Cassidy Drummond

October 7th, 2022

Group Members: Mario and Zakaria

1. What are some of the important factors that influence a person’s wage? Is there a plausible causal link from X variables to Y and not the reverse?

There are many factors including education, family background (family income, their connections, etc.), privilege, work ethic, etc.. There are some plausible causal links that are not reversible; an example is looking at one’s privilege. At certain companies, someone may have a higher wage because of their race or through family connections. When it’s based on this type of variable, the wage is not high because of the person, it’s higher because of the specific variable.

1. Carefully explain the hypothesis tests of each coefficient and all of the coefficients jointly. For each coefficient, explain the t-stat, p-value, and confidence interval.

First, we made a subset that looked at the full-time workforce between the ages of 25-65, their independent income (opposed to their family or household), and if they were college educated. Following, we ran a regression using the data and looked at the individual’s income compared with their age, race, and citizenship status.

Re age: we found that as one gets older, their income increases. The t-value and p-value demonstrate that these results are significant, and thus we reject the null hypothesis.

Re race: we found that identifying as African American had a negative correlation with wages. The p-value demonstrated that it was very statistically significant and the t-value was -11.055 which is significant but less than the t-value of age.

Re citizenship: we found that there was a somewhat positive correlation, however, the t-value and p-value state that it is not statistically significant.

1. Explain your results, giving details about the estimation and providing any relevant graphics. What are the changes from what you’d previously found (with k-nn or averages) and why might this be so? How do changes in specification (e.g. logs) change the estimated coefficients? What are some relevant predicted values? Do those seem sensible? What additional information would be useful?

To be honest, I was not able to figure out how to use the predicted model. When I tried running a summary, I just got various tables with the mean and was unable to see the hypothesis tests. Looking at your examples, research and some of my peer’s work, by using log, it minimizes how wide of a variation there is. In theory, the summary statistics should be much closer together and thus the data should be scaled smaller. The coefficient and p-value are useful values in demonstrating how varying and significant the outputs are. For this data, I think it would be helpful to have a follow up survey conducted as well as more in depth questions regarding wages such as, how much they pay in taxes, are they supporting anyone (not including children), etc..

Output:

> model\_temp1 <- lm(INCWAGE ~ AGE + AfAm + Asian + Amindian + race\_oth + Hispanic + educ\_hs + educ\_somecoll + educ\_college + educ\_advdeg + CITIZEN, data=dat\_use)

> summary(model\_temp1)

Call:

lm(formula = INCWAGE ~ AGE + AfAm + Asian + Amindian + race\_oth +

Hispanic + educ\_hs + educ\_somecoll + educ\_college + educ\_advdeg +

CITIZEN, data = dat\_use)

Residuals:

Min 1Q Median 3Q Max

-156692 -52419 -23688 12937 591430

Coefficients: (3 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 63119.21 3250.37 19.419 < 2e-16 \*\*\*

AGE 1344.43 66.61 20.184 < 2e-16 \*\*\*

AfAm -31425.82 2842.58 -11.055 < 2e-16 \*\*\*

Asian 7633.18 3074.69 2.483 0.013051 \*

Amindian -23835.67 17081.55 -1.395 0.162908

race\_oth -17490.54 3044.82 -5.744 9.35e-09 \*\*\*

Hispanic -11706.49 3032.64 -3.860 0.000114 \*\*\*

educ\_hs NA NA NA NA

educ\_somecoll NA NA NA NA

educ\_college -28041.87 1500.43 -18.689 < 2e-16 \*\*\*

educ\_advdeg NA NA NA NA

CITIZEN 1309.88 885.50 1.479 0.139089

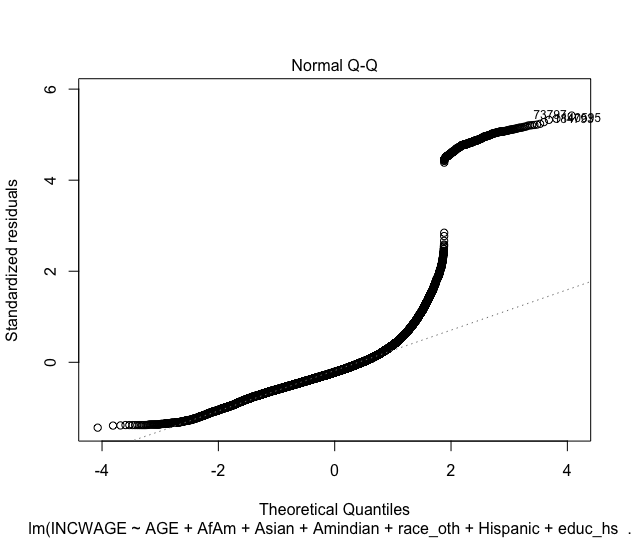
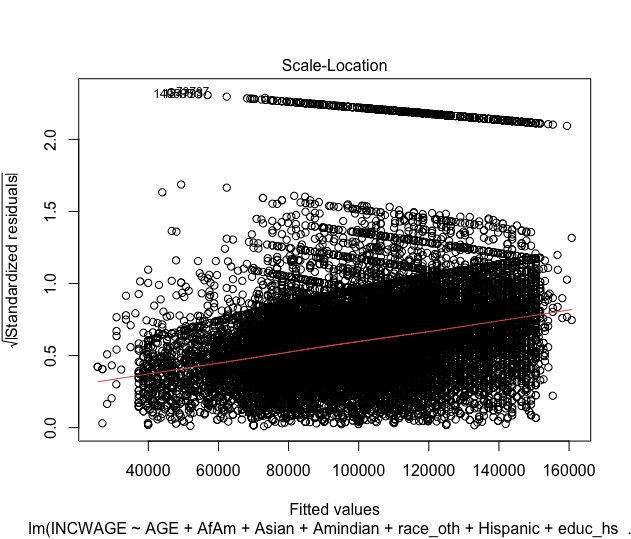
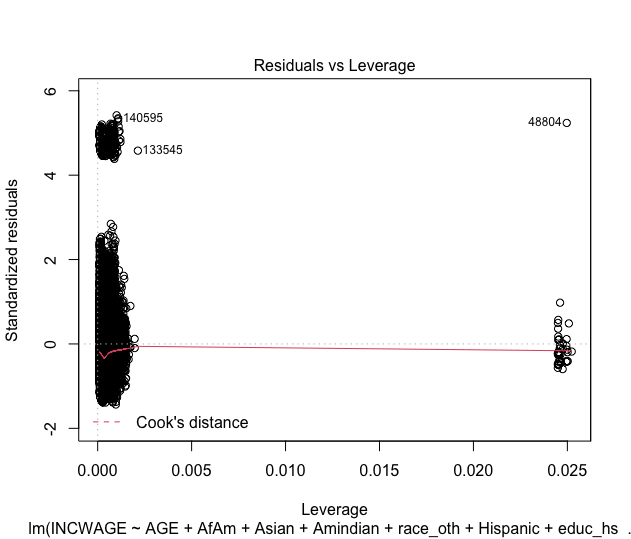
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 109200 on 21714 degrees of freedom

Multiple R-squared: 0.04697, Adjusted R-squared: 0.04661

F-statistic: 133.8 on 8 and 21714 DF, p-value: < 2.2e-16



to\_be\_predicted2 <- data.frame(AGE = 25:65, female = 1, AfAm = 0, Asian = 0, Amindian = 1, race\_oth = 1, Hispanic = 1, educ\_hs = 0, educ\_somecoll = 0, educ\_college = 1, educ\_advdeg = 0)

> model\_temp2 <- lm(INCWAGE ~ AGE + AfAm + Asian + Amindian + race\_oth + Hispanic + educ\_hs + educ\_somecoll + educ\_college + educ\_advdeg, data=dat\_use)

> to\_be\_predicted2$yhat <- predict(model\_temp2, newdata = to\_be\_predicted2)

Warning message:

In predict.lm(model\_temp2, newdata = to\_be\_predicted2) :

prediction from a rank-deficient fit may be misleading

> plot(to\_be\_predicted2)

> summary(to\_be\_predicted2$yhat)

Min. 1st Qu. Median Mean 3rd Qu. Max.

17215 30718 44220 44220 57722 71224

