HW7: Chapter 7, Sections 8.1 to 8.3

Your Name Here

The code below just loads some packages and makes it so that enough digits are printed that you won't get confused by rounding errors.

```
library(dplyr) # functions like summarize
library(ggplot2) # for making plots
library(readr)

options("pillar.sigfig" = 10) # print 10 significant digits in summarize output
```

Crowdedness and GDP

Danielle Vasilescu and Howard Wainer (*Chance*, 2005) used data from the United Nations Center for Human Settlements to investigate aspects of living conditions for several countries. Among the variables they looked at were the country's per capita gross domestic product (GDP, in dollars) and Crowdedness, defined as the average number of persons per room living in homes there. Suppose we want to estimate the relationship between these variables, using GDP as the explanatory variable and Crowdedness as the response.

The following code reads the data in:

```
crowdedness <- read_csv("http://www.evanlray.com/data/sdm4/Crowdedness.csv")
```

```
## Parsed with column specification:
## cols(
##   Country = col_character(),
##   Crowdedness = col_character(),
##   fertility = col_character(),
##   GDP = col_double()
## )
```

```
crowdedness <- crowdedness %>%
  mutate(
    Crowdedness = as.numeric(Crowdedness),
    fertility = as.numeric(fertility)
  ) %>%
  filter(
    !is.na(Crowdedness)
  )
```

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

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

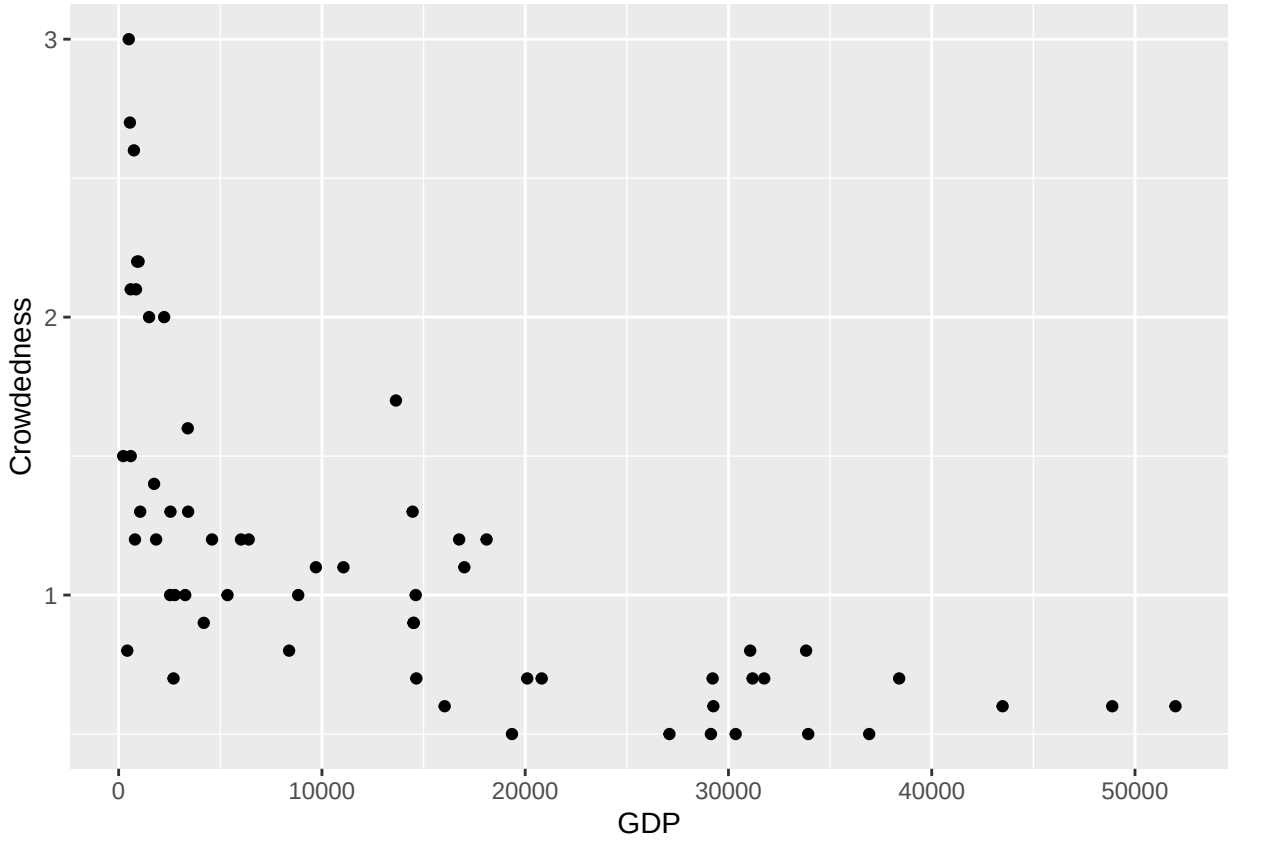
```
head(crowdedness)
```

```
## # A tibble: 6 x 4
##   Country    Crowdedness fertility   GDP
##   <chr>            <dbl>     <dbl> <dbl>
## 1 ARUBA              0.7     NA    20100
## 2 AUSTRIA            0.7      1.28 31187
## 3 AZERBAIJAN         2.1      2.1    853
## 4 BAHAMAS            1.3      2.29 14462
```

```
## 5 BELGIUM            0.6      1.66 29257
## 6 BERMUDA            0.6      1.67 51991
```

(a) Create an appropriate plot of the data.

```
ggplot(data = crowdedness, mapping = aes(x = GDP, y = Crowdedness)) +
  geom_point()
```

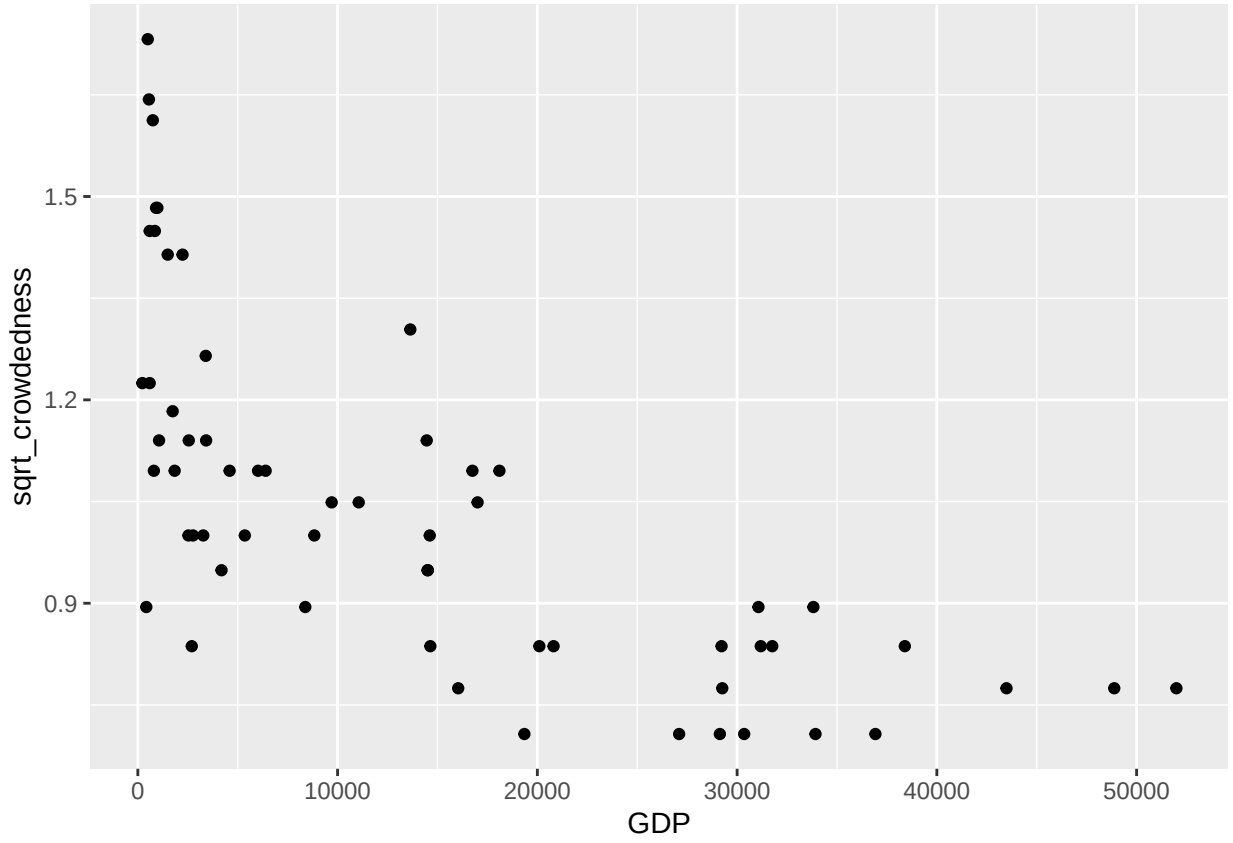


(b) Find a transformation of the data so that the simple linear regression model conditions are as well satisfied as possible. You do not need to show all of the steps in your process; you can just keep your final selected transformation. (It's also fine if you want to keep all of the steps you took for your records.) For your final selected transformation, please create 3 plots: (1) a scatter plot with the transformed variables, (2) a scatter plot of the residuals vs. the transformed explanatory variable, and (3) a histogram or density plot of the residuals. No need to discuss these plots in this part.

I will keep all the steps in my process for you to refer to, but you did not need to do that.

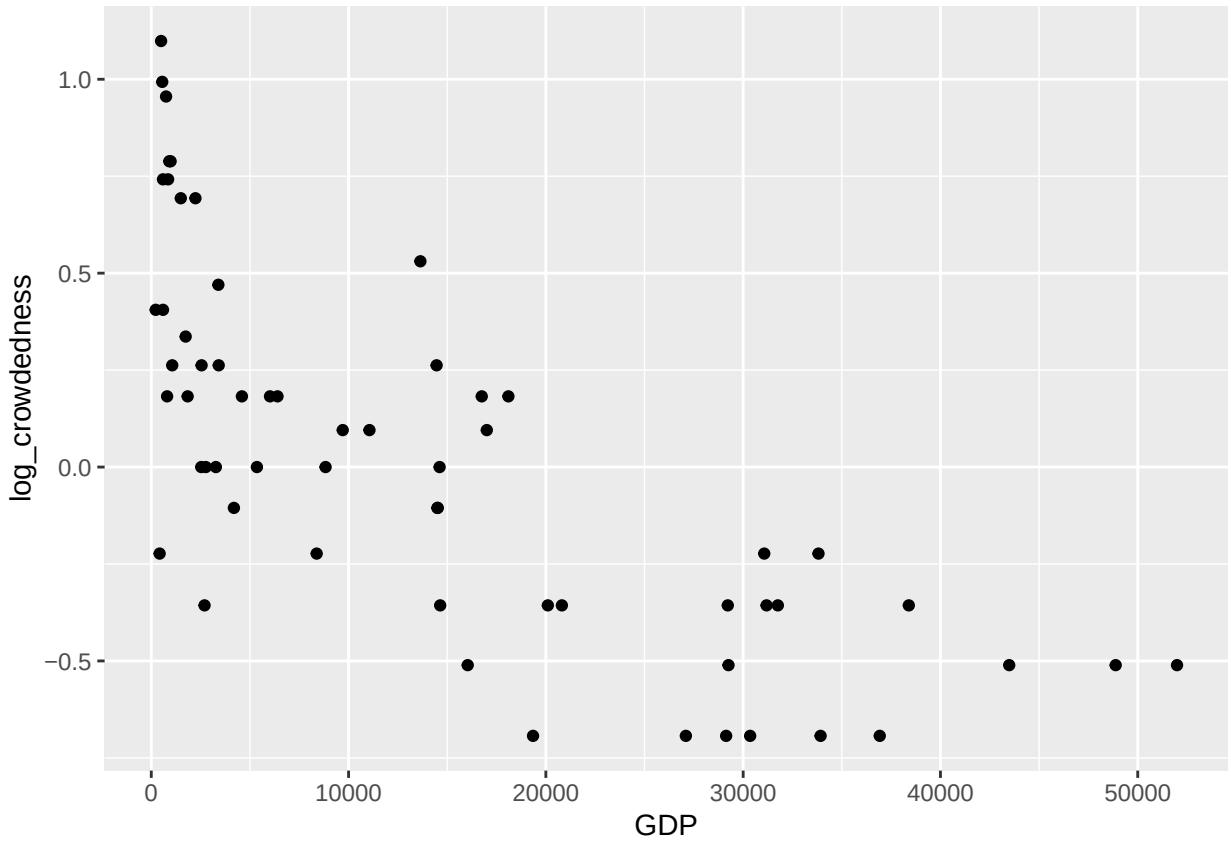
The standard deviation of Crowdedness is larger for small values of GDP than for large values of GDP, which suggests that we should start with a transformation of Crowdedness (the response variable). Since that variable is skewed right, I will move down the ladder of powers.

```
crowdedness <- crowdedness %>%
  mutate(
    sqrt_crowdedness = sqrt(Crowdedness)
  )
ggplot(data = crowdedness, mapping = aes(x = GDP, y = sqrt_crowdedness)) +
  geom_point()
```



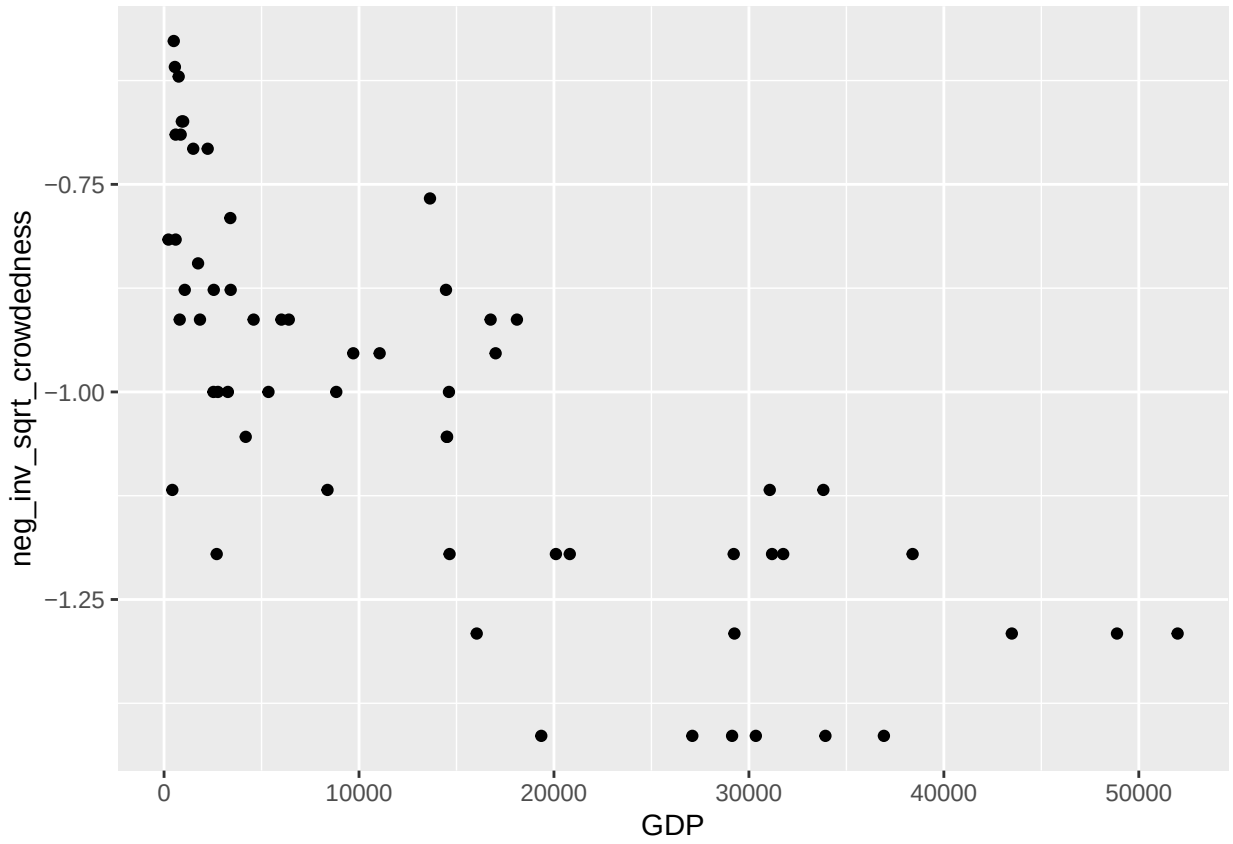
Not good enough. Let's go down another step.

```
crowdedness <- crowdedness %>%
  mutate(
    log_crowdedness = log(Crowdedness)
  )
ggplot(data = crowdedness, mapping = aes(x = GDP, y = log_crowdedness)) +
  geom_point()
```



Still not good enough. Another step down.

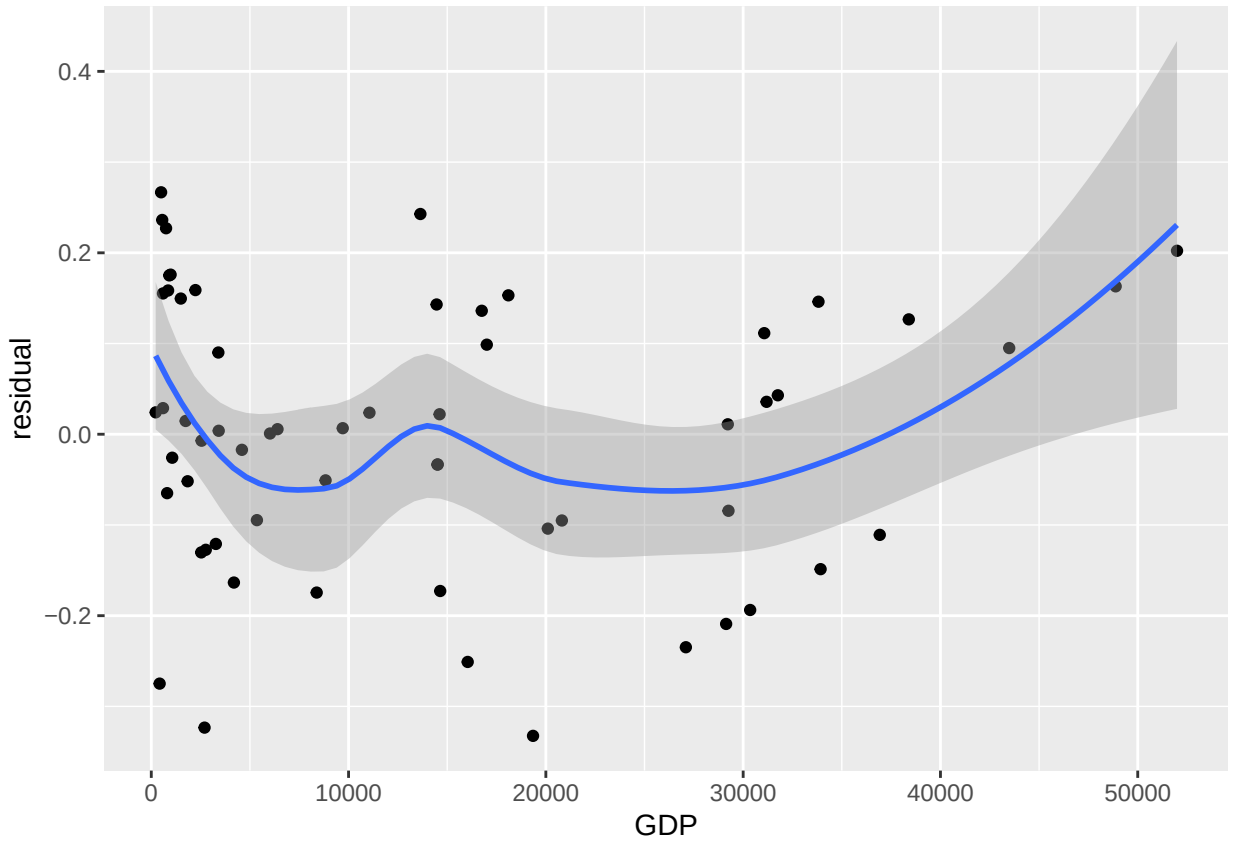
```
crowdedness <- crowdedness %>%
  mutate(
    neg_inv_sqrt_crowdedness = -1/sqrt(Crowdedness)
  )
ggplot(data = crowdedness, mapping = aes(x = GDP, y = neg_inv_sqrt_crowdedness)) +
  geom_point()
```



Much better. Let's look at a residuals plot based on this transformation.

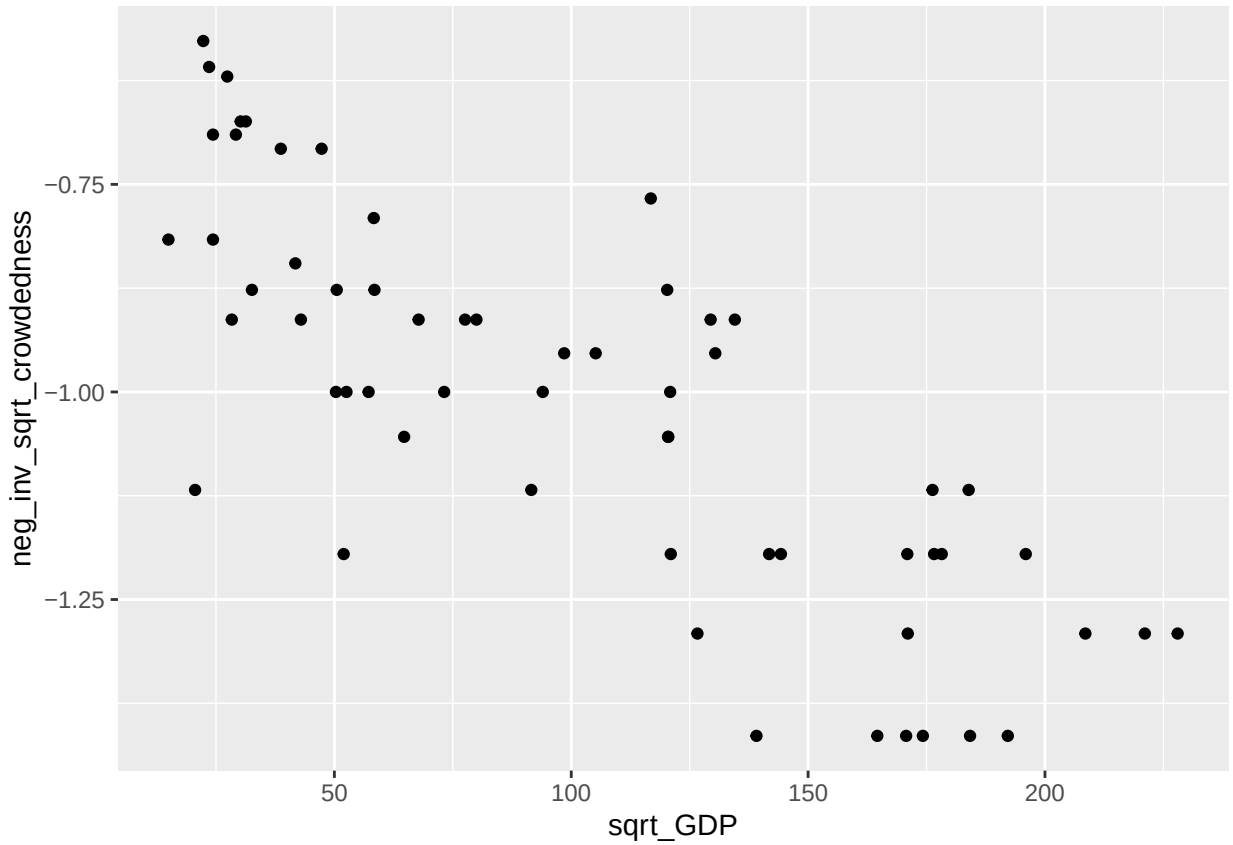
```
lm_fit <- lm(neg_inv_sqrt_crowdedness ~ GDP, data = crowdedness)
crowdedness <- crowdedness %>%
  mutate(
    residual = residuals(lm_fit)
  )

ggplot(data = crowdedness, mapping = aes(x = GDP, y = residual)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

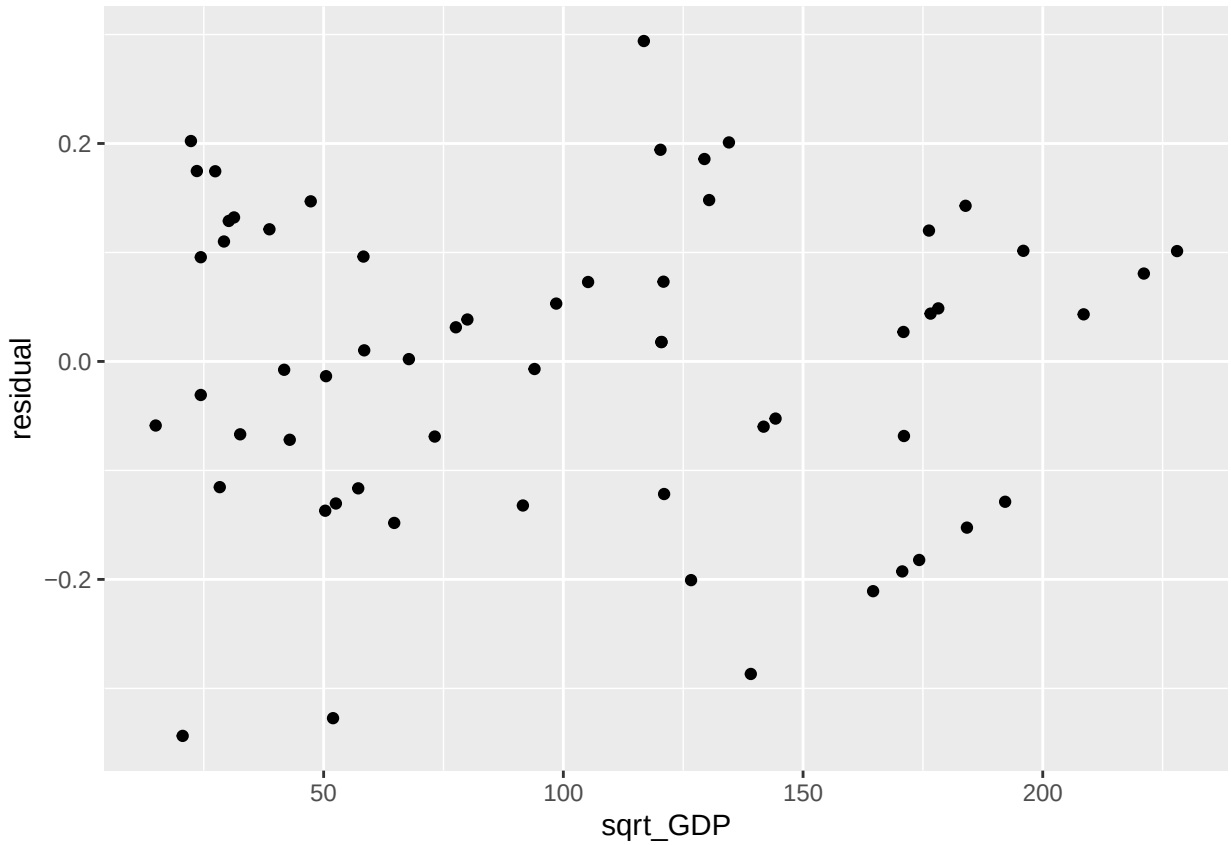


The standard deviations look ok now, but we do still seem to have a bit of a non-linear pattern. The residuals are mostly positive for low GDPs, slightly more negative in the middle, and then mostly positive again. Let's try a transformation of GDP. We will move down the ladder for that variable since it is skewed right.

```
crowdedness <- crowdedness %>%
  mutate(
    sqrt_GDP = sqrt(GDP)
  )
ggplot(data = crowdedness, mapping = aes(x = sqrt_GDP, y = neg_inv_sqrt_crowdedness)) +
  geom_point()
```




```
lm_fit <- lm(neg_inv_sqrt_crowdedness ~ sqrt_GDP, data = crowdedness)
crowdedness <- crowdedness %>%
  mutate(
    residual = residuals(lm_fit)
  )

ggplot(data = crowdedness, mapping = aes(x = sqrt_GDP, y = residual)) +
  geom_point()
```



This looks pretty good. If I wanted to, I could compare the standard deviations of the residuals in the left and right sides of the plot.

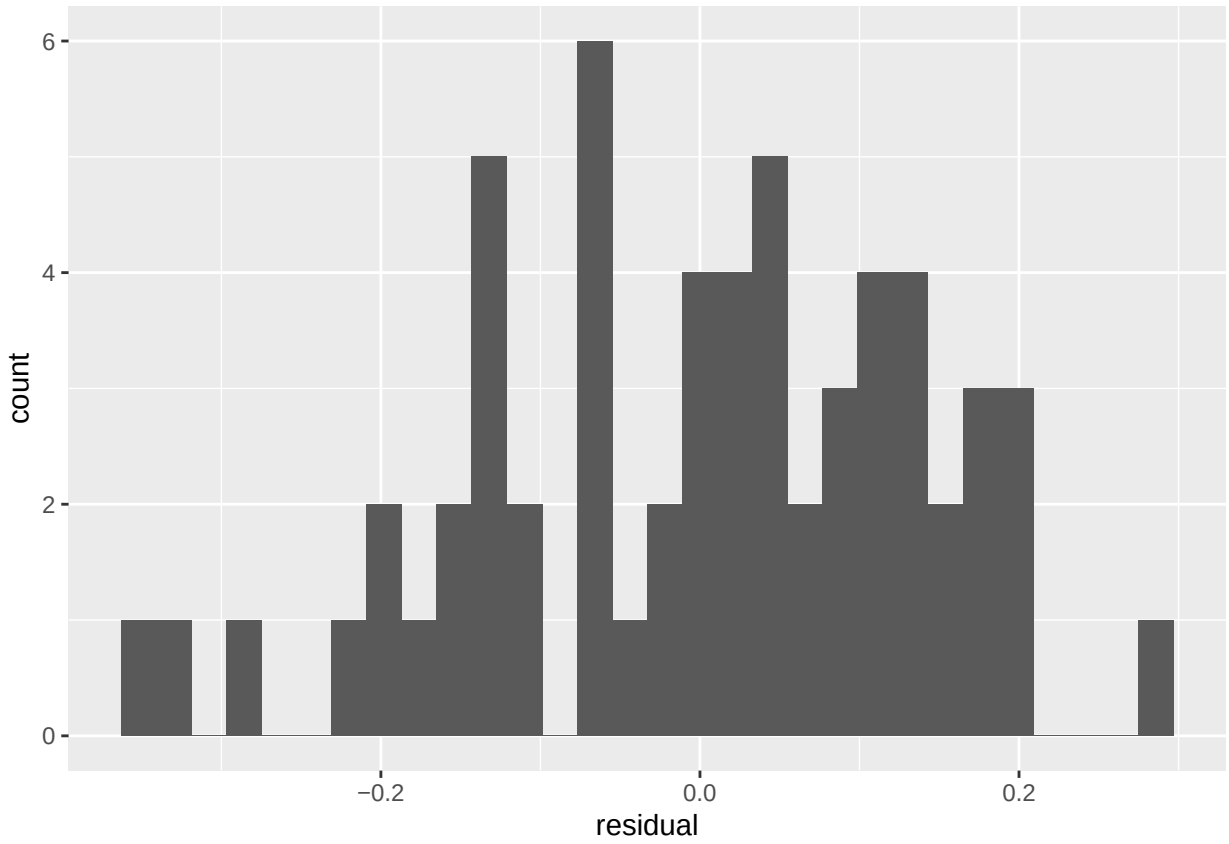
```
crowdedness <- crowdedness %>%
  mutate(
    crowdedness_grouped = ifelse(sqrt_GDP <= 100, "small gdp", "large gdp")
  )

crowdedness %>%
  group_by(crowdedness_grouped) %>%
  summarize(
    sd(residual)
  )
```

```
## # A tibble: 2 x 2
##   crowdedness_grouped `sd(residual)`
##   <chr>                        <dbl>
## 1 large gdp            0.1465930851
## 2 small gdp            0.1356903647
```

Not exactly the same, but quite similar - definitely good enough. Here is a histogram of the residuals.

```
ggplot(data = crowdedness, mapping = aes(x = residual)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



(c) Discuss all of the linear regression model conditions based on your transformed variables. For each condition, you should write a sentence or two describing whether or not the condition is satisfied and why. If your conclusion is based on the plots you made for part (b), please clearly indicate which plot or plots you are looking at and describe a specific characteristic of that plot that your conclusion is based on.

Linear: After transformation, the relationship between $\sqrt{GDP}$ and $-1/\sqrt{Crowdedness}$ is approximately linear. This can be seen in the scatterplot of these variables.

Independence: As usual, it is difficult to assess this condition. I could imagine that the residuals for two culturally similar countries could be similar.

Normally distributed errors around the mean: The histogram of the residuals from this fit shows them to be approximately normally distributed.

Equal standard deviation of residuals for all values of the explanatory variable: The plot of residuals vs. square root of GDP above shows the standard deviations of the residuals to be approximately equal after transformation.

No outliers: The scatter plot of the transformed variables and of the residuals vs. GDP show no outliers.

(d) What are the interpretations of the estimated intercept and slope? Please interpret the coefficient estimates in context on the scale of the *transformed* data.

```
summary(lm_fit)
```

```
##
## Call:
## lm(formula = neg_inv_sqrt_crowdedness ~ sqrt_GDP, data = crowdedness)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34357 -0.11561  0.01783  0.10367  0.29404
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.7131360  0.0351460  -20.29  < 2e-16 ***
## sqrt_GDP    -0.0029784  0.0002967  -10.04  2.7e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1413 on 58 degrees of freedom
## Multiple R-squared:  0.6347, Adjusted R-squared:  0.6284
## F-statistic: 100.8 on 1 and 58 DF,  p-value: 2.698e-14
```

Intercept: We estimate that in the sub-population of countries with a GDP of 0, the mean of $\frac{-1}{\sqrt{Crowdedness}}$ is about -0.713.

Slope: We estimate that in all countries, a 1 unit increase in the square root of GDP is associated with a decrease in $\frac{-1}{\sqrt{Crowdedness}}$ of about -0.003.

(e) Find a set of three Bonferroni-adjusted confidence intervals with familywise confidence level of 95% for the median crowdedness in the "population" for countries with a GDP of $5000, $25000, and $45000. Interpret your intervals in context. You can use the predict function to generate the confidence intervals on the transformed scale, but you will have to then transform back to the original data scale.

To achieve a familywise confidence level of 95% using the Bonferroni adjustment, our three confidence intervals will have individual confidence levels of 98.3%.

My selected transformation for the explanatory variable was a square root. Therefore, I need to first create a data frame with the square root of the specified GDPs to use as an input to the model. The resulting confidence interval will be for the mean of the response variable in the model, $Y = \frac{-1}{\sqrt{Crowdedness}}$, at the given values of GDP. We will then need to transform back. Solving the equation $Y = \frac{-1}{\sqrt{Crowdedness}}$ for Crowdedness in terms of Y, we obtain $\sqrt{Crowdedness} = \frac{-1}{Y}$, or $Crowdedness = \frac{1}{Y^2}$.

```
predict_data <- data.frame(
  GDP = c(5000, 25000, 45000)
) %>%
  mutate(
    sqrt_GDP = sqrt(GDP)
  )

predictions <- predict(lm_fit, newdata = predict_data, level = 0.983, interval = "confidence")
predictions <- as.data.frame(predictions) %>%
  mutate(
    lwr_orig_scale = 1/lwr^2,
    upr_orig_scale = 1/upr^2
  )
predictions
```

```
##          fit        lwr       upr lwr_orig_scale upr_orig_scale
## 1 -0.9237432 -0.9738033 -0.873683      1.0545265      1.3100631
## 2 -1.1840679 -1.2451475 -1.122988      0.6449980      0.7929568
```

```
## 3 -1.3449575 -1.4374221 -1.252493      0.4839844      0.6374550
```

We are 95% confident that in the sub-population of countries with a GDP of $5000, the median crowdedness index is between 1.055 and 1.310, in the sub-population of countries with a GDP of $25000 the median crowdedness index is between 0.645 and 0.793, and in the sub-population of countries with a GDP of $45000 the median crowdedness index is between 0.484 and 0.637.