ECON_104_Project_1

Maritza Jimenez, Titania Le, Aanya Pramanik, Elaine Tran

2023-10-12

Question 1

In this project we will attempt to answer the question: Do experience and education have an impact on a person's wage?

Question 2a

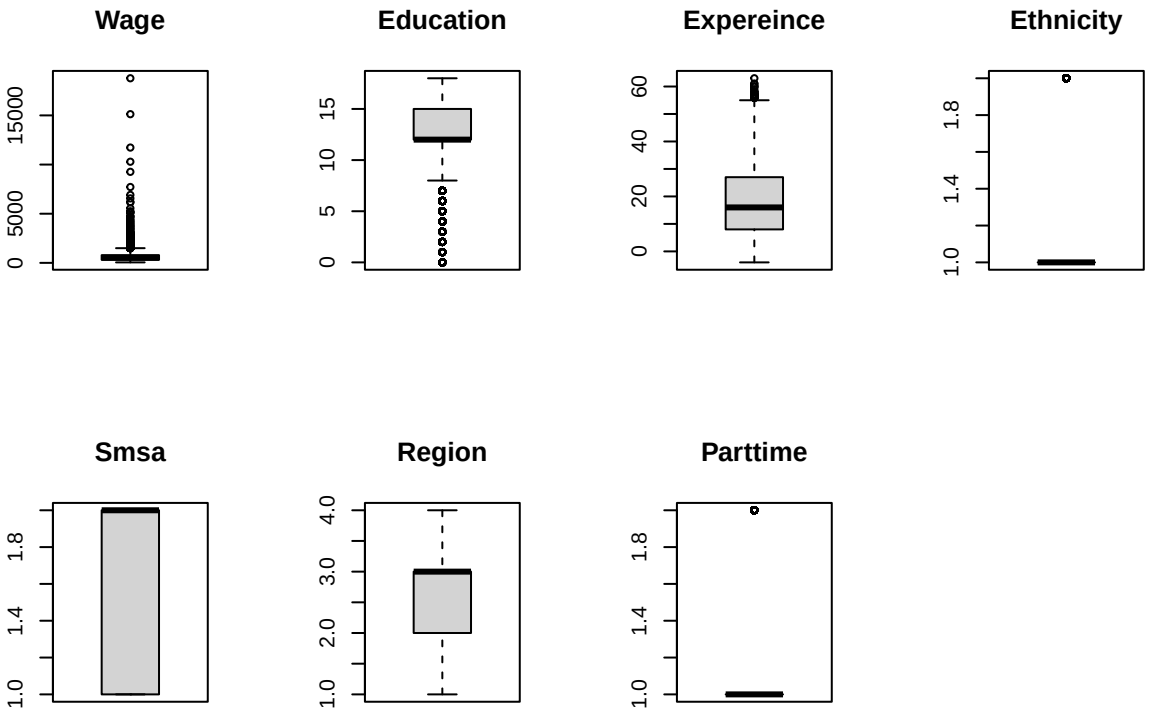
Christian Kleiber and Achim Zeileis (2008). Applied Econometrics with R. New York: Springer-Verlag. ISBN 978-0-387-77316-2. URL https://CRAN.R-project.org/package=AER

Question 2b

The data set used in this project is CPS1988 taken from the March 1988 Current Population Survey from the US census Bureau that can be found in the AER package. The data contains 7 columns/variables with the names: wage, education, experience, ethnicity, smsa, region, and parttime. There are a total of 28155 rows/observations.

Question 2c

```
library(AER)
library(ggplot2)
data("CPS1988")
attach(CPS1988)
par(mfrow = c(2, 4))
boxplot(CPS1988$wage, main = 'Wage')
boxplot(CPS1988$education, main = 'Education')
boxplot(CPS1988$experience, main = 'Expereince')
boxplot(CPS1988$ethnicity, main = 'Ethnicity')
boxplot(CPS1988$smsa, main = 'Smsa')
boxplot(CPS1988$region, main = 'Region')
boxplot(CPS1988$parttime, main = 'Parttime')
```


Wage
0
5000
15000
Education
0
5
10
15
Expereince
0
20
40
60
Ethnicity
1.0
1.4
1.8
Smsa
1.0
1.4
1.8
Region
1.0
2.0
3.0
4.0
Parttime
1.0
1.4
1.8

```r
summary(CPS1988$wage)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    50.05   308.64   522.32   603.73   783.48 18777.20
```

```r
hist(wage, main = "Histogram of Wage", xlab = "wage")
```

The histogram of wage indicated that the data is skewed to the right with the majority of workers making between $0 and $2000 each week

```r
summary(CPS1988$education)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.00   12.00   13.07   15.00   18.00
```

```r
hist(education, main = "Histogram of Education", xlab = "Years of Education")
```

The histogram of education indicates that the data is slightly skewed to the left but the mode is approximately 11 years of education.

```r
summary(CPS1988$experience)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    -4.0     8.0    16.0    18.2    27.0    63.0
```

```r
hist(experience, main = "Histogram of Experience", xlab = "Years of Experience")
```

The histogram of experience indicates data is skewed to the right with the majority of workers having between 0-20 years of experience.

```r
selected_vars <- CPS1988[c("wage", "education", "experience")]
cor_matrix <- cor(selected_vars)
cor_matrix
```

```
##                  wage  education experience
## wage       1.0000000  0.3016440  0.1942204
## education  0.3016440  1.0000000 -0.2867064
## experience 0.1942204 -0.2867064  1.0000000
```

```r
reg.mod <- lm(wage ~ education + experience + ethnicity + smsa + region + parttime ,
              data = CPS1988)

ethnicity_level_counts <- table(CPS1988$ethnicity)
ethnicity_level_percentages <- prop.table(ethnicity_level_counts) * 100
ethnicity_level_percentages
```

```
##
##      cauc      afam
## 92.072456  7.927544
```

92.07% of workers in the data set are Caucasian and 7.93% of workers are African American

```r
smsa_level_counts <- table(CPS1988$region)
smsa_level_percentages <- prop.table(smsa_level_counts) * 100
smsa_level_percentages
```

```
##
## northeast   midwest     south      west
##  22.87693  24.37578  31.11348  21.63381
```

25.65% of workers in the dataset do not reside in a Standard Metropolitan Statistical Area and 74.35% do.

```r
region_level_counts <- table(CPS1988$region)
region_level_percentages <- prop.table(region_level_counts) * 100
region_level_percentages
```

```
##
## northeast   midwest     south      west
##  22.87693  24.37578  31.11348  21.63381
```

22.88% of workers in the dataset are in the northeast, 24.38% are in the midwest, 31.11% are in the south, and 21.63% are in the west.

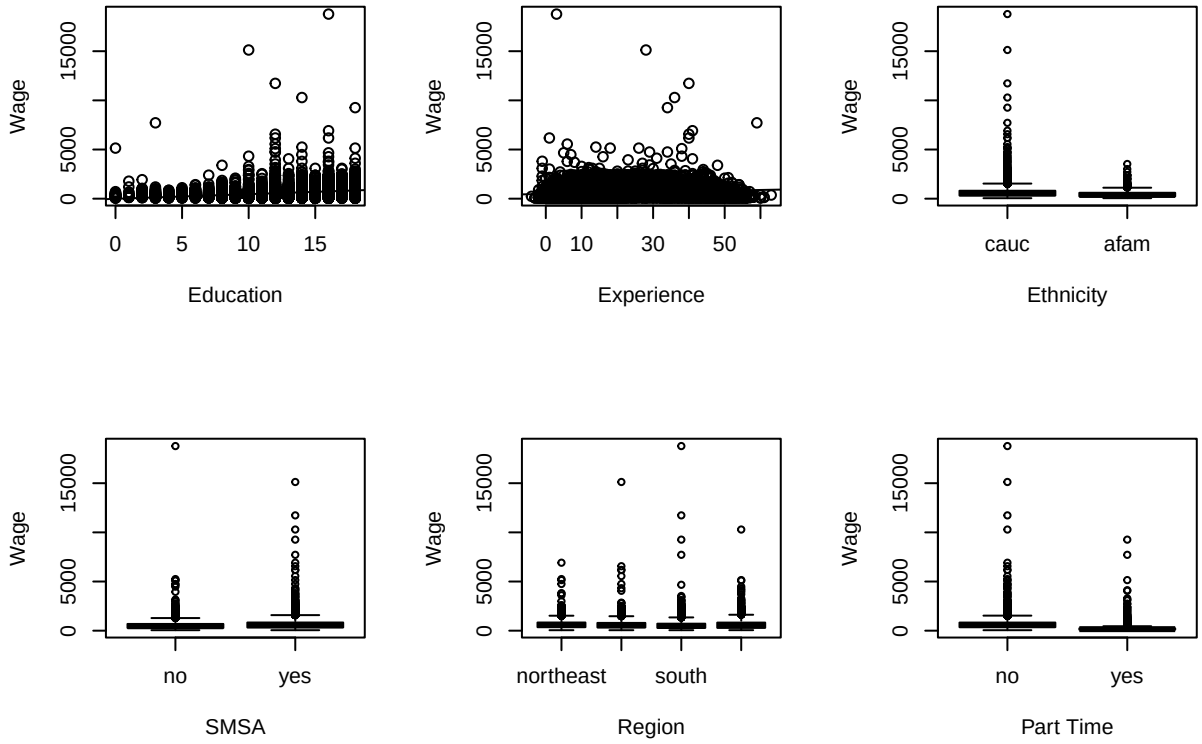
```r
pt_level_counts <- table(CPS1988$parttime)
pt_level_percentages <- prop.table(pt_level_counts) * 100
pt_level_percentages
```

```
##
##       no      yes
## 91.03534  8.96466
```

91.03% of workers in the dataset do not work parttime while 8.96% do work parttime.

Question 2d

```
library(AER)
data("CPS1988")
par(mfrow = c(2, 3))
plot(CPS1988$education, CPS1988$wage, type = "p", xlab = "Education", ylab = "Wage")
abline(lm(wage~education, CPS1988))
plot(CPS1988$experience,CPS1988$wage, type="p", xlab="Experience", ylab="Wage")
abline(lm(wage~experience, CPS1988))
plot(CPS1988$ethnicity,CPS1988$wage, type="p", xlab="Ethnicity", ylab="Wage")
plot(CPS1988$smsa,CPS1988$wage, type="p", xlab="SMSA", ylab="Wage")
plot(CPS1988$region,CPS1988$wage, type="p", xlab="Region", ylab="Wage")
plot(CPS1988$parttime,CPS1988$wage, type="p", xlab="Part Time", ylab="Wage")
```



Some possible regression violations may include:

Non-normality: The skewness of the histograms depict data that is not normally distributed.

Heteroskedasticity: The scatter plots for many of the binary variables indicate heteroskedasticity as there are several outliers in each level.

Multicollinearity: Based on the correlation matrix, the independent (numeric) variables do not appear to have high correlation; thus, multicollinearity is not observed.

Linearity: The data appears to be linear as there are no distinguishable curves on the scatterplot.

Question 3a

```
reg.mod <- lm(wage ~ education + experience + ethnicity + smsa + region + parttime ,
              data = CPS1988)
summary(reg.mod)
```

```
##
## Call:
## lm(formula = wage ~ education + experience + ethnicity + smsa +
##     region + parttime, data = CPS1988)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1042.4  -207.9   -48.8   135.8 18207.6
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -335.8539    14.2762 -23.525  < 2e-16 ***
## education       57.1181     0.8539  66.890  < 2e-16 ***
## experience       9.7961     0.1888  51.878  < 2e-16 ***
## ethnicityafam -121.2918     8.8799 -13.659  < 2e-16 ***
## smsayes         97.6972     5.4639  17.880  < 2e-16 ***
## regionmidwest  -19.2316     6.8981  -2.788  0.00531 **
## regionsouth    -37.8420     6.5783  -5.753 8.88e-09 ***
## regionwest       1.6474     7.0954   0.232  0.81641
## parttimeyes   -357.3211     8.2805 -43.152  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.6 on 28146 degrees of freedom
## Multiple R-squared:  0.2431, Adjusted R-squared:  0.2429
## F-statistic:  1130 on 8 and 28146 DF,  p-value: < 2.2e-16
```

An initial review of the multiple linear regression model shows that all variables beside "regionwest" are statistically significant and different from zero which can indicate that they are significant predictors for the initial model estimated. The "partyimesyes" variable is significantly less than (value = -357.3211) the rest of the estimates which could indicate an anomaly in the set of variables. One can interpret each estimate in the context of the data as follows: Holding all else constant, for 1 unit (year) increase in education, it is associated with a wage increase of 57.1181 dollars and for a 1 unit (year) increase in experience, it is associated with a wage increase of 9.7961 dollars. For the rest of the variables, being African American as one's ethnicity is associated with a decreased wage of 121.2918 dollars; living in a Standard Metropolitan Statistical Area tends to increase wages by 97.6972 dollars; living in the regions midwest, south, or west are associated with a decrease in wages of 19.2316, 37.8420 and increase in wages of 1.6474, respectively. Lastly, working parttime is associated with a decrease of 357.3211 dollars in wages. Economically speaking, knowing the important predictors that influence wage can speak to the current conditions of the labor market and how a consumer may see employment as being attractive or unattractive. This motivation can then reflect the unemployment and labor force participation rates of the economy.

Question 3b

The model appears to be a good fit for the data. Most of the variables (7 out of the 8) are found to be statistically significant, however, the R-squared of the model seems to indicate a counterargument for these variables being good predictors because it is a low number. It suggest that only about 24.29% of the variability in the data can be explained by the model indicating that there may be some variables that can be eliminated to have greater goodness of fit. This could be explained by how the model seems to have slight heteroskedasticity in its fitted values, as discussed in 2(d).

Question 4

```
vif(reg.mod)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## education  1.108388  1        1.052800
## experience 1.102725  1        1.050107
## ethnicity  1.040531  1        1.020064
## smsa       1.029448  1        1.014617
## region     1.055339  3        1.009017
## parttime   1.011660  1        1.005813
```

Based on the VIF findings, none of the variables should be removed, as their VIF values are within the acceptable range (1-5)

Question 5

```
library(leaps)
library(MASS)
full_model <- lm(wage ~ education + experience + ethnicity + smsa + region + parttime,
              data = CPS1988)
backAIC <- step(full_model, direction = 'backward', data = CPS1988)
```

```
summary(backAIC)
```

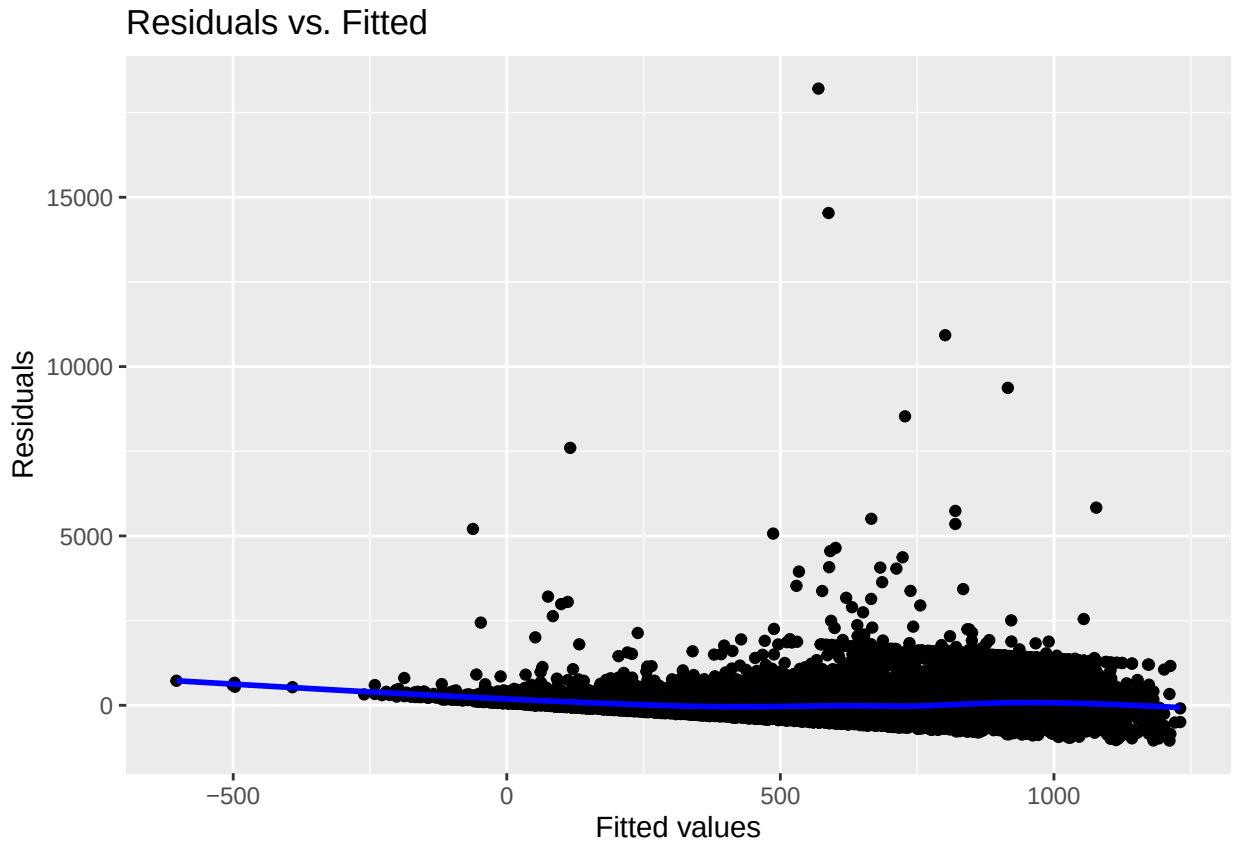
```
##
## Call:
## lm(formula = wage ~ education + experience + ethnicity + smsa +
##     region + parttime, data = CPS1988)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1042.4  -207.9   -48.8   135.8 18207.6
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -335.8539    14.2762 -23.525  < 2e-16 ***
## education       57.1181     0.8539  66.890  < 2e-16 ***
```

```
## experience       9.7961     0.1888  51.878  < 2e-16 ***
## ethnicityafam -121.2918     8.8799 -13.659  < 2e-16 ***
## smsayes         97.6972     5.4639  17.880  < 2e-16 ***
## regionmidwest  -19.2316     6.8981  -2.788  0.00531 **
## regionsouth    -37.8420     6.5783  -5.753 8.88e-09 ***
## regionwest       1.6474     7.0954   0.232  0.81641
## parttimeyes   -357.3211     8.2805 -43.152  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.6 on 28146 degrees of freedom
## Multiple R-squared:  0.2431, Adjusted R-squared:  0.2429
## F-statistic:  1130 on 8 and 28146 DF,  p-value: < 2.2e-16
```

Perfoming backwards AIC on our original model gives us a new model that would be the best fit. In this case, our original model is reproduced, so we keep all our variables since they provide the necessary explanatory power.

Question 6

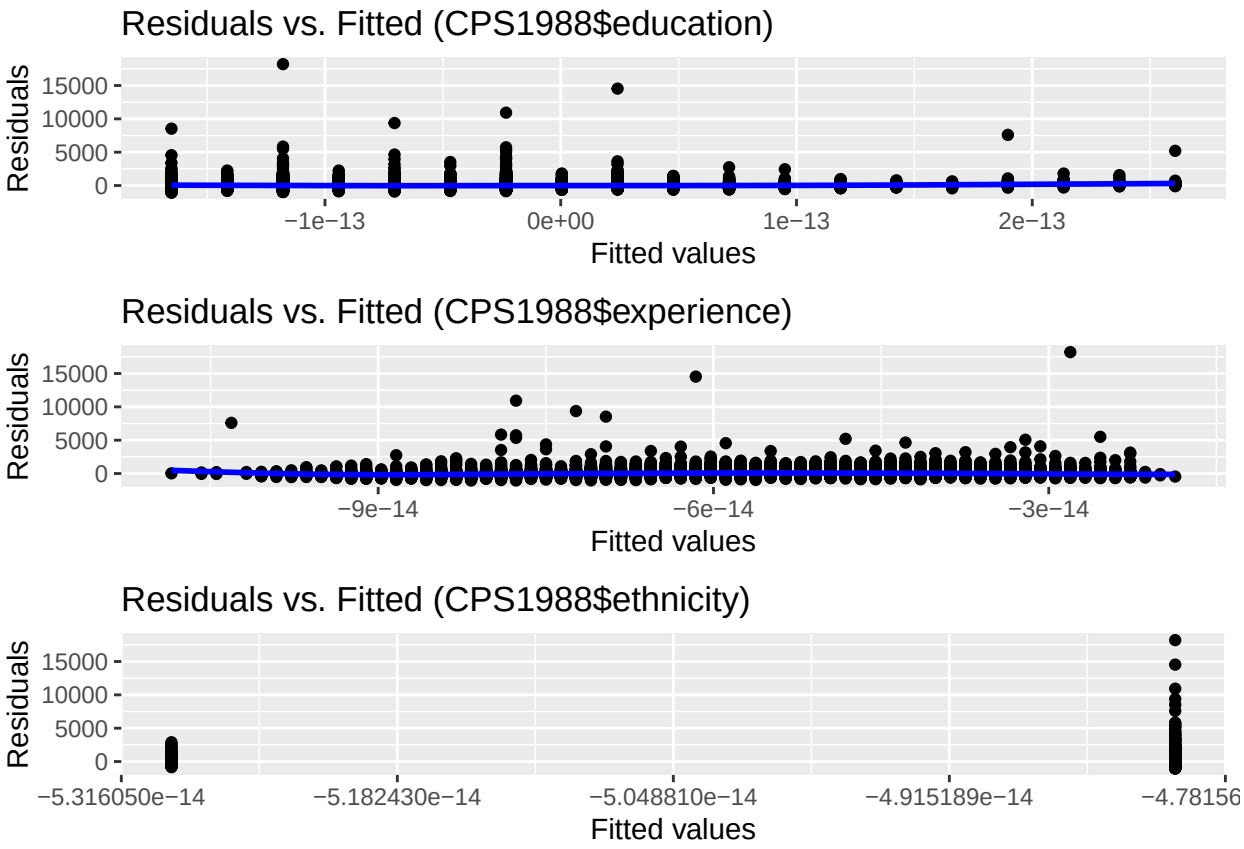
```
plot_lm <- ggplot(full_model, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_smooth(se = FALSE, color = "blue") +
  ggtitle("Residuals vs. Fitted") +
  xlab("Fitted values") +
  ylab("Residuals")
  print(plot_lm)
```



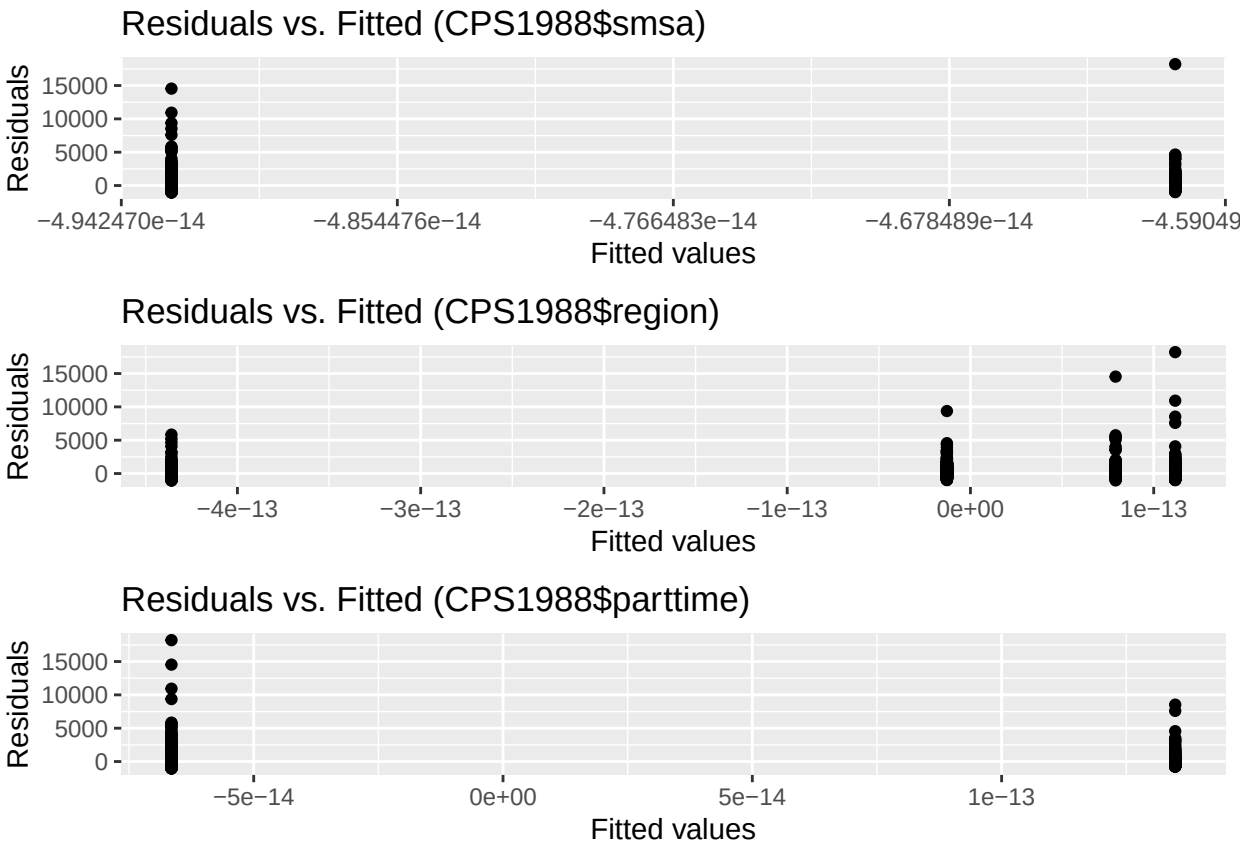

```
#plot residuals against each variable separately

library(gridExtra)

lm_education <- ggplot(lm(resid(full_model) ~ CPS1988$education),
                       aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_smooth(se = FALSE, color = "blue") +
  ggtitle("Residuals vs. Fitted (CPS1988$education)") +
  xlab("Fitted values") +
  ylab("Residuals")


lm_experience <- ggplot(lm(resid(full_model) ~ CPS1988$experience),
```

```
                       aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_smooth(se = FALSE, color = "blue") +
  ggtitle("Residuals vs. Fitted (CPS1988$experience)") +
  xlab("Fitted values") +
  ylab("Residuals")



  lm_ethnicity <- ggplot(lm(resid(full_model) ~ CPS1988$ethnicity),
                         aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_smooth(se = FALSE, color = "blue") +
  ggtitle("Residuals vs. Fitted (CPS1988$ethnicity)") +
  xlab("Fitted values") +
  ylab("Residuals")


  lm_smsa <- ggplot(lm(resid(full_model) ~ CPS1988$smsa),
                    aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_smooth(se = FALSE, color = "blue") +
  ggtitle("Residuals vs. Fitted (CPS1988$smsa)") +
  xlab("Fitted values") +
  ylab("Residuals")


  lm_region <- ggplot(lm(resid(full_model) ~ CPS1988$region),
                      aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_smooth(se = FALSE, color = "blue") +
  ggtitle("Residuals vs. Fitted (CPS1988$region)") +
  xlab("Fitted values") +
  ylab("Residuals")


  lm_parttime <- ggplot(lm(resid(full_model) ~ CPS1988$parttime),
                        aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_smooth(se = FALSE, color = "blue") +
  ggtitle("Residuals vs. Fitted (CPS1988$parttime)") +
  xlab("Fitted values") +
  ylab("Residuals")
```

```
grid.arrange(lm_education, lm_experience, lm_ethnicity, nrow = 3)
```




```
grid.arrange(lm_smsa, lm_region, lm_parttime, nrow = 3)
```



Looking at the residual vs fitted plot we notice that there is evidence of heteroskedasticity, seeing as variance continues to increase as we move further right across the graph. We also note that there are a few outliers and will need to take that into consideration when interpreting our results.

Question 7

```
resettest(full_model, power = 2)
```

```
##
##  RESET test
##
## data:  full_model
## RESET = 422.71, df1 = 1, df2 = 28145, p-value < 2.2e-16
```

The P-value we get from out RESET test is extremely small and less than .05 indicating that we reject the null hypothesis. Rejecting the null hypothesis, means that our regression could benefit from adding squares and/or interactions terms between our variables.

Question 8

```
# White Test
library(broom)
library(knitr)
alpha <- 0.05
ressq <- resid(reg.mod)^2
modres <- lm(ressq ~ experience + I(experience^2))
summary(modres)
```

```
##
## Call:
## lm(formula = ressq ~ experience + I(experience^2))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
##   -368628   -138173   -104478    -51186 331407916
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     103684.66   37686.65   2.751  0.00594 **
## experience        1795.33    3973.30   0.452  0.65138
## I(experience^2)     38.47      84.80   0.454  0.65009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2641000 on 28152 degrees of freedom
## Multiple R-squared:  0.0003099,  Adjusted R-squared:  0.0002389
## F-statistic: 4.363 on 2 and 28152 DF,  p-value: 0.01275
```

```
N <- nobs(modres)
gmodres <- glance(modres)
S <- gmodres$df
chisqcr <- qchisq(1- alpha, S - 1)
Rsqres <- gmodres$r.squared
chisq <- N * Rsqres
pval2 <- 1 - pchisq(chisq, S-1)
print(pval2)
```

```
## [1] 0.003139735
```

```
#Weighted Least Squares
w <- 1/wage
reg.mod.wls <- lm(wage ~ education + experience + ethnicity + smsa + region
                  + parttime, weights = w)

cov1 <- hccm(reg.mod.wls, type = "hc1")
vcv <- coeftest(reg.mod.wls, vcov. = cov1 )
kable(tidy(vcv))
```

term	estimate	std.error	statistic	p.value
(Intercept)	-76.361980	11.0172238	-6.931145	0.0000000
education	32.298325	0.7023788	45.984197	0.0000000
experience	5.650995	0.1547324	36.521080	0.0000000
ethnicityafam	-69.940453	5.1069400	-13.695178	0.0000000
smsayes	54.441250	3.4636363	15.717947	0.0000000
regionmidwest	-19.825312	4.9876683	-3.974866	0.0000706
regionsouth	-38.802556	4.7452778	-8.177088	0.0000000
regionwest	-17.219503	5.1427807	-3.348287	0.0008142
parttimeyes	-283.638861	3.6585201	-77.528304	0.0000000

```
summary(reg.mod.wls)
```

```
##
## Call:
## lm(formula = wage ~ education + experience + ethnicity + smsa +
##     region + parttime, weights = w)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -74.451  -3.176   4.323  11.094 133.975
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -76.3620     8.6591  -8.819  < 2e-16 ***
## education       32.2983     0.5335  60.536  < 2e-16 ***
## experience       5.6510     0.1075  52.586  < 2e-16 ***
## ethnicityafam  -69.9405     4.9389 -14.161  < 2e-16 ***
## smsayes         54.4413     3.2637  16.681  < 2e-16 ***
## regionmidwest  -19.8253     4.4162  -4.489 7.18e-06 ***
## regionsouth    -38.8026     4.1715  -9.302  < 2e-16 ***
## regionwest     -17.2195     4.5306  -3.801 0.000145 ***
## parttimeyes   -283.6389     3.4518 -82.172  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.9 on 28146 degrees of freedom
## Multiple R-squared:  0.3182, Adjusted R-squared:  0.318
## F-statistic:  1642 on 8 and 28146 DF,  p-value: < 2.2e-16
```

The model produces a p value less than our significance level of 0.05, meaning we reject our null hypothesis of homoskedasticity and conclude that heteroskedasticity is present in out model. We must correct for this violation by using Weighted Least Squares.

Question 9

```
backward <- step(reg.mod.wls, direction = "backward")
```

```
## Start:  AIC=143996.6
## wage ~ education + experience + ethnicity + smsa + region + parttime
##
##              Df Sum of Sq     RSS    AIC
## <none>                     4682133 143997
## - region      3     14967 4697100 144080
## - ethnicity   1     33360 4715492 144194
## - smsa        1     46287 4728420 144272
## - experience  1    460005 5142138 146633
## - education   1    609607 5291740 147441
## - parttime    1   1123244 5805377 150049
```

```
summary(backward)
```

```
##
## Call:
## lm(formula = wage ~ education + experience + ethnicity + smsa +
##     region + parttime, weights = w)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -74.451  -3.176   4.323  11.094 133.975
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -76.3620     8.6591  -8.819  < 2e-16 ***
## education       32.2983     0.5335  60.536  < 2e-16 ***
## experience       5.6510     0.1075  52.586  < 2e-16 ***
## ethnicityafam  -69.9405     4.9389 -14.161  < 2e-16 ***
## smsayes         54.4413     3.2637  16.681  < 2e-16 ***
## regionmidwest  -19.8253     4.4162  -4.489 7.18e-06 ***
## regionsouth    -38.8026     4.1715  -9.302  < 2e-16 ***
## regionwest     -17.2195     4.5306  -3.801 0.000145 ***
## parttimeyes   -283.6389     3.4518 -82.172  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.9 on 28146 degrees of freedom
## Multiple R-squared:  0.3182, Adjusted R-squared:  0.318
## F-statistic:  1642 on 8 and 28146 DF,  p-value: < 2.2e-16
```

```
#interaction terms
reg.mod.int <- lm(wage ~ education + experience + ethnicity + smsa + region +
                    parttime + region * ethnicity * smsa +
                    ethnicity * parttime * region + experience * parttime *
                    education, weights = w)
summary(reg.mod.int)
```

```
##
## Call:
## lm(formula = wage ~ education + experience + ethnicity + smsa +
##     region + parttime + region * ethnicity * smsa + ethnicity *
##     parttime * region + experience * parttime * education, weights = w)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -83.784  -3.080   3.945  10.584 134.056
##
## Coefficients:
##                                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              -148.91638   15.77445  -9.440  < 2e-16
## education                                  35.22203    1.04554  33.688  < 2e-16
## experience                                  5.29558    0.47105  11.242  < 2e-16
## ethnicityafam                            -119.65441   56.17029  -2.130 0.033163
## smsayes                                    63.78680    8.31084   7.675 1.71e-14
## regionmidwest                             -33.95149    9.35169  -3.631 0.000283
## regionsouth                               -33.81274    9.27055  -3.647 0.000265
## regionwest                                -19.14632    9.75766  -1.962 0.049751
## parttimeyes                               173.51892   27.65068   6.275 3.54e-10
## ethnicityafam:regionmidwest                21.37219   81.18885   0.263 0.792367
## ethnicityafam:regionsouth                  43.80254   57.38245   0.763 0.445265
## ethnicityafam:regionwest                   74.60013   96.25044   0.775 0.438309
## smsayes:regionmidwest                      17.46134   10.39155   1.680 0.092902
## smsayes:regionsouth                       -23.90783   10.18759  -2.347 0.018944
## smsayes:regionwest                        -10.19961   10.69539  -0.954 0.340271
## ethnicityafam:smsayes                       8.87005   57.08519   0.155 0.876521
## ethnicityafam:parttimeyes                 118.13567   29.14110   4.054 5.05e-05
## regionmidwest:parttimeyes                  24.05896   10.42343   2.308 0.020997
## regionsouth:parttimeyes                    52.92292   10.20215   5.187 2.15e-07
## regionwest:parttimeyes                     52.18115   10.67658   4.887 1.03e-06
## experience:parttimeyes                     -4.73548    0.80548  -5.879 4.17e-09
## education:experience                        0.22592    0.03844   5.877 4.23e-09
## education:parttimeyes                     -31.62189    1.98356 -15.942  < 2e-16
## ethnicityafam:smsayes:regionmidwest       -28.85642   82.40289  -0.350 0.726201
## ethnicityafam:smsayes:regionsouth         -13.87377   58.71619  -0.236 0.813213
## ethnicityafam:smsayes:regionwest          -57.34304   96.85810  -0.592 0.553834
## ethnicityafam:regionmidwest:parttimeyes   -40.81797   37.17640  -1.098 0.272234
## ethnicityafam:regionsouth:parttimeyes     -38.92601   33.06408  -1.177 0.239090
## ethnicityafam:regionwest:parttimeyes      -78.61220   45.98122  -1.710 0.087340
## education:experience:parttimeyes           -0.18318    0.06487  -2.824 0.004750
##
## (Intercept)                              ***
## education                                ***
## experience                               ***
## ethnicityafam                            *
## smsayes                                  ***
## regionmidwest                            ***
## regionsouth                              ***
## regionwest                               *
## parttimeyes                              ***
## ethnicityafam:regionmidwest
## ethnicityafam:regionsouth
```

```
## ethnicityafam:regionwest
## smsayes:regionmidwest                    .
## smsayes:regionsouth                      *
## smsayes:regionwest
## ethnicityafam:smsayes
## ethnicityafam:parttimeyes                ***
## regionmidwest:parttimeyes                *
## regionsouth:parttimeyes                  ***
## regionwest:parttimeyes                   ***
## experience:parttimeyes                   ***
## education:experience                     ***
## education:parttimeyes                    ***
## ethnicityafam:smsayes:regionmidwest
## ethnicityafam:smsayes:regionsouth
## ethnicityafam:smsayes:regionwest
## ethnicityafam:regionmidwest:parttimeyes
## ethnicityafam:regionsouth:parttimeyes
## ethnicityafam:regionwest:parttimeyes     .
## education:experience:parttimeyes         **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.56 on 28125 degrees of freedom
## Multiple R-squared:  0.3534, Adjusted R-squared:  0.3527
## F-statistic: 530.1 on 29 and 28125 DF,  p-value: < 2.2e-16
```

```
# Schwartz/ BIC
mod1 <- lm(wage ~ education + experience + ethnicity + smsa + region + parttime,
           data = CPS1988)
mod2 <- lm(log(wage) ~ education + experience + ethnicity + smsa + region +
             parttime, weights = w)
mod3 <- lm(log(wage) ~ education + experience + ethnicity + smsa + region +
             parttime + region * ethnicity * smsa + ethnicity * parttime *
             education, weights = w)
mod4 <- lm(wage ~ education + experience + ethnicity + smsa + region + parttime
           + region * ethnicity * smsa + ethnicity * parttime * education,
           weights = w)
BIC(mod1)
```

```
## [1] 416612.2
```

```
BIC(mod2)
```

```
## [1] 60153.27
```

```
BIC(mod3)
```

```
## [1] 60061.12
```

```
BIC(mod4)
```

```
## [1] 397397.8
```

When comparing models using the Schwartz criteria, 4 models were looked at – the original model with all explanatory variables, a model with logged wage and all the explanatory variables, a model with logged wage and all the explanatory variables with interaction terms, and a model with un-logged wage and all the explanatory variables with interaction terms. Using Schwartz/BIC, it was found that logging the wage produced a model with a lower BIC than the original, therefore logging the wage did produce the best model after all. The model with the lowest BIC was the model logged wage and interaction terms. This demonstrates that the logged wage interactions between certain variables may account for the skewness and outliers in the data.

Question 10

The data is heteroskedastic, and upon further analysis, all of the explanatory variables contribute to heteroskedasticity to some extent. According to backward selection the model logged wage and interactions between the explanatory variables was the best fit. In order to attempt to estimate a better model, wage was logged and several interactions were introduced. Interaction terms were added between region, ethnicity, and living in a Standard Metropolitan Statistical Area; ethnicity, working part-time, and region; and experience, working part-time and education. Looking at the significance of various variables and interactions, being of African American ethnicity plays a large effect on wage. In the end, selecting a model with interaction terms and logged wage better fits the data, likely due to this observation around ethnicity. We conclude that while education and experience certainly contribute to an individual's wage, they do not represent the sole determinants. It is evident that ethnicity exerts a significant influence on one's wage.