W271 Summer 2022 Lecture Video Question Solutions Week 11

Contents

Week 11

11.3 Using OLS Regression Model on Panel Data

Q: What is wrong with this histogram? Answer this question in one sentence in the following text box.

```
#install.packages("wooldridge")
library(wooldridge)
```

Solution:

```
## Warning: package 'wooldridge' was built under R version 4.1.1
```

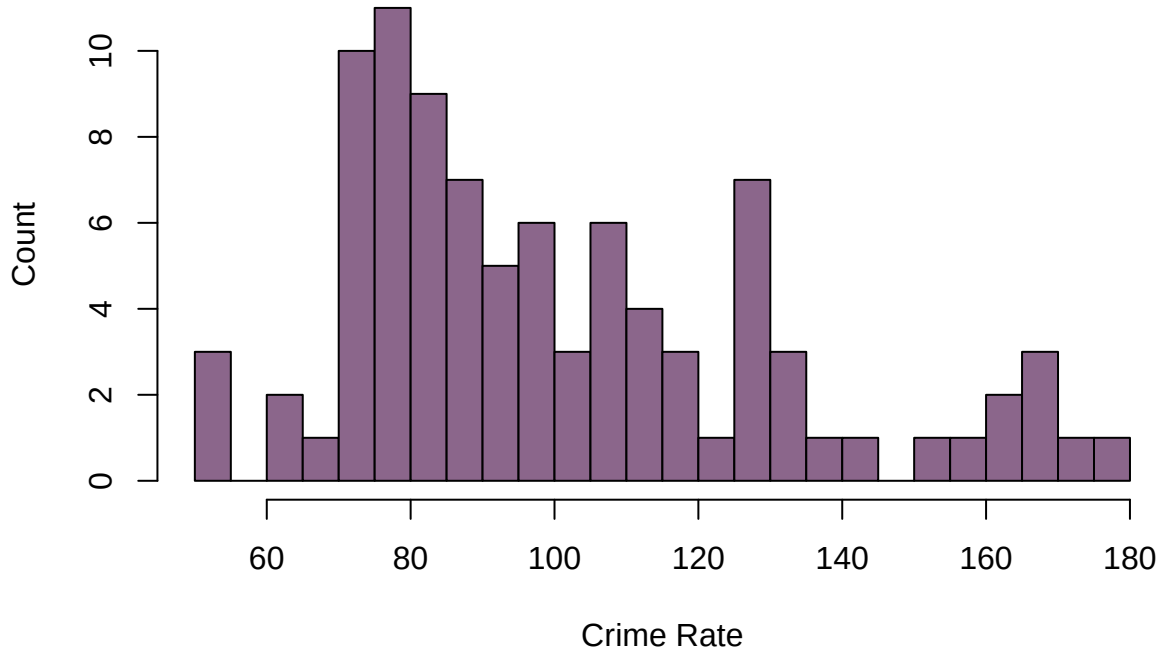
```
data("crime2")

head(crime2)
```

```
##      pop crimes unem officers pcinc west nrtheast south year  area d87   popden
## 1 229528  17136  8.2      326  8532    1        0     0   82  44.6   0 5146.368
## 2 246815  17306  3.7      321 12155    1        0     0   87  44.6   1 5533.969
## 3 814054  75654  8.1     1621  7551    1        0     0   82 375.0   0 2170.811
## 4 933177  83960  5.4     1803 11363    1        0     0   87 375.0   1 2488.472
## 5 374974  31352  9.0      633  8343    1        0     0   82  49.8   0 7529.599
## 6 406297  31364  5.9      685 11729    1        0     0   87  49.8   1 8158.574
##     crmrte   offarea   lawexpc    polpc     lpop   loffic   lpcinc llawexpc
## 1 74.65756  7.309417  850.8599 1.420306 12.34378 5.786897 9.051579 6.746247
## 2 70.11729  7.197309 2262.4399 1.300569 12.41639 5.771441 9.405496 7.724199
## 3 92.93487  4.322667  875.0800 1.991268 13.60978 7.390799 8.929436 6.774315
## 4 89.97221  4.808000 1069.6400 1.932109 13.74635 7.497207 9.338118 6.975078
## 5 83.61113 12.710844 1121.8999 1.688117 12.83461 6.450470 9.029179 7.022779
```

```
## 6 77.19476 13.755020 1545.6000 1.685959 12.91484 6.529419 9.369820 7.343167
##    lpopden   lcrimes    larea  lcrmrte     clcrimes      clpop    clcrmrte
## 1 8.546046  9.748937 3.797734 4.312912           NA         NA          NA
## 2 8.618661  9.758808 3.797734 4.250169 0.0098714828 0.07261372 -0.06274271
## 3 7.682856 11.233926 5.926926 4.531899           NA         NA          NA
## 4 7.819424 11.338096 5.926926 4.499501 0.1041698456 0.13656807 -0.03239822
## 5 8.926597 10.353033 3.908015 4.426177           NA         NA          NA
## 6 9.006824 10.353416 3.908015 4.346332 0.0003833771 0.08022785 -0.07984495
##      lpolpc      clpolpc  cllawexp cunem   clpopden lcrmrt_1   ccrmrte
## 1 0.3508723           NA        NA    NA         NA       NA        NA
## 2 0.2628021 -0.088070214 0.9779520  -4.5 0.07261467 4.312912 -4.540268
## 3 0.6887718           NA        NA    NA         NA       NA        NA
## 4 0.6586123 -0.030159533 0.2007623  -2.7 0.13656807 4.531899 -2.962654
## 5 0.5236138           NA        NA    NA         NA       NA        NA
## 6 0.5223344 -0.001279354 0.3203883  -3.1 0.08022785 4.426177 -6.416374
```

```
hist(crime2$crmrte, col = "plum4", xlab = "Crime Rate", ylab = "Count", main = "", breaks = 30)
```

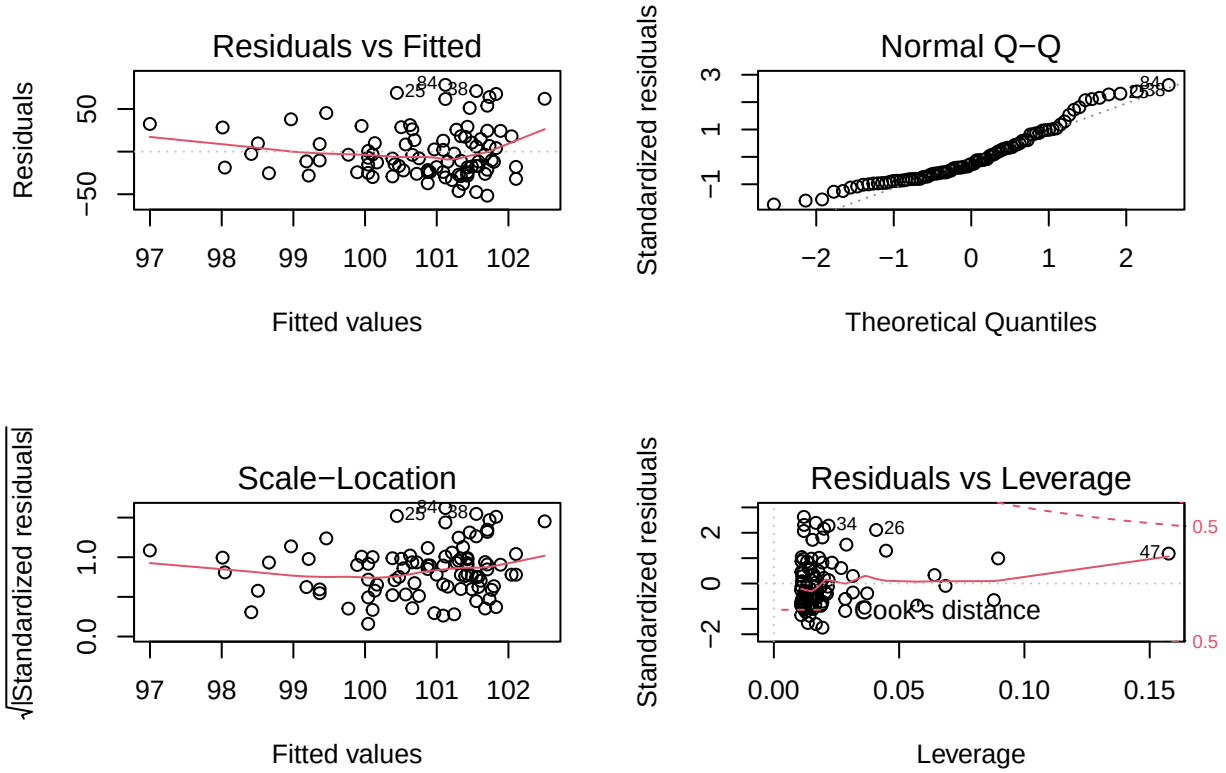


The histogram is effectively treating each observation as independent and from the same distribution, but we know this data is a panel data set. It has structure both over time and also location. Hence, the histrogram is lumping different cities together, which likely have different means of crime rates, making the single histogram plot not super useful in actually examining the variation within various locations.

Q: What is wrong with this histogram? Answer this question in one sentence in the following text box.

```
model <- lm(crmrte ~ unem, data = crime2)

par(mfrow = c(2, 2))
plot(model)
```



Solution:

The residuals vs. fitted plot shows a slight trend in the residuals towards the ends of the distribution of fitted values, but it is not super concerning. A qq plot of the residuals shows deviation from normality for residuals at the extremes, but this is also quite common in regression. There are also no few points with a particularly high leverage, suggesting there are not many outliers.

However, given the panel data structure, we know that this model is not correct despite the lack of obvious signs from the diagnostic plots. This highlights that even if the diagnostic plots look ok you should make sure you understand the structure of the data to be able to design an appropriate model.

11.4 Exploratory Panel Data Analysis: A Two-Period Panel

Q: Perform regression diagnostics, and interpret the results. Comment on the regression results and whether this regression makes sense.

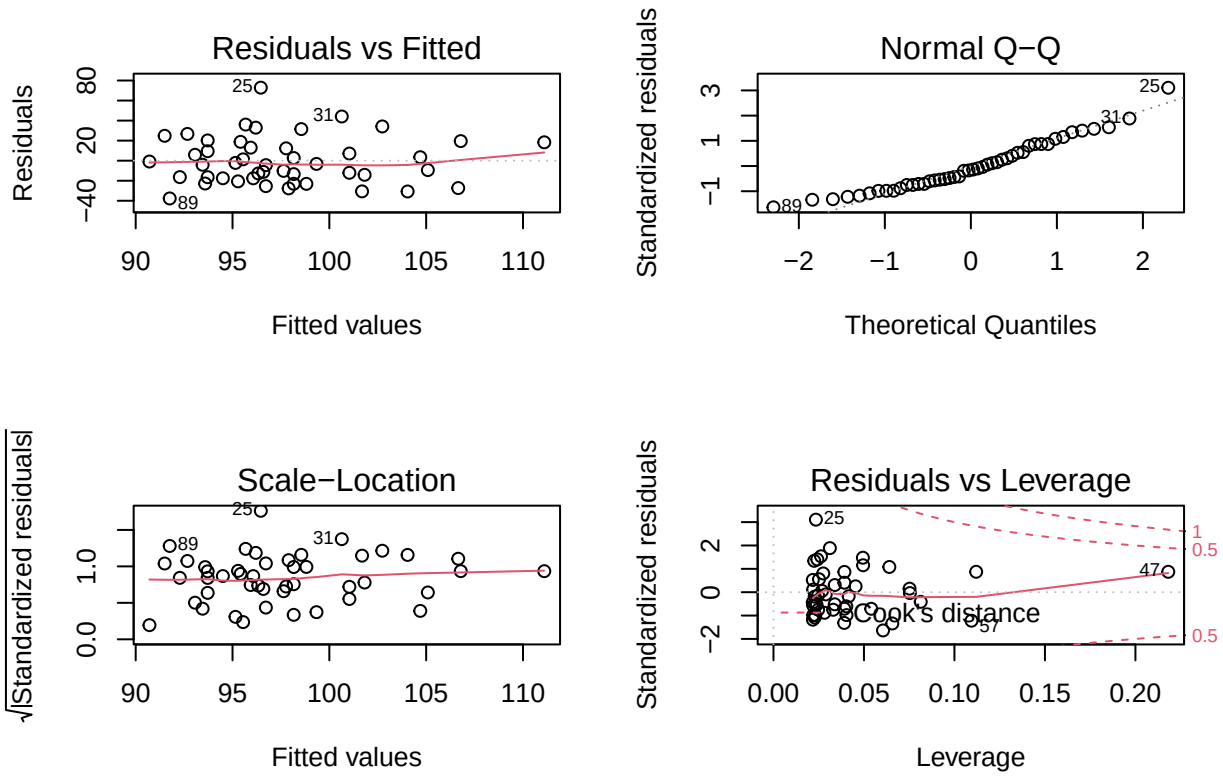
```
crime.82 <- crime2[crime2$year == 82,]

model <- lm(crmrte ~ unem, data = crime.82)
summary(model)
```

Solution:

```
##
## Call:
## lm(formula = crmrte ~ unem, data = crime.82)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.693 -17.292  -3.671  16.994  72.854
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   84.569     10.904   7.756 9.07e-10 ***
## unem           1.307      1.027   1.272     0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.74 on 44 degrees of freedom
## Multiple R-squared:  0.03549,    Adjusted R-squared:  0.01357
## F-statistic: 1.619 on 1 and 44 DF,  p-value: 0.2099
```

```
par(mfrow = c(2, 2))
plot(model)
```



Focusing on the 1982 data first, we see no evidence of a relationship between unemployment rate and crime rate since the coefficient on the unemployment rate is not statistically significant. It does has the intuitive sign since a higher unemployment rate should be positively correlated with a higher crime rate.

The residual diagnostic plots also look fine with little trend in the residuals the qq plot is generally linear. There are potentially some high leverage points that it would be good to test the inclusion of, but again we also know this model does not fit the data structure well.

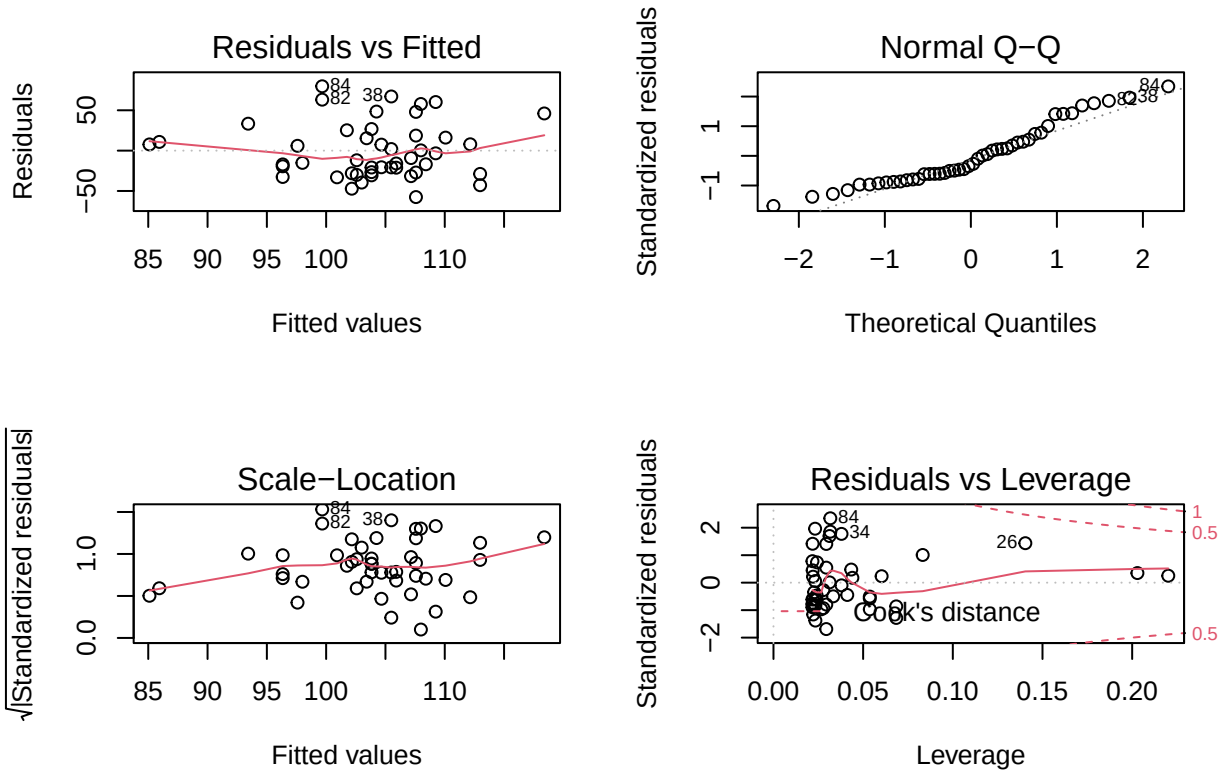
```
crime.87 <- crime2[crime2$year == 87,]

model <- lm(crmrte ~ unem, data = crime.87)
summary(model)
```

```
##
## Call:
## lm(formula = crmrte ~ unem, data = crime.87)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -57.55 -27.01 -10.56  18.01  79.75
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  128.378     20.757   6.185  1.8e-07 ***
## unem          -4.161      3.416  -1.218     0.23
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.6 on 44 degrees of freedom
## Multiple R-squared:  0.03262,    Adjusted R-squared:  0.01063
## F-statistic: 1.483 on 1 and 44 DF,  p-value: 0.2297
```

```
par(mfrow = c(2, 2))
plot(model)
```

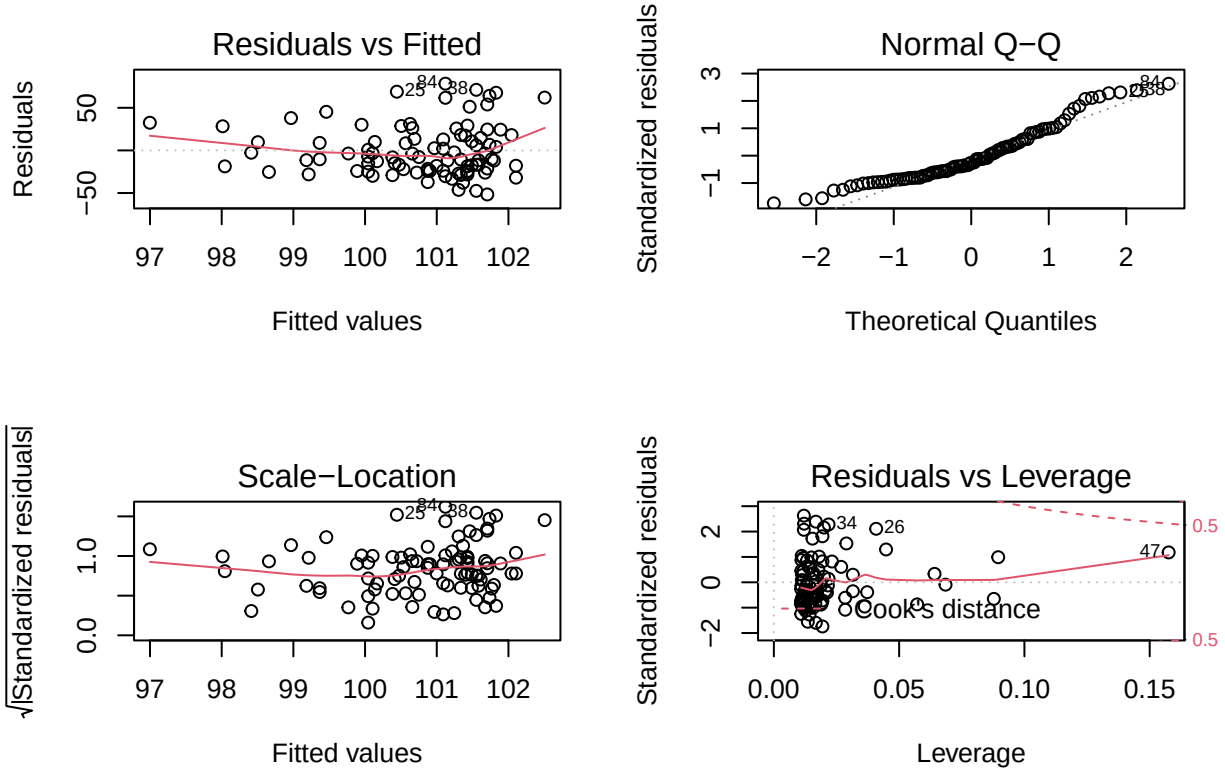


Focusing on the 1987 data, we see the same thing of statistically insignificant results. But this time the sign on the unemployment rate coefficient is opposite what we would intuitively expect. The diagnostic plots mostly look fine however.

Q: What is wrong with this histogram? Answer this question in one sentence in the following text box.

```
model <- lm(crmrte ~ unem, data = crime2)

par(mfrow=c(2,2))
plot(model)
```



Solution:

The residuals vs. fitted plot shows a slight trend in the residuals towards the ends of the distribution of fitted values, but it is not super concerning. A qq plot of the residuals shows deviation from normality for residuals at the extremes, but this is also quite common in regression. There are also no few points with a particularly high leverage, suggesting there are not many outliers.

However, given the panel data structure, we know that this model is not correct despite the lack of obvious signs from the diagnostic plots. This highlights that even if the diagnostic plots look ok you should make sure you understand the structure of the data to be able to design an appropriate model.

11.5 Unobserved Effect Models and Pooled OLS and First-Difference Models

Q: In the model below what are the intercepts when $t = 1$ and when $t = 2$?

Solution: The model is $y_{i,t} = \beta_0 + \delta_0 d2_t + \beta_1 x_{i,t} + a_i + \epsilon_{i,t}$

Note the panel data structure in this model where i indexes location/group and t indexes time period

$d2_t$ is an indicator that is one when $t = 2$ and 0 when $t = 1$ (assuming we only have two periods here).

Hence when $t = 1$, the model returns:

$y_{i,t} = \beta_0 + \beta_1 x_{i,t} + a_i + \epsilon_{i,t}$ since $d2_t = 0$

And when $t = 2$, the model returns:

$y_{i,t} = \beta_0 + \delta_0 + \beta_1 x_{i,t} + a_i + \epsilon_{i,t}$ since $d2_t = 1$

The intercept when $t = 1$ is $\beta_1 + a_i$, and when $t = 2$ the intercept is $\beta_0 + \delta_0 + a_i$.