CDC 500 Cities: Healthcare Access, Behaviors, and Health Outcomes

Stat 198 Final Project

Maya Ghanem and Isabelle Xiong

11/1/2021

Description of Data

(Include description of how you edited the data)

Research Questions

1) Do cities with a greater lack of healthcare access have poorer mental health and/or physical health outcomes?

2)Does healthcare access, mental health, and/or physical health outcomes vary by state?

Variables of Interest

Explanatory Variables:

1) Healthcare Access for Adults (18+): Percent of City Population that Lacks Insurance, Percent of City Population with visits to doctor for routine checkup within the past year, Percent of City Population who have high blood pressure and are taking medicine for high blood pressure control.
2) Geographic Distribution by State

Response Variables:

1) Behavior for Adults (18+): Percent of city population currently smoking, percent of city population currently reporting binge drinking habits, percent of city population reporting No leisure-time physical activity
2) Health Outcomes for Adults (18+): Percent of city population with coronary heart disease, percent of population diagnosed with diabetes, percent of city population with kidney disease

Linear Regressions

NOTE: Create regressions first between the explanatory (access) variables– this can indicate what kind of interactions are needed.

–> insurance vs. visits to doctor –> insurance vs. medicine –> visits to doctor vs. medicine

NOTE: We will not do third order interactions because they are beyond the scope of this course

Regressions for Healthcare Access and Behaviors Variables

Fit with Interaction Variables

1) Access Variables vs. Smoking

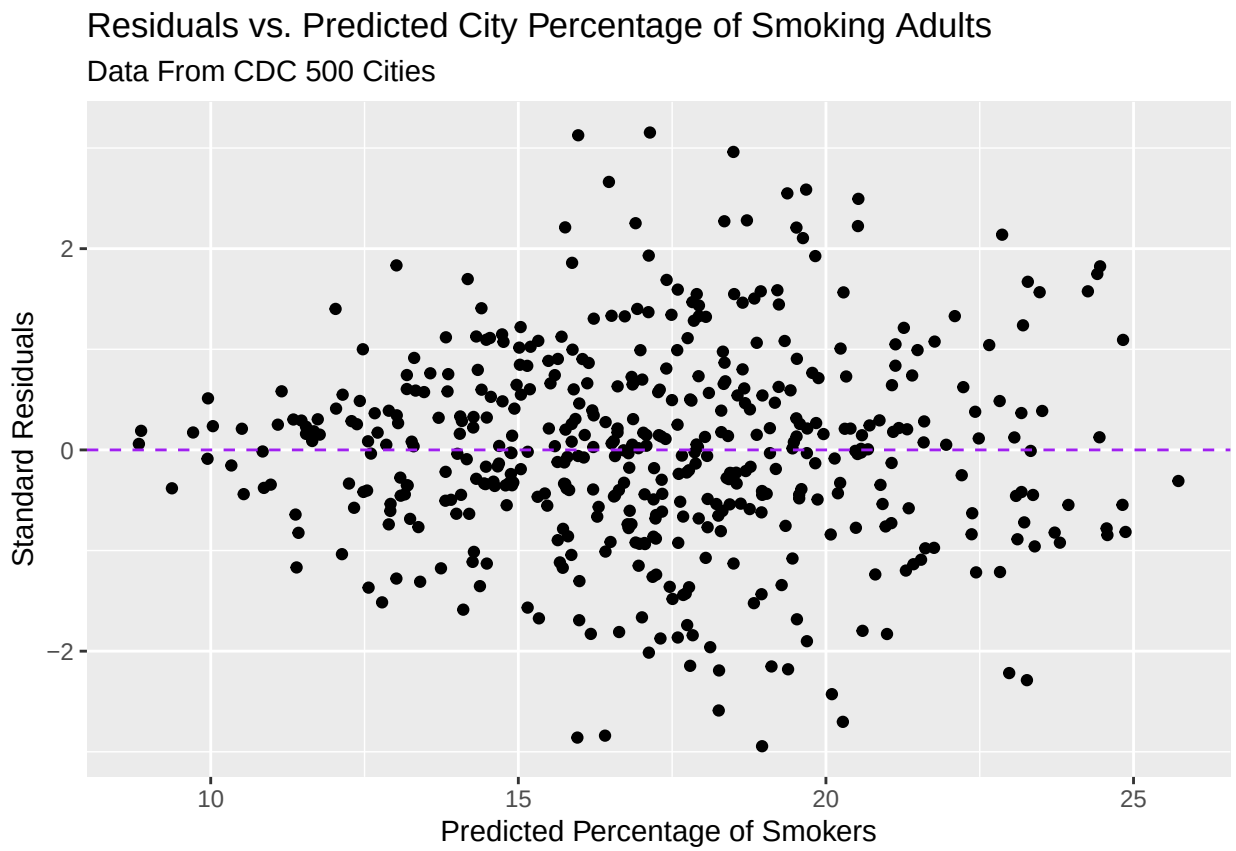
```
int_access_smoking_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(smoking ~ insurance + visits_to_doctor + medicine_high_bp + (insurance * visits_to_doctor) + (insu
int_access_smoking_fit_aug <- augment(int_access_smoking_fit$fit)
tidy(int_access_smoking_fit) %>%
  print()
```

```
## # A tibble: 7 x 5
##   term                              estimate std.error statistic  p.value
##   <chr>                                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                         88.9     24.0          3.70 2.41e- 4
## 2 insurance                            0.872    0.417        2.09 3.71e- 2
## 3 visits_to_doctor                    -2.13     0.362       -5.90 6.95e- 9
## 4 medicine_high_bp                    -0.756    0.463       -1.63 1.03e- 1
## 5 insurance:visits_to_doctor           0.0227   0.00634      3.59 3.69e- 4
## 6 insurance:medicine_high_bp          -0.0414   0.00628     -6.58 1.25e-10
## 7 visits_to_doctor:medicine_high_bp    0.0299   0.00667      4.48 9.60e- 6
```

```
glance(int_access_smoking_fit)$adj.r.squared %>%
  print()
```

```
## [1] 0.5691301
```

```
int_access_smoking_fit_aug %>%
  ggplot(mapping = aes(x = .fitted, y = .std.resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "purple", lty = "dashed") +
  labs(
    title = "Residuals vs. Predicted City Percentage of Smoking Adults",
    subtitle = "Data From CDC 500 Cities",
    x = "Predicted Percentage of Smokers",
    y = "Standard Residuals"
  )
```



2) Access Variables vs. Binge Drinking

```
int_access_binge_drinking_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(binge_drinking ~ insurance + visits_to_doctor + medicine_high_bp + (insurance * visits_to_doctor)
int_access_binge_drinking_fit_aug <- augment(int_access_binge_drinking_fit$fit)
tidy(int_access_binge_drinking_fit) %>%
  print()
```

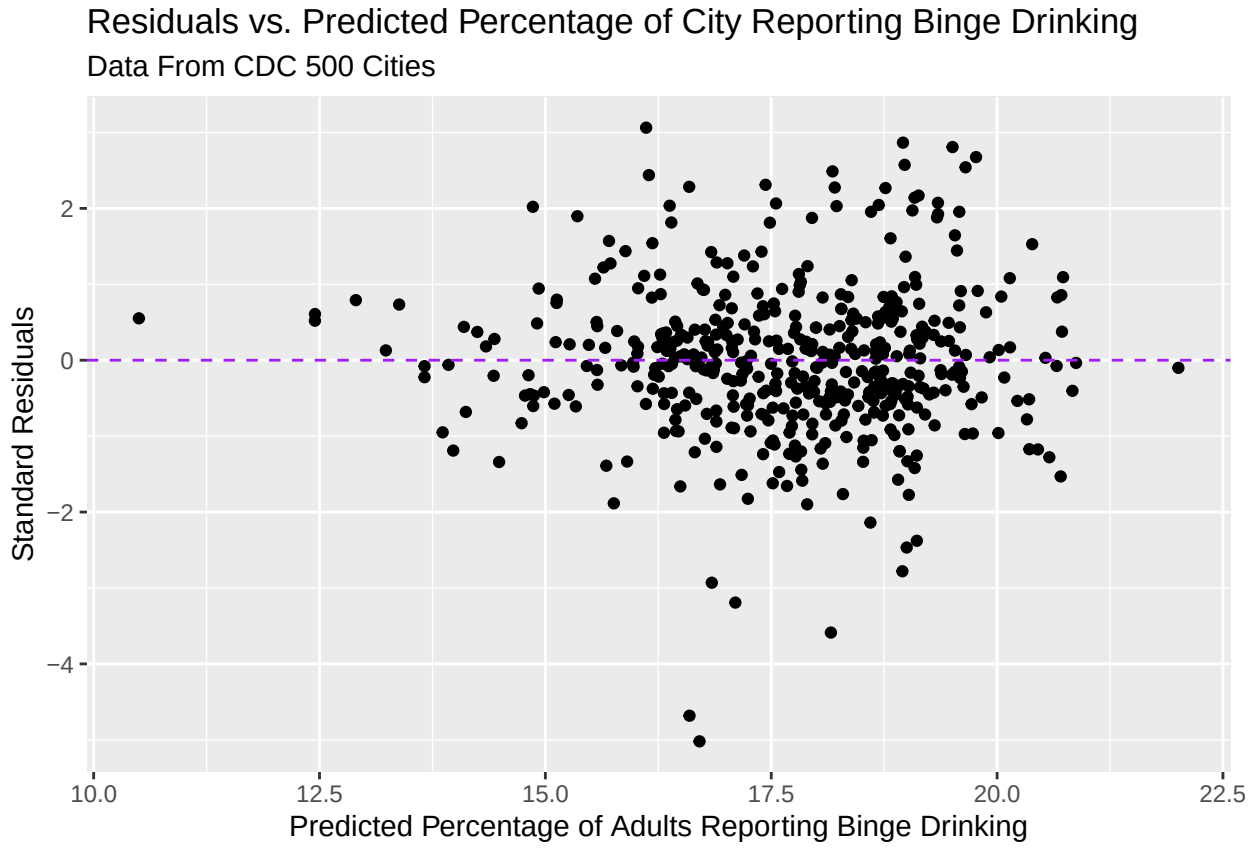
```
## # A tibble: 7 x 5
##   term                                estimate std.error statistic  p.value
##   <chr>                                  <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                         -132.       17.8       -7.40  6.26e-13
## 2 insurance                             -0.125     0.309     -0.406 6.85e- 1
## 3 visits_to_doctor                       2.41      0.268      8.98  6.70e-18
## 4 medicine_high_bp                       2.54      0.344      7.38  7.12e-13
## 5 insurance:visits_to_doctor            -0.00655   0.00470   -1.39  1.64e- 1
## 6 insurance:medicine_high_bp             0.00686   0.00466    1.47  1.42e- 1
## 7 visits_to_doctor:medicine_high_bp     -0.0401    0.00495   -8.10  4.93e-15
```

```
glance(int_access_binge_drinking_fit)$adj.r.squared %>%
  print()
```

```
## [1] 0.3488416
```

```
int_access_binge_drinking_fit_aug %>%
  ggplot(mapping = aes(x = .fitted, y = .std.resid)) +
  geom_point() +
```

```
  geom_hline(yintercept = 0, color = "purple", lty = "dashed") +
  labs(
    title = "Residuals vs. Predicted Percentage of City Reporting Binge Drinking",
    subtitle = "Data From CDC 500 Cities",
    x = "Predicted Percentage of Adults Reporting Binge Drinking",
    y = "Standard Residuals"
  )
```



3) Access Variables vs. Physical Activity

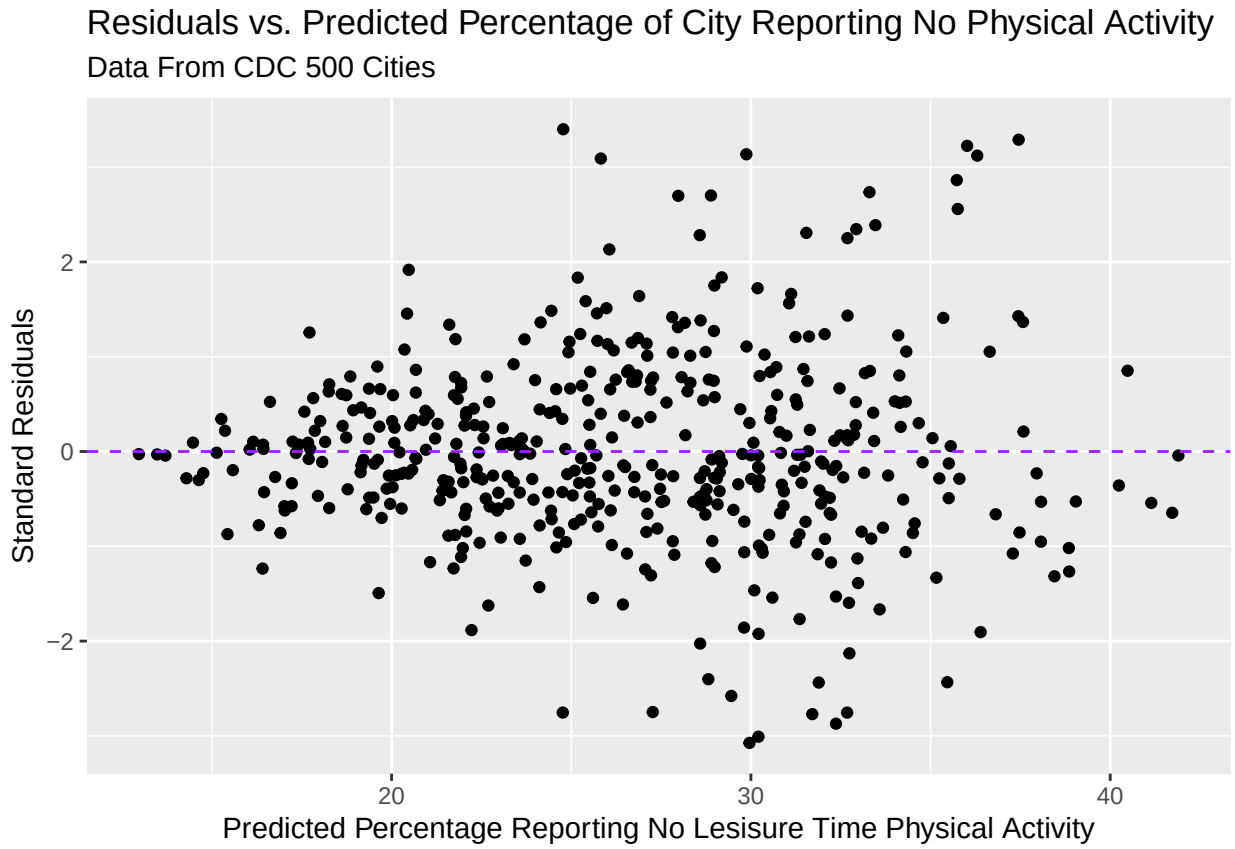
```
int_access_physical_activity_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(physical_activity ~ insurance + visits_to_doctor + medicine_high_bp + (insurance * visits_to_docto
int_access_physical_activity_fit_aug <- augment(int_access_physical_activity_fit$fit)
tidy(int_access_physical_activity_fit) %>%
  print()
```

```
## # A tibble: 7 x 5
##   term                              estimate std.error statistic      p.value
##   <chr>                                <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)                       55.1       20.8         2.64  0.00845
## 2 insurance                          1.96       0.361       5.42  0.0000000972
## 3 visits_to_doctor                  -1.47       0.313      -4.69  0.00000361
## 4 medicine_high_bp                  -0.744      0.402      -1.85  0.0646
## 5 insurance:visits_to_doctor         0.000790   0.00549     0.144 0.886
## 6 insurance:medicine_high_bp        -0.0257     0.00545    -4.72  0.00000317
## 7 visits_to_doctor:medicine_high_bp  0.0271     0.00578     4.68  0.00000373
```

```
glance(int_access_physical_activity_fit)$adj.r.squared %>%
  print()
```

```
## [1] 0.8488063
```

```
int_access_physical_activity_fit_aug %>%
  ggplot(mapping = aes(x = .fitted, y = .std.resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "purple", lty = "dashed") +
  labs(
    title = "Residuals vs. Predicted Percentage of City Reporting No Physical Activity",
    subtitle = "Data From CDC 500 Cities",
    x = "Predicted Percentage Reporting No Lesisure Time Physical Activity",
    y = "Standard Residuals"
  )
```



Regressions for Healthcare Access and Health Outcomes

Fit with Interaction Variables

4) Access Variables vs. Heart Disease

```
int_access_heart_disease_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(heart_disease ~ insurance + visits_to_doctor + medicine_high_bp + (insurance * visits_to_doctor) +
int_access_heart_disease_fit_aug <- augment(int_access_heart_disease_fit$fit)
tidy(int_access_heart_disease_fit) %>%
```

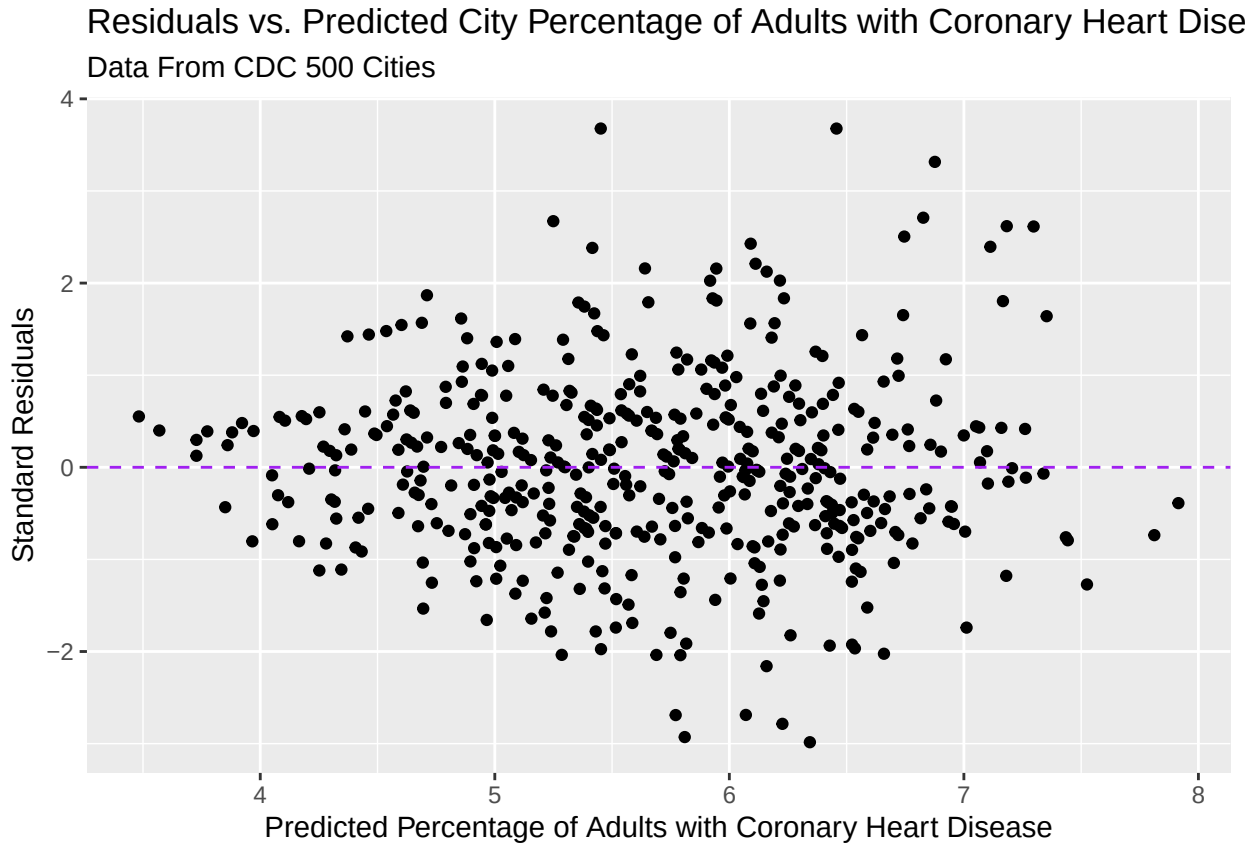
```
  print()
```

```
## # A tibble: 7 x 5
##   term                              estimate std.error statistic  p.value
##   <chr>                                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                       23.9       4.94         4.84 1.74e- 6
## 2 insurance                          0.352     0.0857       4.10 4.79e- 5
## 3 visits_to_doctor                  -0.480     0.0743      -6.46 2.70e-10
## 4 medicine_high_bp                  -0.289     0.0952      -3.04 2.52e- 3
## 5 insurance:visits_to_doctor         0.00239   0.00130      1.84 6.67e- 2
## 6 insurance:medicine_high_bp        -0.00780   0.00129     -6.04 3.19e- 9
## 7 visits_to_doctor:medicine_high_bp  0.00767   0.00137      5.59 3.80e- 8
```

```
glance(int_access_heart_disease_fit)$adj.r.squared %>%
  print()
```

```
## [1] 0.6667498
```

```
int_access_heart_disease_fit_aug %>%
  ggplot(mapping = aes(x = .fitted, y = .std.resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "purple", lty = "dashed") +
  labs(
    title = "Residuals vs. Predicted City Percentage of Adults with Coronary Heart Disease",
    subtitle = "Data From CDC 500 Cities",
    x = "Predicted Percentage of Adults with Coronary Heart Disease",
    y = "Standard Residuals"
  )
```



NOTE: A linear regression is not fitting for this relationship because there is a significant pattern in the residual plot.

5) Access Variables vs. Diabetes

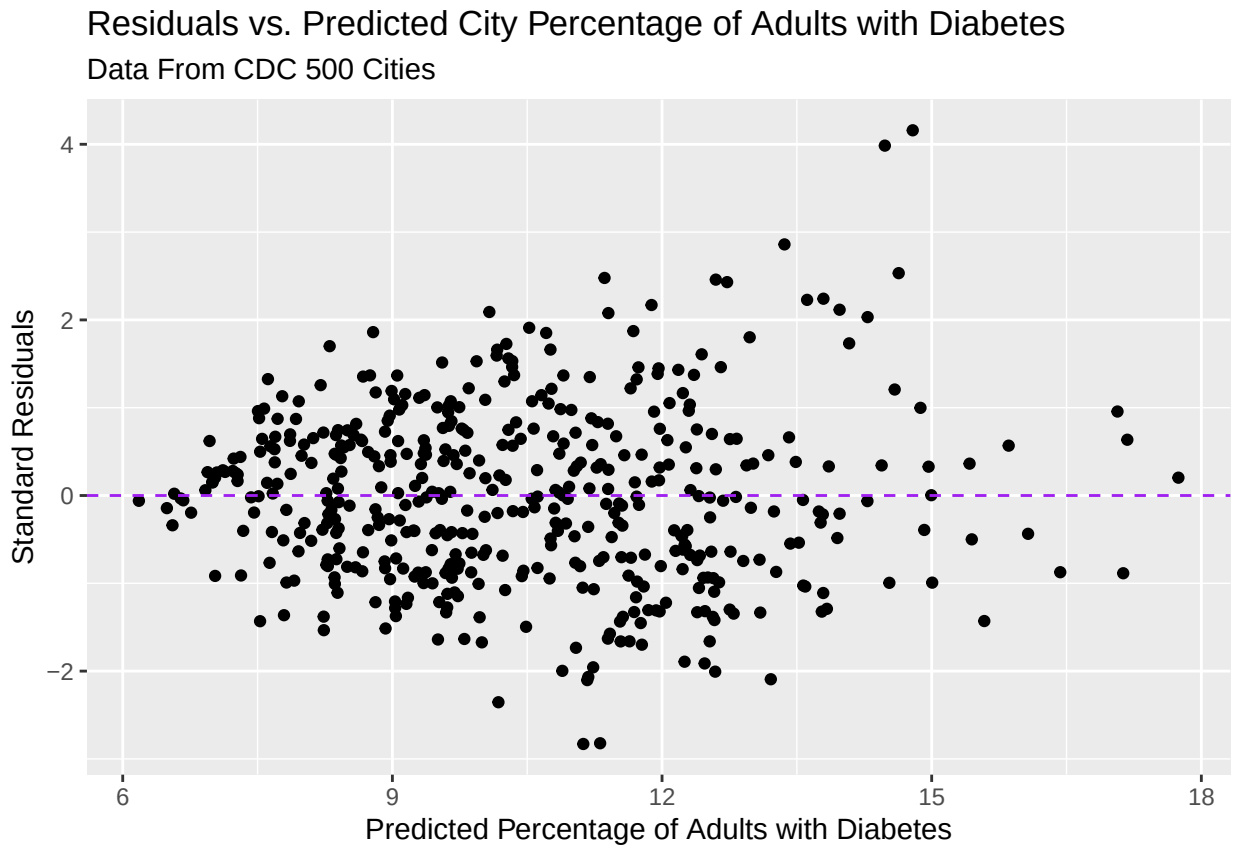
```
int_access_diabetes_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(diabetes ~ insurance + visits_to_doctor + medicine_high_bp + (insurance * visits_to_doctor) + (ins
int_access_diabetes_fit_aug <- augment(int_access_diabetes_fit$fit)
tidy(int_access_diabetes_fit) %>%
  print()
```

```
## # A tibble: 7 x 5
##   term                              estimate std.error statistic  p.value
##   <chr>                                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                       69.9      11.4         6.12  1.97e- 9
## 2 insurance                          0.975     0.198       4.92  1.22e- 6
## 3 visits_to_doctor                  -1.07      0.172      -6.25  9.40e-10
## 4 medicine_high_bp                  -1.40      0.220      -6.36  4.72e-10
## 5 insurance:visits_to_doctor        -0.00935   0.00301    -3.10  2.03e- 3
## 6 insurance:medicine_high_bp        -0.00147   0.00299    -0.493 6.22e- 1
## 7 visits_to_doctor:medicine_high_bp  0.0230    0.00317     7.24  1.87e-12
```

```
glance(int_access_diabetes_fit)$adj.r.squared %>%
  print()
```

```
## [1] 0.7110294
```

```
int_access_diabetes_fit_aug %>%
  ggplot(mapping = aes(x = .fitted, y = .std.resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "purple", lty = "dashed") +
  labs(
    title = "Residuals vs. Predicted City Percentage of Adults with Diabetes",
    subtitle = "Data From CDC 500 Cities",
    x = "Predicted Percentage of Adults with Diabetes",
    y = "Standard Residuals"
  )
```



6) Access Variables vs. Kidney Disease

```
int_access_kidney_disease_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(kidney_disease ~ insurance + visits_to_doctor + medicine_high_bp + (insurance * visits_to_doctor)
int_access_kidney_disease_fit_aug <- augment(int_access_kidney_disease_fit$fit)
tidy(int_access_kidney_disease_fit) %>%
  print()
```

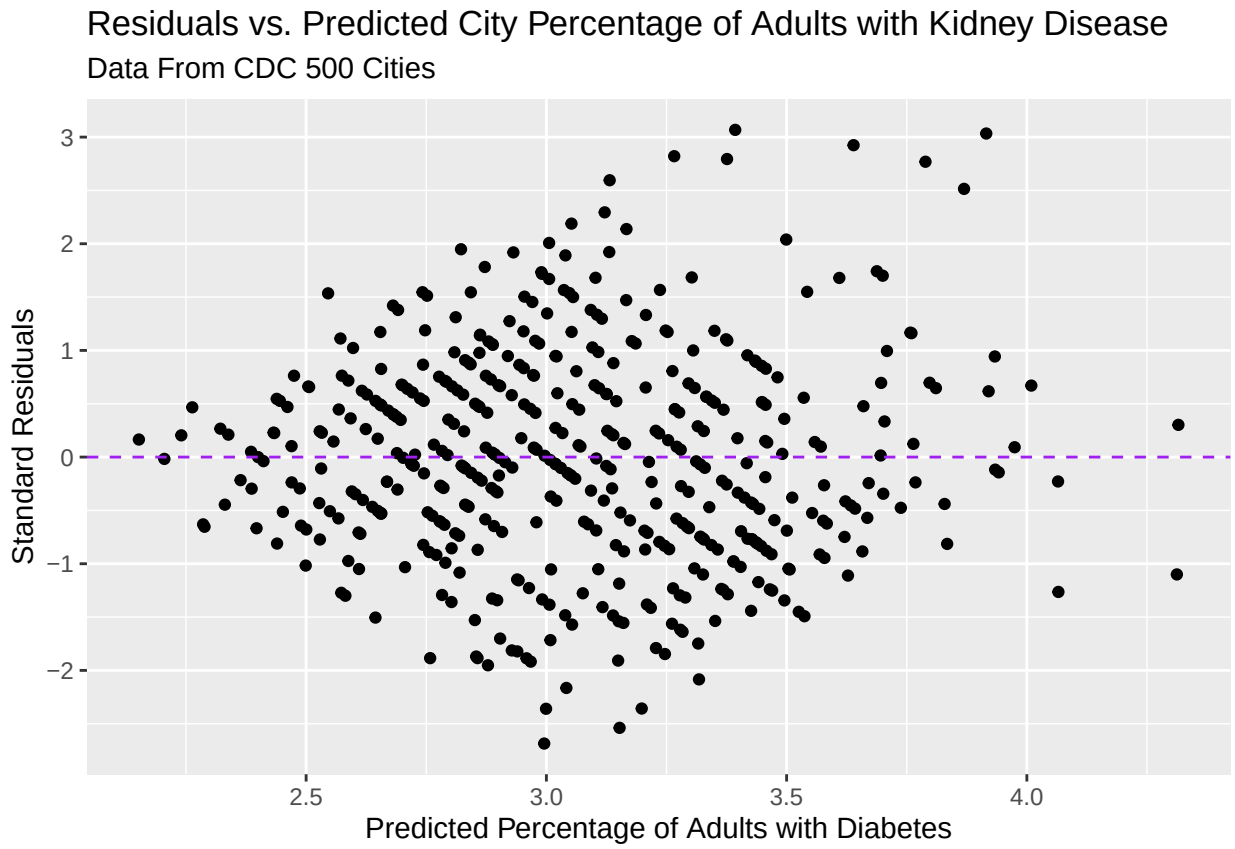
```
## # A tibble: 7 x 5
##   term                          estimate std.error statistic  p.value
##   <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                     22.9       2.50       9.16  1.63e-18
## 2 insurance                        0.198     0.0435     4.56  6.44e- 6
## 3 visits_to_doctor                -0.361     0.0377    -9.57  6.10e-20
## 4 medicine_high_bp                -0.372     0.0483    -7.70  8.53e-14
```

```
## 5 insurance:visits_to_doctor         0.000243  0.000661     0.368 7.13e- 1
## 6 insurance:medicine_high_bp        -0.00297   0.000655    -4.53  7.40e- 6
## 7 visits_to_doctor:medicine_high_bp  0.00646   0.000696     9.28  6.23e-19
```

```
glance(int_access_kidney_disease_fit)$adj.r.squared %>%
  print()
```

```
## [1] 0.6193093
```

```
int_access_kidney_disease_fit_aug %>%
  ggplot(mapping = aes(x = .fitted, y = .std.resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "purple", lty = "dashed") +
  labs(
    title = "Residuals vs. Predicted City Percentage of Adults with Kidney Disease",
    subtitle = "Data From CDC 500 Cities",
    x = "Predicted Percentage of Adults with Diabetes",
    y = "Standard Residuals"
  )
```



NOTE: A linear regression is not fitting for this relationship because there is a significant pattern in the residual plot.

Regression With Most Correlated Variables

ANOVA Testing

Initial Visualizations

Does (Insert Variable) Have Variation Across States?