Unemployment Analysis in the United States (2000-2019)

Laura Yuan

```
# Packages used in analysis:
library(readr)
library(dplyr)
library(ggplot2)
library(scales)
library(car)
library(tidyr)
library(GGally)
library(mctest)
library(sandwich)
library(stargazer)
library(corpcor)
library(ppcor)
library(AER)
```

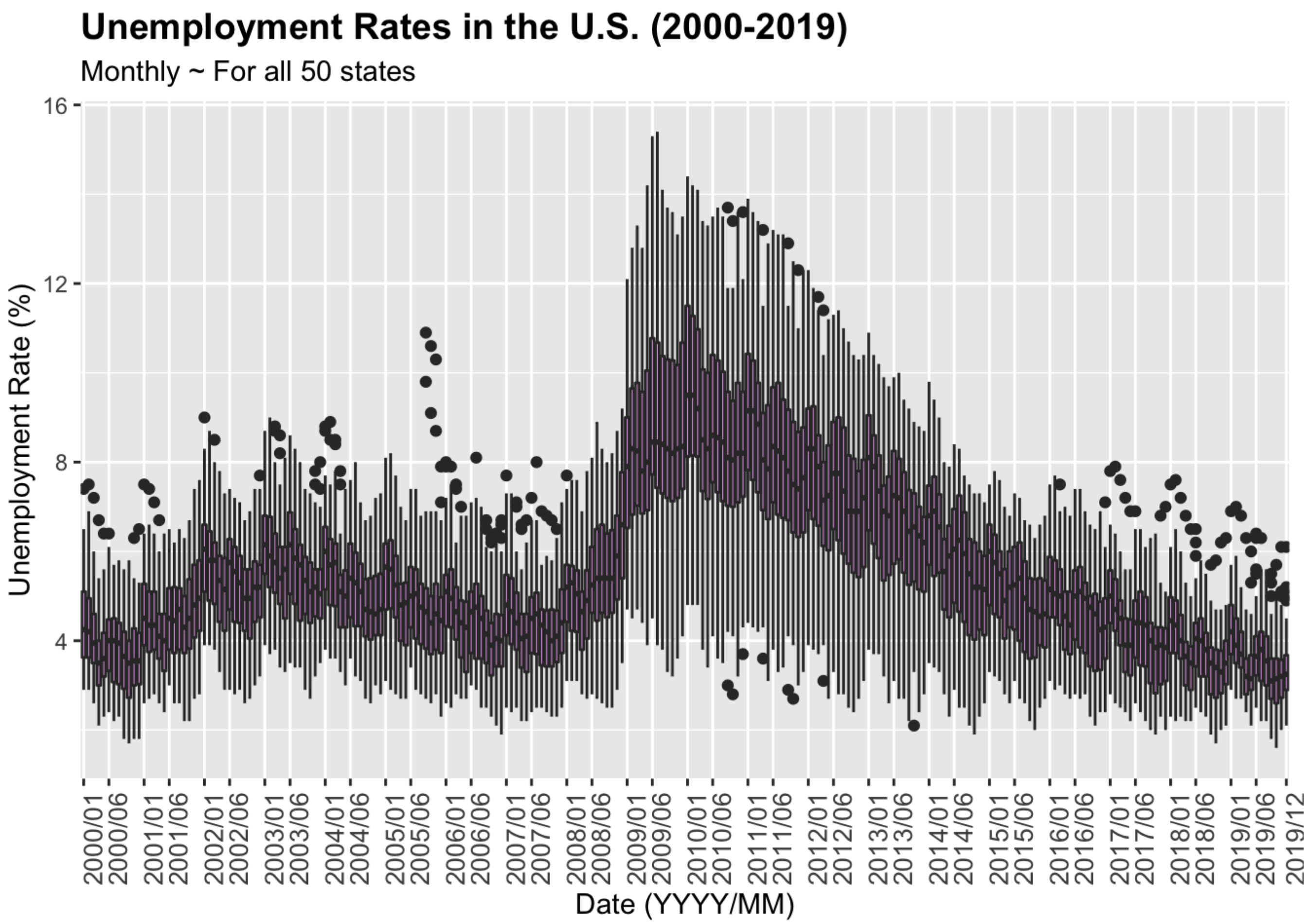
Data Visualization

Box plot of Unemployment Rates

We can first create a simple box plot of all the unemployment rates in the United States to see the general trends and rates over time.

```
library(ggplot2)

# Box Plot
ggplot(unemploymentrates, aes(Year_Month, Rate))+
    geom_boxplot(varwidth=T, fill="plum") +
    labs(title="Unemployment Rates in the U.S. (2000-2019)",
    subtitle = "Monthly ~ For all 50 states",
    x="Date (YYYY/MM)",
    y="Unemployment Rate (%)") +
    theme(axis.text.x = element_text(angle=90, size=10)) +
    theme(plot.title = element_text(size=14, face="bold")) +
    scale_x_discrete(breaks=c("2000/01","2000/06","2001/01","2001/06","2002/01","2002/0
6","2003/01","2003/06","2004/01","2004/06","2005/01",
                              "2005/06","2006/01","2006/06","2007/01","2007/06","2008/0
1","2008/06","2009/01","2009/06","2010/01","2010/06",
                              "2011/01","2011/06","2012/01","2012/06","2013/01","2013/0
6","2014/01","2014/06","2015/01","2015/06","2016/01",
                              "2016/06","2017/01","2017/06","2018/01","2018/06","2019/0
1","2019/06","2019/12"))
```



Scatterplot of Unemployment Rates

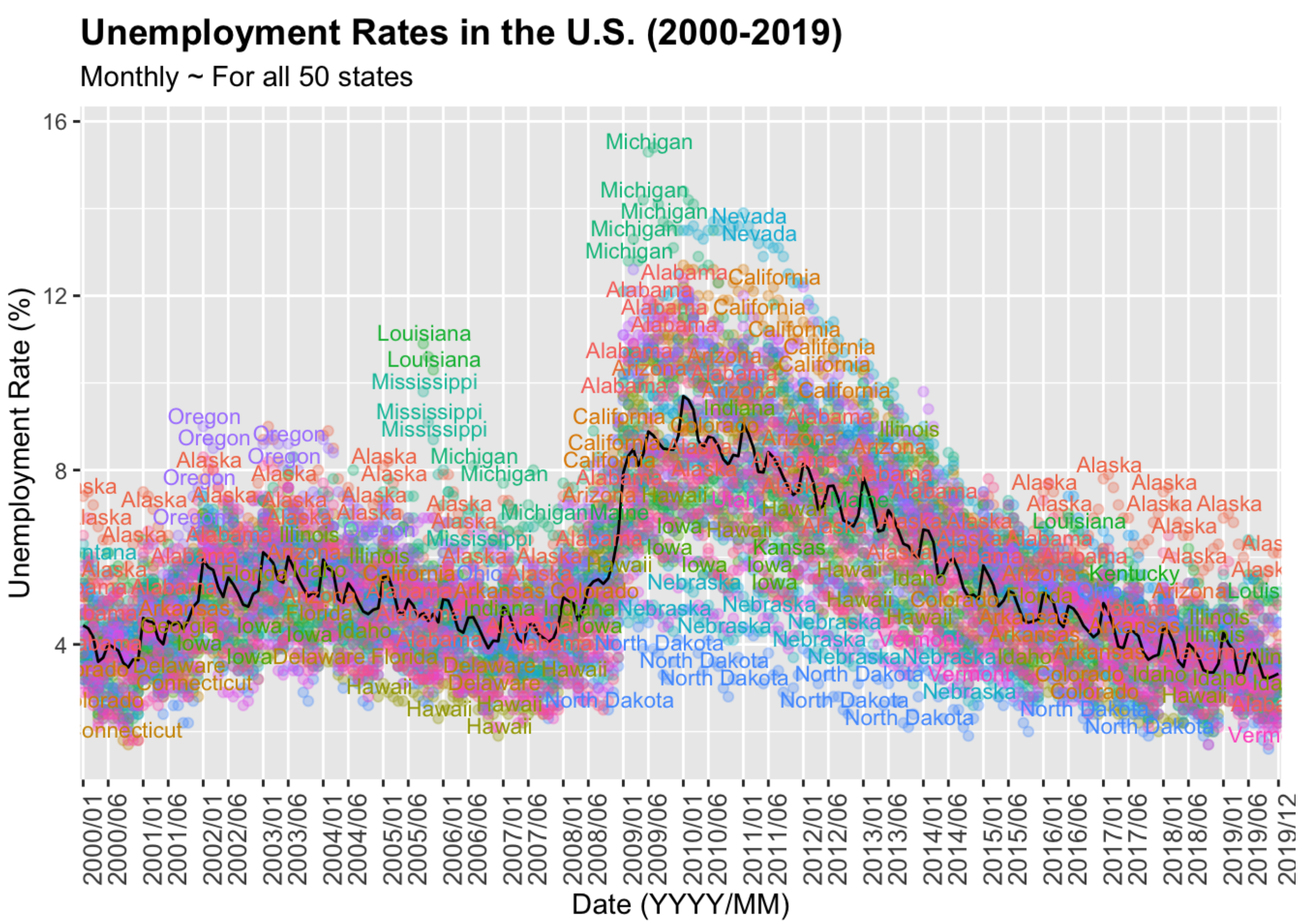
If we would like to assess the unemployment rates by state, we can create a unique scatter plot with state names as labels. This will enable us to evaluate which states seem to have higher/lower unemployment in comparison to other states (outliers). The box plot above and the scatterplot below should show similar trends in unemployment rates over time.

```
library(ggplot2)
library(scales)
library(dplyr)

# Creating a mean column for the trend line
new_unemploymentrates <- unemploymentrates %>% group_by(Year_Month) %>% summarize(Mean_R
ate=mean(Rate))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
# Scatter plot  with individual state labels
ggplot(data = unemploymentrates, aes(x = Year_Month, y = Rate, color = State)) +
    geom_point(alpha=0.3) +
    geom_line(data = new_unemploymentrates, aes(x=Year_Month, y = Mean_Rate, group=1), c
olor="Black") +
    theme(axis.text.x = element_text(angle=90, size=10)) +
    theme(plot.title = element_text(size=14, face="bold")) +
    theme(legend.position = "none") +
    geom_text(aes(label=State), size=3, nudge_x = 0.25, nudge_y = 0.25, check_overlap =
T) +
    labs(title="Unemployment Rates in the U.S. (2000-2019)",
         subtitle = "Monthly ~ For all 50 states",
         x="Date (YYYY/MM)",
         y="Unemployment Rate (%)") +
    guides(col = guide_legend(nrow = 25)) +
    scale_x_discrete(breaks=c("2000/01","2000/06","2001/01","2001/06","2002/01","2002/0
6",
    "2003/01","2003/06","2004/01","2004/06","2005/01","2005/06","2006/01","2006/06","200
7/01","2007/06","2008/01","2008/06",
    "2009/01","2009/06","2010/01","2010/06","2011/01","2011/06","2012/01","2012/06","201
3/01","2013/06","2014/01","2014/06",
    "2015/01","2015/06","2016/01","2016/06","2017/01","2017/06","2018/01","2018/06","201
9/01","2019/06","2019/12"))
```



From the scatterplot above, we can observe that some states such as Alaska, California, Louisiana, Michigan, etc. tend to have higher unemployment rates than other states. We can also observe that states such as North Dakota, Hawaii, Nebraska, etc. tend to have lower unemployment rates. Although this scatterplot helps us point out the states that seem to have lower/higher unemployment rates, there are multiple factors that we have not accounted for such as population, income, education, etc. This will be explored further in the regression models below.

Geographical Map of Unemployment Rates

To take a step further in visualizing this type of panel data, we can create a geographical map showing the changes in unemployment rates by state over time.

```
# Source: https://socviz.co/maps.html
library(mapproj)
```

```
## Loading required package: maps
```

```
library(viridis)
```

```
## Loading required package: viridisLite
```

```
##
## Attaching package: 'viridis'
```

```
## The following object is masked from 'package:scales':
##
##     viridis_pal
```

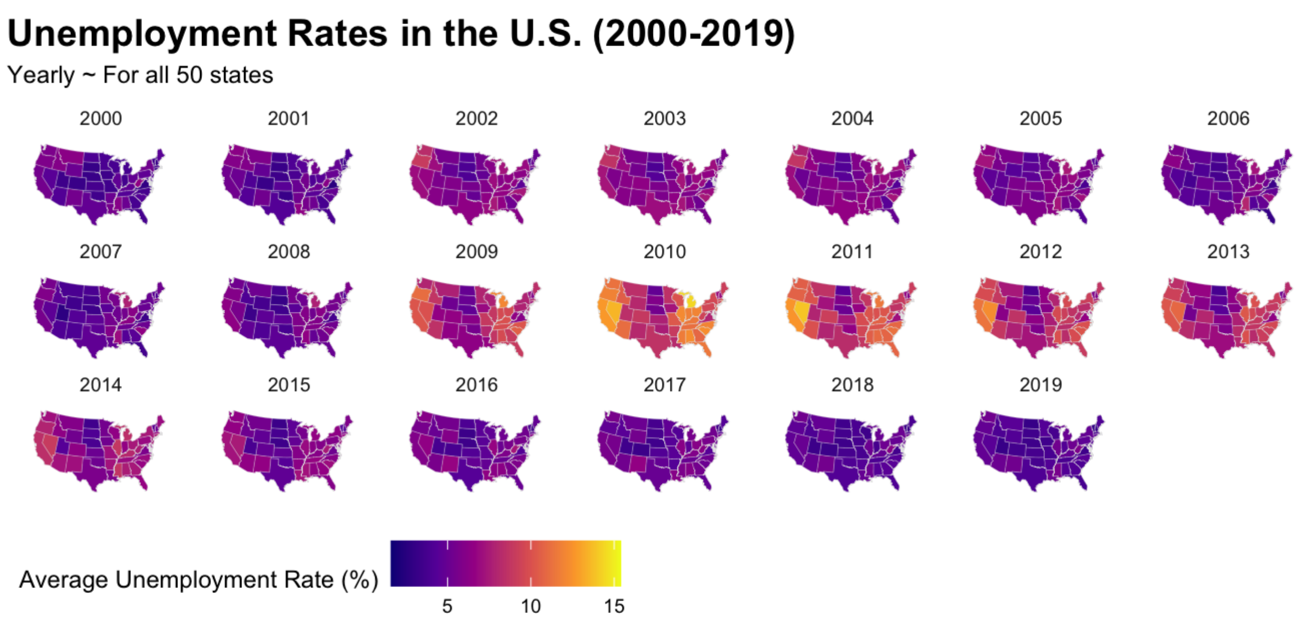
```
library(dplyr)
library(ggplot2)
library(maps)
library(ggthemes)

# Creating a blank U.S. map
us_states <- map_data("state")

# Merging data with map
unemploymentrates$region <- tolower(unemploymentrates$State)
us_states_unemp <- left_join(us_states,unemploymentrates)
```

```
## Joining, by = "region"
```

```
# Geographical map
ggplot(data = subset(us_states_unemp, Year >= 2000), mapping = aes(x = long, y = lat, gr
oup = group, fill = Rate)) +
    geom_polygon(color = "gray90", size = 0.05) +
    coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
    scale_fill_viridis_c(option = "plasma") +
    theme_map() +
    facet_wrap(~ Year, ncol = 7) +
    theme(legend.position = "bottom", strip.background = element_blank()) +
    labs(fill = "Average Unemployment Rate (%)",
         title = "Unemployment Rates in the U.S. (2000-2019)",
         subtitle = "Yearly ~ For all 50 states") +
    theme(plot.title = element_text(size=14, face="bold"))
```

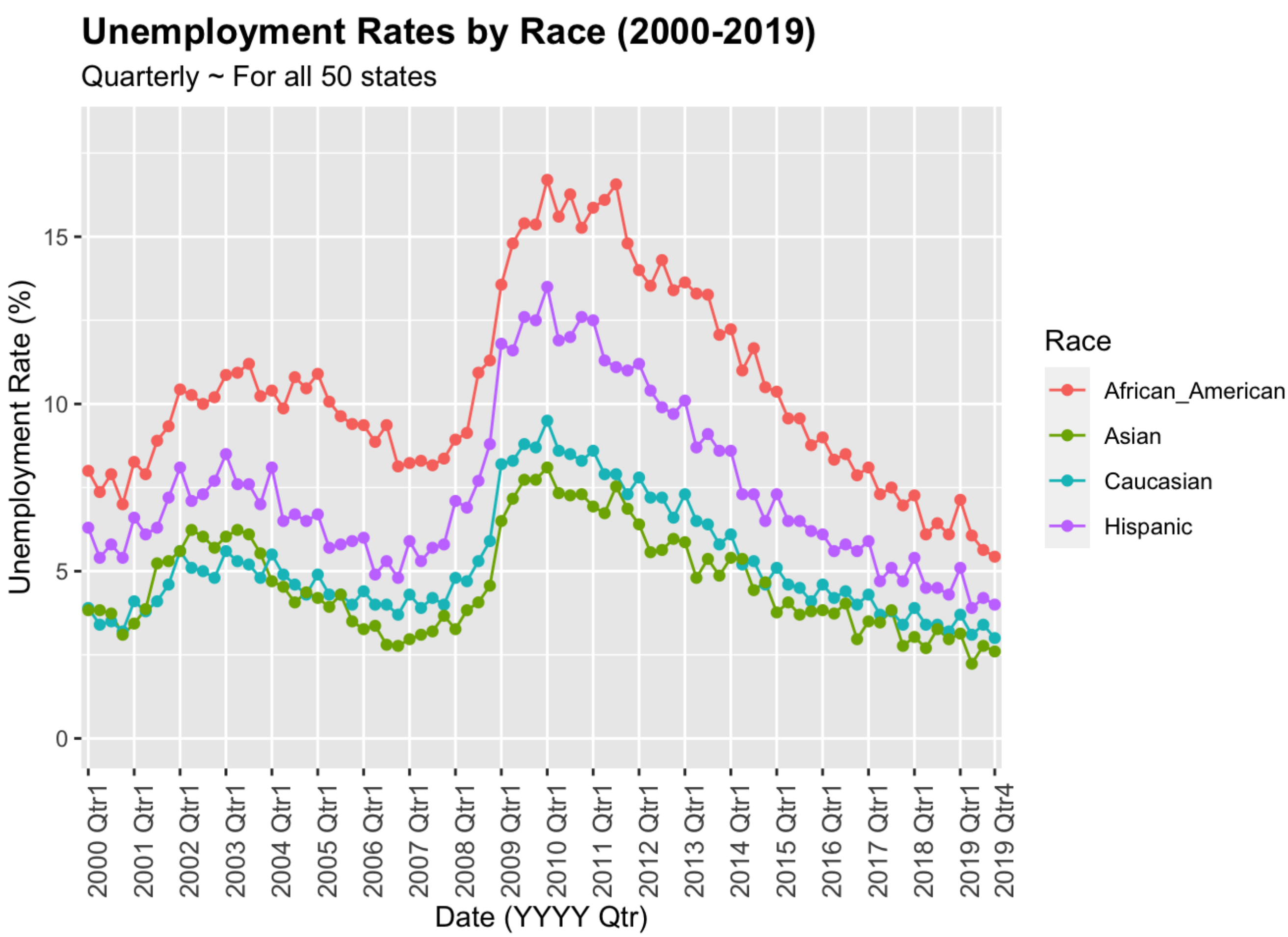


From the geographical map above, we can observe that from 2009 to around 2013, states in the West and East coasts tend to experience higher unemployment rates than other states. There are many factors that may be correlated with these trends, such as the industries that dominate these states, education, financial market health, etc. These factors are not observable from the graphs itself but can be explored further by performing a regression analysis.

Line plot of Unemployment Rates by Race

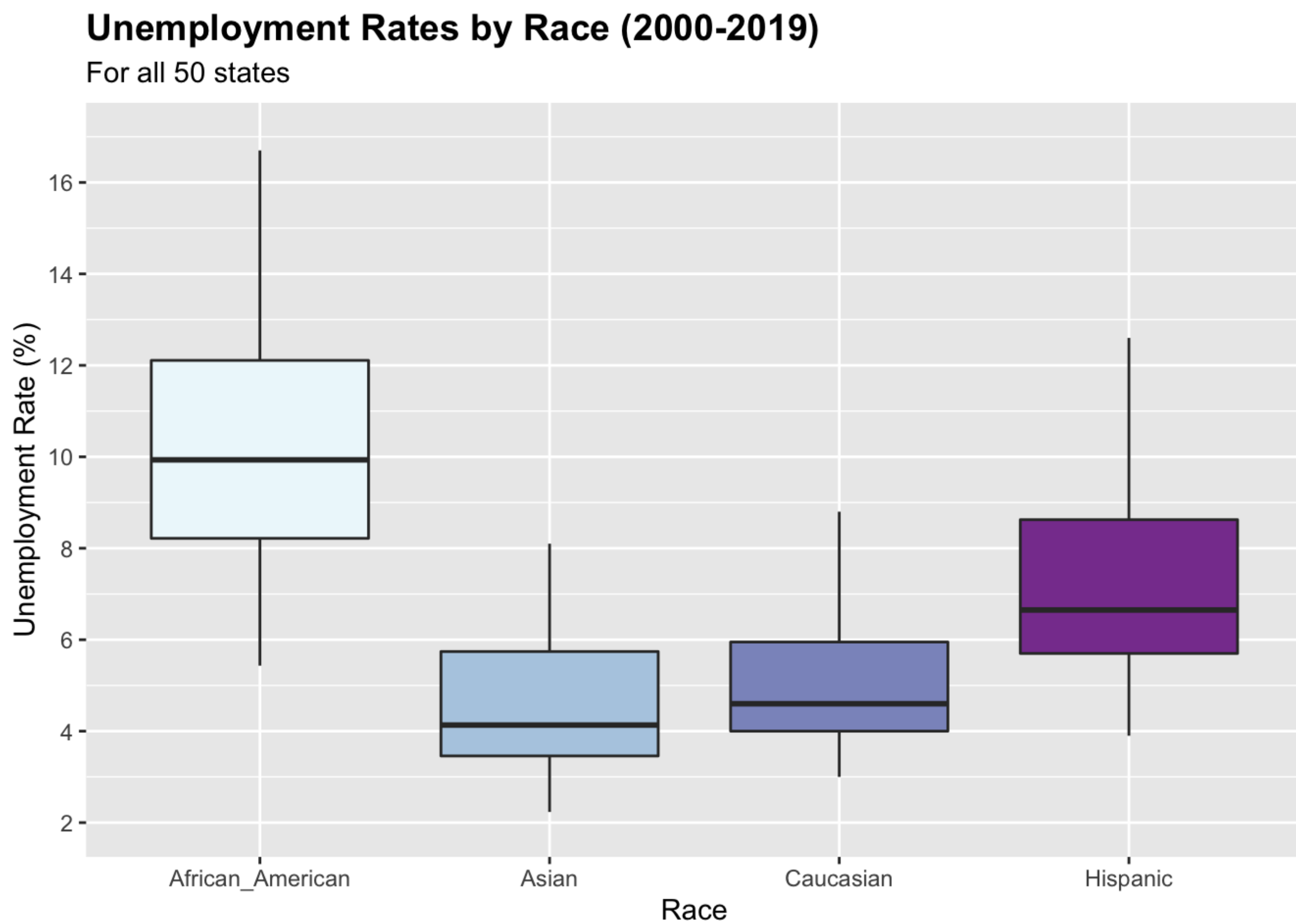
```
library(ggplot2)

ggplot(data = subset(race, Year >= 2000), aes(x = Year_Qtr, y = Rate, group = Race)) +
  geom_line(aes(color=Race)) +
  geom_point(aes(color=Race)) +
  labs(title="Unemployment Rates by Race (2000-2019)") +
  labs(subtitle="Quarterly ~ For all 50 states") +
  labs(x="Date (YYYY Qtr)", y="Unemployment Rate (%)") +
  theme(axis.text.x = element_text(size=10,angle=90)) +
  theme(plot.title = element_text(size=14, face="bold")) +
  scale_y_continuous(limits=c(0,18)) +
  scale_x_discrete(breaks=c("2000 Qtr1", "2001 Qtr1","2002 Qtr1", "2003 Qtr1", "2004 Qtr
1", "2005 Qtr1", "2006 Qtr1", "2007 Qtr1",
                            "2008 Qtr1", "2009 Qtr1", "2010 Qtr1", "2011 Qtr1", "2012 Qt
r1", "2013 Qtr1", "2014 Qtr1", "2015 Qtr1",
                            "2016 Qtr1", "2017 Qtr1", "2018 Qtr1", "2019 Qtr1", "2019 Qt
r4"))
```



```r
library(ggplot2)

ggplot(data = subset(race, Year >= 2000), aes(x=Race, y=Rate,fill=Race)) +
  geom_boxplot(outlier.shape = NA) +
  labs(title="Unemployment Rates by Race (2000-2019)") +
  labs(subtitle="For all 50 states") +
  theme(plot.title = element_text(size=14, face="bold")) +
  labs(x="Race", y="Unemployment Rate (%)") +
  scale_y_continuous(breaks=c(0,2,4,6,8,10,12,14,16),limits=c(2,17))+
  theme(legend.position="none") +
  scale_fill_brewer(palette="BuPu")
```



#

Understanding the Dataset

```r
# Data head
head(state_model)
```

Year_Month <chr>	State <chr>	Year <dbl>	Month <dbl>	Rate <dbl>	HVI <dbl>	West <dbl>	Northeast <dbl>	South <dbl>	Midwest <dbl>
2000/01	Alabama	2000	1	5.1	100017	0	0	1	0
2000/02	Alabama	2000	2	5.1	100282	0	0	1	0
2000/03	Alabama	2000	3	4.5	100557	0	0	1	0

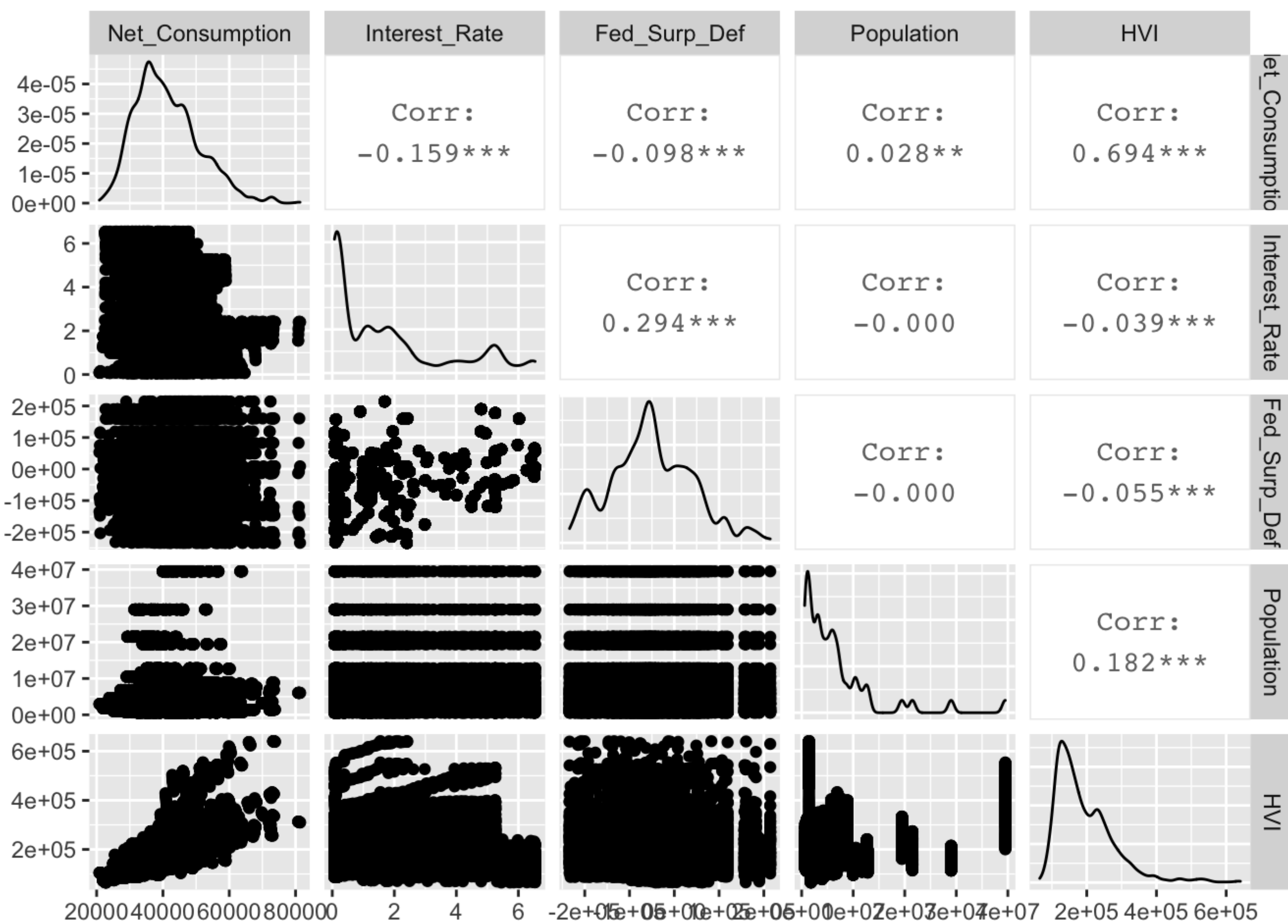
Year_Month	State	Year	Month	Rate	HVI	West	Northeast	South	Midwest
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2000/04	Alabama	2000	4	3.8	100953	0	0	1	0
2000/05	Alabama	2000	5	4.1	101306	0	0	1	0
2000/06	Alabama	2000	6	4.9	101703	0	0	1	0

6 rows | 1-10 of 20 columns

Multicollinearity Analysis

```
library(GGally)
library(ppcor)

test1 <- subset(state_model, select = c(Net_Consumption, Interest_Rate, Fed_Surp_Def, Po
pulation, HVI))
ggpairs(test1)
```




```
pcor(test1, method = "pearson")
```

```
## $estimate
##                 Net_Consumption Interest_Rate Fed_Surp_Def   Population
## Net_Consumption      1.00000000   -0.16790000 -0.030906349 -0.140913444
## Interest_Rate       -0.16790000    1.00000000  0.282589332 -0.019238139
## Fed_Surp_Def        -0.03090635    0.28258933  1.000000000  0.003924064
## Population          -0.14091344   -0.01923814  0.003924064  1.000000000
## HVI                  0.70394218    0.10082803 -0.010759257  0.227643093
##                         HVI
## Net_Consumption  0.70394218
## Interest_Rate    0.10082803
## Fed_Surp_Def    -0.01075926
## Population       0.22764309
## HVI              1.00000000
##
## $p.value
##                 Net_Consumption Interest_Rate  Fed_Surp_Def    Population
## Net_Consumption    0.000000e+00  1.426507e-76  7.100894e-04  2.959431e-54
## Interest_Rate      1.426507e-76  0.000000e+00 4.304697e-219  3.510506e-02
## Fed_Surp_Def       7.100894e-04 4.304697e-219  0.000000e+00  6.673682e-01
## Population         2.959431e-54  3.510506e-02  6.673682e-01  0.000000e+00
## HVI                0.000000e+00  1.737509e-28  2.386445e-01 8.002902e-141
##                           HVI
## Net_Consumption  0.000000e+00
## Interest_Rate    1.737509e-28
## Fed_Surp_Def     2.386445e-01
## Population      8.002902e-141
## HVI              0.000000e+00
##
## $statistic
##                 Net_Consumption Interest_Rate Fed_Surp_Def  Population
## Net_Consumption        0.000000    -18.653495   -3.3865333 -15.5886227
## Interest_Rate        -18.653495      0.000000   32.2647394  -2.1073834
## Fed_Surp_Def          -3.386533     32.264739    0.0000000   0.4297734
## Population           -15.588623     -2.107383    0.4297734   0.0000000
## HVI                  108.547899     11.099420   -1.1784402  25.6041006
##                       HVI
## Net_Consumption 108.54790
## Interest_Rate    11.09942
## Fed_Surp_Def     -1.17844
## Population       25.60410
## HVI               0.00000
##
## $n
## [1] 12000
##
## $gp
## [1] 3
##
## $method
## [1] "pearson"
```

Tests for Heteroskedasticity in Error Terms

```r
# Median Income
hetero <- state_model %>% group_by(State,Year) %>% summarize(Rate = mean(Rate), Median_I
ncome = mean(Median_Income))
```

```
## `summarise()` regrouping output by 'State' (override with `.groups` argument)
```

```r
reg <- lm(Rate ~ Median_Income, data= hetero)
summary(reg)
```

```
##
## Call:
## lm(formula = Rate ~ Median_Income, data = hetero)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6636 -1.3610 -0.4742  0.9585  8.0212
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.653e+00  2.951e-01  25.930  < 2e-16 ***
## Median_Income -4.198e-05  5.599e-06  -7.498 1.43e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.934 on 998 degrees of freedom
## Multiple R-squared:  0.05333,    Adjusted R-squared:  0.05238
## F-statistic: 56.22 on 1 and 998 DF,  p-value: 1.43e-13
```

```r
confint(reg)
```

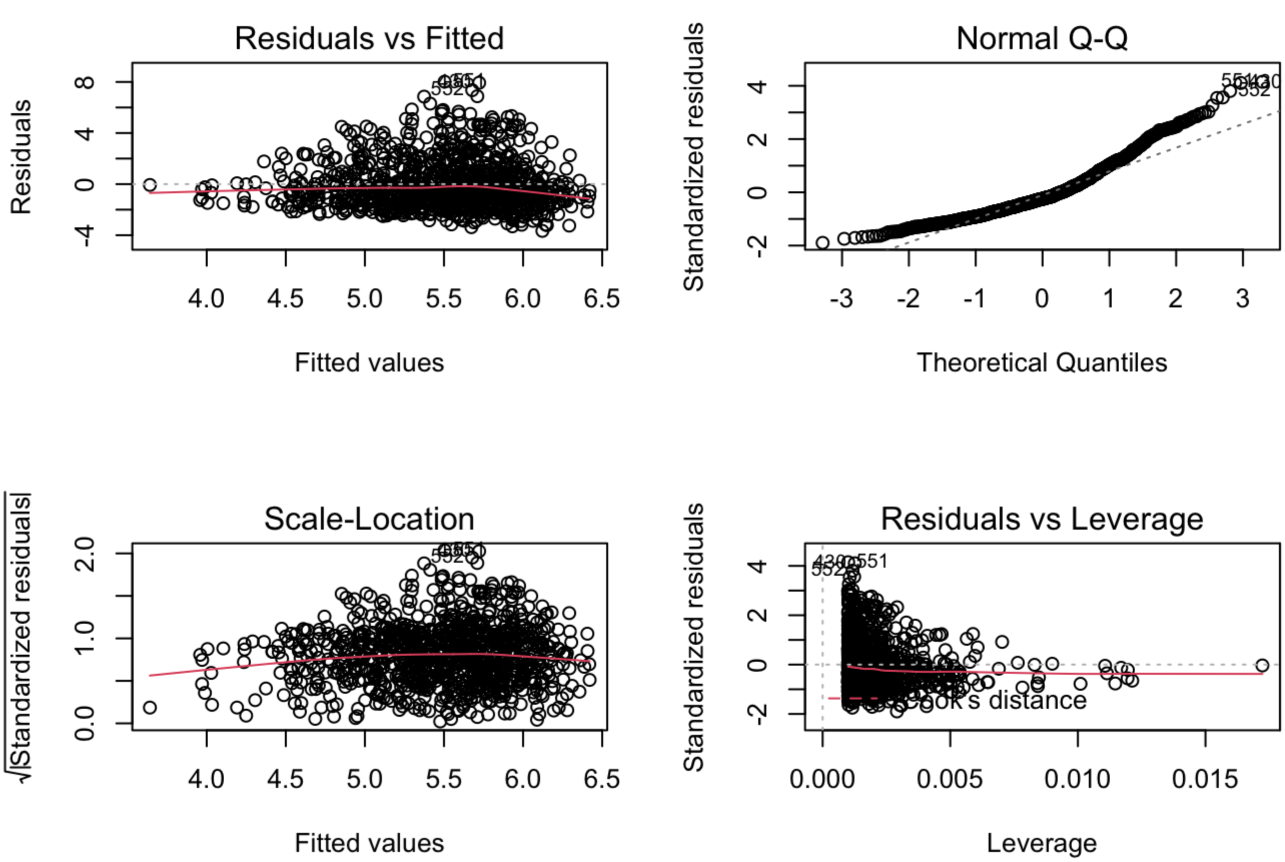
```
##                       2.5 %        97.5 %
## (Intercept)    7.073945e+00  8.232296e+00
## Median_Income -5.296677e-05 -3.099267e-05
```

```r
cse = function(reg) {
rob = sqrt(diag(vcovHC(reg, type = "HC1")))
return(rob)
}

stargazer(reg, se=list(cse(reg)), title="Effect of Median Income on Unemployment Rate",
type="text", df=FALSE, digits=3)
```

```
##
## Effect of Median Income on Unemployment Rate
## ===============================================
##                           Dependent variable:
##                       ---------------------------
##                                  Rate
## -----------------------------------------------
## Median_Income                -0.00004***
##                                (0.00000)
##
## Constant                       7.653***
##                                 (0.255)
##
## -----------------------------------------------
## Observations                     1,000
## R2                               0.053
## Adjusted R2                      0.052
## Residual Std. Error              1.934
## F Statistic                    56.217***
## ===============================================
## Note:                 *p<0.1; **p<0.05; ***p<0.01
```

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



```r
# Fed_Surp_Def
hetero_1 <- state_model %>% group_by(Year,Month) %>% summarize(Rate = mean(Rate), Fed_Su
rp_Def = mean(Fed_Surp_Def))
```

```
## `summarise()` regrouping output by 'Year' (override with `.groups` argument)
```

```r
reg1 <- lm(Rate ~ Fed_Surp_Def, data= hetero_1)
summary(reg1)
```

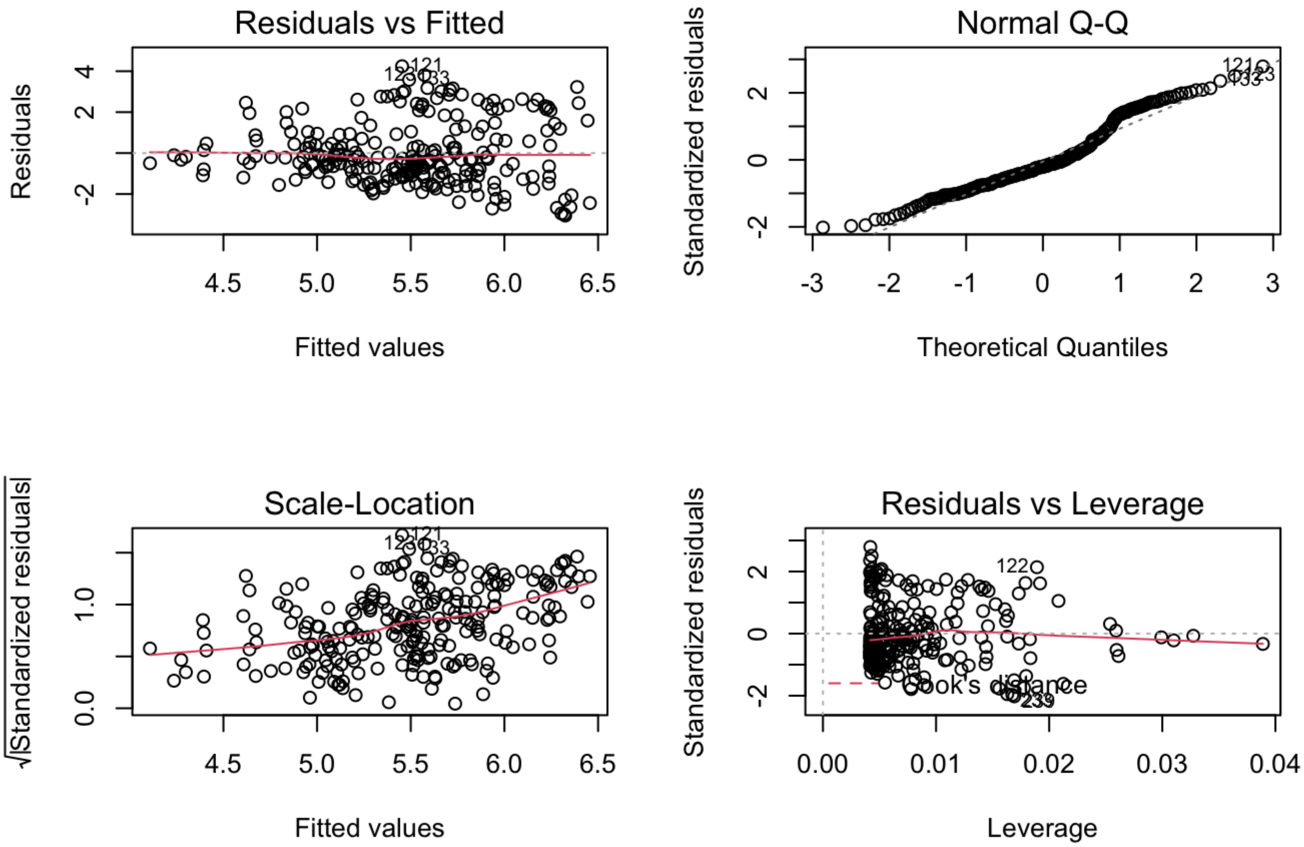
```
##
## Call:
## lm(formula = Rate ~ Fed_Surp_Def, data = hetero_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0525 -1.0839 -0.3192  0.9260  4.2423
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.230e+00  1.118e-01   46.79  < 2e-16 ***
## Fed_Surp_Def -5.239e-06  1.078e-06   -4.86 2.13e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.524 on 238 degrees of freedom
## Multiple R-squared:  0.09029,    Adjusted R-squared:  0.08647
## F-statistic: 23.62 on 1 and 238 DF,  p-value: 2.129e-06
```

```r
cse = function(reg1) {
rob = sqrt(diag(vcovHC(reg1, type = "HC1")))
return(rob)
}

stargazer(reg1, se=list(cse(reg1)), title="Effect of Median Income on Unemployment Rate"
,
type="text", df=FALSE, digits=3)
```

```
##
## Effect of Median Income on Unemployment Rate
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                Rate
## -----------------------------------------------
## Fed_Surp_Def                -0.00001***
##                              (0.00000)
##
## Constant                     5.230***
##                               (0.088)
##
## -----------------------------------------------
## Observations                   240
## R2                            0.090
## Adjusted R2                   0.086
## Residual Std. Error           1.524
## F Statistic                  23.622***
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

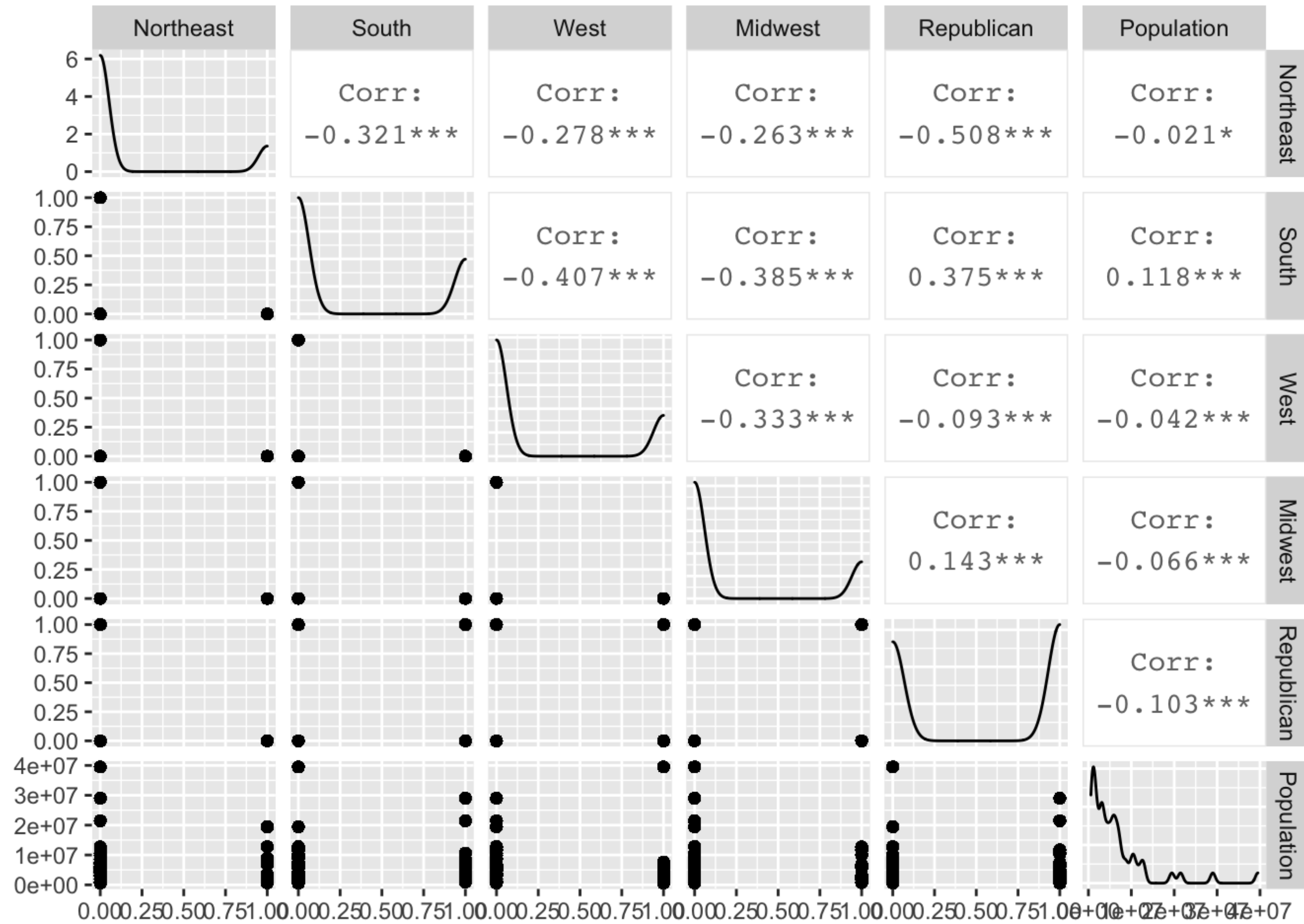
```
par(mfrow=c(2,2))
plot(reg1)
```



#

Multicollinearity between States

```
test2 <- subset(state_model, select = c(Northeast, South, West, Midwest, Republican, Pop
ulation))
ggpairs(test2)
```



#

Regression Models

Logit Models

```
logit <- glm(Republican ~ Median_Income, family = binomial(link="probit"), data = state_
model)

# Confidence Intervals
confint(logit)
```

```
## Waiting for profiling to be done...
```

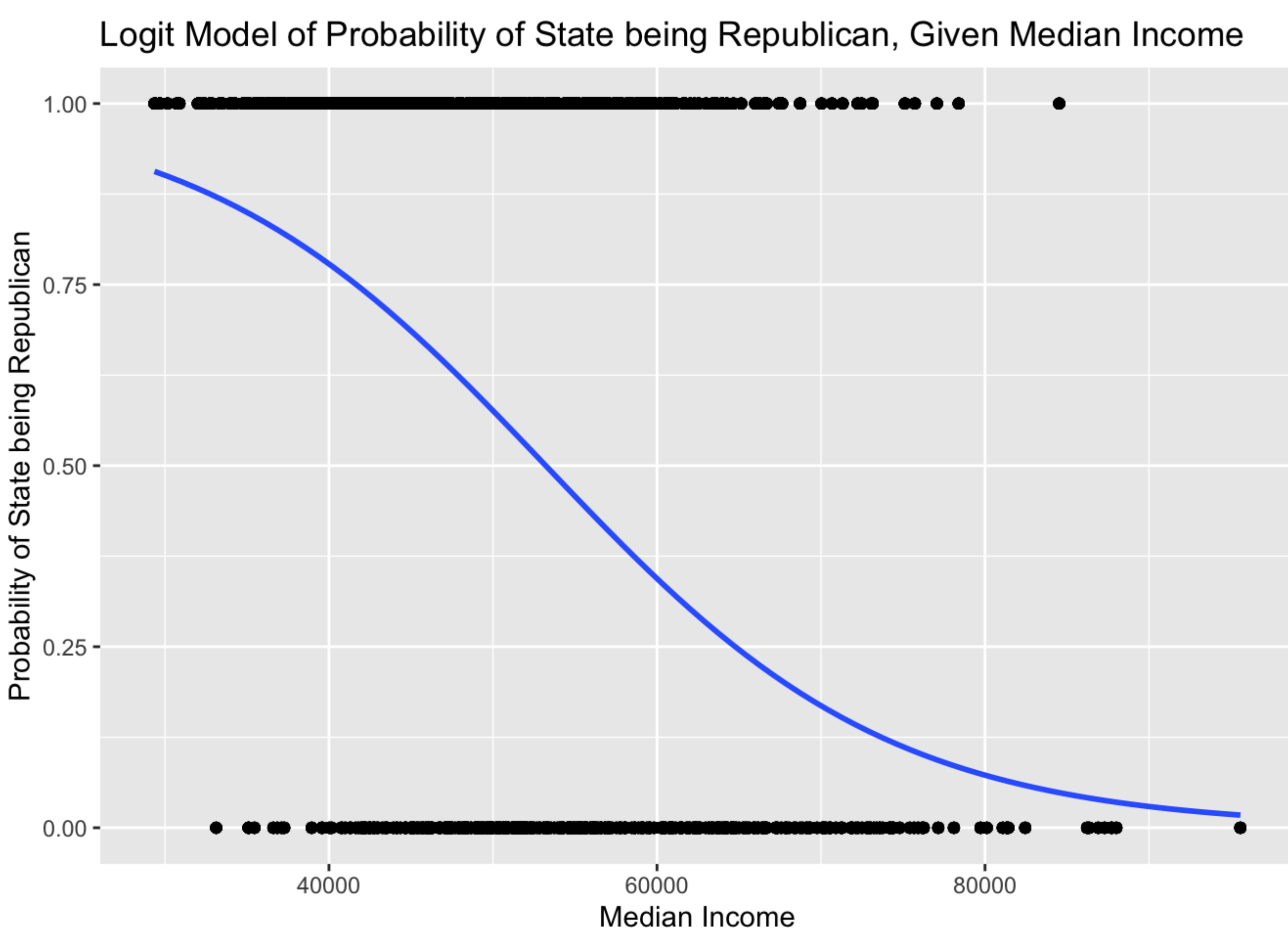
```
##                        2.5 %        97.5 %
## (Intercept)    2.930175e+00  3.190700e+00
## Median_Income -5.978273e-05 -5.481469e-05
```

```
# Z test - SIGNIFICANT!!
coeftest(logit, vcov. = vcovHC, type = "HC1")
```

```
##
## z test of coefficients:
##
##                  Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)    3.0600e+00  6.7559e-02  45.294 < 2.2e-16 ***
## Median_Income -5.7290e-05  1.3301e-06 -43.071 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Plot
ggplot(data=state_model, aes(x=Median_Income, y=Republican)) +
  geom_point() +
  labs(title="Logit Model of Probability of State being Republican, Given Median Income"
) +
  labs(x="Median Income", y="Probability of State being Republican")+
  stat_smooth(method="glm", method.args=list(family="binomial"), se=FALSE)
```

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




```
###### Adding another variable into the model (Region)
logit1 <- glm(Republican ~ Median_Income + South, family = binomial(link="probit"), data
= state_model)
summary(logit1)
```

```
##
## Call:
## glm(formula = Republican ~ Median_Income + South, family = binomial(link = "probit"),
##     data = state_model)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9154  -1.0022   0.3665   0.9328   2.5303
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.381e+00  7.119e-02   33.45   <2e-16 ***
## Median_Income -4.879e-05  1.323e-06  -36.89   <2e-16 ***
## South          9.155e-01  2.954e-02   30.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 16559  on 11999  degrees of freedom
## Residual deviance: 13195  on 11997  degrees of freedom
## AIC: 13201
##
## Number of Fisher Scoring iterations: 5
```

```
# Z test - SIGNIFICANT!!
coeftest(logit1, vcov. = vcovHC, type = "HC1")
```

```
##
## z test of coefficients:
##
##                  Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)    2.3810e+00  7.4085e-02  32.138 < 2.2e-16 ***
## Median_Income -4.8785e-05  1.4121e-06 -34.549 < 2.2e-16 ***
## South          9.1546e-01  2.6831e-02  34.120 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Finding the change
predictions <- predict(logit1,
                       newdata = data.frame("South" = c(0,1),
                                            "Median_Income" = c(35700,35700)),
                       type = "response")

predictions
```

```
##         1         2
## 0.7387007 0.9400037
```

```r
diff(predictions)
```

```
##         2
## 0.2013031
```

```r
# CONCLUSION: We find that non-Southern states have a 73.8% probability of being Republi
can, while Southern states have a 94% probability of being Republican.
```

```r
# Home Value Index of State vs. Unemployment Rate
logit <- glm(High_Unemp ~ HVI, family = binomial(link="probit"), data = state_model)
summary(logit)
```

```
##
## Call:
## glm(formula = High_Unemp ~ HVI, family = binomial(link = "probit"),
##     data = state_model)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4410  -1.2617   0.9861   1.0614   1.5407
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.035e-01  2.915e-02   17.27   <2e-16 ***
## HVI         -1.927e-06  1.377e-07  -13.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 16510  on 11999  degrees of freedom
## Residual deviance: 16312  on 11998  degrees of freedom
## AIC: 16316
##
## Number of Fisher Scoring iterations: 3
```

```r
# Confidence Intervals
confint(logit)
```

```
## Waiting for profiling to be done...
```

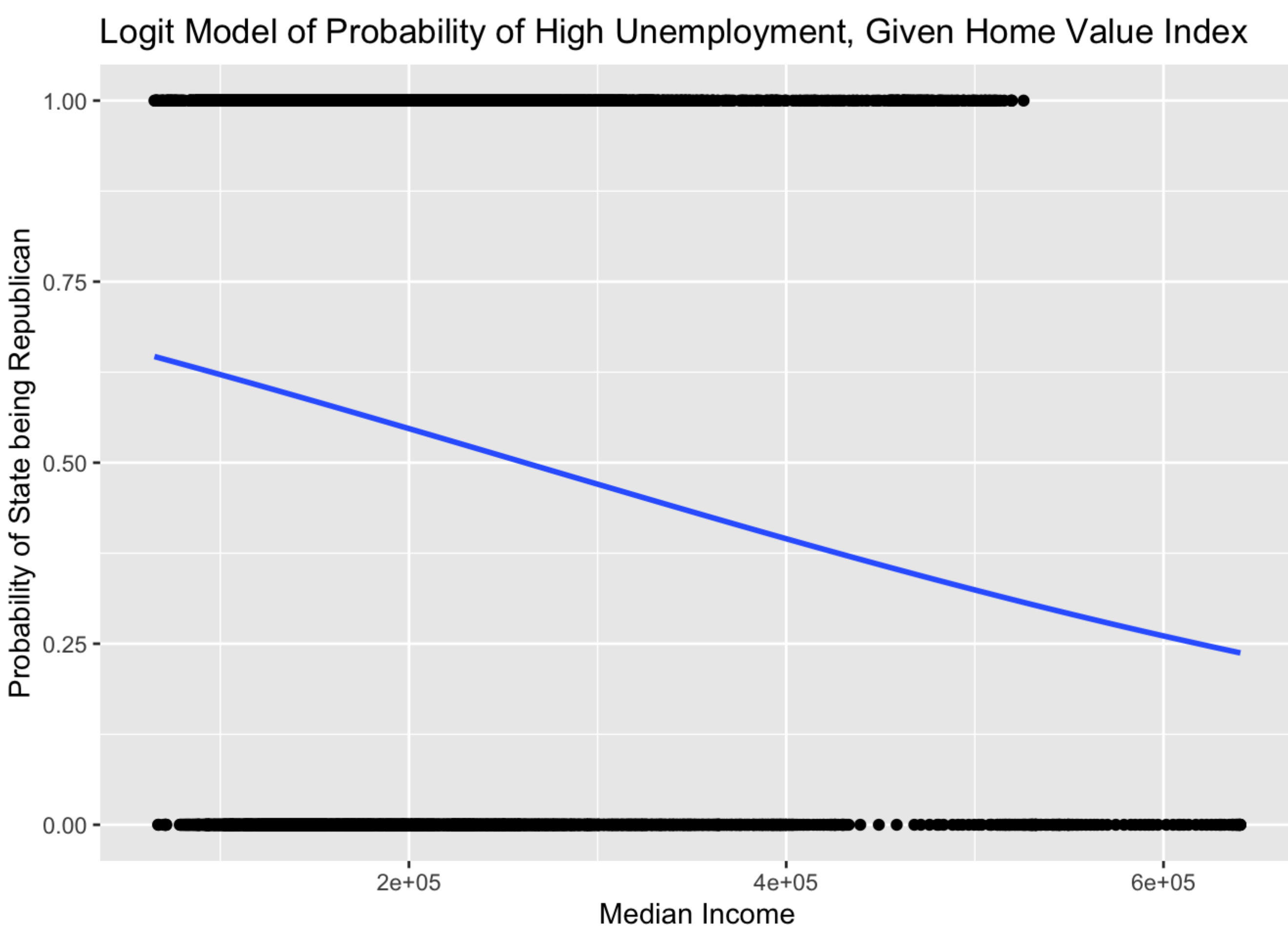
```
##                     2.5 %        97.5 %
## (Intercept)  4.462498e-01  5.609513e-01
## HVI         -2.198184e-06 -1.655979e-06
```

```
# Z test - SIGNIFICANT!!
coeftest(logit, vcov. = vcovHC, type = "HC1")
```

```
##
## z test of coefficients:
##
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)  5.0353e-01  2.8601e-02  17.605 < 2.2e-16 ***
## HVI         -1.9266e-06  1.3484e-07 -14.288 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Plot
ggplot(data=state_model, aes(x=HVI, y=High_Unemp)) +
  geom_point() +
  labs(title="Logit Model of Probability of High Unemployment, Given Home Value Index")
 +
  labs(x="Median Income", y="Probability of State being Republican")+
  stat_smooth(method="glm", method.args=list(family="binomial"), se=FALSE)
```

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



```
###### Adding another variable into the model (Region)
logit1 <- glm(High_Unemp ~ Population + West, family = binomial(link="probit"), data = s
tate_model)

# Z test - SIGNIFICANT!!
coeftest(logit1, vcov. = vcovHC, type = "HC1")
```

```
##
## z test of coefficients:
##
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -1.0556e-01  1.8583e-02 -5.6805 1.343e-08 ***
## Population   2.9398e-08  2.0663e-09 14.2273 < 2.2e-16 ***
## West         1.8224e-01  2.7019e-02  6.7449 1.532e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Finding the change
predictions <- predict(logit1,
                       newdata = data.frame("West" = c(0,1),
                                            "HVI" = c(35700,35700),
                                            "Population" = c(200000,200000)),
                       type = "response")

predictions
```

```
##         1         2
## 0.4603000 0.5328988
```

```
diff(predictions)
```

```
##          2
## 0.07259886
```

```
# CONCLUSION: We find that non-Western states have a 46% probability of experiencing hig
h unemployment while Western states have a 53% probability of experiencing high unemploy
ment.
```

```
# Playing around with other regions
logit1 <- glm(High_Unemp ~ Median_Income + Midwest, family = binomial(link="probit"), da
ta = state_model)

# Z test - SIGNIFICANT!!
coeftest(logit1, vcov. = vcovHC, type = "HC1")
```

```
##
## z test of coefficients:
##
##                  Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)    1.7463e+00  5.7998e-02  30.110 < 2.2e-16 ***
## Median_Income -2.9474e-05  1.0725e-06 -27.482 < 2.2e-16 ***
## Midwest       -3.9464e-01  2.7059e-02 -14.585 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Finding the change
predictions <- predict(logit1,
                       newdata = data.frame("Midwest" = c(0,1),
                                            "Median_Income" = c(35700,35700),
                       type = "response"))

predictions
```

```
##         1         2
## 0.6941062 0.2994616
```

```
diff(predictions)
```

```
##          2
## -0.3946446
```

```
# CONCLUSION: We find that non-Midwestern states have a 69.4% probability of experiencin
g high unemployment while Midwestern states have only a 30% of experiencing high unemplo
yment. This is a large difference, which shows us that Midwestern states are less likely
to experience high unemployment.
```

Multiple Linear Regression Models

```
# Income, Population, West, Republican
model1 <- lm(Rate ~ 0 + Median_Income + West + Northeast + Midwest + Republican , data =
state_model)
summary(model1)
```

```
##
## Call:
## lm(formula = Rate ~ 0 + Median_Income + West + Northeast + Midwest +
##     Republican, data = state_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8802 -1.4032  0.0345  1.6242 11.7867
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## Median_Income  8.217e-05  1.012e-06  81.164  < 2e-16 ***
## West           5.717e-01  6.237e-02   9.166  < 2e-16 ***
## Northeast      6.159e-01  7.833e-02   7.863 4.09e-15 ***
## Midwest       -1.658e-01  6.104e-02  -2.717   0.0066 **
## Republican     1.464e+00  4.645e-02  31.506  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.443 on 11995 degrees of freedom
## Multiple R-squared:  0.8263, Adjusted R-squared:  0.8262
## F-statistic: 1.141e+04 on 5 and 11995 DF,  p-value: < 2.2e-16
```

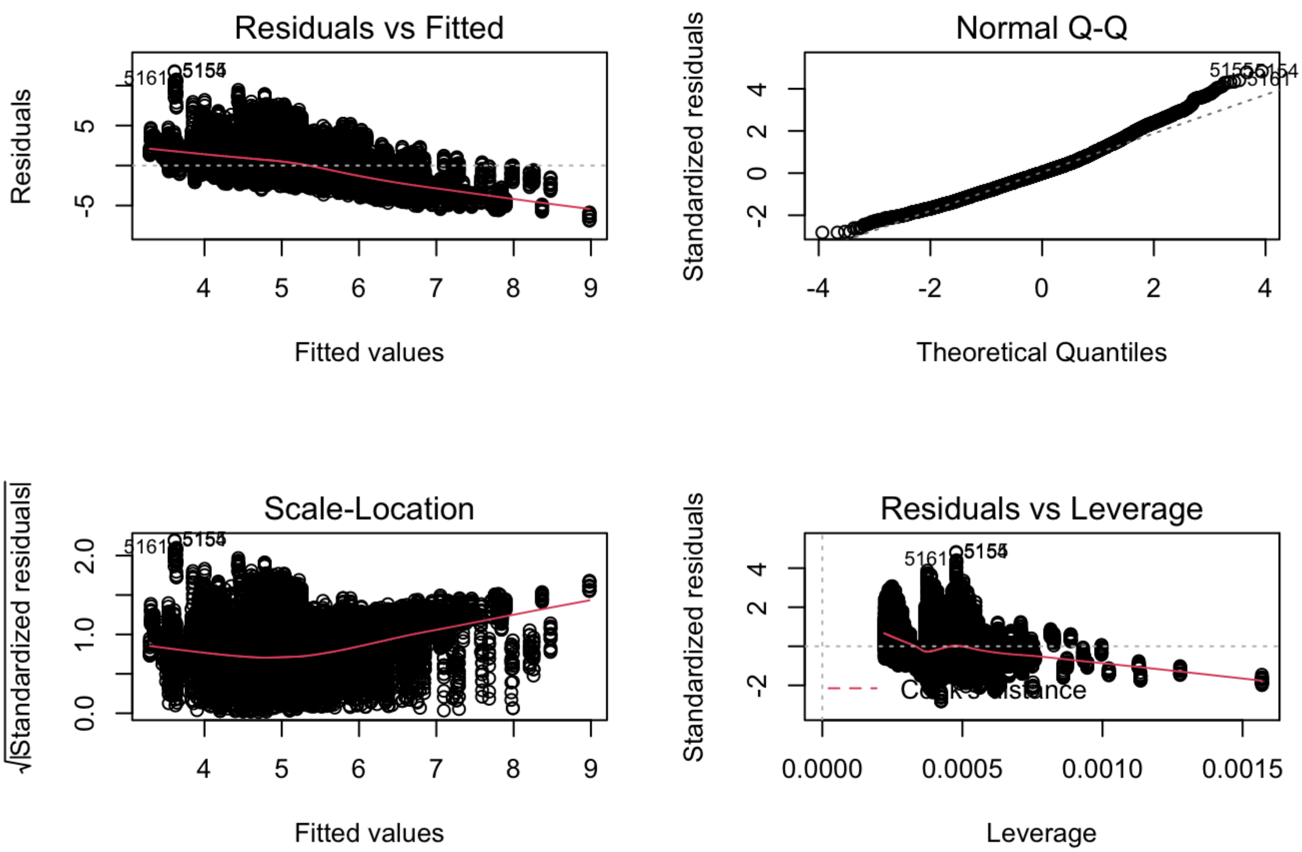
```
sigma(model1)/mean(state_model$Rate)
```

```
## [1] 0.4450813
```

```
stargazer(model1, type="text", median =TRUE, digits = 2, title= "Descriptive Statistics
 on the relationship between Experience (x) and Weekly Wage (y)")
```

```
##
## Descriptive Statistics on the relationship between Experience (x) and Weekly Wage (y)
## ================================================
##                          Dependent variable:
##                      ----------------------------
##                                  Rate
## ------------------------------------------------
## Median_Income                  0.0001***
##                                 (0.0000)
##
## West                            0.57***
##                                  (0.06)
##
## Northeast                       0.62***
##                                  (0.08)
##
## Midwest                         -0.17***
##                                  (0.06)
##
## Republican                      1.46***
##                                  (0.05)
##
## ------------------------------------------------
## Observations                     12,000
## R2                                0.83
## Adjusted R2                       0.83
## Residual Std. Error      2.44 (df = 11995)
## F Statistic         11,409.99*** (df = 5; 11995)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

```
# First, we check for heteroskedasticity in the error terms
par(mfrow=c(2,2))
plot(model1)
```




```
# FG test
omcdiag(model1)
```

```
##
## Call:
## omcdiag(mod = model1)
##
##
## Overall Multicollinearity Diagnostics
##
##                        MC Results detection
## Determinant |X'X|:         0.4657         0
## Farrar Chi-Square:      9169.1036         1
## Red Indicator:             0.3012         0
## Sum of Lambda Inverse:     6.1926         0
## Theil's Method:            1.3045         1
## Condition Number:          4.7314         0
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
```

```
imcdiag(model1)
```

```
##
## Call:
## imcdiag(mod = model1)
##
##
## All Individual Multicollinearity Diagnostics Result
##
##               VIF    TOL       Wi       Fi Leamer   CVIF Klein  IND1   IND2
## West       1.4834 0.6741 1933.014 2899.763 0.8210 1.4769     1 2e-04 0.9447
## Northeast  1.8658 0.5360 3461.934 5193.333 0.7321 1.8576     1 1e-04 1.3451
## Midwest    1.3533 0.7390 1412.611 2119.093 0.8596 1.3474     1 2e-04 0.7567
## Republican 1.4901 0.6711 1959.800 2939.945 0.8192 1.4836     1 2e-04 0.9535
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## * all coefficients have significant t-ratios
##
## R-square of y on all x: 0.0251
##
## * use method argument to check which regressors may be the reason of collinearity
## ===================================
```

```
# Function to calculate corrected SEs for regression
cse = function(model1) {
rob = sqrt(diag(vcovHC(model1, type = "HC1")))
return(rob)
}

stargazer(model1, se=list(cse(model1)), title="Model 1 - Multiple Linear Regression",
type="text", df=FALSE, digits=3)
```

```
##
## Model 1 - Multiple Linear Regression
## ===============================================
##                          Dependent variable:
##                      ---------------------------
##                                 Rate
## -----------------------------------------------
## Median_Income                 0.0001***
##                               (0.00000)
##
## West                          0.572***
##                                (0.059)
##
## Northeast                     0.616***
##                                (0.068)
##
## Midwest                       -0.166***
##                                (0.060)
##
## Republican                    1.464***
##                                (0.046)
##
## -----------------------------------------------
## Observations                   12,000
## R2                              0.826
## Adjusted R2                     0.826
## Residual Std. Error             2.443
## F Statistic                 11,410.000***
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
# Confidence model
confint(model1)
```

```
##                        2.5 %        97.5 %
## Median_Income  8.018167e-05  8.415041e-05
## West           4.494591e-01  6.939768e-01
## Northeast      4.623300e-01  7.694080e-01
## Midwest       -2.854700e-01 -4.617381e-02
## Republican     1.372483e+00  1.554591e+00
```

```
# Income, Population, West, Republican
model1 <- lm(Rate ~ 0 + log(Median_Income) + West + Northeast + Midwest + Republican + P
opulation + AboveHigh_Rank , data = state_model)
summary(model1)
```

```
##
## Call:
## lm(formula = Rate ~ 0 + log(Median_Income) + West + Northeast +
##     Midwest + Republican + Population + AboveHigh_Rank, data = state_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9761 -1.3952 -0.4172  1.0517 10.1246
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## log(Median_Income)  3.535e-01  7.020e-03  50.351  < 2e-16 ***
## West                7.318e-01  5.665e-02  12.917  < 2e-16 ***
## Northeast           3.559e-01  7.010e-02   5.077 3.89e-07 ***
## Midwest             3.621e-01  5.946e-02   6.091 1.16e-09 ***
## Republican         -1.565e-01  4.377e-02  -3.577 0.000349 ***
## Population          4.555e-09  2.839e-09   1.605 0.108570
## AboveHigh_Rank      5.365e-02  1.747e-03  30.705  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.95 on 11993 degrees of freedom
## Multiple R-squared:  0.8893, Adjusted R-squared:  0.8893
## F-statistic: 1.377e+04 on 7 and 11993 DF,  p-value: < 2.2e-16
```

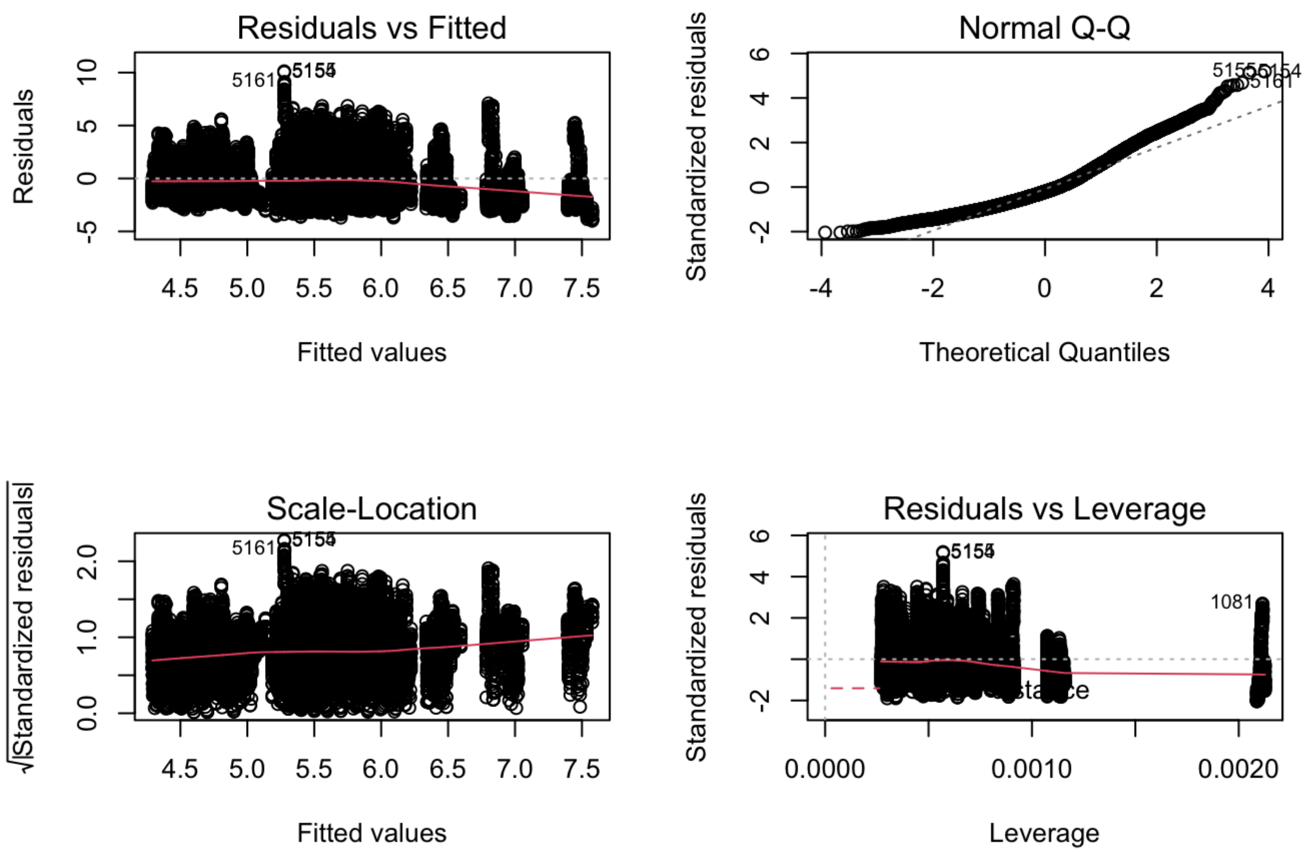
```
sigma(model1)/mean(state_model$Rate)
```

```
## [1] 0.3552444
```

```
stargazer(model1, type="text", median =TRUE, digits = 2, title= "Descriptive Statistics
 on the relationship between Experience (x) and Weekly Wage (y)")
```

```
##
## Descriptive Statistics on the relationship between Experience (x) and Weekly Wage (y)
## ================================================
##                         Dependent variable:
##                     ----------------------------
##                                Rate
## ------------------------------------------------
## log(Median_Income)           0.35***
##                               (0.01)
##
## West                         0.73***
##                               (0.06)
##
## Northeast                    0.36***
##                               (0.07)
##
## Midwest                      0.36***
##                               (0.06)
##
## Republican                   -0.16***
##                               (0.04)
##
## Population                     0.00
##                               (0.00)
##
## AboveHigh_Rank               0.05***
##                              (0.002)
##
## ------------------------------------------------
## Observations                  12,000
## R2                             0.89
## Adjusted R2                    0.89
## Residual Std. Error      1.95 (df = 11993)
## F Statistic         13,769.82*** (df = 7; 11993)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

```
# First, we check for heteroskedasticity in the error terms
par(mfrow=c(2,2))
plot(model1)
```




```
# FG test
omcdiag(model1)
```

```
##
## Call:
## omcdiag(mod = model1)
##
##
## Overall Multicollinearity Diagnostics
##
##                            MC Results detection
## Determinant |X'X|:             0.2142         0
## Farrar Chi-Square:         18483.3712         1
## Red Indicator:                 0.2530         0
## Sum of Lambda Inverse:        11.2745         0
## Theil's Method:                2.1567         1
## Condition Number:             10.0933         0
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
```

```
imcdiag(model1)
```

```
##
## Call:
## imcdiag(mod = model1)
##
##
## All Individual Multicollinearity Diagnostics Result
##
##                   VIF    TOL        Wi       Fi Leamer   CVIF Klein  IND1
## West           1.9504 0.5127 2279.7029 2849.866 0.7161 2.0234     1 2e-04
## Northeast      2.2950 0.4357 3106.4952 3883.443 0.6601 2.3810     1 2e-04
## Midwest        2.0489 0.4881 2516.0934 3145.379 0.6986 2.1256     1 2e-04
## Republican     1.5355 0.6513 1284.4993 1605.758 0.8070 1.5930     1 3e-04
## Population     1.3631 0.7336  871.0549 1088.909 0.8565 1.4142     1 3e-04
## AboveHigh_Rank 2.0816 0.4804 2594.5697 3243.483 0.6931 2.1596     1 2e-04
##                  IND2
## West           1.0835
## Northeast      1.2548
## Midwest        1.1384
## Republican     0.7755
## Population     0.5924
## AboveHigh_Rank 1.1554
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## * all coefficients have significant t-ratios
##
## R-square of y on all x: 0.1083
##
## * use method argument to check which regressors may be the reason of collinearity
## ===================================
```

```
# Function to calculate corrected SEs for regression
cse = function(model1) {
rob = sqrt(diag(vcovHC(model1, type = "HC1")))
return(rob)
}

stargazer(model1, se=list(cse(model1)), title="Model 2 - Multiple Linear Regression",
type="text", df=FALSE, digits=3)
```

```
##
## Model 2 - Multiple Linear Regression
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                Rate
## -----------------------------------------------
## log(Median_Income)           0.353***
##                               (0.007)
##
## West                         0.732***
##                               (0.059)
##
## Northeast                    0.356***
##                               (0.068)
##
## Midwest                      0.362***
##                               (0.061)
##
## Republican                   -0.157***
##                               (0.044)
##
## Population                     0.000
##                               (0.000)
##
## AboveHigh_Rank               0.054***
##                               (0.002)
##
## -----------------------------------------------
## Observations                  12,000
## R2                             0.889
## Adjusted R2                    0.889
## Residual Std. Error            1.950
## F Statistic                 13,769.820***
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
# Confidence model
confint(model1)
```

```
##                            2.5 %        97.5 %
## log(Median_Income)  3.397029e-01  3.672236e-01
## West                6.207577e-01  8.428606e-01
## Northeast           2.184936e-01  4.933082e-01
## Midwest             2.456011e-01  4.786932e-01
## Republican         -2.423355e-01 -7.075350e-02
## Population         -1.008792e-09  1.011927e-08
## AboveHigh_Rank      5.022044e-02  5.706958e-02
```

```
# Income, Population, West, Republican
model3 <- lm(Rate ~ 0 + log(Median_Income) + Interest_Rate + West + Northeast + Midwest
 + Republican + AboveHigh_Rank , data = state_model)
summary(model3)
```

```
##
## Call:
## lm(formula = Rate ~ 0 + log(Median_Income) + Interest_Rate +
##     West + Northeast + Midwest + Republican + AboveHigh_Rank,
##     data = state_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9375 -1.2306 -0.1602  0.9758  9.3852
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## log(Median_Income)  0.422182   0.006309  66.917  < 2e-16 ***
## Interest_Rate      -0.475362   0.008184 -58.082  < 2e-16 ***
## West                0.785340   0.049756  15.784  < 2e-16 ***
## Northeast           0.420088   0.061822   6.795 1.13e-11 ***
## Midwest             0.425100   0.051729   8.218 2.28e-16 ***
## Republican         -0.134207   0.038088  -3.524 0.000427 ***
## AboveHigh_Rank      0.056655   0.001353  41.873  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.723 on 11993 degrees of freedom
## Multiple R-squared:  0.9136, Adjusted R-squared:  0.9136
## F-statistic: 1.812e+04 on 7 and 11993 DF,  p-value: < 2.2e-16
```

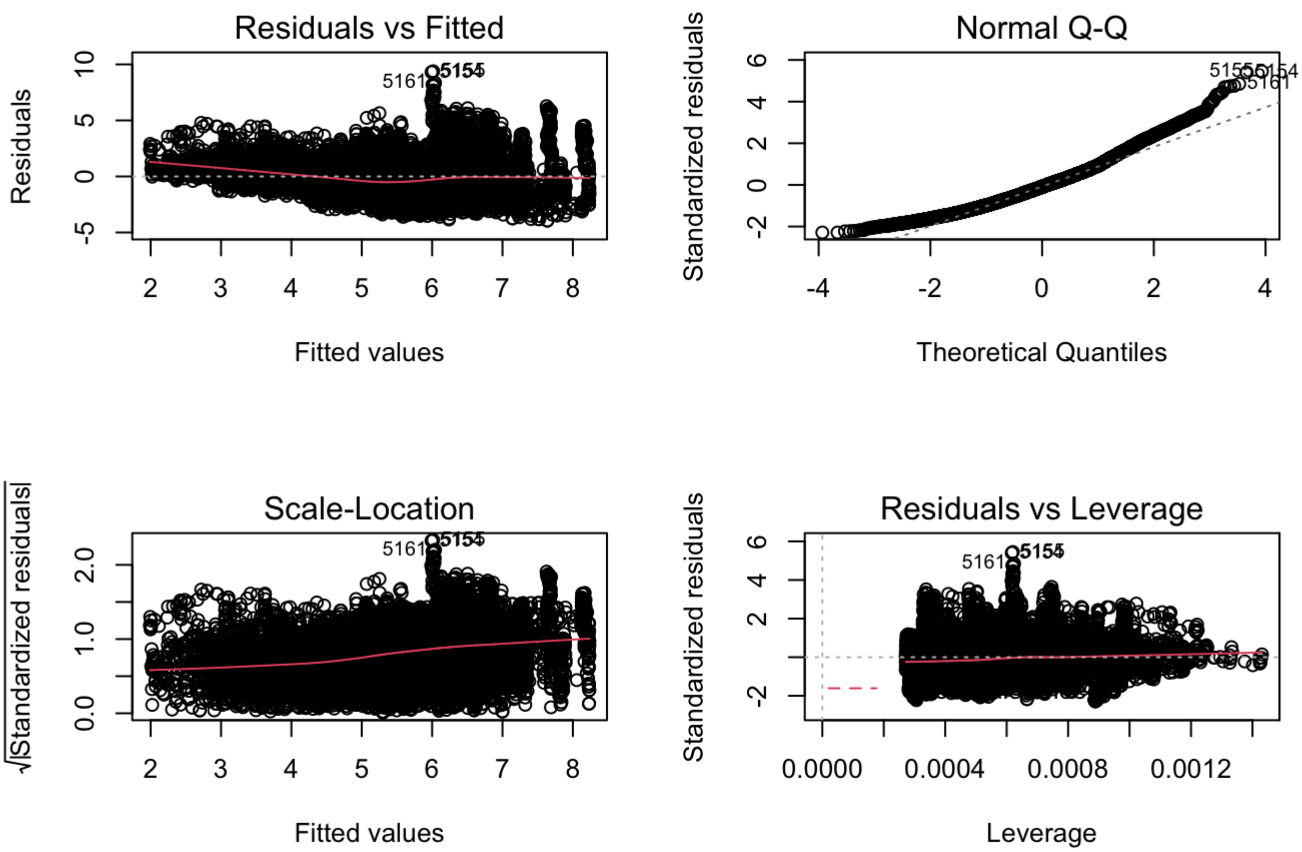
```
sigma(model1)/mean(state_model$Rate)
```

```
## [1] 0.3552444
```

```
stargazer(model1, type="text", median =TRUE, digits = 2, title= "Descriptive Statistics
 on the relationship between Experience (x) and Weekly Wage (y)")
```

```
##
## Descriptive Statistics on the relationship between Experience (x) and Weekly Wage (y)
## ===============================================
##                         Dependent variable:
##                     ----------------------------
##                                 Rate
## -----------------------------------------------
## log(Median_Income)           0.35***
##                               (0.01)
##
## West                         0.73***
##                               (0.06)
##
## Northeast                    0.36***
##                               (0.07)
##
## Midwest                      0.36***
##                               (0.06)
##
## Republican                   -0.16***
##                               (0.04)
##
## Population                     0.00
##                               (0.00)
##
## AboveHigh_Rank               0.05***
##                              (0.002)
##
## -----------------------------------------------
## Observations                  12,000
## R2                             0.89
## Adjusted R2                    0.89
## Residual Std. Error      1.95 (df = 11993)
## F Statistic         13,769.82*** (df = 7; 11993)
## ===============================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

```
# First, we check for heteroskedasticity in the error terms
par(mfrow=c(2,2))
plot(model3)
```




```
# FG test
omcdiag(model3)
```

```
##
## Call:
## omcdiag(mod = model3)
##
##
## Overall Multicollinearity Diagnostics
##
##                        MC Results detection
## Determinant |X'X|:         0.2920         0
## Farrar Chi-Square:     14767.2347         1
## Red Indicator:             0.2217         0
## Sum of Lambda Inverse:    10.2734         0
## Theil's Method:            0.6489         1
## Condition Number:         10.0196         0
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
```

```
imcdiag(model3)
```

```
##
## Call:
## imcdiag(mod = model3)
##
##
## All Individual Multicollinearity Diagnostics Result
##
##                    VIF    TOL       Wi       Fi Leamer   CVIF Klein  IND1   IND2
## Interest_Rate  1.0000 1.0000    0.000    0.000 1.0000 1.0018     0 4e-04 0.0000
## West           1.9222 0.5202 2212.121 2765.382 0.7213 1.9256     1 2e-04 1.2849
## Northeast      2.2829 0.4380 3077.316 3846.966 0.6619 2.2869     1 2e-04 1.5051
## Midwest        1.9794 0.5052 2349.470 2937.082 0.7108 1.9829     1 2e-04 1.3252
## Republican     1.4942 0.6692 1185.566 1482.081 0.8181 1.4969     1 3e-04 0.8859
## AboveHigh_Rank 1.5947 0.6271 1426.631 1783.437 0.7919 1.5975     1 3e-04 0.9988
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## * all coefficients have significant t-ratios
##
## R-square of y on all x: 0.3183
##
## * use method argument to check which regressors may be the reason of collinearity
## ===================================
```

```
# Function to calculate corrected SEs for regression
cse = function(model1) {
rob = sqrt(diag(vcovHC(model1, type = "HC1")))
return(rob)
}

stargazer(model1, se=list(cse(model1)), title="Model 3 - Multiple Linear Regression",
type="text", df=FALSE, digits=3)
```

```
##
## Model 3 - Multiple Linear Regression
## ===============================================
##                          Dependent variable:
##                      ---------------------------
##                                 Rate
## -----------------------------------------------
## log(Median_Income)            0.353***
##                                (0.007)
##
## West                          0.732***
##                                (0.059)
##
## Northeast                     0.356***
##                                (0.068)
##
## Midwest                       0.362***
##                                (0.061)
##
## Republican                    -0.157***
##                                (0.044)
##
## Population                      0.000
##                                (0.000)
##
## AboveHigh_Rank                0.054***
##                                (0.002)
##
## -----------------------------------------------
## Observations                   12,000
## R2                              0.889
## Adjusted R2                     0.889
## Residual Std. Error             1.950
## F Statistic                 13,769.820***
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
# Confidence model
confint(model1)
```

```
##                           2.5 %        97.5 %
## log(Median_Income)  3.397029e-01  3.672236e-01
## West                6.207577e-01  8.428606e-01
## Northeast           2.184936e-01  4.933082e-01
## Midwest             2.456011e-01  4.786932e-01
## Republican         -2.423355e-01 -7.075350e-02
## Population         -1.008792e-09  1.011927e-08
## AboveHigh_Rank      5.022044e-02  5.706958e-02
```