813_PS5

X. Zhang

February 9, 2019

```
library(haven)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
problem_set_5 <- "https://instructure-uploads.s3.amazonaws.com/account_83960000000000001/attachments/178
problem_set_5_data <- read_dta(problem_set_5)

head(problem_set_5_data)
```

```
## # A tibble: 6 x 3
##   state pcrimer  povr
##   <chr>   <dbl> <dbl>
## 1 AL      3502.  16.2
## 2 AK      2739.  11.4
## 3 AZ      3539.  18.3
## 4 AR      3660.  18
## 5 CA      2759.  16.4
## 6 CO      2685.  12.5
```

```
m1 <- lm(pcrimer ~ povr, data = problem_set_5_data)

summary(m1)
```

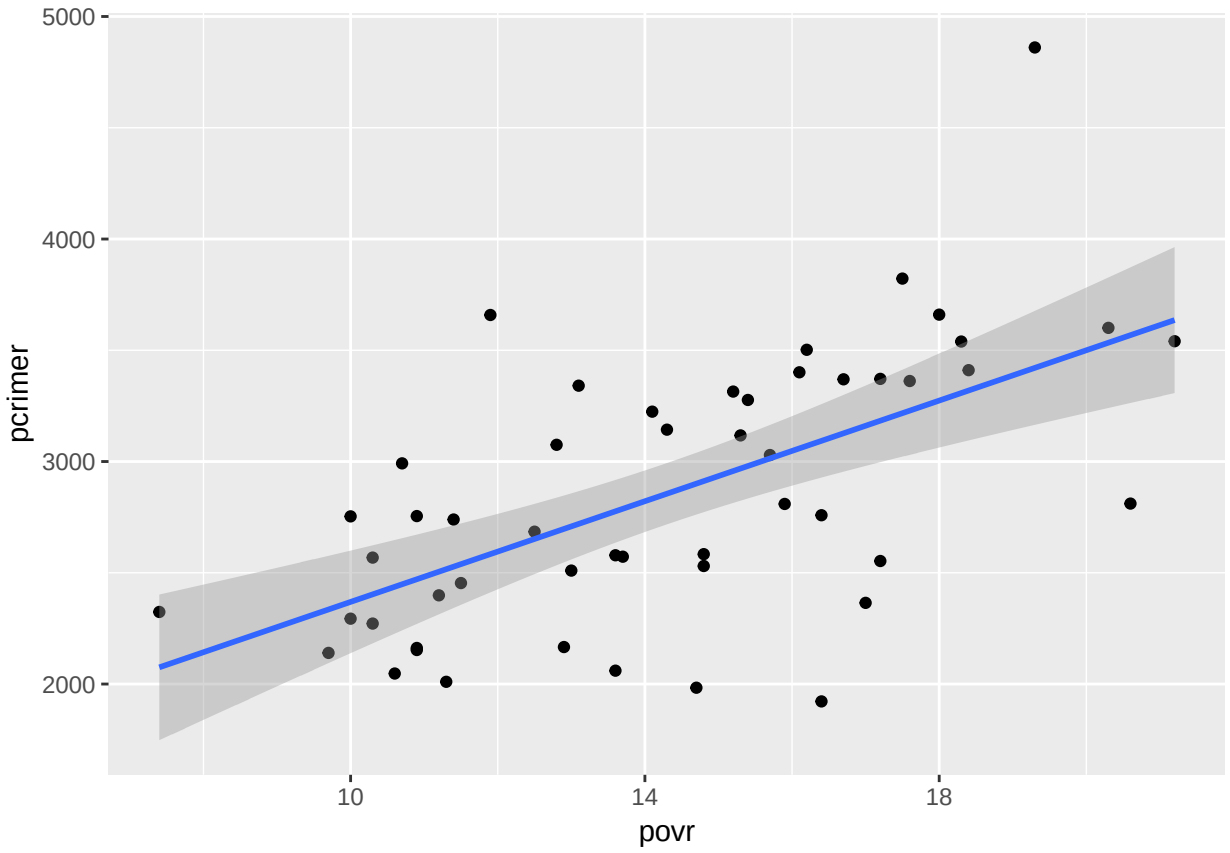
```
##
## Call:
## lm(formula = pcrimer ~ povr, data = problem_set_5_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1170.93  -313.38    32.82   292.31  1439.90
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1238.23     313.93   3.944 0.000255 ***
## povr          113.09      21.44   5.275 2.99e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 489.5 on 49 degrees of freedom
## Multiple R-squared:  0.3622, Adjusted R-squared:  0.3492
## F-statistic: 27.83 on 1 and 49 DF,  p-value: 2.991e-06
```

1.(a) (b)

```
sm1 <- summary(m1)
m1_mse <- mean(sm1$residuals^2)

ggplot(problem_set_5_data, aes(x = povr, y = pcrimer)) + geom_point() +
  geom_smooth(method='lm')
```



the independent variable percentage of population under poverty line (poverty level) have a coefficient of 113.09, suggesting one percent increase in poverty level is associated with 113.09 increase in property crime rate per 100,000 population.

The intercept is 1238.23, suggesting that that when there is zero precent population living under poverty line, there will be 1237.23 property crimes per 100,000 population.

1. (c)

poverty level variable has a standard error 21.44, offering a measurement of the standard deviation of the coefficient.

The poverty level variable has a t value 0of 5.275, indicating that assuming the error term follows normal distribution, the estimated coefficient is 5.275 times the standard error.

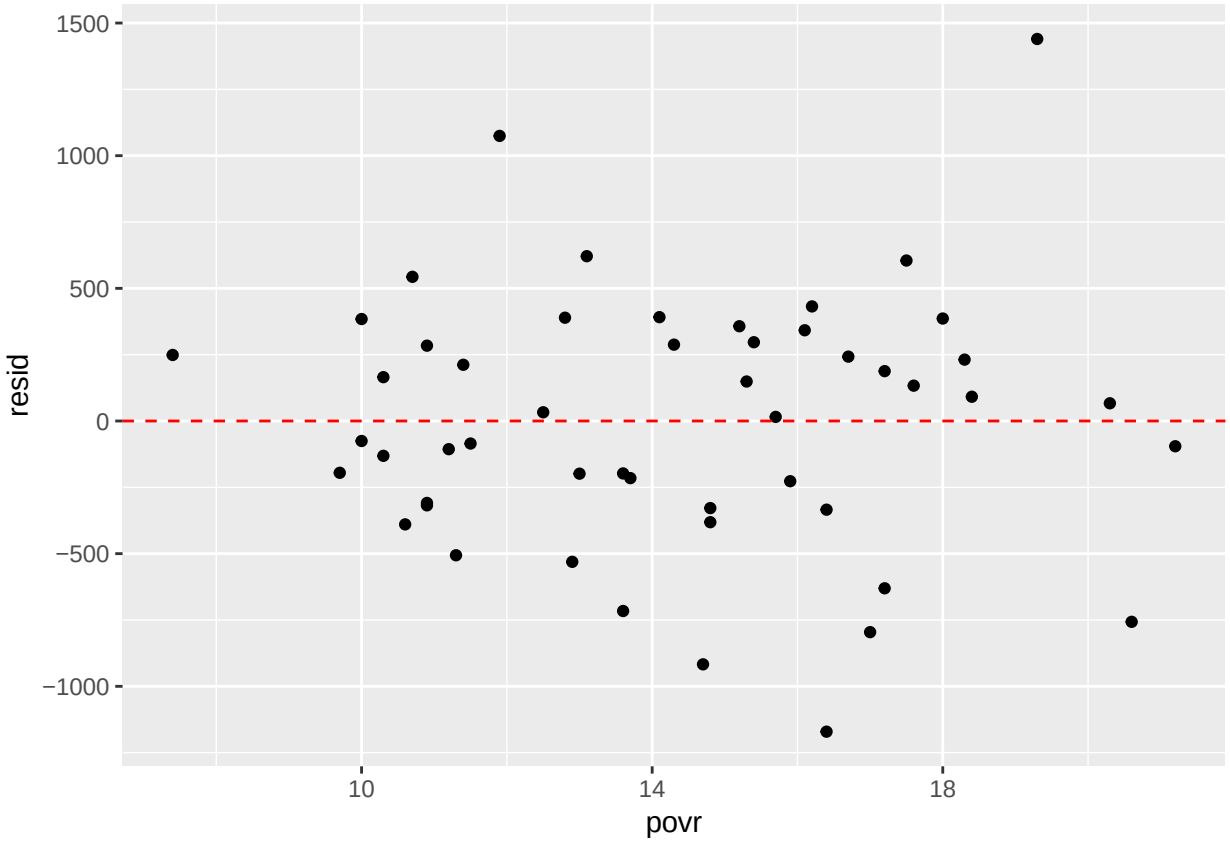
poverty level has a p-value of 2.99e-06 suggesting it is unlikely that the associaiton is due to chance.

R-squared: 0.3622, suggesting that 36% of the dependent variable variation can be explained by this linear model.

1.(d)

```
problem_set_5_data$resid <- resid(m1)

ggplot(problem_set_5_data, aes(x = povr, y =resid))+geom_point()+
  geom_hline(yintercept = 0, linetype = "dashed", col = "red")
```



Ideally, residual should be evenly and randomly spreaded at around horizontal line (residual = 0). Most of the residual, particularly states with poverty level of 10-18% does distributed evenly and randomly around the 0 residual line, suggest the data is a good fit. One outlier is DC which as a high residual of 1439.90.

2.(a) I suspect property crime rate is also influenced by rate of uranization. Thus, i obtained state level urban population ratio from 2010 US Census. https://www.census.gov/prod/cen2010/cph-2-1.pdf

```
problem_set_5_data$urbanization_rate<- c(0.59, 0.66, 0.90,0.56,0.95,0.86,0.88,0.83,1.00,0.91,0.75,0.92,(

m2 <- lm(pcrimer ~ povr + urbanization_rate, data = problem_set_5_data)

summary(m2)
```

```
##
## Call:
## lm(formula = pcrimer ~ povr + urbanization_rate, data = problem_set_5_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1290.67  -369.35    36.45   293.73  1214.98
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         594.88     464.72   1.280   0.2067
## povr                114.79      20.95   5.478 1.55e-06 ***
## urbanization_rate   835.49     453.60   1.842   0.0717 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 478 on 48 degrees of freedom
## Multiple R-squared:  0.4043, Adjusted R-squared:  0.3795
## F-statistic: 16.29 on 2 and 48 DF,  p-value: 3.985e-06
```

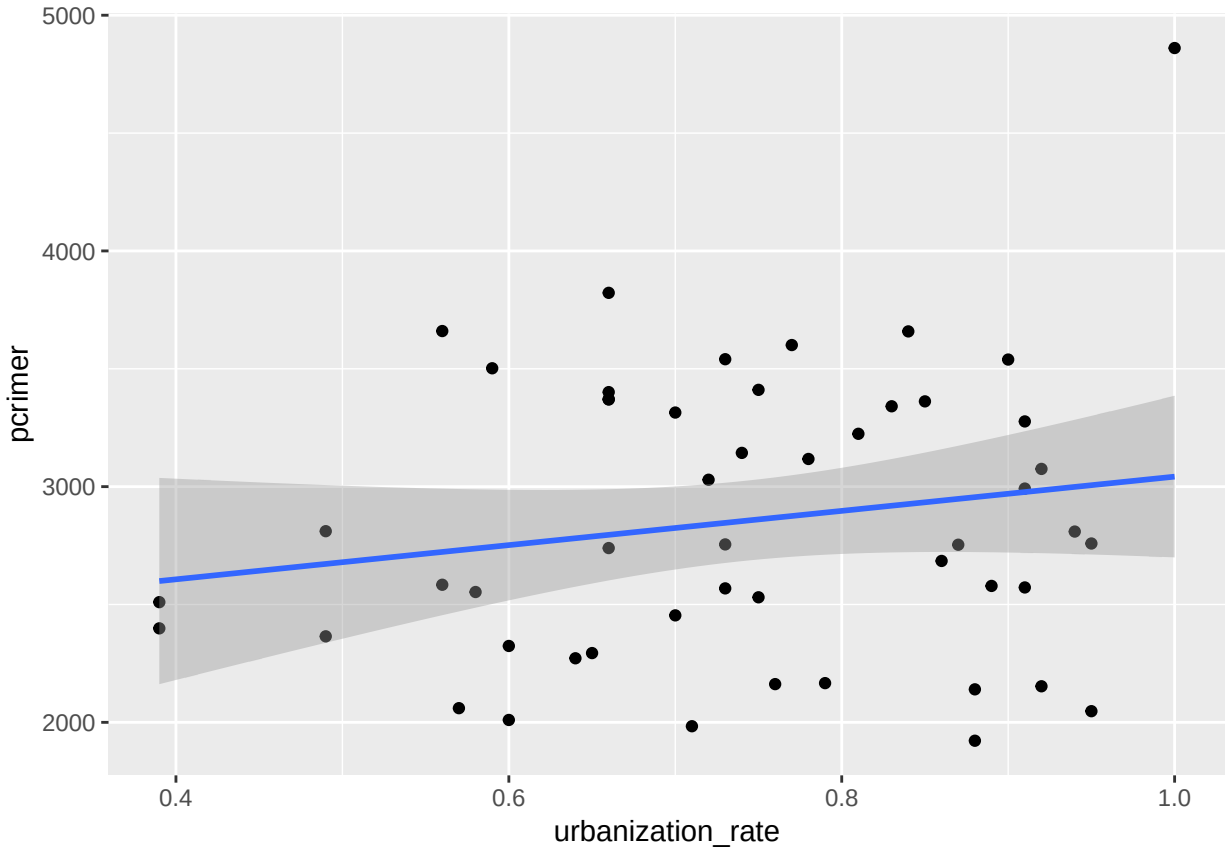
```
sm2<- summary(m2)

m2_mse <- mean(sm2$residuals^2)

m2_mse < m1_mse
```

```
## [1] TRUE
```

```
ggplot(problem_set_5_data, aes(x = urbanization_rate, y = pcrimer)) + geom_point() +
  geom_smooth(method='lm')
```



c. How well does the model fit the data? d. Interpret the estimated parameters substantively and statistically.

R-squared: 0.4043 suggest 40% of the variation in property crime rate is explained by the second model.R-squareed has increased with added variable.

The mean sqaured error of model model 2 is smaller than the mse of model 1, suggesting a slightly better fit to the data.

The urbanization term has a coefficient of 835.49, indicating that 1 percent increase in the state's urbanization level is associated with a 835.49 increase in property crimes per 100,000 population.

Urbanization variable has a large standard error of 453.6. t statistics of 1.842 suggest the estimated coefficient is 1.842 times the standard deviation.

The urbanization variable has a p-value of 0.0717 suggesting a reasonable probability that this association is due to chance.

The interception in model 2 is 594.88. Sugessting that if both urbanization rate and poverty level are zero, the number of property crime per 100,000 population is 594.88. However, given that urbanization rate cannot be 0, this interpcet does not have substantive meaning.

The coefficient and standard error of poverty level variable did not change signficantly after adding the urbanization variable to the model.