20-12

Nikhil Gopal and Kerem Tuncer

4/8/2022

```
rm(list = ls())
library(rstanarm)
library(ggplot2)
```

Unis: nsg2127, kt2716

20.12

Working through your own example: Continuing the example from the final exercises of the earlier chapters, consider a treatment effect that can only be estimated using an observational study. Using your data, assess issues of imbalance and lack of overlap. Use matching if necessary to get comparable treatment and control groups, then perform a regression analysis adjusting for matching variables or the propensity score and estimate the treatment effect. Graph the data and fitted model and assess your assumptions.

```
data <- read.csv("final.csv")

data$mortality <- as.numeric(data$mortality)
```

```
## Warning: NAs introduced by coercion
```

```
data$GDP.annual.growth.rate <- as.numeric(data$GDP.annual.growth.rate)
```

```
## Warning: NAs introduced by coercion
```

```
data$GDP.per.capita <- as.numeric(data$GDP.per.capita)
```

```
## Warning: NAs introduced by coercion
```

This is a dataset taken from the World Bank (https://data.worldbank.org/indicator). We wanted to investigate if there was a relationship between a country's infant mortality rate and a country's wealth. It would be hypothesized that richer countries would have lower infant mortality rates. This type of experiment would have to be observational as one single researcher does not have the capacity to change the wealth of a country randomly for experimental purposes. To measure wealth, we chose GDP per capita. GDP is not a perfect estimator of country wealth, as certain countries have lots of undeclared wealth for example, or weird distributions of wealth, but for now it is an easy and quick approximation

We also decided to add a change GDP annual growth rate variable, and its interaction. The rational was certain countries that are in the process of becoming wealthier might invest in health care before other types of infrastructure, which would help to lower infant mortality rate.

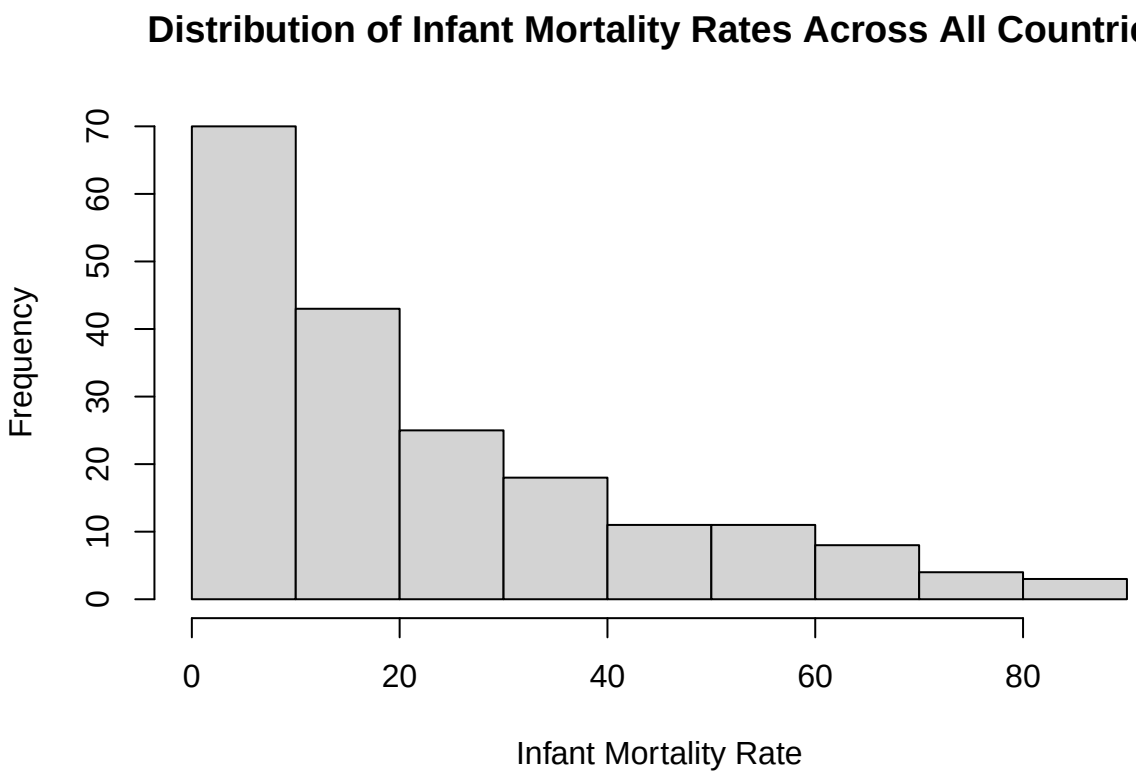
```r
fit <- stan_glm(mortality~GDP.per.capita*GDP.annual.growth.rate, data = data, refresh = 0)

print(fit)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      mortality ~ GDP.per.capita * GDP.annual.growth.rate
##  observations: 184
##  predictors:   4
## ------
##                                       Median MAD_SD
## (Intercept)                           36.7    2.0
## GDP.per.capita                         0.0    0.0
## GDP.annual.growth.rate                -0.7    0.5
## GDP.per.capita:GDP.annual.growth.rate  0.0    0.0
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 16.1    0.8
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

We found that the wealth variable did not have any effect on infant mortality rate, while the growth rate did have a coefficient of -0.7, a small effect given the distribution of mortality rates:

```r
hist(data$mortality, main = "Distribution of Infant Mortality Rates Across All Countries", xlab = "Infa
```

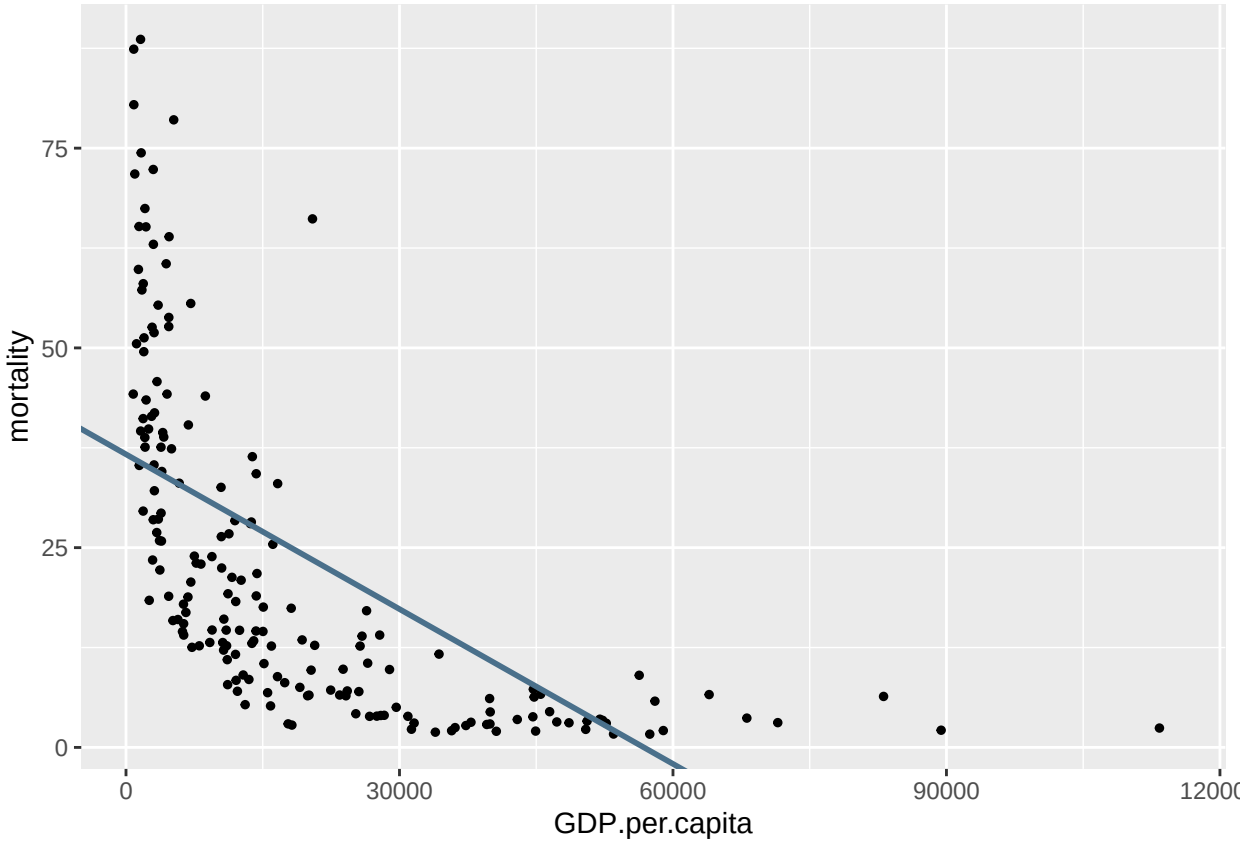


Now lets asses our model fit. Below we will graph a scatter plot of the wealth variable vs the mortality variable, with the best fit line overlayed:

```
base <- ggplot(data, aes(x = GDP.per.capita, y = mortality)) +
  geom_point(size = 1, position = position_jitter(height = 0.05, width = 0.1))

base + geom_abline(intercept = coef(fit)[1], slope = coef(fit)[2],
                   color = "skyblue4", size = 1)
```

```
## Warning: Removed 33 rows containing missing values (geom_point).
```



We observe that there is a poor fit, consistent with the low treatment effect that we observed. There appears to be a non-linear relationship between the variables, although we do notice from the graph (not model) that wealthier countries do have lower mortality rates as hypothesized. Due to the non-linear relationship between the variables, we did not meet the linear relationship assumption of regression and a different type of model would have been better suited to this data.