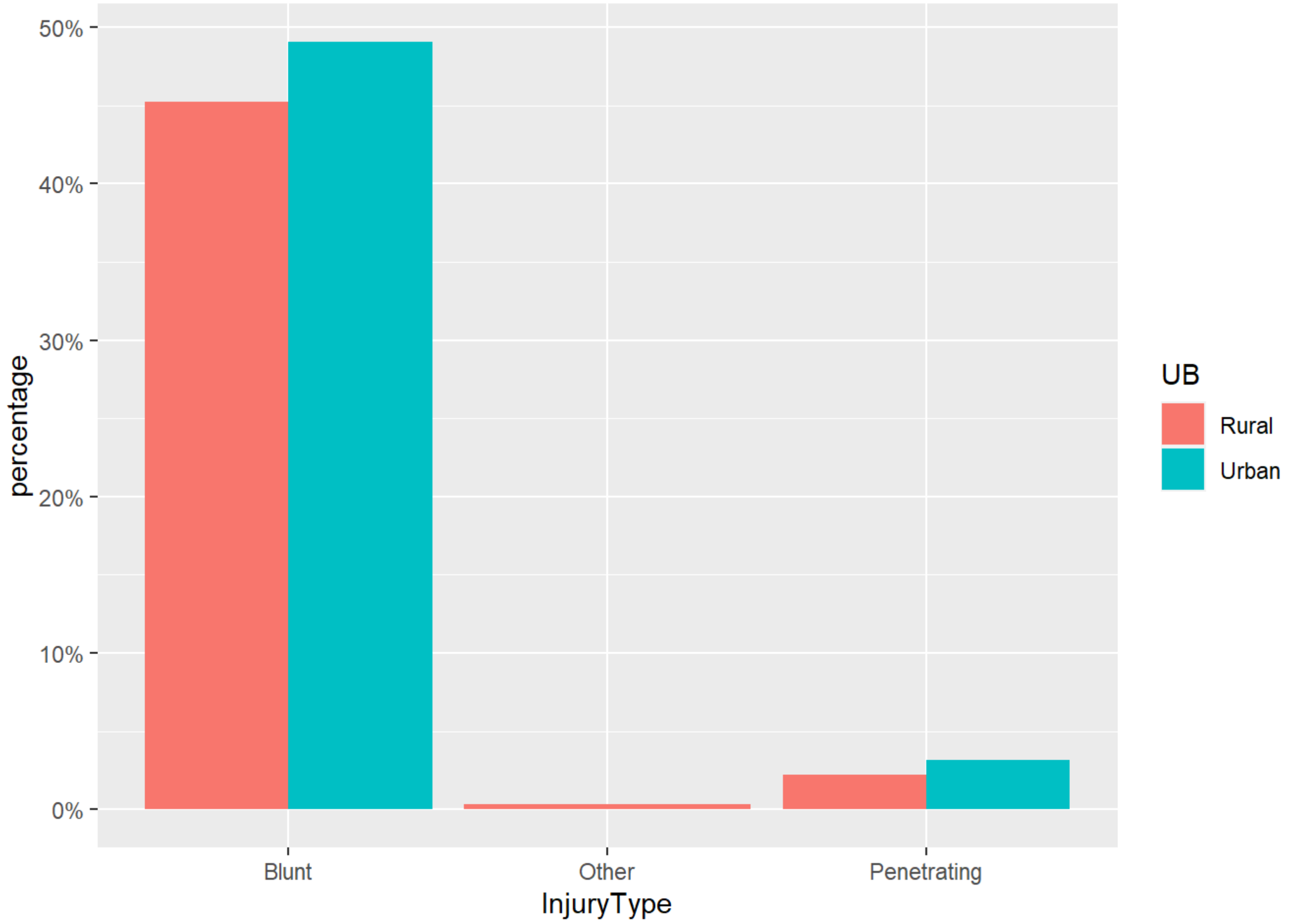


Relationship between ISS and GCS





Violin plots for ISS and GCS between urban or rural. For ISS, urban or rural patients are likely to have similar ISS, the means are both around 15, and the urban patients' ISS is more concentrated on the mean. We see that most of the data is concentrated near 15, which corresponds to less-severe injury. The concentration is somewhat larger for Urban compared to Rural. One possible explanation, if we were to speculate, could be that people close to Maine Medical Center might go to Maine Medical Center even if they have less severe injuries, while people further away might not bother traveling that distance if they have less severe injuries. The top of the boxes are high up to the top of the plot and did not show the mean of the data, the violins are also in a strange shape, which means most patients have the highest GCS. So for the GCS, we need to use Kolmogorov–Smirnov test to find the correlation between urban and rural.



Method

```
##
##  Welch Two Sample t-test
##
## data:  df$ISS and df$GCS
## t = 8.6026, df = 441.76, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.231263 6.737092
## sample estimates:
## mean of x mean of y
##  17.37975  11.89557
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$ISS
## W = 0.94956, p-value = 6.119e-09
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$GCS
## W = 0.64812, p-value < 2.2e-16
```

The Shapiro-Wilk test is a way to tell if a random sample comes from a normal distribution. From the output, the p-value < 0.05 implying that the distribution of the data are significantly different from normal distribution. In other words, we can not assume the normality.

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  df$ISS[df$UB == "Urban"] and df$ISS[df$UB == "Rural"]
## D = 0.12844, p-value = 0.1483
## alternative hypothesis: two-sided
```

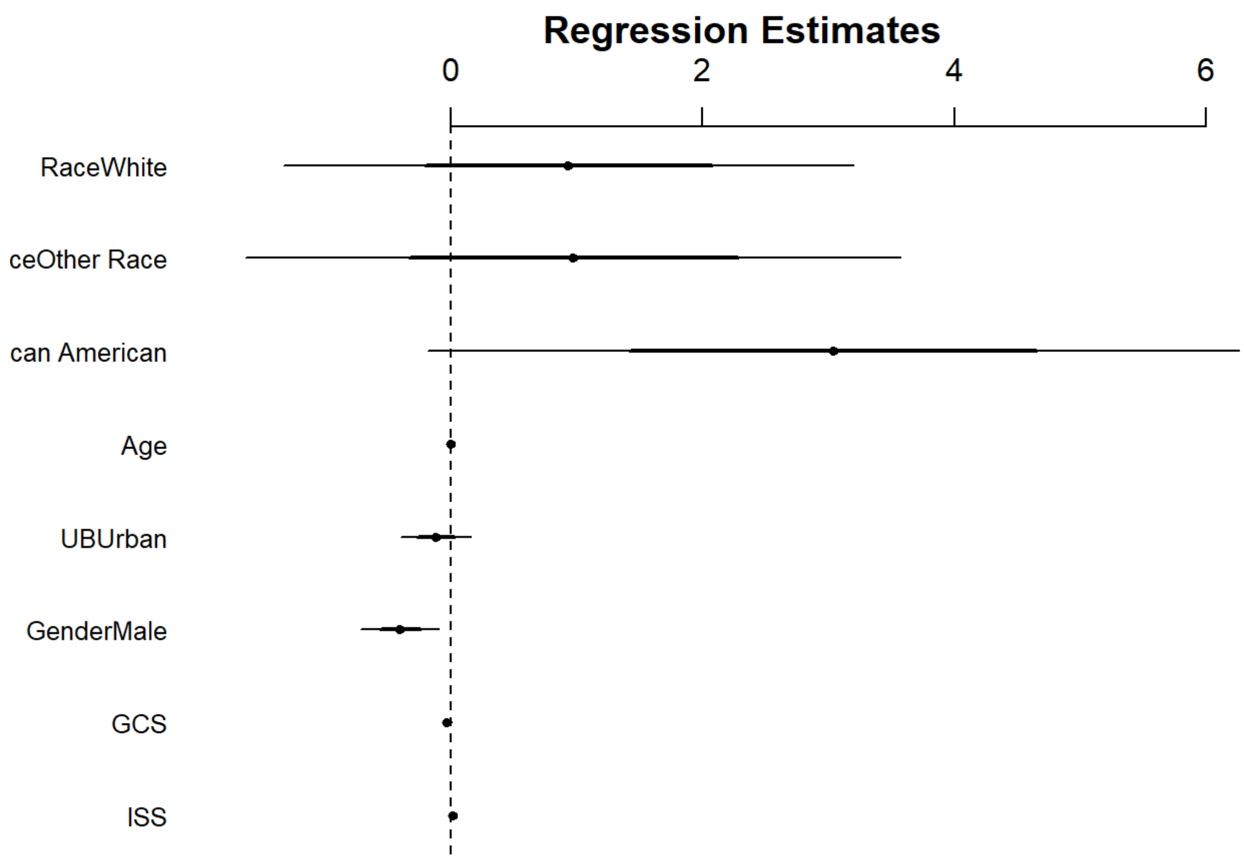
```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  df$GCS[df$UB == "Urban"] and df$GCS[df$UB == "Rural"]
## D = 0.10351, p-value = 0.3669
## alternative hypothesis: two-sided
```

The Kolmogorov-Smirnov (K-S) Test is one method to be used for determining if the underlying distribution of the Rural GCS vs Urban GCS, and Rural GCS vs Urban GCS are in fact different. As the above table shows, Since the p-value is greatter than .05, we fail to reject the null hypothesis. We do not have sufficient evidence to say that there is a difference between these distributions

Fit In Model

Model 1

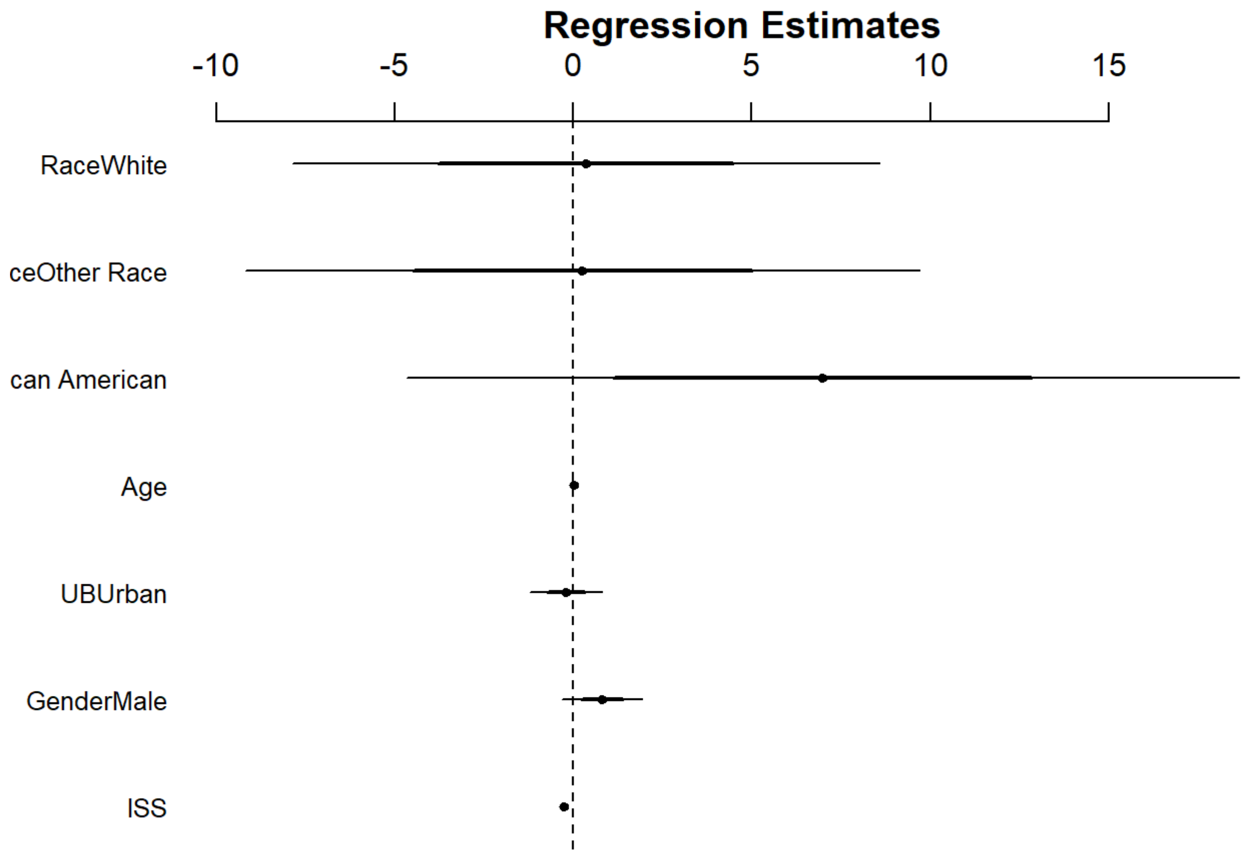
```
##
## Call:
## lm(formula = log(as.numeric(HospitalDays)) ~ ISS + GCS + Gender +
##     UB + Age + Race, data = Age_Cleaned_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.56988 -0.85104 -0.09372  0.73605  3.13557
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.795710   1.188324   0.670  0.50370
## ISS                         0.020141   0.008063   2.498  0.01312 *
## GCS                        -0.028663   0.017097  -1.677  0.09485 .
## GenderMale                 -0.403173   0.155054  -2.600  0.00985 **
## UBUrban                    -0.117728   0.140459  -0.838  0.40271
## Age                         0.001356   0.003571   0.380  0.70447
## RaceBlack or African American 3.038075   1.610997   1.886  0.06044 .
## RaceOther Race              0.971672   1.300425   0.747  0.45562
## RaceWhite                   0.934458   1.132605   0.825  0.41010
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.124 on 259 degrees of freedom
## Multiple R-squared:  0.1129, Adjusted R-squared:  0.08545
## F-statistic: 4.118 on 8 and 259 DF,  p-value: 0.0001208
```



```
##            GVIF Df GVIF^(1/(2*Df))
## ISS    1.476704  1        1.215197
## GCS    1.438842  1        1.199518
## Gender 1.060465  1        1.029789
## UB     1.044431  1        1.021974
## Age    1.066899  1        1.032908
## Race   1.065625  3        1.010650
```

Model 2

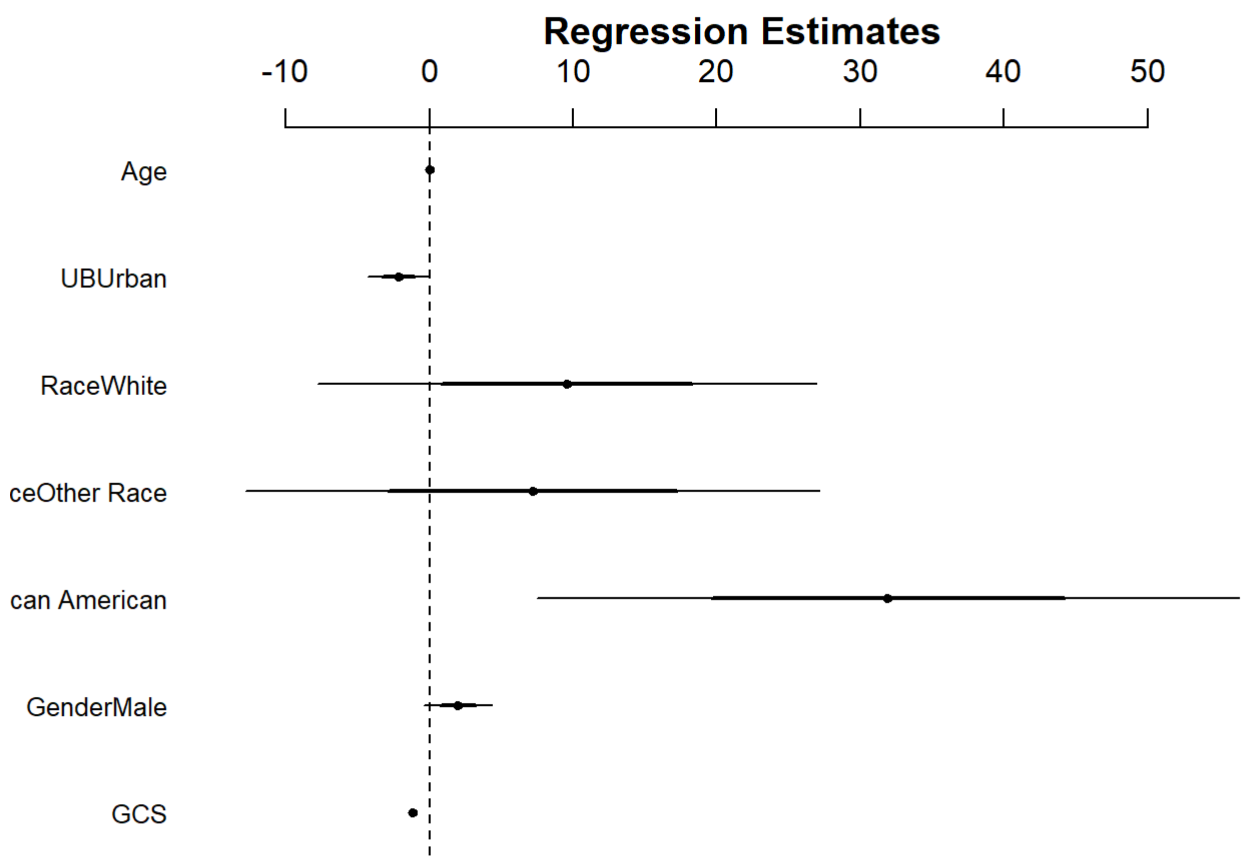
```
##
## Call:
## lm(formula = GCS ~ ISS + Gender + UB + Age + Race, data = Age_Cleaned_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.724  -2.034   1.017   2.666   7.766
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  13.91990    4.22321   3.296  0.00112 **
## ISS                          -0.25315    0.02468 -10.258  < 2e-16 ***
## GenderMale                    0.80528    0.56022   1.437  0.15180
## UBUrban                      -0.20103    0.50935  -0.395  0.69340
## Age                           0.02706    0.01284   2.107  0.03607 *
## RaceBlack or African American  6.98605    5.82766   1.199  0.23171
## RaceOther Race                0.26178    4.71715   0.055  0.95579
## RaceWhite                     0.35278    4.10836   0.086  0.93164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.077 on 260 degrees of freedom
## Multiple R-squared:  0.305,  Adjusted R-squared:  0.2863
## F-statistic:  16.3 on 7 and 260 DF,  p-value: < 2.2e-16
```




```
##             GVIF Df GVIF^(1/(2*Df))
## ISS     1.051232  1        1.025296
## Gender  1.052104  1        1.025721
## UB      1.043805  1        1.021668
## Age     1.048986  1        1.024200
## Race    1.055112  3        1.008981
```

Model 3

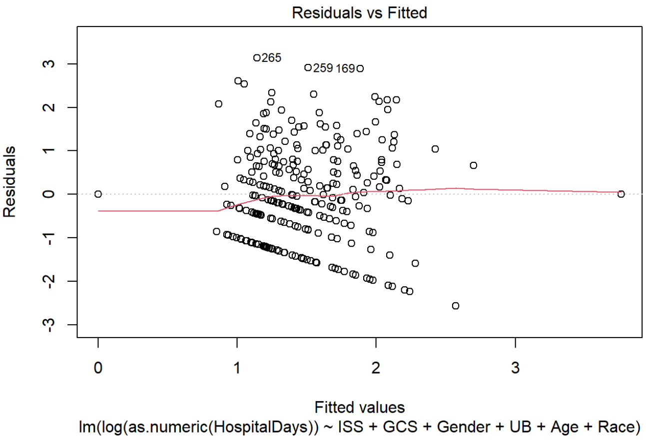
```
##
## Call:
## lm(formula = ISS ~ GCS + Gender + Race + UB + Age, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.1458  -5.1815  -0.3672   4.9750  24.4773
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    20.50043    9.05099   2.265  0.02434 *
## GCS                            -1.13814    0.11095 -10.258  < 2e-16 ***
## GenderMale                      1.97280    1.18629   1.663  0.09752 .
## RaceBlack or African American 31.92065   12.23162   2.610  0.00959 **
## RaceOther Race                  7.17127    9.99217   0.718  0.47359
## RaceWhite                       9.55838    8.69110   1.100  0.27244
## UBUrban                        -2.17476    1.07187  -2.029  0.04348 *
## Age                             0.01407    0.02745   0.512  0.60878
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.644 on 260 degrees of freedom
##   (因为不存在，48个观察量被删除了)
## Multiple R-squared:  0.3228, Adjusted R-squared:  0.3046
## F-statistic: 17.71 on 7 and 260 DF,  p-value: < 2.2e-16
```

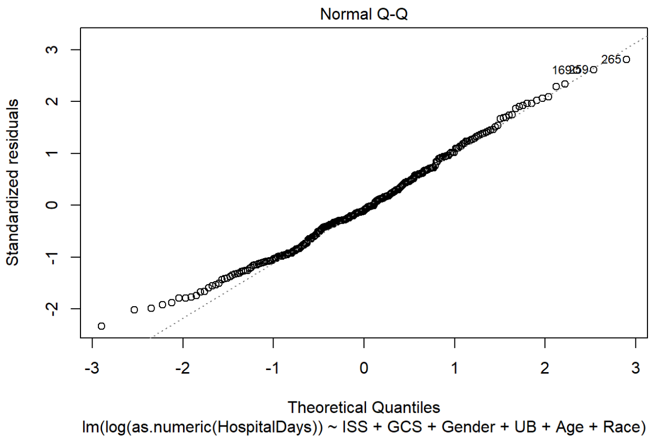


```
##             GVIF Df GVIF^(1/(2*Df))
## GCS    1.024279  1        1.012066
## Gender 1.049303  1        1.024355
## Race   1.033291  3        1.005473
## UB     1.028152  1        1.013978
## Age    1.065822  1        1.032387
```

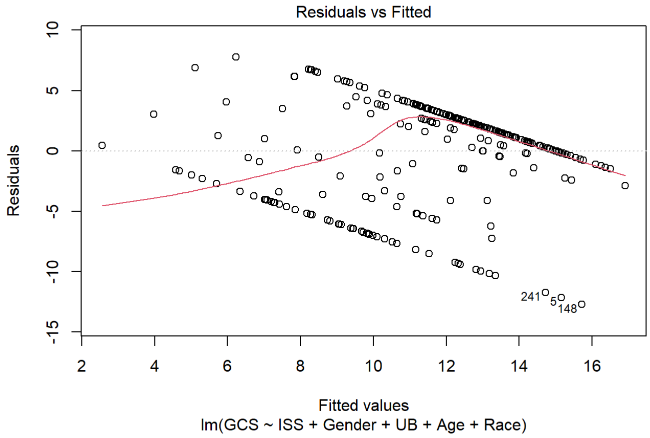
Check the Model

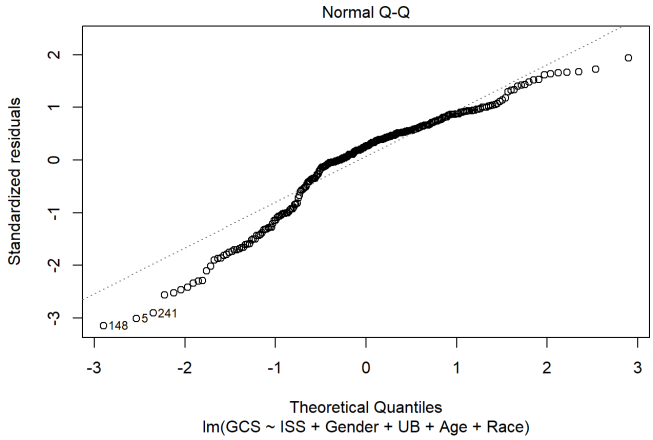
Model 1





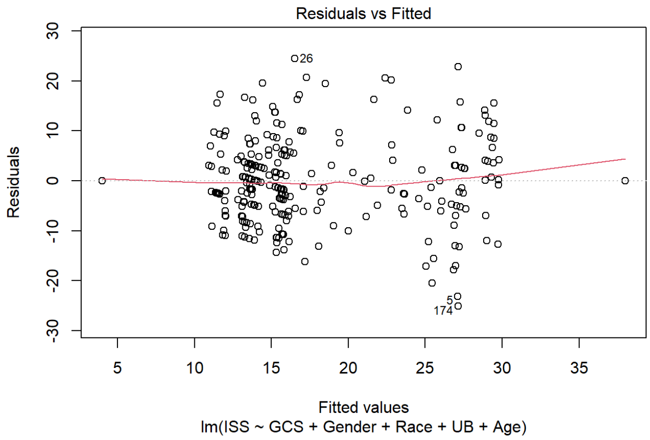
Model 2

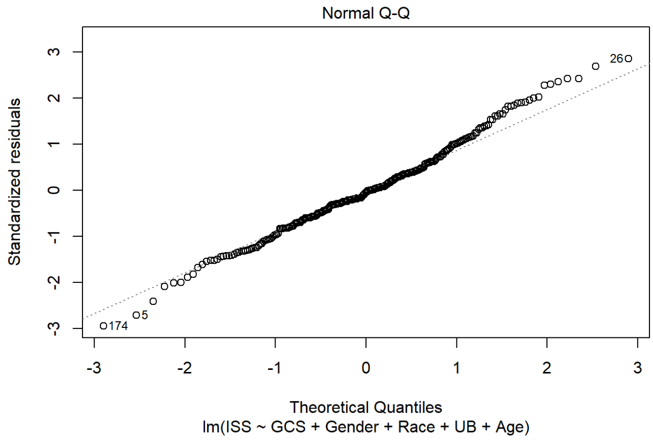




By summary of the model2, we can see that differ from model1, `ISS` and `GCS` now impact the `Age` of the patients more.

Model 3





By summary of the model3, when we are trying to find the influence of `Gender` , `Race` , and `UB` (which is the living area of the patients) on the `ISS` , we can see since the p-value of the `UB` is smaller than 0.05, only `UB` do not have significance influence on the `ISS` of the patients.

Discussion

First, based on the results of the Kolmogorov-Smirnov test, there is insufficient evidence to conclude that there is a difference in distribution of either ISS or GCS scores between rural and urban patient origins. However, the linear model with ISS score as a response suggests that when additional variables are taken into account that patients with urban origins are associated with lower ISS scores. Neither of the other models suggest that there is evidence of association between patient origin and either hospital days or GCS.

There are two major limitations in this analysis. The first is the concern for confounding that could occur if other variables explain ISS, GCS, or days in the hospital. For example, it could be that distance to the hospital is important as patients may only be transported long distances if they have more severe injuries. This could make it appear as though urban patients have less severe injuries even if they don’t in practice as the most local patients may be considered urban. In this way, the existing analysis could be criticized as misleading as the association may disappear with the addition of a distance variable. The second major limitation concerns the GCS model–based on diagnostic plots, some model assumptions do not appear reasonable (such as the normality of the residuals), which means that the p-values and confidence intervals may not be trustworthy. A model with more accurate assumptions may have different results.

Conclusion

We find that there is evidence from our linear model of an association between ISS and whether a patient is from a rural or urban origin–with the caution that this association may be due to a confounding variable like distance to hospital. We do not find evidence of an association between rural/urban origin and either days in hospital or GCS. For future work, we strongly recommend an analysis which includes this confounding variable and which uses a more robust model, such as quantile regression or a censored linear model, on the GCS outcome.

Contribution

Tianjian Xie: Data cleaning, part of EDA(violin plot and box plot of ISS and GCS, scatter plot of ISS against Total Hospital Days and ISS vs GCS, barplot of Injury Types, models fitting), T test and KS test, writing report. Haocheng Zhu: Data cleaning, combining the EDAs for writing slides and report, presentation Nuo Chen: Data cleaning, part of EDA(Visualization of ISS against AGE and ISS vs GCS, analysis of Injury Types, models fitting), Presentation for client Haochen Pan: Data cleaning, part of EDA(Scatter plot of ISS against Total Hospital Days and ISS vs GCS, models analysis for linear model), writing the slides and report Shengbo Wang: Data cleaning, part of EDA(Analysis for injury type, models fitting),KS test analysis, Model analysis for problems of linear model, presentation and writing report