STAT108 Project Update 4 - Analysis

Yijia Sun

2022-11-28

Contents

1 Research Question

We want to examine the associations of state-level abortion restrictions on abortions rate and cross-state movement to obtain abortion care in the United States, 2017 - 2019.

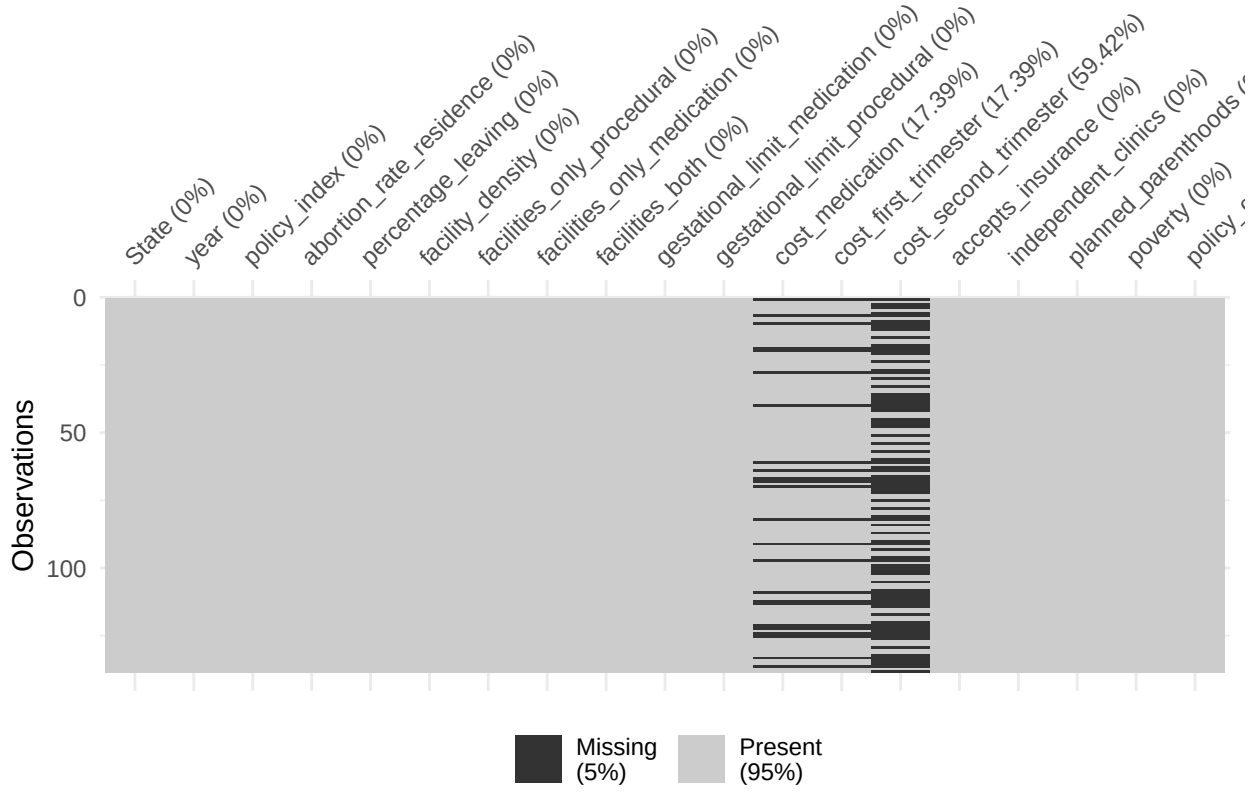
2 Missing Value Evaluation

Before data analysis, we first conduct missing value evaluation.

The plot below provides a visualization of the amount of missing data, showing in black the location of missing values. The percentage of missing values for each variable is shown. It indicates that `cost_medication` and `cost_first_trimester` have 17.39% missing value, and `cost_second_trimester` has 59.42% missing value. Therefore `cost_second_trimester` has the most missing values.

```
vis_miss(df)
```

```
## Warning: `gather_()` was deprecated in tidyr 1.2.0.
## Please use `gather()` instead.
```

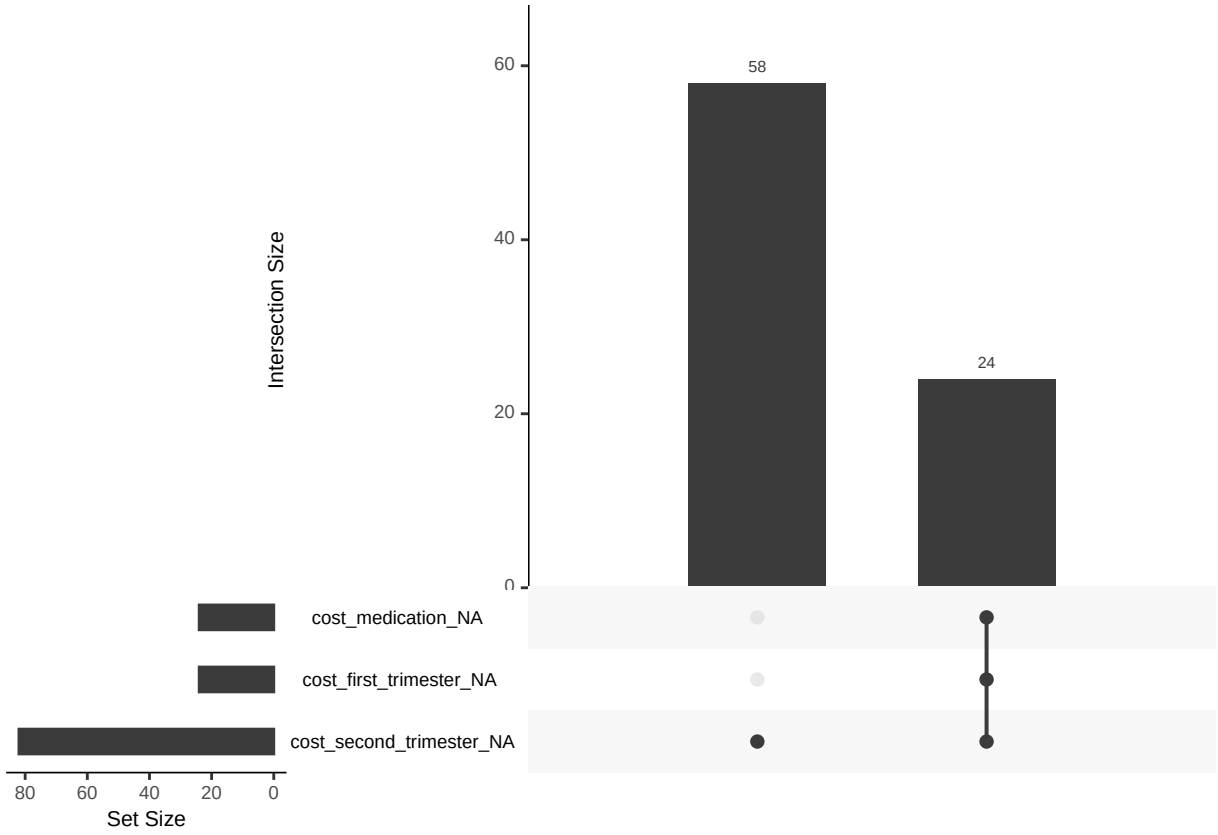


2.1 Exploring patterns of missingness

After identifying missing value, we have to check for the patterns to determine which way of missing imputation is needed.

We used an upset plot from the `UpSetR` package to visualize the intersections of missingness. The plot below shows that in most (58) cases, only `cost_second_trimester` has missing value. But there are 24 cases where `cost_medication`, `cost_first_trimester`, and `cost_second_trimester` have missing values together.

```
gg_miss_upset(df)
```



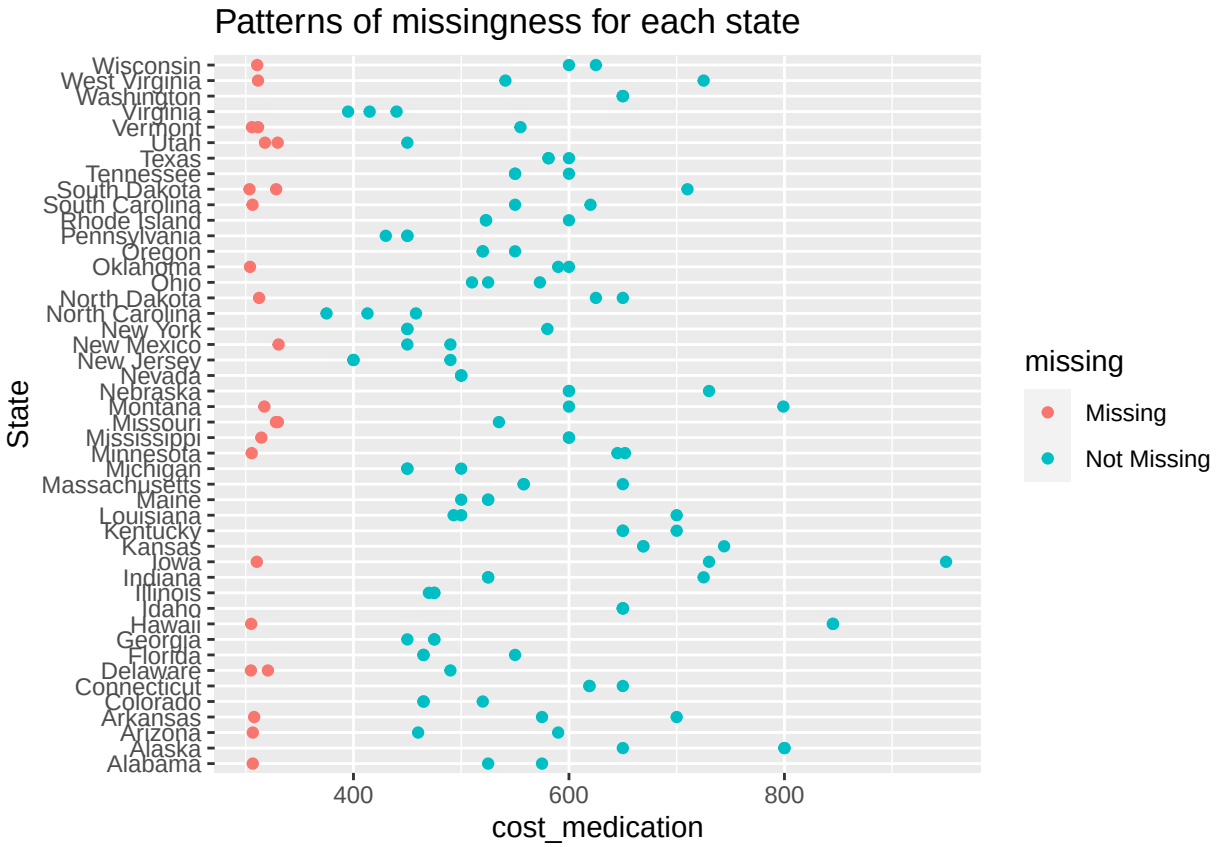
Therefore, since the missing rate of `cost_second_trimester` (59.42%) is too large to analyze, we'll not include this variable for the rest of our analysis.

```
df_drop <- df[, -14]
```

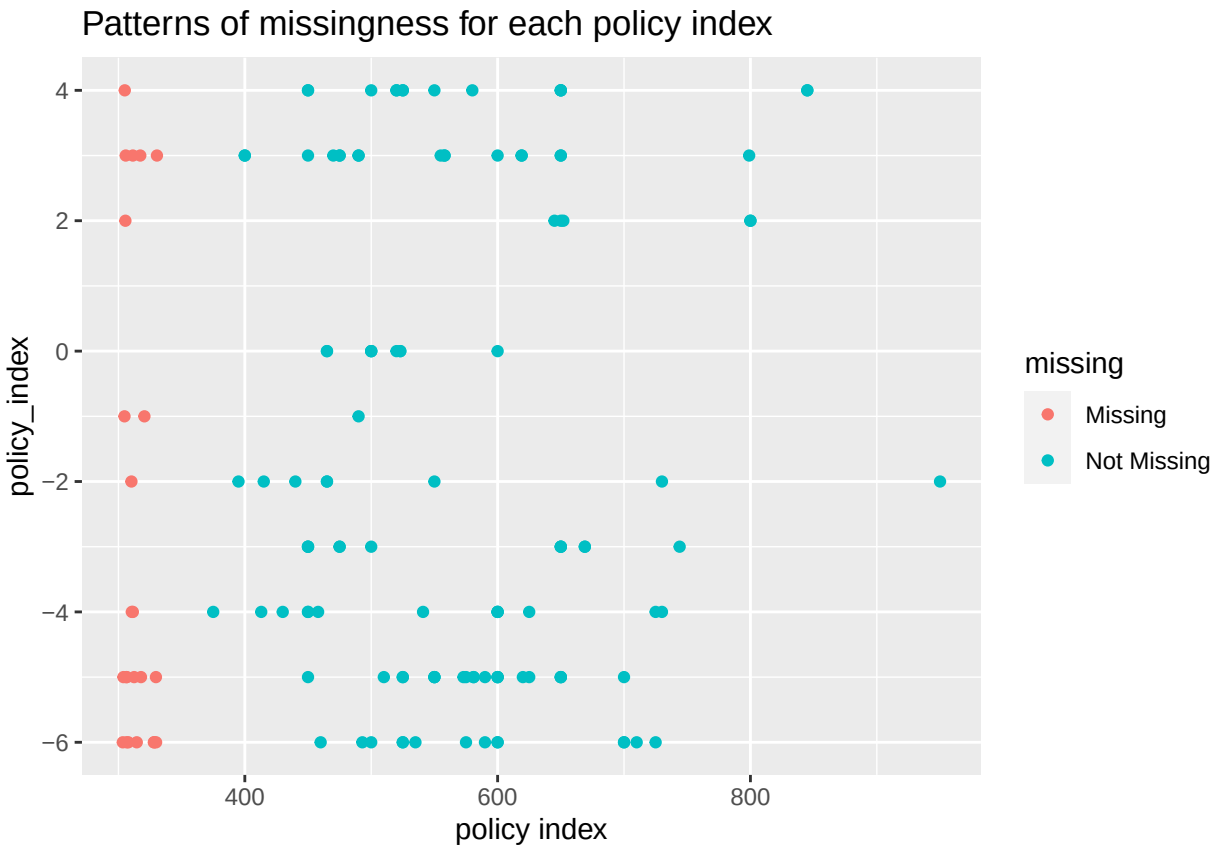
For the missingness of `cost_medication` and `cost_first_trimester`, we explore more patterns related to its abortion policy `policy_index`. Since missingness of these two variables are together in all cases, we only visualize `cost_medication` here.

From the plots below, we find that in most states, missingness only occurs for one year. For Vermont, Utah, South Dakota, and Delaware, missingness occurs for two years. Also, missingness occurs more often in states with policy index 3, -5, and -6.

```
ggplot(df_drop,
       aes(x = cost_medication,
           y = State)) +
 geom_miss_point() + ggtitle("Patterns of missingness for each state")
```




```
ggplot(df_drop,
       aes(x = cost_medication,
           y = policy_index)) +
 geom_miss_point() + ggtitle("Patterns of missingness for each policy index") + xlab("policy index")
```



From the observations we found, there's no obvious pattern between missingness and abortion policy. Therefore we think it's reasonable to impute missing values with the mean for the corresponding state. For instance, in Alabama, `cost_medication` for 2017 is missing, but for 2018 and 2019 is 525 and 575 respectively. We'll impute the `cost_medication` for 2017 with the mean of 2018 and 2019 in Alabama, 550.

2.2 Missing Value Imputation

```
df_impute <- df_drop %>%
  group_by(State) %>%
  mutate(cost_medication = ifelse(is.na(cost_medication), mean(cost_medication,na.rm=TRUE), cost_medica
  mutate(cost_first_trimester = ifelse(is.na(cost_first_trimester), mean(cost_first_trimester,na.rm=TRU

head(df_impute)
```

```
## # A tibble: 6 x 18
## # Groups:   State [2]
##   State    year policy~1 abort~2 perce~3 facil~4 facil~5 facil~6 facil~7 gesta~8
##   <chr>   <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 Alabama  2017       -5    6.65    29.2   0.451      20       0      80       9
## 2 Alabama  2018       -5    7.15    31.2   0.451      20       0      80       9
## 3 Alabama  2019       -5    7.13    37.0   0.271      33       0      67       9
## 4 Alaska   2017        2    8.48    11.3   3.65        0      17      83      10
## 5 Alaska   2018        2    8.43     8.16  3.68        0      17      83      10
## 6 Alaska   2019        2    8.38     7.98  3.72        0      17      83      10
```

```
## # ... with 8 more variables: gestational_limit_procedural <dbl>,
## #   cost_medication <dbl>, cost_first_trimester <dbl>, accepts_insurance <dbl>,
## #   independent_clinics <dbl>, planned_parenthoods <dbl>, poverty <dbl>,
## #   policy_catog <chr>, and abbreviated variable names 1: policy_index,
## #   2: abortion_rate_residence, 3: percentage_leaving, 4: facility_density,
## #   5: facilities_only_procedural, 6: facilities_only_medication,
## #   7: facilities_both, 8: gestational_limit_medication
```

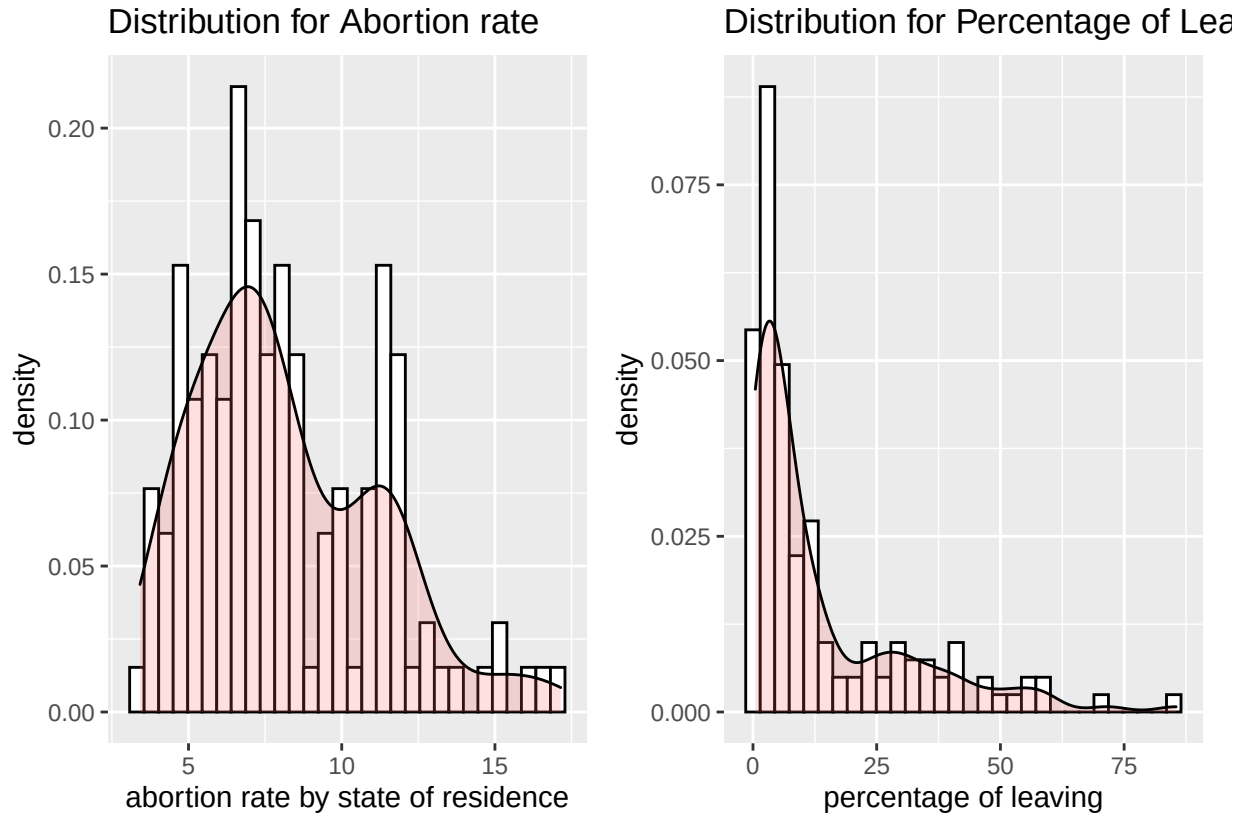
3 Summary of EDA

From last project update, we conducted univariate, bivariate, and multivariate data analysis, without missing value imputation. Therefore, we include a summary and updated version of EDA here.

First, we examined the distribution of outcome variables. The histograms below indicate that both distribution is right-skewed.

```
his1 <- ggplot(df_impute, aes(x = abortion_rate_residence)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white") +
  geom_density(alpha=.2, fill="#FF6666") +
  #facet_grid(policy_catog ~ ., scales = "free") +
  xlab("abortion rate by state of residence") + ggtitle("Distribution for Abortion rate")

his2 <- ggplot(df_impute, aes(x = percentage_leaving)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white") +
  geom_density(alpha=.2, fill="#FF6666") +
  #facet_grid(policy_catog ~ ., scales = "free") +
  xlab("percentage of leaving") + ggtitle("Distribution for Percentage of Leaving")

his1 + his2
```

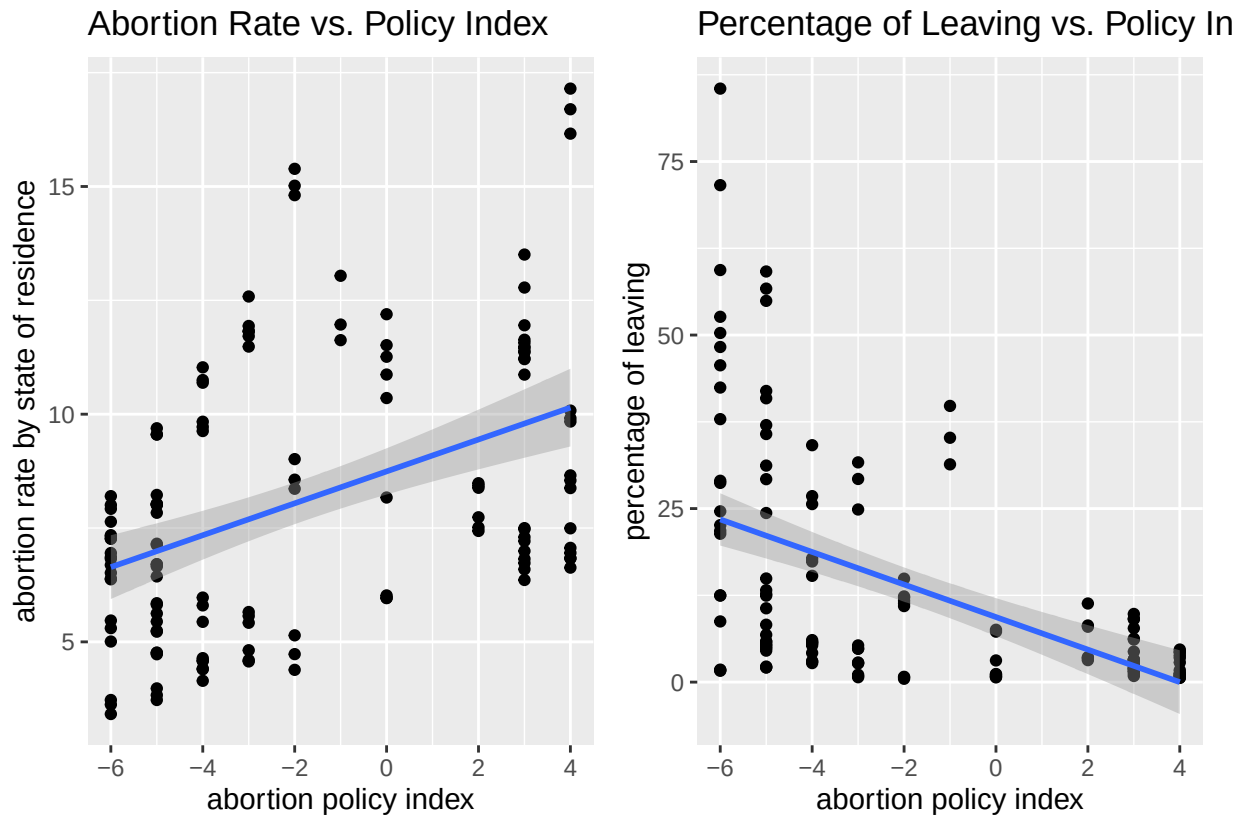
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Then, we check the relationship between each outcome variables and the main covariates of interest `policy_index`. Dotplots below indicate the linear relationship between each outcome variables and policy index.

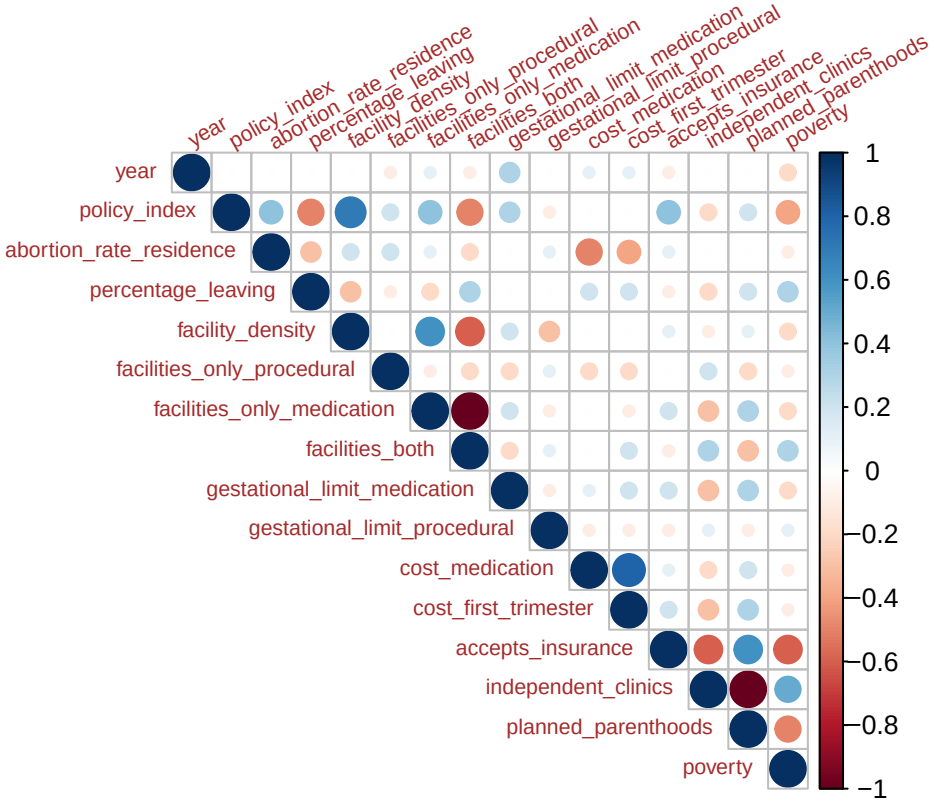
```
dot1 <- ggplot(df_impute, aes(x = policy_index, y = abortion_rate_residence)) +
  geom_point() +
  xlab("abortion policy index") +
  ylab("abortion rate by state of residence") +
  ggtitle("Abortion Rate vs. Policy Index") +
  geom_smooth(method ="lm")

dot2 <- ggplot(df_impute, aes(x = policy_index, y = percentage_leaving)) +
  geom_point() +
  xlab("abortion policy index") +
  ylab("percentage of leaving") +
  ggtitle("Percentage of Leaving vs. Policy Index") +
  geom_smooth(method ="lm")

dot1 + dot2
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



Then, we use correlation plots to check the level of interaction between the variables. Based on the correlation matrix below, we find four pairs of variables to be significantly correlated: `policy_index` & `facility_density`, `facilities_only_medication` & `facilities_both`, `independent_clinics` & `planned_parenthoods`, and `cost_medication` & `cost_first_trimester`.

```
correlation <- df_impute[, -c(1,18)]
corr <- round(cor(correlation,use="pairwise.complete.obs"), 1)
corrplot(corr, tl.col = "brown", bg = "White", tl.srt=30, tl.cex =0.7,type = "upper")
```



4 Building the Model

We will use multiple linear regression model to generate inference to the general association between abortion policy and outcome in the United States. Based on the dotplots, we can identify some linear relationship between abortion rate/percentage of leaving and policy index. Therefore, we start building out model with multiple linear regression.

For each regression model, we will first conduct model selection with AIC by removing the insignificant variables and adding new interaction terms.

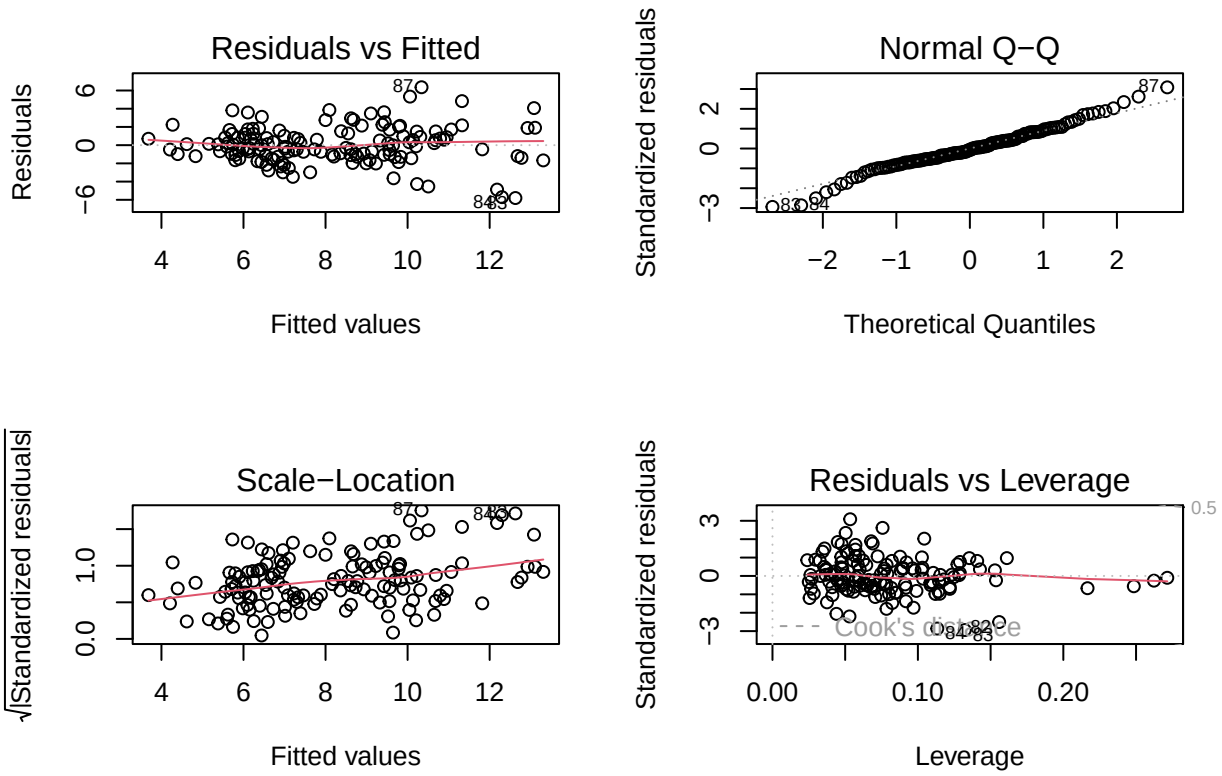
4.1 Multiple Linear Regression for Abortion Rate

```
# exclude State, policy_catog, percentage_leaving
df1 <- df_impute[, -c(1, 5, 18)]

linear_reg <- lm(abortion_rate_residence
                 ~ .
                 + policy_index * facility_density
                 + facilities_only_medication * facilities_both
                 + independent_clinics * planned_parenthoods
                 + cost_medication * cost_first_trimester,
                 data = df1)
```

```
lm <- step(linear_reg, k = 2, trace=0)

# summary(lm)
tidy(lm) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-1125.671	473.282	-2.378	0.019
year	0.457	0.231	1.979	0.050
policy_index	0.582	0.084	6.912	0.000
facility_density	0.796	0.403	1.977	0.050
facilities_only_procedural	2.166	0.602	3.599	0.000
facilities_only_medication	2.135	0.601	3.551	0.001
facilities_both	2.142	0.602	3.560	0.001
gestational_limit_procedural	0.095	0.069	1.378	0.171
cost_medication	-0.011	0.002	-6.291	0.000
poverty	0.155	0.074	2.082	0.039
policy_index:facility_density	-0.306	0.096	-3.193	0.002

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



4.1.1 Discussion of the assumptions for the model

In the residual plots, no outliers and no obvious pattern is found and variance of residuals are constant. In the Normal Q-Q plot, we find that residuals from our model forms a nice normal distribution. Therefore, based on these residuals, we can conclude that our model meets the assumption.

4.1.2 Interpretations and findings from the model coefficients

In our model, we found significant positive correlation (p-value < 0.05) between abortion rate and policy index. Specifically, we found a 0.582 unit increase in abortion rate for every one unit increase in policy index (assume other variables don't change), that is, a more supportive abortion policy is correlated with an increase in residence abortion rate.

Also, we found significant positive correlation (p-value < 0.05) of abortion rate with percentage of facilities with only procedural abortion, only medication abortion, and both. One unit increase in percentage of facilities with only procedural, medication, or both is correlated with approximately 2 unit increase in abortion rate (assume other variables don't change).

Moreover, cost of medication abortion is negatively correlated (estimate: -0.011, p-value < 0.05) with abortion rate. This means that increase in cost of medication abortion is correlated with decrease in abortion rate. Percentage of poverty is positively correlated with abortion rate, that is, increase in percentage of poverty is correlated with increase in abortion rate.

In this model, we include the interaction between policy index and facility density. We found that each additional one unit of facility density decrease the effectiveness of policy index on abortion rate by 0.306 unit. This implies that with more facility density, abortion policy tends to have less correlation with abortion rate.

4.2 Multiple Linear Regression for Percentage of Leaving

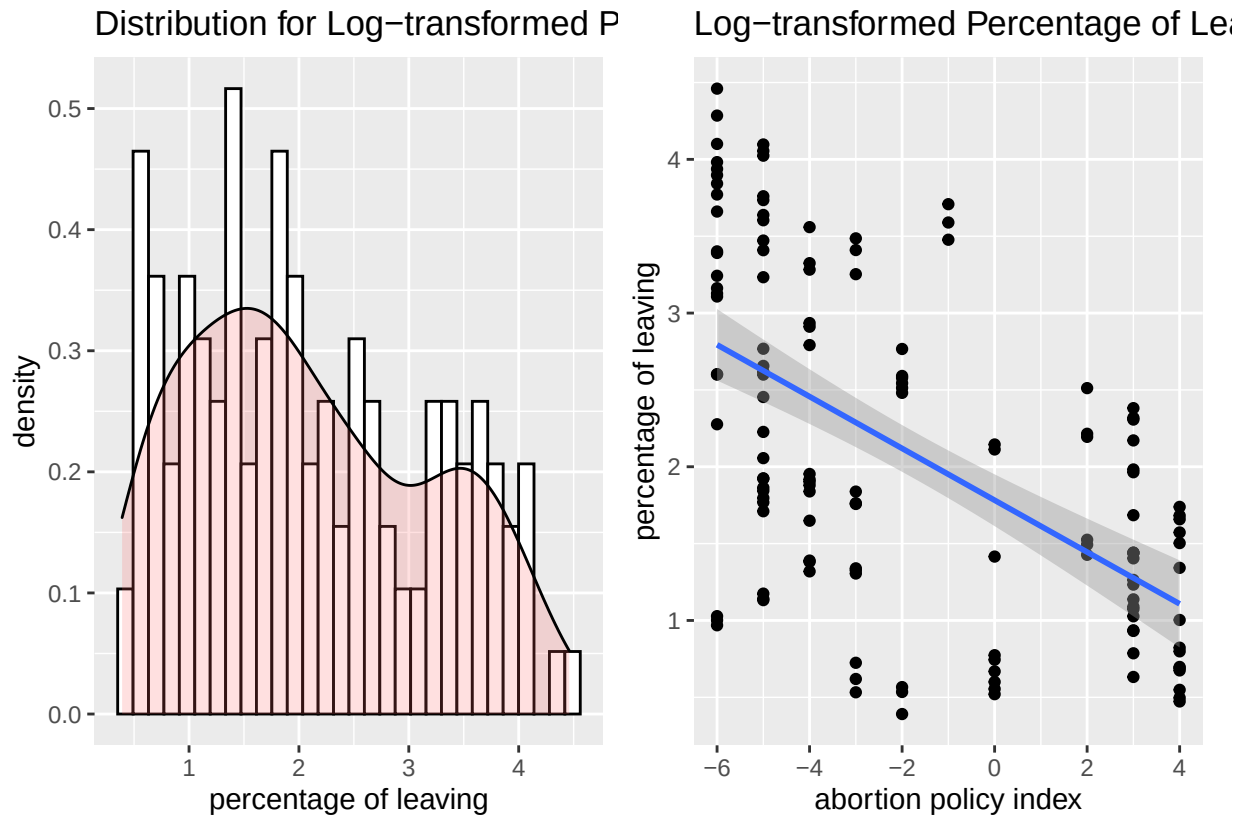
From the previous EDA, we identify the strongly skewed distribution in percentage of leaving, which might influence our model accuracy. Therefore, we adapt log transformation on percentage of leaving.

4.3 Log Transformation on Percentage of Leaving

Through the histogram and dotplots below, we find that after log transformation, the distribution becomes more normal.

```
log1 <- ggplot(df_impute, aes(x = log1p(percentage_leaving))) +
  geom_histogram(aes(y=..density..), colour="black", fill="white") +
  geom_density(alpha=.2, fill="#FF6666") +
  #facet_grid(policy_catog ~ ., scales = "free") +
  xlab("percentage of leaving") + ggtitle("Distribution for Log-transformed Percentage of Leaving")

log2 <- ggplot(df_impute, aes(x = policy_index, y = log1p(percentage_leaving))) +
  geom_point() +
  xlab("abortion policy index") +
  ylab("percentage of leaving") +
  ggtitle("Log-transformed Percentage of Leaving vs. Policy Index") +
  geom_smooth(method ="lm")

log1+log2
```

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

## `geom_smooth()` using formula 'y ~ x'
```



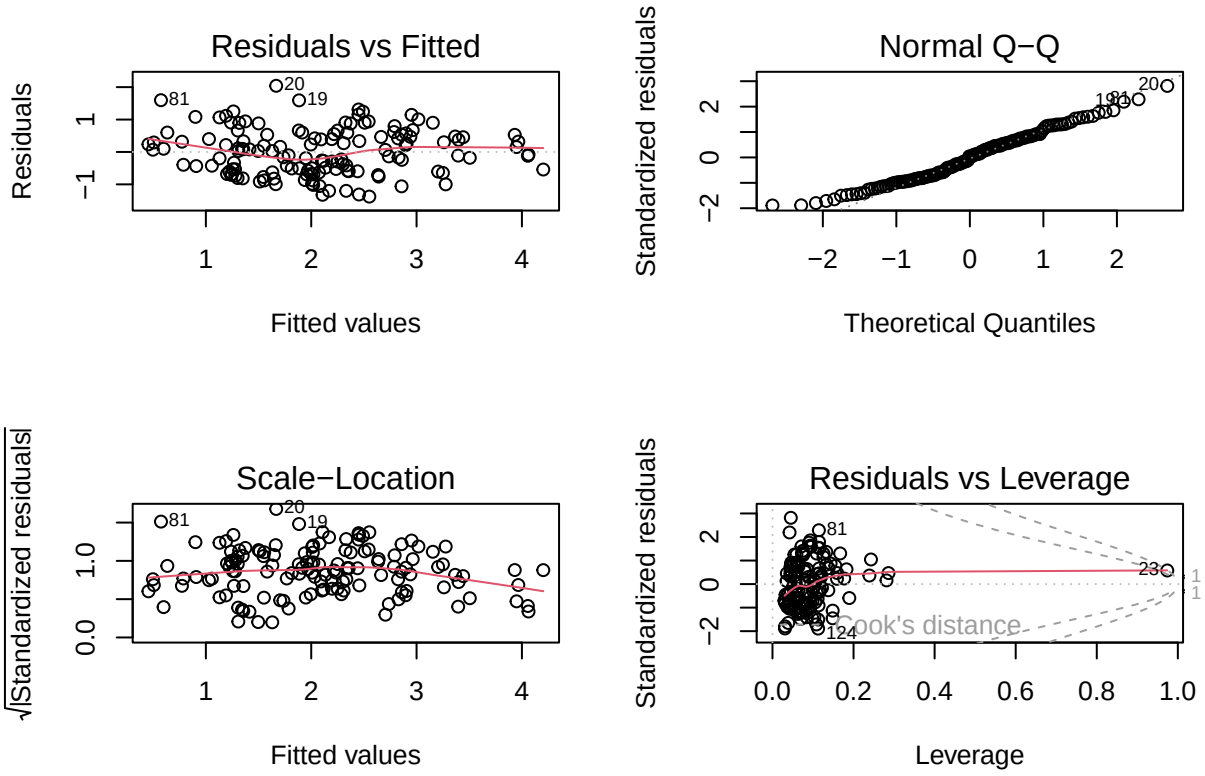
4.4 Log-transformed Multiple Linear Regression

```
# exclude State, policy_catog, abortion_rate_residence
df2 <- df_impute[, -c(1, 4, 18)]

linear_reg2_log <- lm(log1p(percentage_leaving)
                  ~ .
                  + policy_index * facility_density
                  + facilities_only_medication * facilities_both
                  + independent_clinics * planned_parenthoods
                  + cost_medication * cost_first_trimester,
                  data = df2)

lm2_log <- step(linear_reg2_log, k = 2, trace = 0)

#summary(lm2_log)
tidy(lm2_log) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	48.336	22.407	2.157	0.033
policy_index	-0.158	0.036	-4.429	0.000
facility_density	-0.421	0.145	-2.912	0.004
facilities_only_procedural	-0.391	0.215	-1.822	0.071
facilities_only_medication	-0.401	0.215	-1.866	0.064
facilities_both	-0.403	0.215	-1.874	0.063
gestational_limit_procedural	-0.073	0.025	-2.968	0.004
accepts_insurance	0.006	0.003	2.055	0.042
independent_clinics	-0.059	0.131	-0.453	0.651
planned_parenthoods	-0.050	0.131	-0.381	0.704
poverty	0.083	0.029	2.829	0.005
policy_index:facility_density	0.066	0.037	1.776	0.078
independent_clinics:planned_parenthoods	0.000	0.000	-4.163	0.000

```
par(mfrow = c(2, 2))
plot(lm2_log)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



4.4.1 Discussion of the assumptions for the model

In the residual plots, no onvious pattern is found and variance is constant. In the Normal Q-Q plot, we found that residuals from our model forms a nice normal distribution. Therefore, we can conclude that our model meet the assumption. However, one outlier with high leverage is found.

4.4.2 Interpretations and findings from the model coefficients

In our model, we found significant negative correlation (p-value < 0.05) between percentage of leaving and policy index. Specifically, we found a 15.7 percent decrease in percentage of leaving for every one unit increase in policy index (assume other variables don't change), that is, a more supportive abortion policy is correlated with an decrease in percentage of leaving.

We found significant negative correlation (p-value < 0.05) between percentage of leaving and facility density. With one unit increase in facility density, there are 42.1 percent decrease in percentage of leaving, assuming other variables don't change. Also, we found negative correlation (p-value < 0.05) between percentage of leaving and gestational limit for procedural abortion. With one year increase in gestational limit for procedural abortion, there are 7.3 percent decrease in percentage of leaving.

Moreover, percentage of facilities accepting insurance is positively correlated (estimate: 0.006, p-value $<$ 0.05) with percentage of leaving. This means that increase in percentage of facilities accepting insurance is correlated with increase in percentage of leaving. Also, percentage of poverty is positively correlated with percentage of leaving, that is, increase in percentage of poverty is correlated with percentage of leaving.

In this model, we include the interaction between percentage of independent clinics and percentage of planned parenthoods. We found that each additional one unit of percentage of independent clinics decrease the effectiveness of percentage of planned parenthoods on percentage of leaving by 0.03 percent.

5 Additional work

5.1 Multiple Linear Regression for Percentage of Leaving without Log Transformation

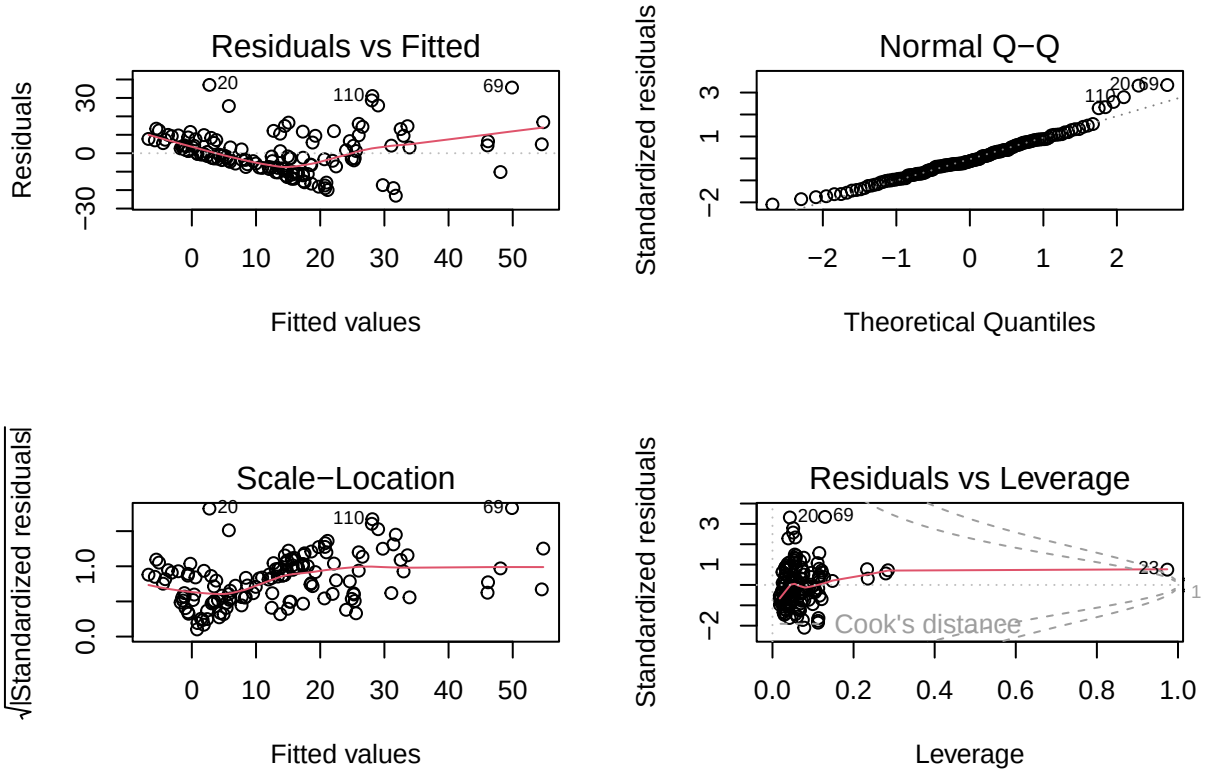
Before log transformation on percentage of leaving, I tried with raw data. However, this model didn't meet the assumption.

```
linear_reg2 <- lm(percentage_leaving
                  ~ .
                  + policy_index * facility_density
                  + facilities_only_medication * facilities_both
                  + independent_clinics * planned_parenthoods
                  + cost_medication * cost_first_trimester,
                  data = df2)

lm2 <- step(linear_reg2, k = 2, trace = 0)

#summary(lm2)
tidy(lm2) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	42.376	194.334	0.218	0.828
policy_index	-2.288	0.483	-4.736	0.000
facility_density	-6.414	2.171	-2.955	0.004
facilities_only_medication	-0.370	0.179	-2.066	0.041
facilities_both	-0.318	0.171	-1.860	0.065
independent_clinics	-0.115	1.947	-0.059	0.953
planned_parenthoods	0.178	1.950	0.091	0.927
poverty	1.159	0.430	2.696	0.008
policy_index:facility_density	1.483	0.555	2.672	0.009
independent_clinics:planned_parenthoods	-0.007	0.001	-5.368	0.000

```
par(mfrow = c(2, 2))
plot(lm2)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



5.1.1 Discussion of the assumptions for the model

In the residual plots, no outliers but specific pattern is found. Fitted values less than 20 tend to have a decreasing trend in residuals. In the Normal Q-Q plot, we found that residuals from our model forms a nice

normal distribution. Therefore, based on these residuals, we can conclude that our model doesn't meet the assumption of homoscedasticity and we might need to transform data.

5.2 Multiple Linear Regression with abortion policy category

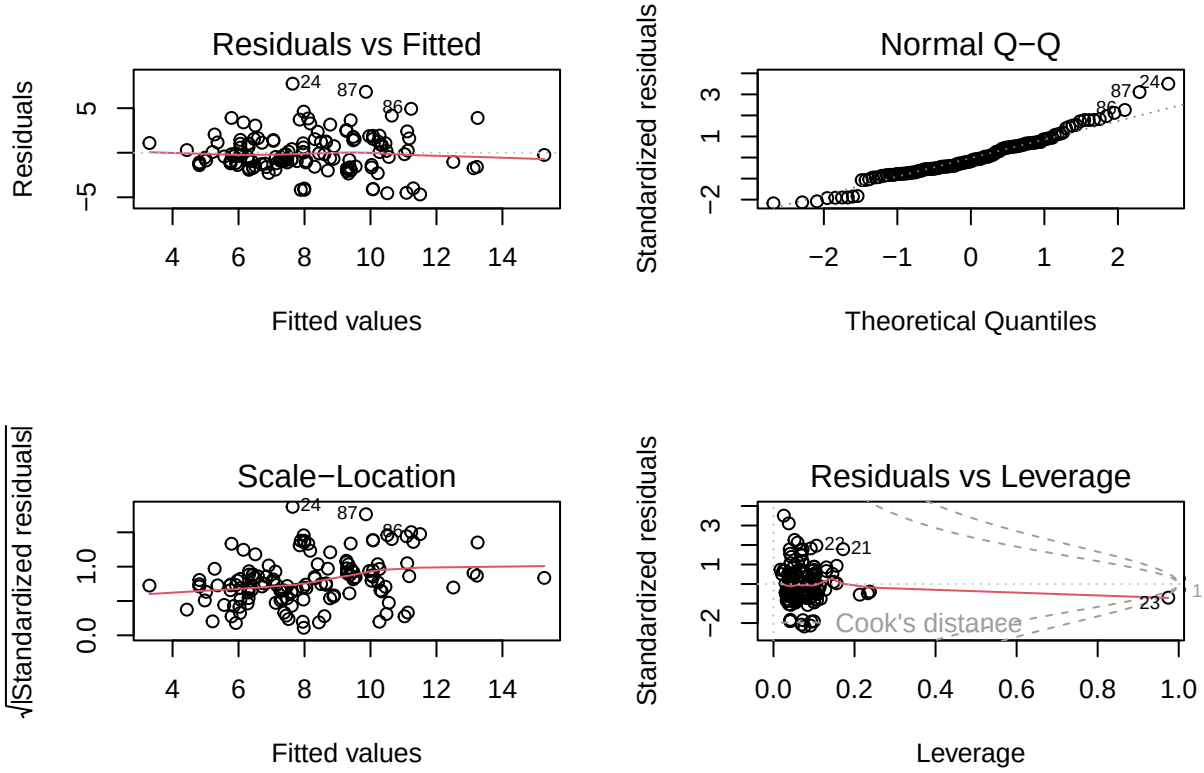
I tried regression model with only policy category, however,

```
df1_cat <- df_impute[, -c(1, 3, 5)]
linear_reg_cat <- lm(abortion_rate_residence
                  ~ .
                  + facilities_only_medication * facilities_both
                  + independent_clinics * planned_parenthoods
                  + cost_medication * cost_first_trimester,
                  data = df1_cat)

lm_cat <- step(linear_reg_cat, k = 2, trace = 0)

# summary(lm_cat)
tidy(lm_cat) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-267.990	69.836	-3.837	0.000
facilities_only_procedural	2.004	0.671	2.985	0.003
facilities_only_medication	1.941	0.676	2.871	0.005
facilities_both	1.975	0.673	2.932	0.004
gestational_limit_procedural	0.108	0.070	1.540	0.126
cost_medication	-0.010	0.002	-5.345	0.000
independent_clinics	0.813	0.385	2.113	0.037
planned_parenthoods	0.811	0.385	2.109	0.037
policy_catogneutral	2.104	0.762	2.763	0.007
policy_catogsupportive	2.587	0.490	5.280	0.000
facilities_only_medication:facilities_both	0.001	0.000	1.993	0.048

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

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```




```
df2_cat <- df_impute[, -c(1, 3, 4)]
linear_reg2_log_cat <- lm(log1p(percentage_leaving)
                  ~ .
                  + facilities_only_medication * facilities_both
                  + independent_clinics * planned_parenthoods
                  + cost_medication * cost_first_trimester,
                  data = df2_cat)

lm2_log_cat <- step(linear_reg2_log_cat, k = 2, trace = 0)

#summary(lm2_log_cat)
tidy(lm2_log_cat) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	54.881	23.390	2.346	0.021
facility_density	-0.304	0.090	-3.366	0.001
facilities_only_procedural	-0.378	0.226	-1.673	0.097
facilities_only_medication	-0.385	0.226	-1.703	0.091
facilities_both	-0.386	0.226	-1.706	0.090
gestational_limit_procedural	-0.079	0.026	-3.084	0.003
independent_clinics	-0.132	0.135	-0.980	0.329
planned_parenthoods	-0.119	0.134	-0.889	0.376
poverty	0.086	0.028	3.012	0.003

term	estimate	std.error	statistic	p.value
policy_catogneutral	-0.356	0.271	-1.313	0.191
policy_catogsupportive	-0.587	0.209	-2.809	0.006
independent_clinics:planned_parenthoods	0.000	0.000	-5.370	0.000

```
par(mfrow = c(2, 2))
plot(lm2_log_cat)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```