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 without Interaction Variables

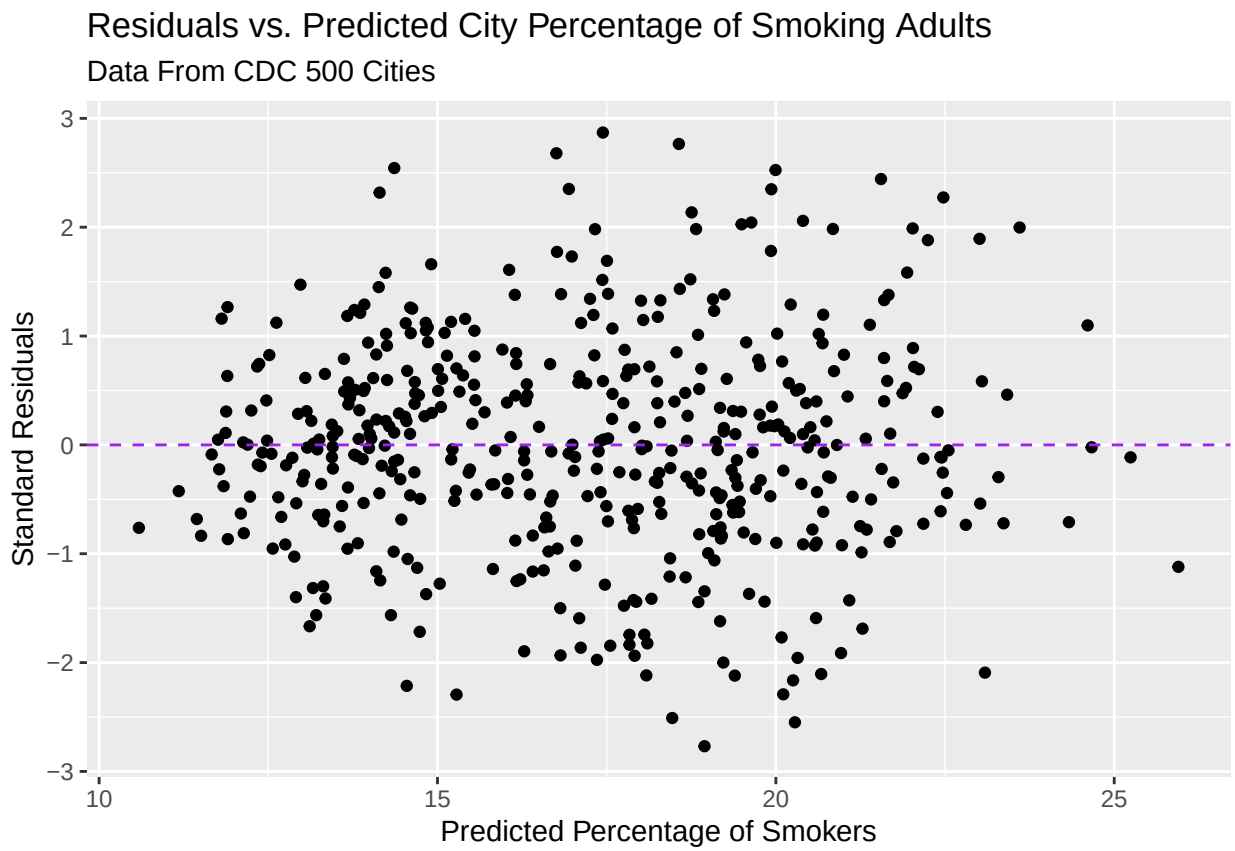
```
access_smoking_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(smoking ~ insurance + visits_to_doctor + medicine_high_bp, data = data_500_cities)
access_smoking_fit_aug <- augment(access_smoking_fit$fit)
tidy(access_smoking_fit) %>%
  print()
```

```
## # A tibble: 4 x 5
##   term             estimate std.error statistic  p.value
##   <chr>               <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)      -15.0       2.08       -7.23 1.99e-12
## 2 insurance          0.0523    0.0237      2.21 2.79e- 2
## 3 visits_to_doctor  -0.0966    0.0446     -2.17 3.08e- 2
## 4 medicine_high_bp   0.674     0.0438     15.4  1.59e-43
```

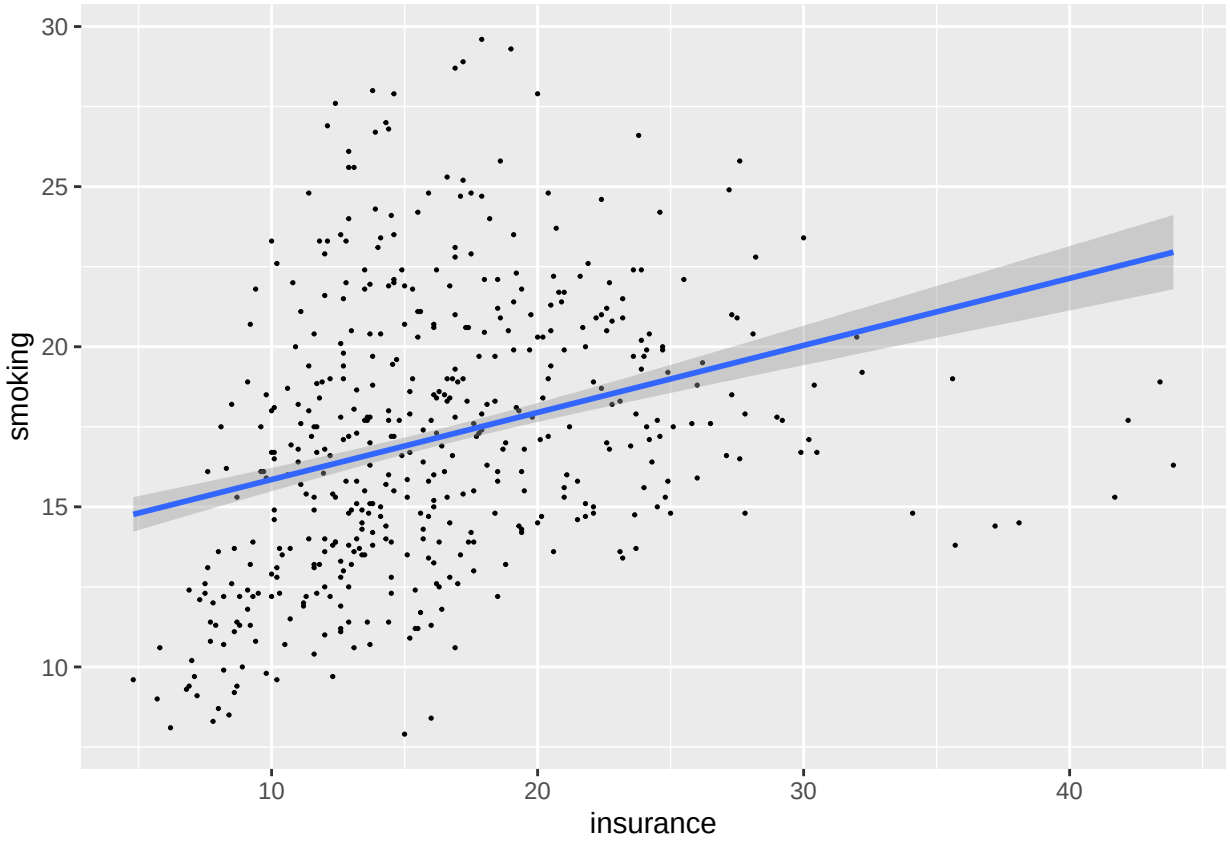
```
glance(access_smoking_fit)$adj.r.squared %>%
  print()
```

```
## [1] 0.5150724
```

```
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"
  )
```




```
data_500_cities %>%
ggplot( mapping = aes(x = insurance, y = smoking)) +
geom_point(size = 0.25) +
geom_smooth(method = "lm", data = access_smoking_fit_aug, mapping = aes(x = insurance, y = .fitted))
```




```
access_binge_drinking_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(binge_drinking ~ insurance + visits_to_doctor + medicine_high_bp, data = data_500_cities)
access_binge_drinking_fit_aug <- augment(access_binge_drinking_fit$fit)
tidy(access_binge_drinking_fit) %>%
  print()
```

2) Access Variables vs. Binge Drinking

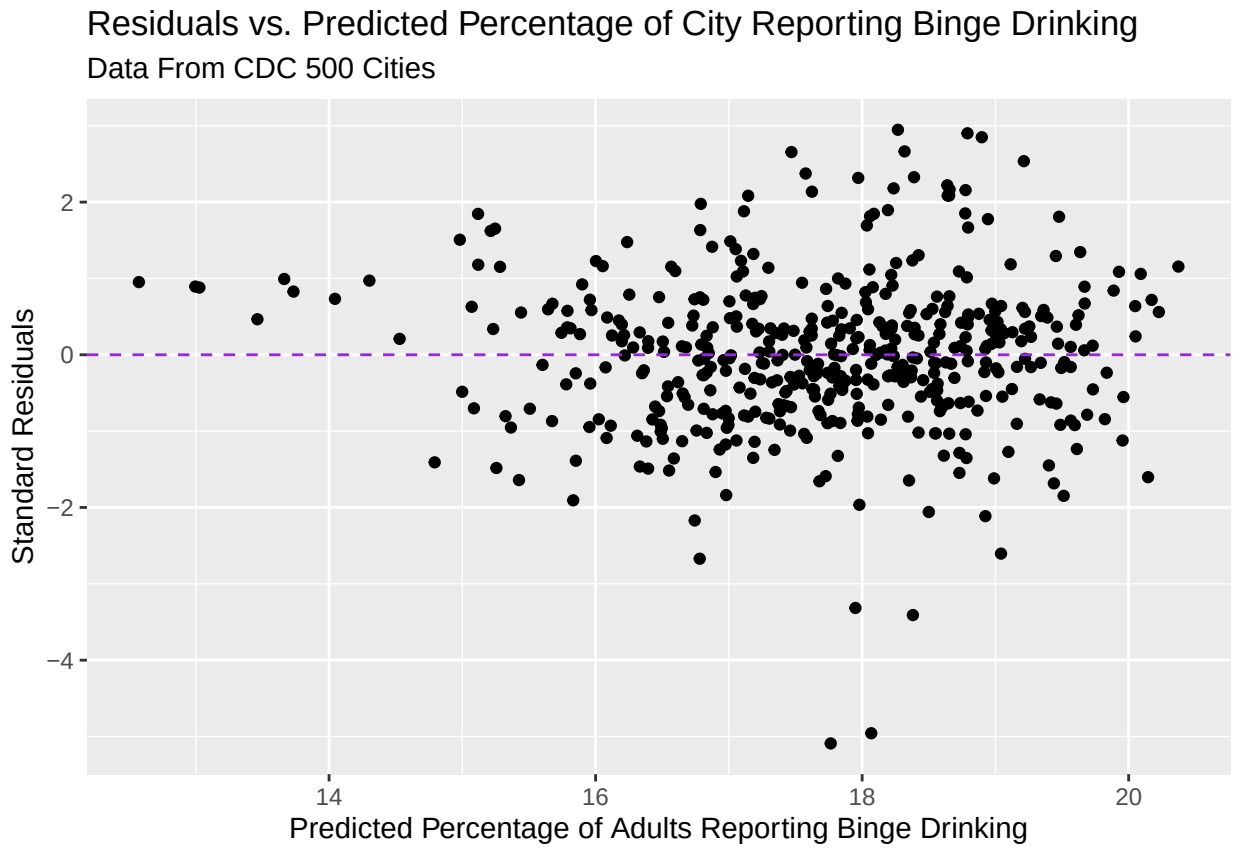
```
## # A tibble: 4 x 5
##   term             estimate std.error statistic  p.value
##   <chr>               <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       24.2       1.58       15.3  2.65e-43
## 2 insurance         -0.162     0.0179     -9.02 4.74e-18
## 3 visits_to_doctor   0.0565    0.0337      1.68 9.45e- 2
## 4 medicine_high_bp  -0.137     0.0331     -4.13 4.39e- 5
```

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

```
## [1] 0.2367489
```

```
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"
  )
```



Fit with Interaction Variables

```
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()
```

1) Access Variables vs. Smoking

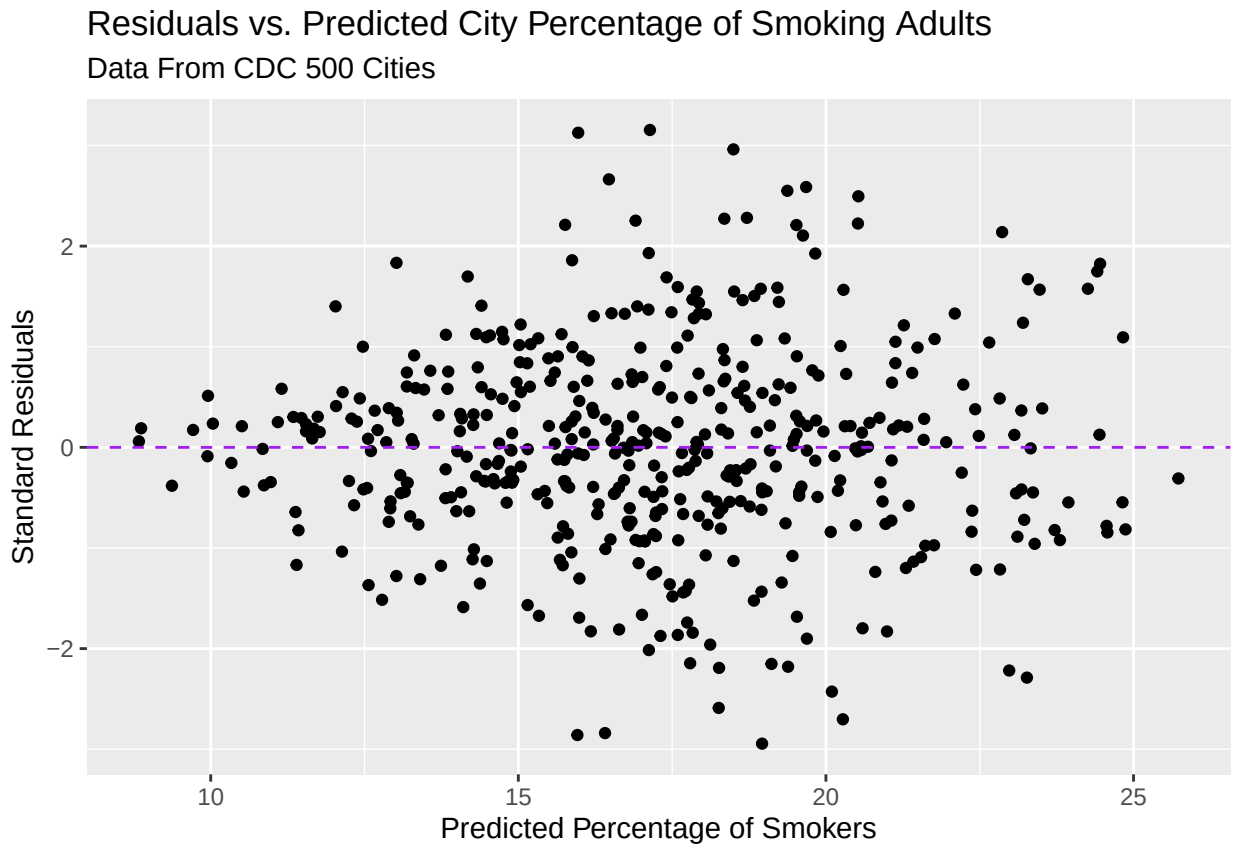
```
## # A tibble: 7 x 5
##   term                          estimate std.error statistic  p.value
##   <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                     88.9     24.0          3.70 2.41e- 4
## 2 insurance                        0.872    0.417        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
```

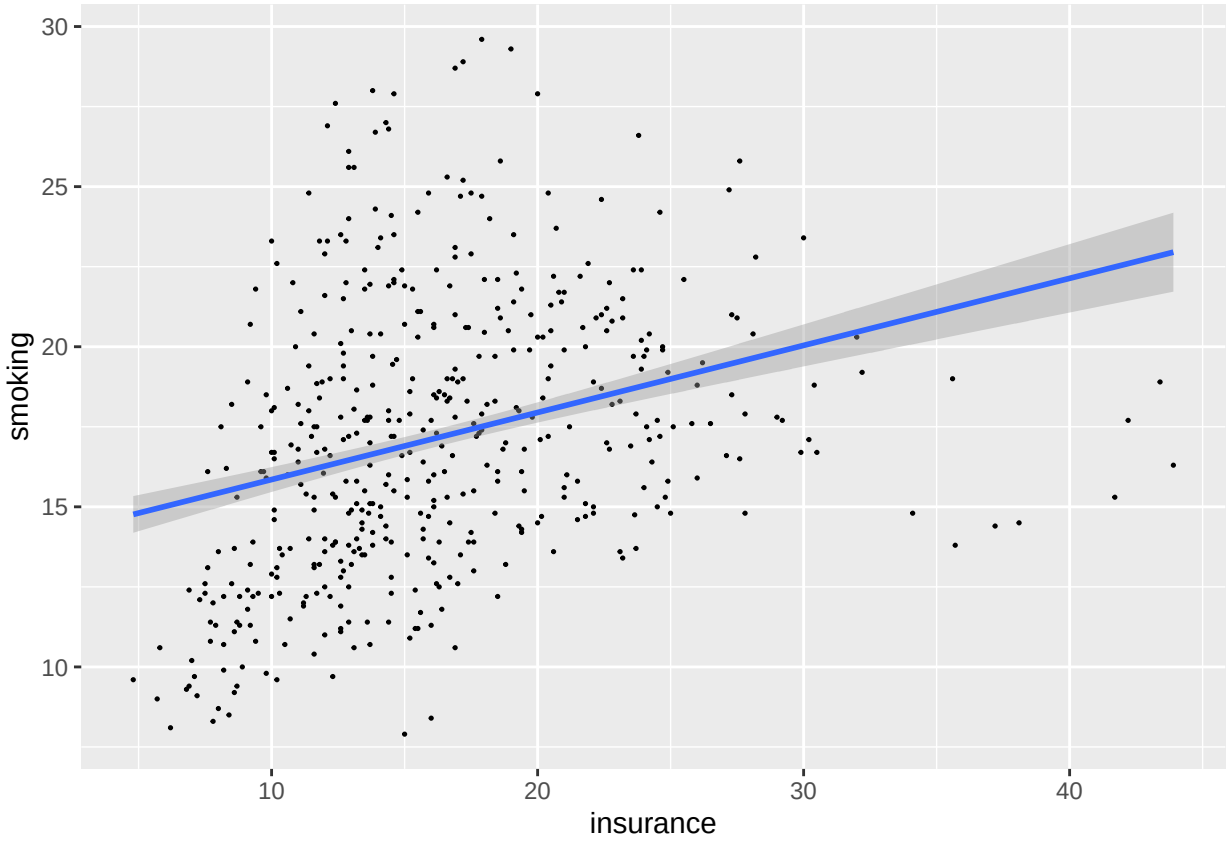
```
glance(int_access_smoking_fit)$adj.r.squared %>%
  print()
```

```
## [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"
  )
```




```
data_500_cities %>%
ggplot( mapping = aes(x = insurance, y = smoking)) +
geom_point(size = 0.25) +
geom_smooth(method = "lm", data = int_access_smoking_fit_aug, mapping = aes(x = insurance, y = .fitted)
```




```
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()
```

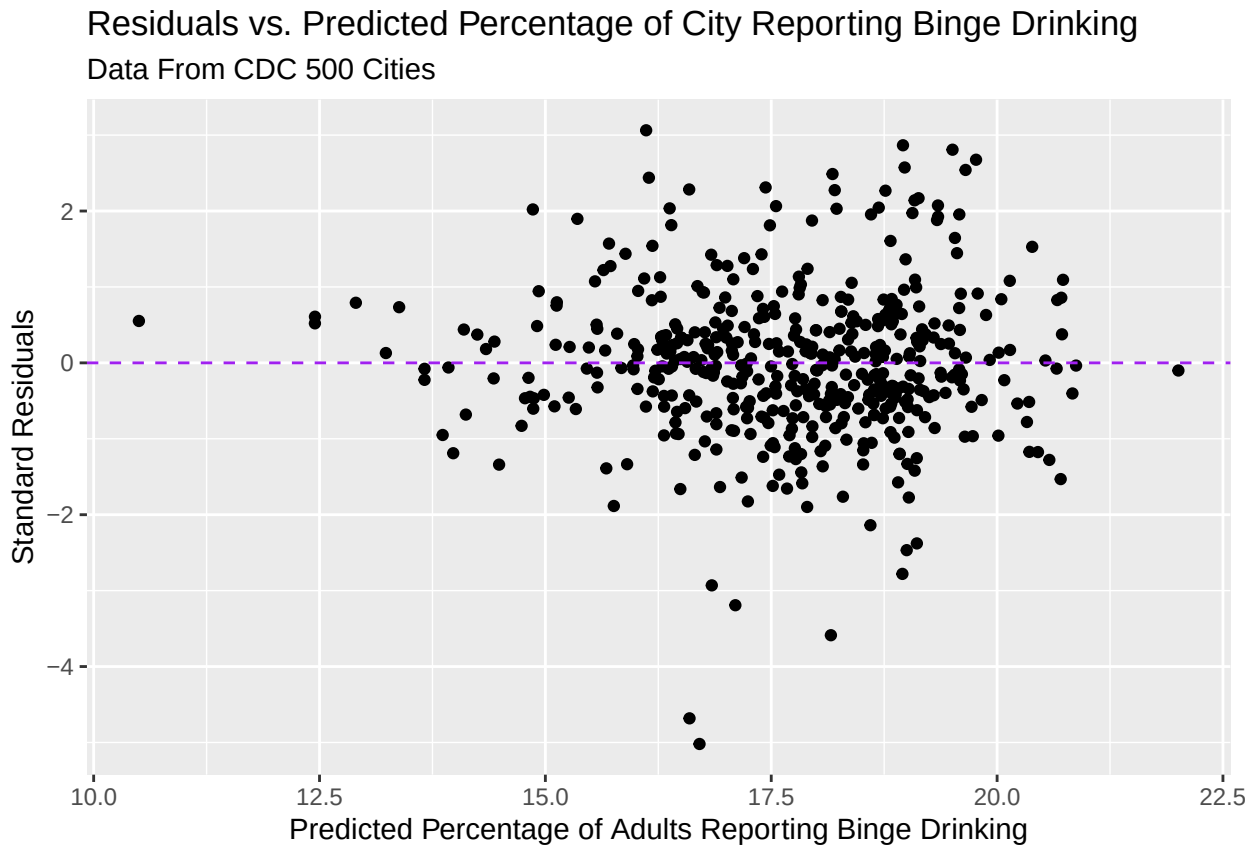
2) Access Variables vs. Binge Drinking

```
## # 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"
  )
```



Why is this graph not working? Ask at some point

```
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()
```

3) Access Variables vs. Physical Activity

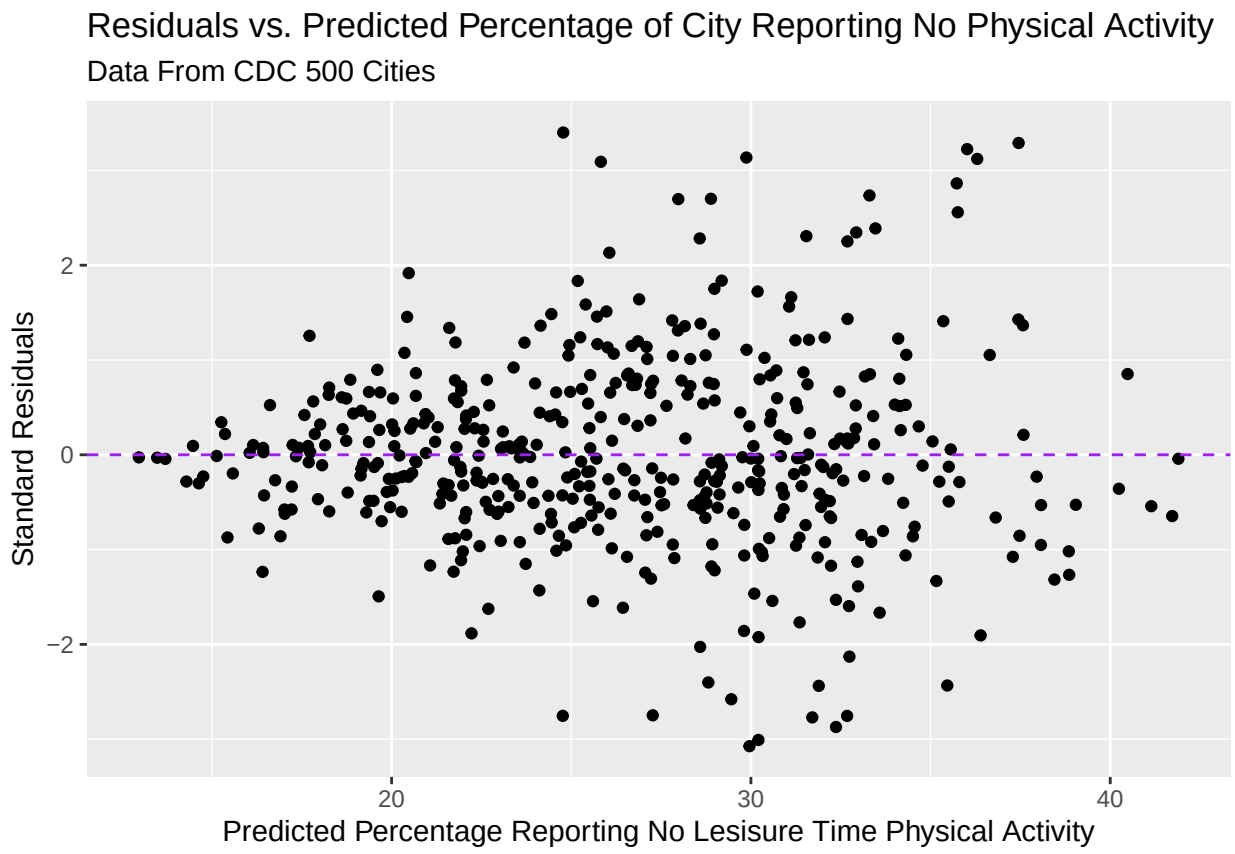
```
## # A tibble: 7 x 5
##   term                              estimate std.error statistic      p.value
##   <chr>                                <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)                         55.1       20.8         2.64  0.00845
```

```
## 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
```

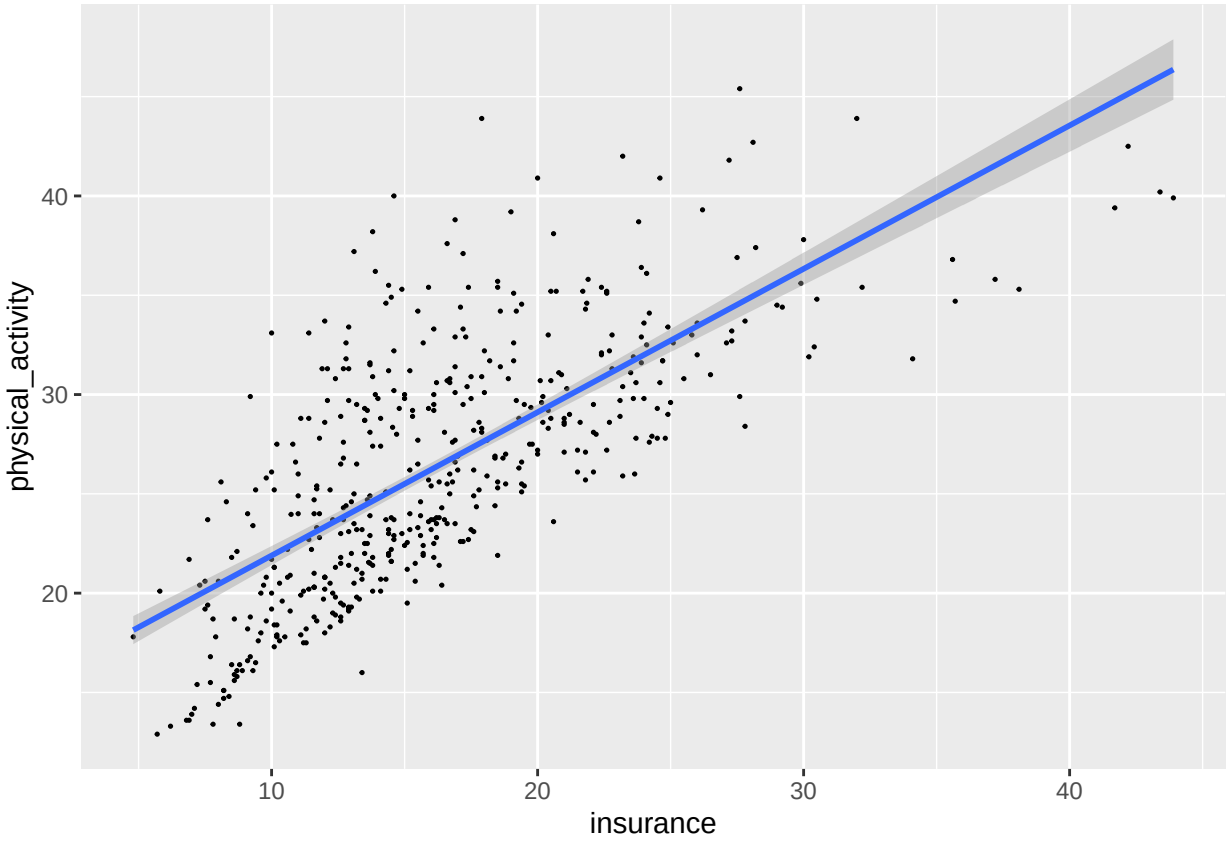
```
glance(int_access_physical_activity_fit)$adj.r.squared %>%
  print()
```

```
## [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"
  )
```




```
data_500_cities %>%
ggplot( mapping = aes(x = insurance, y = physical_activity)) +
geom_point(size = 0.25) +
geom_smooth(method = "lm", data = int_access_physical_activity_fit_aug, mapping = aes(x = insurance, y =
```



Regressions for Healthcare Access and Health Outcomes

Fit with Interaction Variables

```
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()
```

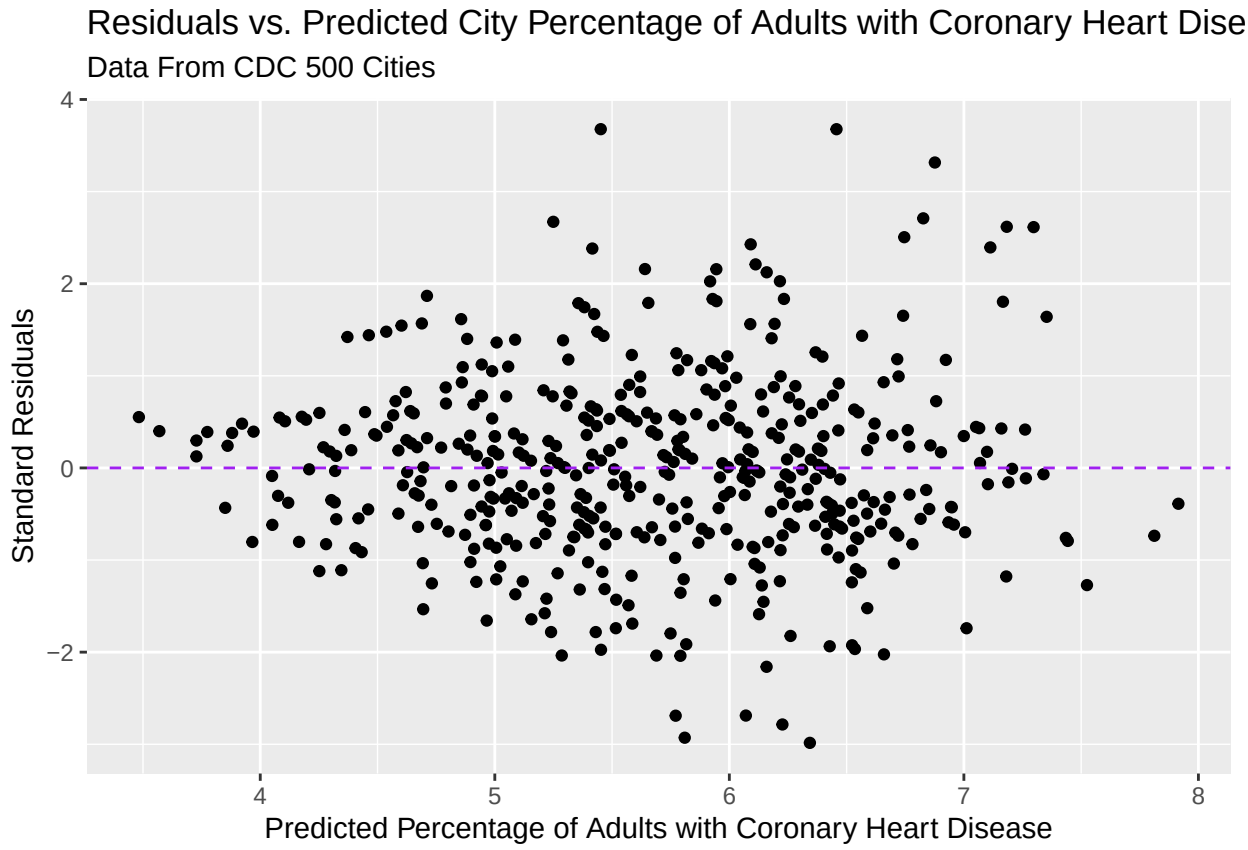
4) Access Variables vs. Heart Disease

```
## # 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.

```
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()
```

5) Access Variables vs. Diabetes

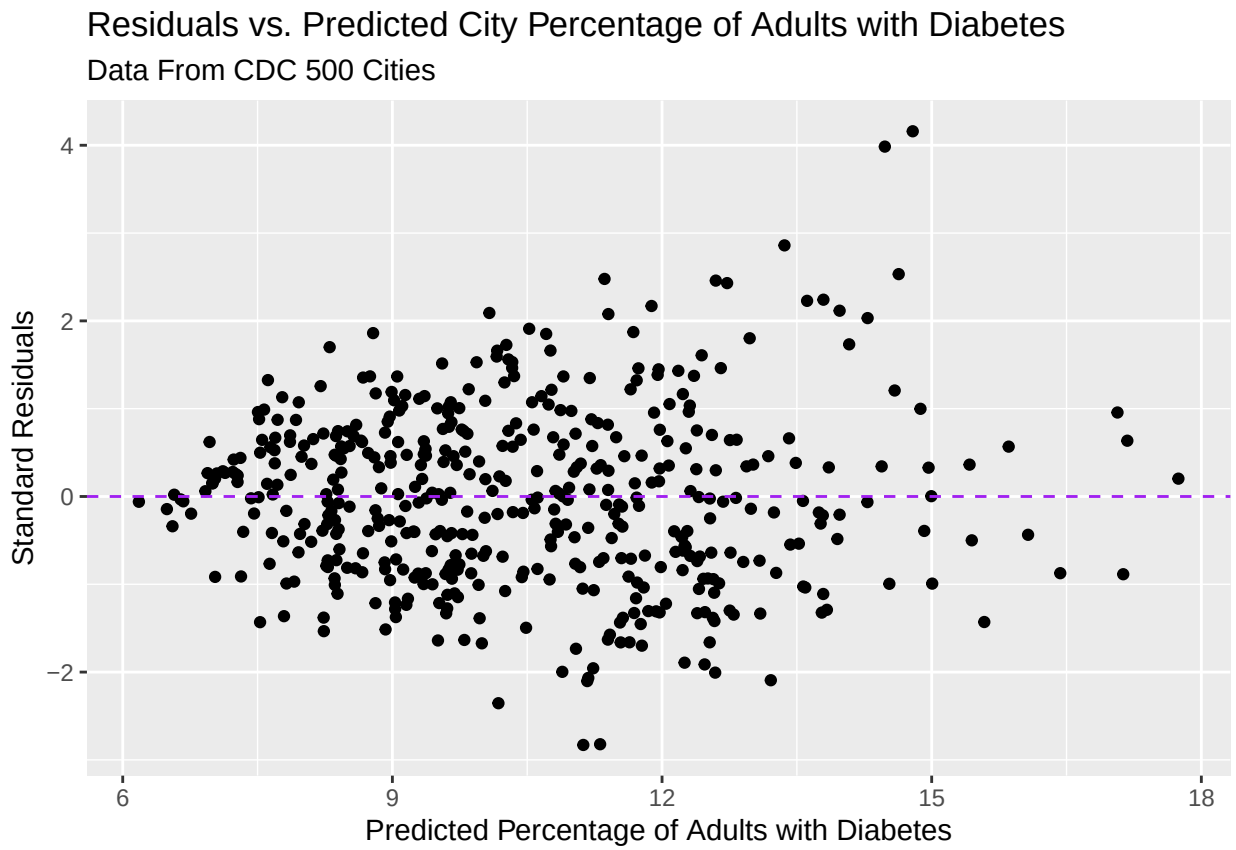
```
## # A tibble: 7 x 5
##   term                              estimate std.error statistic  p.value
```

```
##   <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"
  )
```




```
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()
```

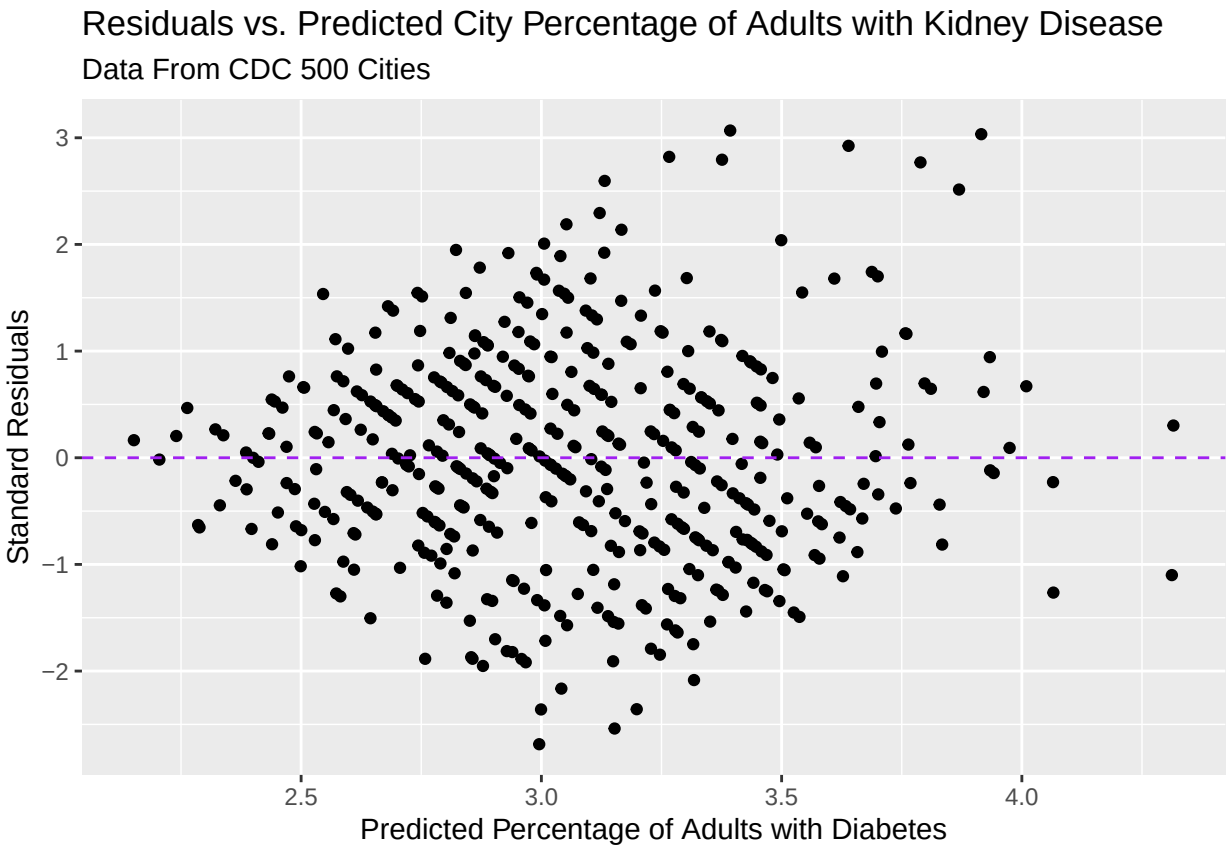
6) Access Variables vs. Kidney Disease

```
## # 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

NOTE: Use initial visualizations to check if assumptions of ANOVA are met!

Overall Tests

```
summary(aov(insurance~StateDesc,data=data_500_cities)) %>%
  print()
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## StateDesc    50   9260  185.20   8.487 <2e-16 ***
## Residuals   424   9252   21.82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(visits_to_doctor~StateDesc,data=data_500_cities)) %>%
  print()
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## StateDesc    50   8395  167.90   44.01 <2e-16 ***
```

```
## Residuals   421   1606    3.81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 3 observations deleted due to missingness
```

```
summary(aov(medicine_high_bp~StateDesc,data=data_500_cities)) %>%
  print()
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## StateDesc    50   9541  190.82   44.25 <2e-16 ***
## Residuals   422   1820    4.31
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

```
summary(aov(smoking~StateDesc,data=data_500_cities)) %>%
  print()
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## StateDesc    50   4752   95.03   9.747 <2e-16 ***
## Residuals   420   4095    9.75
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 4 observations deleted due to missingness
```

```
summary(aov(binge_drinking~StateDesc,data=data_500_cities)) %>%
  print()
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## StateDesc    50   1719   34.37   9.579 <2e-16 ***
## Residuals   421   1511    3.59
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 3 observations deleted due to missingness
```

```
summary(aov(physical_activity~StateDesc,data=data_500_cities)) %>%
  print()
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## StateDesc    50  10657  213.13   10.76 <2e-16 ***
## Residuals   421   8343   19.82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 3 observations deleted due to missingness
```

```
summary(aov(heart_disease~StateDesc,data=data_500_cities)) %>%
  print()
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## StateDesc    50  201.1   4.021   5.974 <2e-16 ***
## Residuals   421  283.4   0.673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 3 observations deleted due to missingness
```

```
summary(aov(diabetes~StateDesc,data=data_500_cities)) %>%
  print()
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## StateDesc    50   1061   21.21   4.632 <2e-16 ***
## Residuals   421   1928    4.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 3 observations deleted due to missingness
```

```
summary(aov(kidney_disease~StateDesc,data=data_500_cities)) %>%
  print()
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## StateDesc    50  21.77  0.4354   2.102 4.48e-05 ***
## Residuals   422  87.41  0.2071
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

It seems like pretty much all the overall tests indicate significant variance across the groups.

Step Down Tests

```
insurance_state_pair <- pairwise.t.test(data_500_cities$insurance,    data_500_cities$StateDesc, p.adj =
sig_ins_state_pairs <- broom::tidy(insurance_state_pair) %>%
  filter(p.value<0.05) %>%
  arrange(group1,group2)
nrow(sig_ins_state_pairs)
```

```
## [1] 0
```

```
print(sig_ins_state_pairs)
```

```
## # A tibble: 0 x 3
## # ... with 3 variables: group1 <chr>, group2 <chr>, p.value <dbl>
```

The overall ANOVA test says there is a significance but the Step Down stests show no significant pairs?

```
doctor_state_pair <- pairwise.t.test(data_500_cities$visits_to_doctor,    data_500_cities$StateDesc, p.a
sig_doctor_state_pairs <- broom::tidy(doctor_state_pair) %>%
  filter(p.value<0.05) %>%
  arrange(group1,group2)
nrow(sig_doctor_state_pairs)
```

```
## [1] 0
```

```
print(sig_doctor_state_pairs)
```

```
## # A tibble: 0 x 3
## # ... with 3 variables: group1 <chr>, group2 <chr>, p.value <dbl>
```

Ok so there is clearly an issue. I think I am interpreting the F-statistic incorrectly?

```
smoking_state_pair <- pairwise.t.test(data_500_cities$smoking,    data_500_cities$StateDesc, p.adj = "ho
sig_smoking_state_pairs <- broom::tidy(smoking_state_pair) %>%
  filter(p.value<0.05) %>%
  arrange(group1,group2)
nrow(sig_smoking_state_pairs)
```

```
## [1] 0
```

```
print(sig_smoking_state_pairs)
```

```
## # A tibble: 0 x 3
## # ... with 3 variables: group1 <chr>, group2 <chr>, p.value <dbl>
```