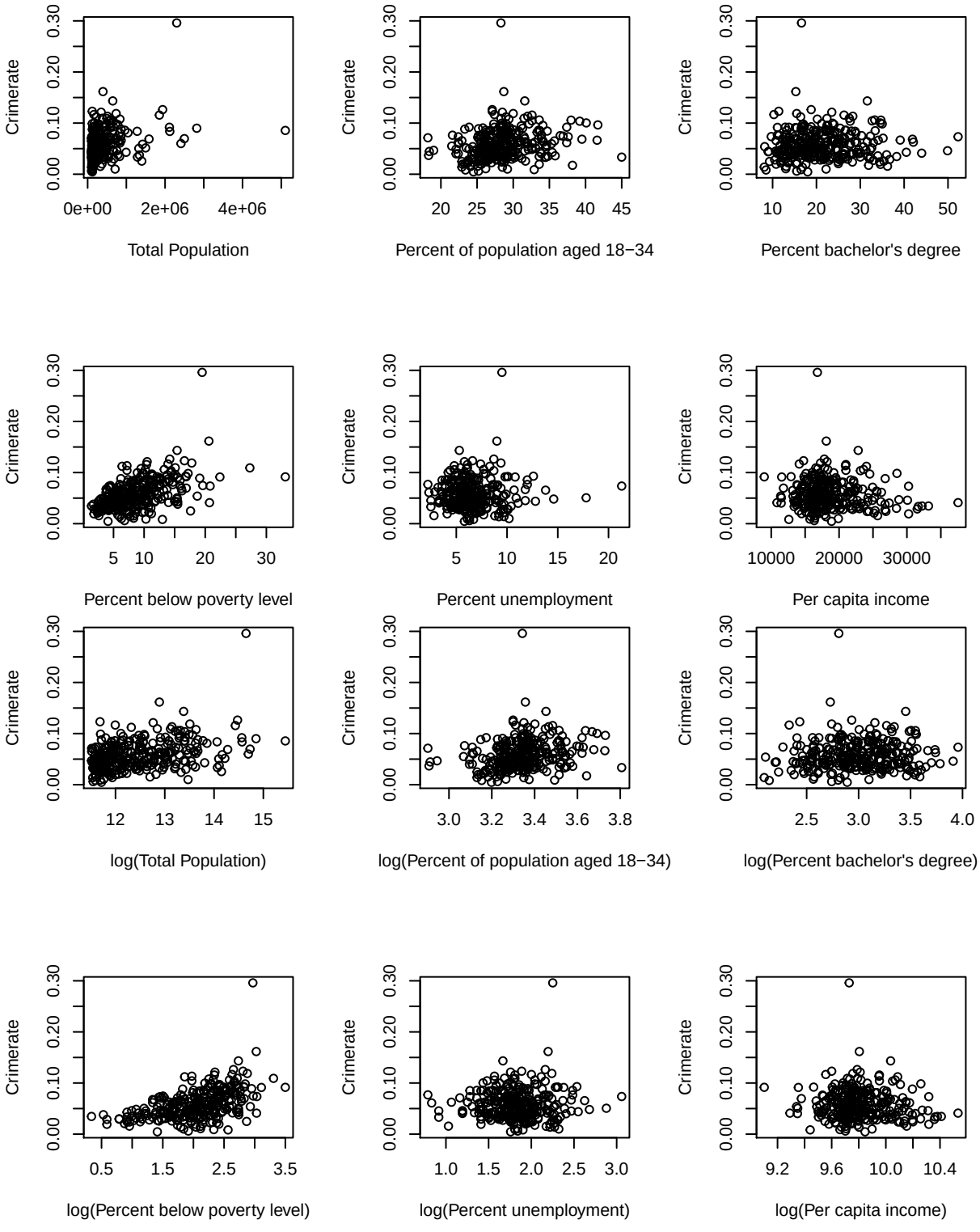
5291 Final Project

Chen Chen, Yang Meng, and Zhichao Hu

Introduction

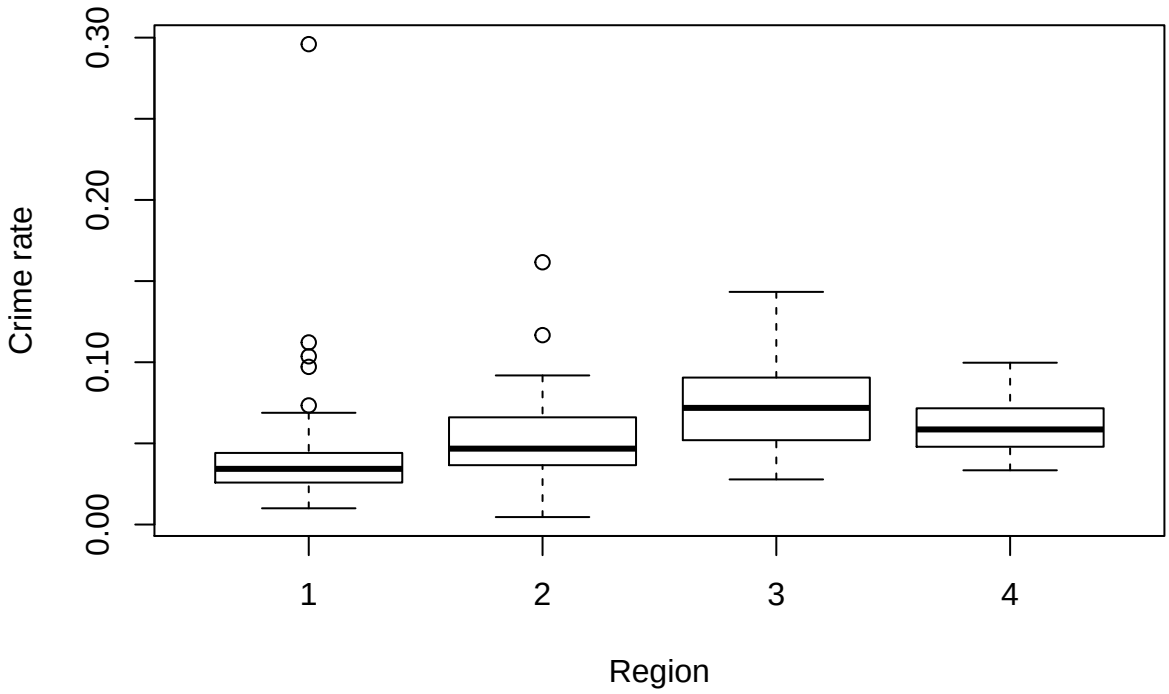
Goal

Dataset



Crime Rate Among Different Regions

Boxplot



Regression Model

```
##
## Call:
## lm(formula = crimerate ~ region)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.046920 -0.015937 -0.005512  0.012357  0.254398
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.041589   0.003116  13.346  < 2e-16 ***
## region2     0.009933   0.004485   2.215   0.0275 *
## region3     0.031474   0.004105   7.668 2.54e-13 ***
## region4     0.019654   0.004843   4.059 6.32e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02699 on 296 degrees of freedom
## Multiple R-squared:  0.1785, Adjusted R-squared:  0.1702
## F-statistic: 21.44 on 3 and 296 DF,  p-value: 1.36e-12
```

$$crimerate = 0.0416 + 0.0099 * NC + 0.0315 * S + 0.0197 * W$$

Least Significant Difference Method

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  crimerate and region
##
##   1       2       3
## 2 0.028   -       -
## 3 2.5e-13 4.9e-07 -
## 4 6.3e-05 0.049   0.010
##
## P value adjustment method: none
```

Bonferroni Method

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  crimerate and region
##
##   1       2       3
## 2 0.16523 -       -
## 3 1.5e-12 3.0e-06 -
## 4 0.00038 0.29289 0.06099
##
## P value adjustment method: bonferroni
```

Tukey Method

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = crimerate ~ region)
##
## $region
##             diff          lwr           upr     p adj
## 2-1  0.009932831 -0.001654555  2.152022e-02 0.1215900
## 3-1  0.031474173  0.020868529  4.207982e-02 0.0000000
## 4-1  0.019653873  0.007142138  3.216561e-02 0.0003668
## 3-2  0.021541342  0.010719623  3.236306e-02 0.0000029
## 4-2  0.009721042 -0.002974368  2.241645e-02 0.1985026
## 4-3 -0.011820300 -0.023626468 -1.413155e-05 0.0495982
```

Multicollinearity

```
##    Variables      VIF       r2
## 1        pop 1.241106 1.114049
## 2      young 2.563906 1.601220
## 3        old 1.985576 1.409105
## 4   highgrad 4.196878 2.048628
## 5   bachelor 7.232377 2.689308
```

```
## 6       poor 3.900997 1.975094
## 7      unemp 1.856206 1.362426
## 8     income 5.181625 2.276318
## 9    avgarea 1.372664 1.171607
## 10   avgphys 3.228907 1.796916
## 11   avgbeds 3.258653 1.805174
```

Model Selection

Full Model

$$\begin{aligned} crimerate \sim & avgarea + log(pop) + log(young) + log(old) + \\ & avgphys + avgbeds + log(highgrad) + log(bachelor) + \\ & log(poor) + log(unemp) + log(income) + region \end{aligned}$$

Reduced Model

```
## Analysis of Variance Table
##
## Response: crimerate
##                Df   Sum Sq  Mean Sq F value    Pr(>F)
## avgarea         1 0.003663 0.003663  9.1645  0.002693 **
## log(pop)        1 0.036952 0.036952 92.4619 < 2.2e-16 ***
## log(young)      1 0.009495 0.009495 23.7580 1.817e-06 ***
## log(old)        1 0.000854 0.000854  2.1372  0.144867
## avgphys         1 0.006774 0.006774 16.9507 5.029e-05 ***
## avgbeds         1 0.031874 0.031874 79.7544 < 2.2e-16 ***
## log(highgrad)   1 0.017311 0.017311 43.3164 2.229e-10 ***
## log(bachelor)   1 0.001558 0.001558  3.8988  0.049287 *
## log(poor)       1 0.025607 0.025607 64.0738 3.058e-14 ***
## log(unemp)      1 0.000616 0.000616  1.5424  0.215285
## log(income)     1 0.002051 0.002051  5.1326  0.024231 *
## region          3 0.011744 0.003915  9.7949 3.606e-06 ***
## Residuals     285 0.113900 0.000400
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\begin{aligned} crimerate \sim & avgarea + log(pop) + log(young) + \\ & avgphys + avgbeds + log(highgrad) + \\ & log(poor) + log(income) + region \end{aligned}$$

```
## Analysis of Variance Table
##
## Model 1: crimerate ~ avgarea + log(pop) + log(young) + log(old) + avgphys +
##     avgbeds + log(highgrad) + log(bachelor) + log(poor) + log(unemp) +
##     log(income) + region
## Model 2: crimerate ~ avgarea + log(pop) + log(young) + avgphys + avgbeds +
##     log(highgrad) + log(poor) + log(income) + region
##   Res.Df     RSS Df   Sum of Sq      F Pr(>F)
## 1    285 0.11390
## 2    288 0.11444 -3 -0.00054439 0.4541 0.7146
```

Interaction with region

By backward selection

$$\begin{aligned} crimerate \sim & avgarea + log(pop) + log(young) + avgphys + \\ & avgbeds + log(highgrad) + log(poor) + log(income) + \\ & region + log(pop) * region + log(income) * region \end{aligned}$$

```
## Analysis of Variance Table
##
## Model 1: crimerate ~ avgarea + log(pop) + log(young) + avgphys + avgbeds +
##     log(highgrad) + log(poor) + log(income) + region
## Model 2: crimerate ~ avgarea + log(pop) + log(young) + avgphys + avgbeds +
##     log(highgrad) + log(poor) + log(income) + region + log(pop) *
##     region + log(income) * region
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    288 0.11444
## 2    282 0.11043  6 0.0040122 1.7076 0.1191
```

Interactions with all other variables

By backward selection

$$\begin{aligned} crimerate \sim & avgarea + log(pop) + log(young) + avgphys + \\ & avgbeds + log(highgrad) + log(poor) + log(income) + \\ & region + log(pop) * region + log(income) * region + \\ & log(pop) * log(poor) + log(pop) * log(income) + \\ & log(poor) * log(income) + log(poor) * region \end{aligned}$$

```
## Analysis of Variance Table
##
## Model 1: crimerate ~ avgarea + log(pop) + log(young) + avgphys + avgbeds +
##     log(highgrad) + log(poor) + log(income) + region
## Model 2: crimerate ~ avgarea + log(pop) + log(young) + avgphys + avgbeds +
##     log(highgrad) + log(poor) + log(income) + region + log(pop) *
##     region + log(income) * region + log(pop) * log(poor) + log(pop) *
##     log(income) + log(poor) * log(income) + log(poor) * region
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1    288 0.114444
## 2    276 0.099822 12  0.014623 3.3692 0.0001319 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = crimerate ~ avgarea + log(pop) + log(young) + avgphys +
##     avgbeds + log(highgrad) + log(poor) + log(income) + region +
##     log(pop) * region + log(income) * region + log(pop) * log(poor) +
##     log(pop) * log(income) + log(poor) * log(income) + log(poor) *
##     region)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.053267 -0.009991  0.000172  0.008316  0.158580
```

```
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -2.576614   1.498009  -1.720   0.0865 .
## avgarea                -0.034739   0.228742  -0.152   0.8794
## log(pop)                0.144911   0.119150   1.216   0.2249
## log(young)              0.039536   0.009986   3.959 9.58e-05 ***
## avgphys                -1.127952   1.215090  -0.928   0.3541
## avgbeds                 2.259507   1.013839   2.229   0.0266 *
## log(highgrad)          -0.036111   0.020792  -1.737   0.0835 .
## log(poor)               0.008273   0.101099   0.082   0.9348
## log(income)             0.269159   0.145470   1.850   0.0653 .
## region2                -0.015448   0.356392  -0.043   0.9655
## region3                 0.287122   0.251085   1.144   0.2538
## region4                 0.468921   0.267847   1.751   0.0811 .
## log(pop):region2       -0.009271   0.005987  -1.548   0.1226
## log(pop):region3       -0.002956   0.005249  -0.563   0.5738
## log(pop):region4       -0.010074   0.005309  -1.897   0.0588 .
## log(income):region2     0.014973   0.039424   0.380   0.7044
## log(income):region3    -0.020462   0.027783  -0.736   0.4621
## log(income):region4    -0.029827   0.028636  -1.042   0.2985
## log(pop):log(poor)      0.010233   0.004206   2.433   0.0156 *
## log(pop):log(income)   -0.015311   0.011441  -1.338   0.1819
## log(poor):log(income)  -0.010510   0.010186  -1.032   0.3030
## log(poor):region2      -0.003181   0.011155  -0.285   0.7758
## log(poor):region3      -0.012680   0.009510  -1.333   0.1835
## log(poor):region4      -0.015927   0.012162  -1.310   0.1914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01902 on 276 degrees of freedom
## Multiple R-squared:  0.6196, Adjusted R-squared:  0.5879
## F-statistic: 19.54 on 23 and 276 DF,  p-value: < 2.2e-16
```