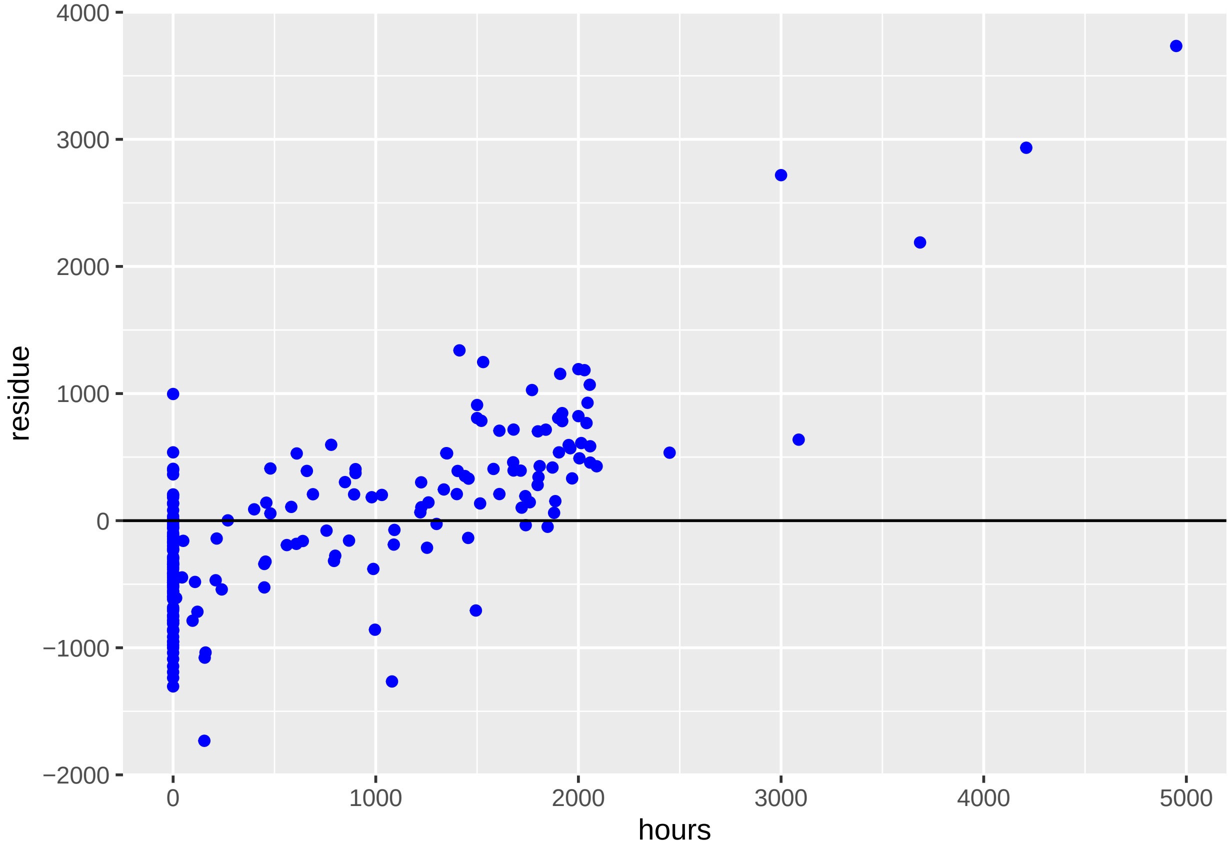
Part A

The dataset Hours is a subset of the 1976 Panel Study of Income Dynamics (PSID) that contains 19 variables. In this project we are to build a model that explains the number of hours worked by women in 1975. The dependent variable here is therefore “hours”. Models with dependent variables that contain many 0’s are not usually estimated by OLS, but we ignore it for this project.

Weighted Least Squares (WLS) can be used when the data is heteroscedastic (but uncorrelated).



Here since the errors are spread equally around the regression line we can see that it is not heterscedastic and so OLS is fine here.

Before I can build our model, I need to decide which variables to include. I want the model to explain the number of hours worked by women in 1975. So I would need to consider whether the women had kids. If she had young kids, clearly she would probably have had to stay home to take care of them so would not have been able to work or work less hours. So I need to include the first variable “youngkids”. Note that the second variable also talks about the number of kids she had but the only difference is that the “youngkids” is the number of children less than 6 years old in household whereas “oldkids” is the number of children between ages 6 and 18 in household. Having a kid younger than 6 years old would definitely require the mother to stay home more but so would a 8 year old using same logic. So including these two variable seem redundnt and instead I will create a dummy variable of our own that is 1 if she had kids younger than 18 and 0 otherwise. We will call this variable “haveKids”. I added both youngkids and “havekids” since “youngkids” is giving the number of under 6 year old kids whereas “havekids” is a dummby variable for is she had kids at all or not.

We will have 2 groups - Woman with kids Vs Woman with no kids

Age of the woman clearly has an effect on the hours she can work. Certain jobs could be labour intensive, so the younger the woman is, the more huours she can put in. Moreover, An older woman might be more likely to be working full-time whereas an young woman might be doing a part-time job. So I decided to include this in my model.

If the woman is highly educated, she will clearly more likely to have a stable job, whereas, if the woman was not that well-educated, she would probably not be working at a stable job with “good” hours. So I thought of including “eductaion” in my model since education seem to have an effect on the number of hours worked. However, there is another varable called “college” which says “yes” is she attended college and “no” otherwise. Both are good variables to indicate the likely of the woman getting a job. However, we are more concerned with how many hours she worked, not trying to find if she had a job. So I excluded them both.

“experience” is the actual years of wife’s previous labor market experience. Higher expeirence means she is more likely to have a job and probably a good one with stable hours. So I decided to include this in my model.

“unemp” is the unemployment rate in county of residence, in percentage points. So this would gives us an idea of the demand and supply of the labor force. “city” tells us if the individual lives in a large city or not. Both “city” and “unemp” tell us if the wife had a job or not -“emp” more so than “city”. We are more concerned with how many hours she works once she already had the job, so theses two variables were not included.

“fincome” is the family income, in 1975 dollars. This would include the amount earned by both the husband and the wife. A higher “fincome” could be due to higher wage earned by the husband or the wife. “fincome” is clealry correlated with “wage”,“hhours”,“hwage”. Including all these will create some multicollinearity. “fincome” is correlated with number of hours worked by the woman too, since, the more she worked the more the “fincome” would be. We need each variable to be independant so we cannot include this in our model since we have “hwage”,“hhours” and “wage”. We must include “hhours” since clearly if the husband is woking more hours, the wife needs to stay home to take care of the family, specially in 1975. If the family didnot have kids, the wife could still decide to work less or no hours if the husband is already working a lot hours or has high wage and thus earning enough for the family. Moreover, a small “hhour” and “hwage” value might imply that wife needs to work more hours to support the family. So including all the terms as well as fincome seems redundant. I decided to skip fincome and include "hwage" and "hhours".

fincome seemed to have some measurement error.

As for "wage", it doesnot really teel us much about the hours she would work since if the wife's wage was high, its not like she would necessarily work less hours.

"hage","meducation" and "feduction" donot really effect the woman's ability to work and so I excluded them.

"tax" is the Marginal tax rate facing the wife. Since it is the marginal rate, it doesnot really help us with the hours she worked. So I excluded that as well.

To summarise, I included "havekids","age","youngkids","experience","hwage" and "hhours" in my model.

Once I have selected my variables, I will try 3 different models.

If the woman had kids younger than 18 at home, she might decide to stay home to take care of them or work less hours.This is also dependant on whether the husband is working long hours. If "hhours" is high the wife with kids might have to work less or no hours since you need atleast one parent to be there for the kids. So I want to introduce the interaction between "havekids" and "hhours" since I believe there is a relationship between them.

I want to try expereinceˆ2. Clearly if the woman had exprience equal to sample average she will probably be working a stable job with good hours.

if the "hwage" is high the wife might decide to stay home since the husband is already providing enough money to support the family. This further increased if the husband is working more hours. So I want to see the effect of adding an interaction for hhour and hwage.

I added the interaction $havekids \times hhours$ since having kids and the husband working long hours will have an added effect on the wife's ability to work.

I tried $log(hours)$ to make sure the residuals are not too skewed. I have both $havekids \times hhours^2$ and hhours, since there is likely to things how the husband's working hours effects the wife's: 1. Having kids and housband working long hours will result in wife deciding to work less hours to stay home and take care of the kids, 2. Husband woring long hours and so earning enough to support the family.

$$\widehat{hours} = \underset{(594.58)^{***}}{2939.6} - \underset{(292.87)^{***}}{1041.9}\ youngkids + \underset{(152.16)}{228.12}\ I(youngkids^2) + \underset{(0.2127)}{0.008155}\ I(havekids * hhours)$$
$$- \underset{(474.49)}{83.993}\ havekids - \underset{(9.451)^{***}}{47.45}\ age + \underset{(7.3114)^{***}}{40.049}\ experience$$
$$- \underset{(0.16684)}{0.093896}\ hhours - \underset{(13.482)}{21}\ hwage$$
$$n = 194, \quad R^2 = 0.27741, SSR = 118907739, \bar{R}^2 = 0.24616$$

*pv < 0.1; **pv < 0.05; ***pv < 0.01

$$\widehat{log(hours)} = \underset{(2.4807)^{***}}{12.907} - \underset{(0.23208)^{***}}{0.98748}\ youngkids + \underset{(0.00031382)}{0.00028193}\ I(havekids * hhours) + \underset{(0.012199)^{***}}{0.032683}\ experience$$
$$- \underset{(0.7097)}{0.67896}\ havekids - \underset{(0.61343)^{**}}{1.4902}\ log(age) - \underset{(0.0002301)}{0.0002149}\ hhours$$
$$- \underset{(1.3084e-06)}{1.2126e-06}\ I(hhours * hwage^2) - \underset{(0.29406)}{0.033125}\ log(hwage)$$
$$n = 110, \quad R^2 = 0.24651, SSR = 78.48553, \bar{R}^2 = 0.18683$$

*pv < 0.1; **pv < 0.05; ***pv < 0.01

$$\widehat{hours} = \underset{(595.65)^{***}}{2890.1} - \underset{(132.39)^{***}}{649.94}\ youngkids + \underset{(0.21235)}{0.039836}\ I(havekids * hhours) - \underset{(9.4819)^{***}}{47.283}\ age$$
$$+ \underset{(7.3295)^{***}}{40.507}\ experience - \underset{(471.91)}{177.97}\ havekids - \underset{(0.16732)}{0.086035}\ hhours$$
$$- \underset{(13.428)}{18.547}\ hwage$$
$$n = 194, \quad R^2 = 0.26863, SSR = 120352363, \bar{R}^2 = 0.2411$$

*pv < 0.1; **pv < 0.05; ***pv < 0.01

We first estimate the models and test the homoscedasticity using the short White test:

```
##
##  studentized Breusch-Pagan test
##
## data:  model_1
## BP = 0.99586, df = 2, p-value = 0.6078
```

Since p-value is greater than 0.05, we do not reject the homoscedasticity assumption at 5%, so we do not need to use robust tests.

```
##
```

```
##  studentized Breusch-Pagan test
##
## data:  model_2
## BP = 6.3626, df = 2, p-value = 0.04153
```

Since p-value is less than 0.05, we reject the homoscedasticity assumption at 5%, so we need to use robust tests.

```
##
##  studentized Breusch-Pagan test
##
## data:  model_3
## BP = 0.57246, df = 2, p-value = 0.7511
```

Since p-value is greater than 0.05, we do not reject the homoscedasticity assumption at 5%, so we do not need to use robust tests.

Model_1 and Model_3 are nested models.So we can just use indirect-t test to see if the extra "youngkids^2" term is significant or not in model 1.

```
##
## z test of coefficients:
##
##                         Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)           2.9396e+03  7.8389e+02  3.7500 0.0001768 ***
## youngkids            -1.0419e+03  1.8884e+02 -5.5175 3.438e-08 ***
## I(youngkids^2)        2.2812e+02  7.7847e+01  2.9304 0.0033851 **
## I(havekids * hhours)  8.1550e-03  2.2954e-01  0.0355 0.9716598
## havekids             -8.3993e+01  5.4374e+02 -0.1545 0.8772364
## age                  -4.7450e+01  1.0671e+01 -4.4465 8.729e-06 ***
## experience            4.0049e+01  8.3556e+00  4.7931 1.642e-06 ***
## hhours               -9.3896e-02  2.0464e-01 -0.4588 0.6463586
## hwage                -2.1000e+01  1.1016e+01 -1.9063 0.0566117 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value is less than 0.05 we reject the Null hypothesis that the coefficient of "youngkids^2" is 0. So we conclude that "youngkids^2" is significant. So we choose Model 1 instead.

Now we need to choose between the two non-nested models: Model_1 and Model_2. We need to use robust J-test for that since model 2 is heteroscedastic:

```
## J test
##
## Model 1: hours ~ youngkids + I(youngkids^2) + I(havekids * hhours) + havekids +
##     age + experience + hhours + hwage
## Model 2: log(hours) ~ youngkids + I(havekids * hhours) + experience +
##     havekids + log(age) + hhours + I(hhours * hwage^2) + log(hwage)
##                 Estimate Std. Error z value Pr(>|z|)
## M1 + fitted(M2)   0.6909    0.91397  0.7559   0.4497
## M2 + fitted(M1)   1.7307    1.77742  0.9737   0.3302
```

Both have p-value greater than 0.05 so we can see that both models cannot be rejected. I decided to go with mdoel 1 because choosing model 2 will result in us losing a lot of data (we lose all the data corresponding to hours less than or equal to 0).

Now I will test the null hypothesis that the model is correctly specified at 5% using the RESET test.

Since it is homoscedastic, we can just use the non-robust test.

```
##
##  RESET test
##
## data:  model_1
## RESET = 1.7242, df1 = 2, df2 = 183, p-value = 0.1812
```

Since the p-value is greater than 5% we cannot reject the hypothesis that the model is correctly specified.

The summary for Model 1 is as follows:

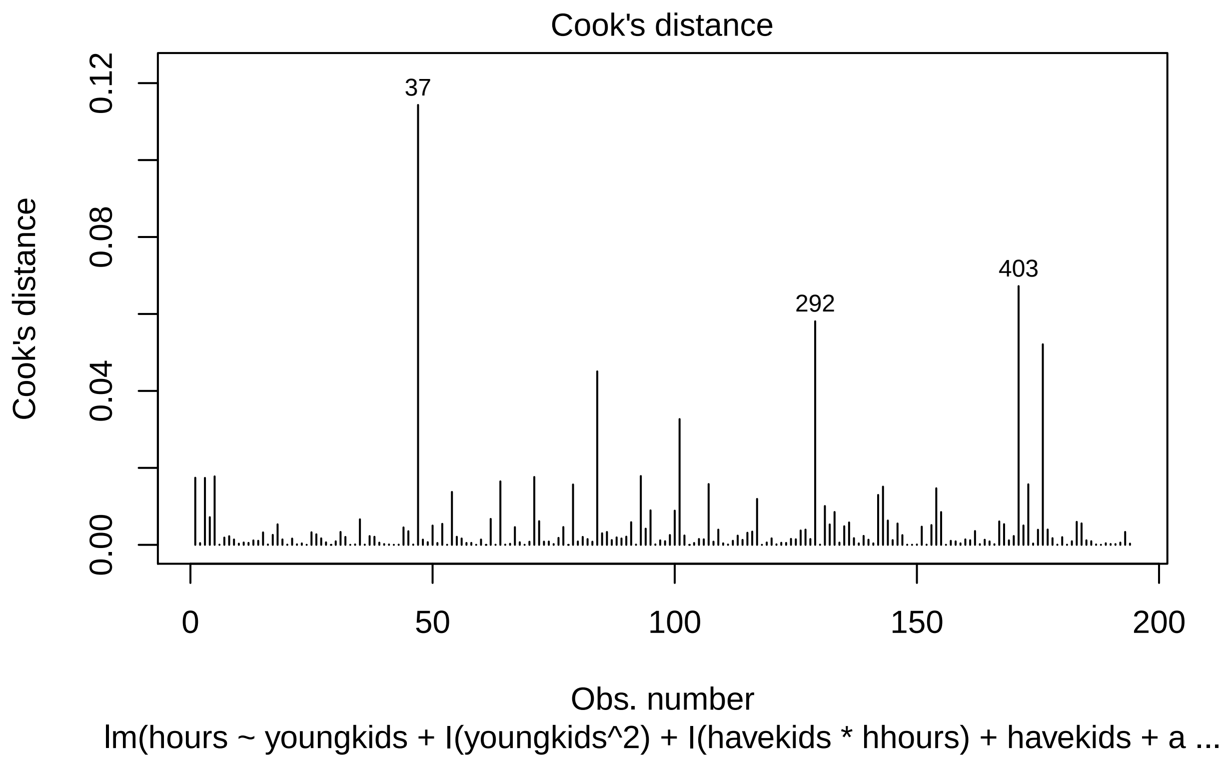
```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                 hours
## ------------------------------------------------
```

```
## youngkids                    -1,041.921***
##                               (292.869)
##
## I(youngkids2)                  228.124
##                               (152.164)
##
## I(havekids * hhours)            0.008
##                                (0.213)
##
## havekids                       -83.993
##                               (474.494)
##
## age                           -47.450***
##                                (9.451)
##
## experience                     40.049***
##                                (7.311)
##
## hhours                         -0.094
##                                (0.167)
##
## hwage                          -21.000
##                               (13.482)
##
## Constant                    2,939.597***
##                               (594.577)
##
## ------------------------------------------------
## Observations                     194
## R2                              0.277
## Adjusted R2                     0.246
## Residual Std. Error      801.713 (df = 185)
## F Statistic            8.878*** (df = 8; 185)
```

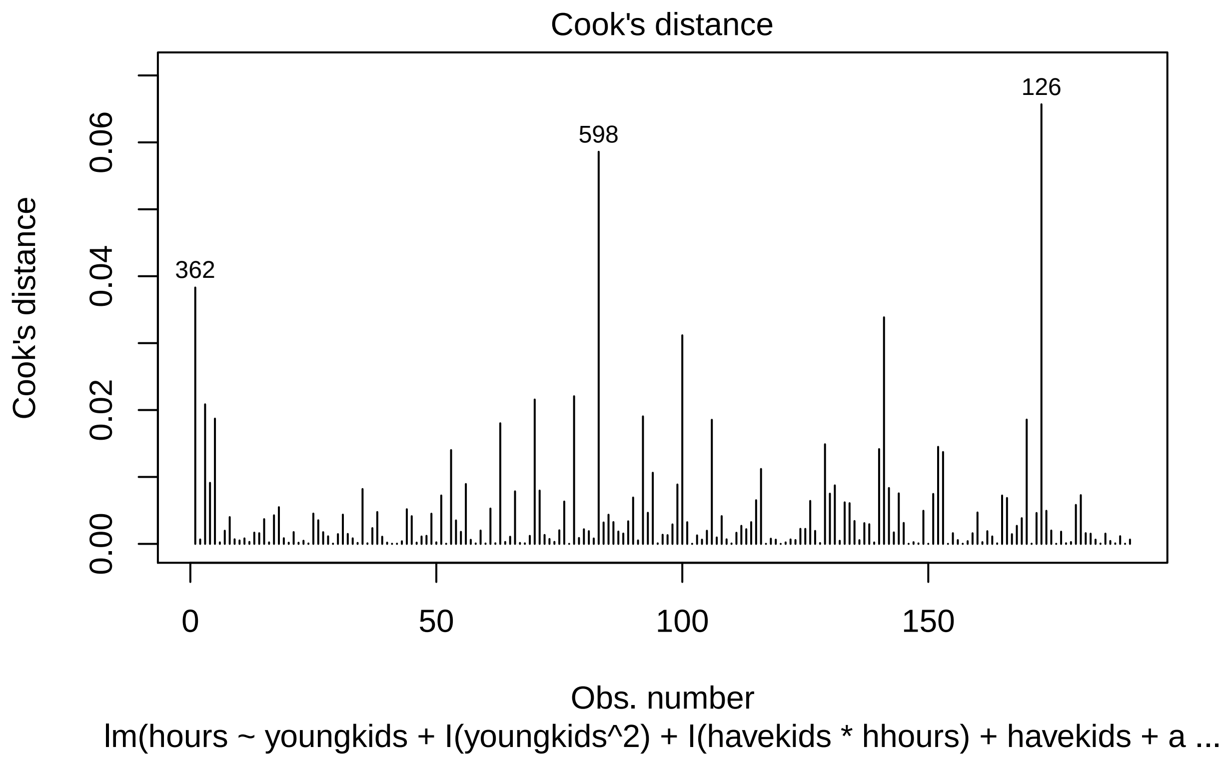
```
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Removing Outliers

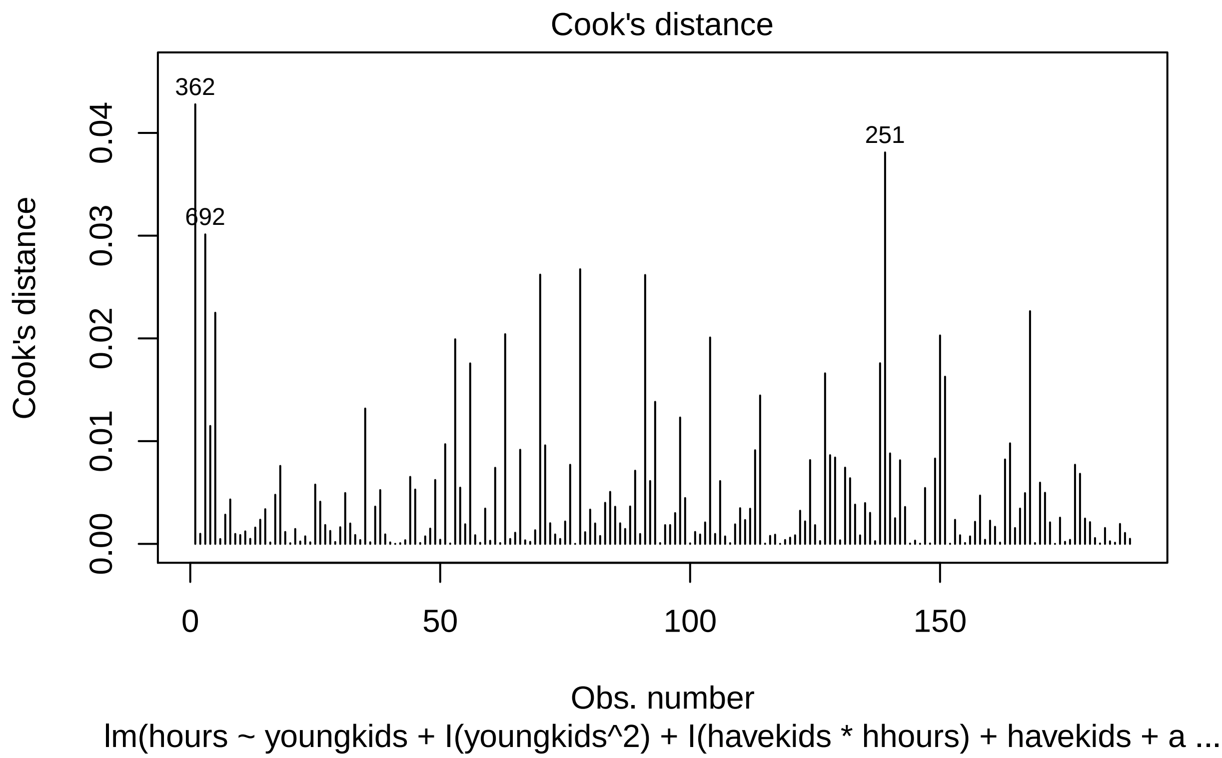
The Cook's distance for detecting outliers is plotted by the plot method for lm objects.



These are the names of the rows (37,403,126).So we can remove observations 126, 403 and 37.



There seem to still be some outliers. Removing those too:

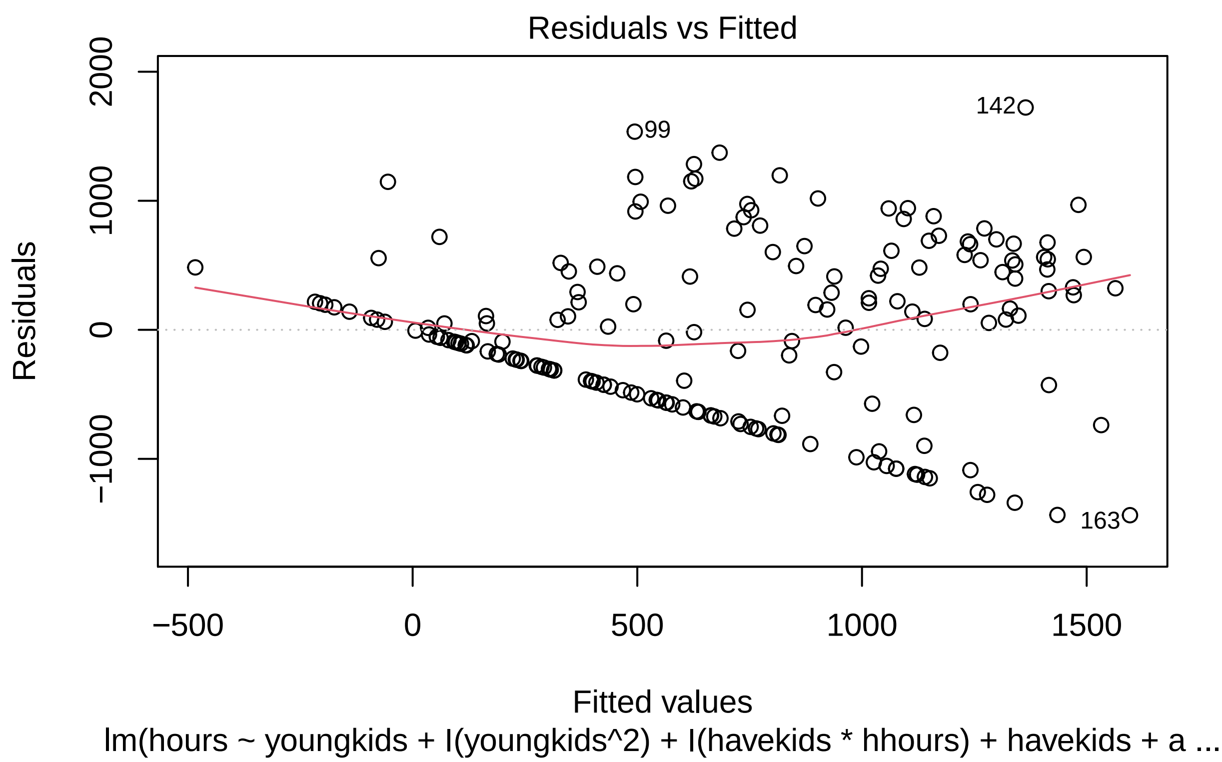


Now, the Cook's distance are closer to being equal. We can compare the results to see what is the impact of dropping these three observations:

```
##
## ===================================================================
##                                       Dependent variable:
##                             ---------------------------------------------
##                                              hours
##                                     (1)                    (2)
## -------------------------------------------------------------------
## youngkids                    -1,041.921***           -942.865***
##                                (292.869)              (242.737)
##
## I(youngkids2)                   228.124                208.111*
##                                (152.164)              (125.783)
##
```

```
## I(havekids * hhours)          0.008                  -0.150
##                              (0.213)                 (0.179)
##
## havekids                     -83.993                 380.870
##                             (474.494)               (402.467)
##
## age                         -47.450***              -40.211***
##                              (9.451)                 (8.260)
##
## experience                   40.049***               45.245***
##                              (7.311)                 (6.323)
##
## hhours                       -0.094                   0.051
##                              (0.167)                 (0.143)
##
## hwage                        -21.000                 -16.769
##                              (13.482)                (11.191)
##
## Constant                   2,939.597***            2,080.710***
##                             (594.577)               (526.131)
##
## -------------------------------------------------------------------
## Observations                   194                     188
## R2                            0.277                   0.345
## Adjusted R2                   0.246                   0.315
## Residual Std. Error    801.713 (df = 185)     662.532 (df = 179)
## F Statistic          8.878*** (df = 8; 185) 11.773*** (df = 8; 179)
## ===================================================================
## Note:                                   *p<0.1; **p<0.05; ***p<0.01
```

Interpretation and Analysis



Since model 1 is homoscedastic we can just use summary() to check the coefficients:

```
##
## Call:
## lm(formula = hours ~ youngkids + I(youngkids^2) + I(havekids *
##     hhours) + havekids + age + experience + hhours + hwage, data = PartA,
##     subset = !(rownames(PartA) %in% c(37, 126, 403, 598, 734,
##         292)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1436.42  -447.30     4.65   482.75  1722.86
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)          2080.7100   526.1310   3.955 0.000110 ***
## youngkids            -942.8650   242.7372  -3.884 0.000144 ***
## I(youngkids^2)        208.1110   125.7830   1.655 0.099774 .
## I(havekids * hhours)   -0.1504     0.1794  -0.839 0.402720
## havekids              380.8701   402.4665   0.946 0.345251
## age                   -40.2112     8.2603  -4.868 2.47e-06 ***
## experience             45.2454     6.3235   7.155 2.07e-11 ***
## hhours                  0.0510     0.1426   0.358 0.721043
## hwage                 -16.7687    11.1911  -1.498 0.135793
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 662.5 on 179 degrees of freedom
## Multiple R-squared:  0.3448, Adjusted R-squared:  0.3155
## F-statistic: 11.77 on 8 and 179 DF,  p-value: 2.059e-13
```

Interpreting the coefficients

Intercept: Here the intercept has no meaning since age would have be 0 which has no meaning.
youngkids and $youngkids^2$ need to be explained using partial effect which is done in the next section. age: women who are 1 more year older will have a difference in hours worked of 40.2112 less when holding all the other variables constant. experience: women who have 1 more year of experience will have a difference in hours worked of 45.2454 more when holding all the other variables constant. hwage: women whose husbands earn 1 more dollar in wage will have a difference in hours worked of 16.7687 less when holding all the other variables constant.

For hhours and $havekids \times hhours$ If the woman had no kids the difference in hours worked will be 0.0510 more when holding all the other variables constant. But, if the woman had kids the difference in hours worked will be 0.0994 less when holding all the other variables constant.

Significance of the coefficients

Firstly, p-value of I(youngkids^2) is greater than 0.05, so we cannot reject Null Hypothesis. So we can say that the relationship is linear.

youngkids,age and experience are the only statisticall significant features in this model since they have p-value less than 0.05 (so we can reject Null Hypothesis.)

Confidence Interval of the coefficients

```
##                             2.5 %       97.5 %
## (Intercept)           1042.4929436 3118.9270931
## youngkids            -1421.8595694 -463.8705005
## I(youngkids^2)         -40.0973524  456.3192898
## I(havekids * hhours)    -0.5043688    0.2034895
## havekids              -413.3193116 1175.0595644
## age                    -56.5112827  -23.9110902
## experience              32.7672287   57.7235865
## hhours                  -0.2303860    0.3323772
## hwage                  -38.8521533    5.3147953
```

These intervals are where we will find its corresponding variables 95% of the time.

Average partial effects:

average partial effect of youngkids on hours is $\beta_1 + 2\beta_2\overline{youngkids}$

```
## youngkids
## -857.0461
```

```
## youngkids youngkids
## -1247.563  -466.529
```

The average hours worked by the woman with a kid yonger than 6 and the same level of education, hhours,age and hwage is 857.0461 hours less, when their initial number of young kids is equal to the sample average.

The true APE is between -1247.563 and -466.529 95% of the time.

Does having children less than 6 have the same effect as having children between ages 6 and 18?

The model is:

$$hours = \beta_0 + \beta_1 youngkids + \beta_2 youngkids^2 + \beta_3 (havekids \times hhours) + \beta_4 age + \beta_5 experience + \beta_6 hhours + \beta_7 hwage$$

if the woman has children less than 6, then we can use this model for inference however, if the woman doe not have children less than 6 but instead only has children between ages 6 and 18, then youngkids = 0 but havekids is still 1. So our new model would be:

$$hours = \beta_0 + \beta_3(havekids \times hhours) + \beta_4 age + \beta_5 experience + \beta_6 hhours + \beta_7 hwage$$

So using

$$hours = \beta_0 + \beta_1 youngkids + \beta_2 youngkids^2 + \beta_3(havekids \times hhours) + \beta_4 age + \beta_5 experience + \beta_6 hhours + \beta_7 hwage$$

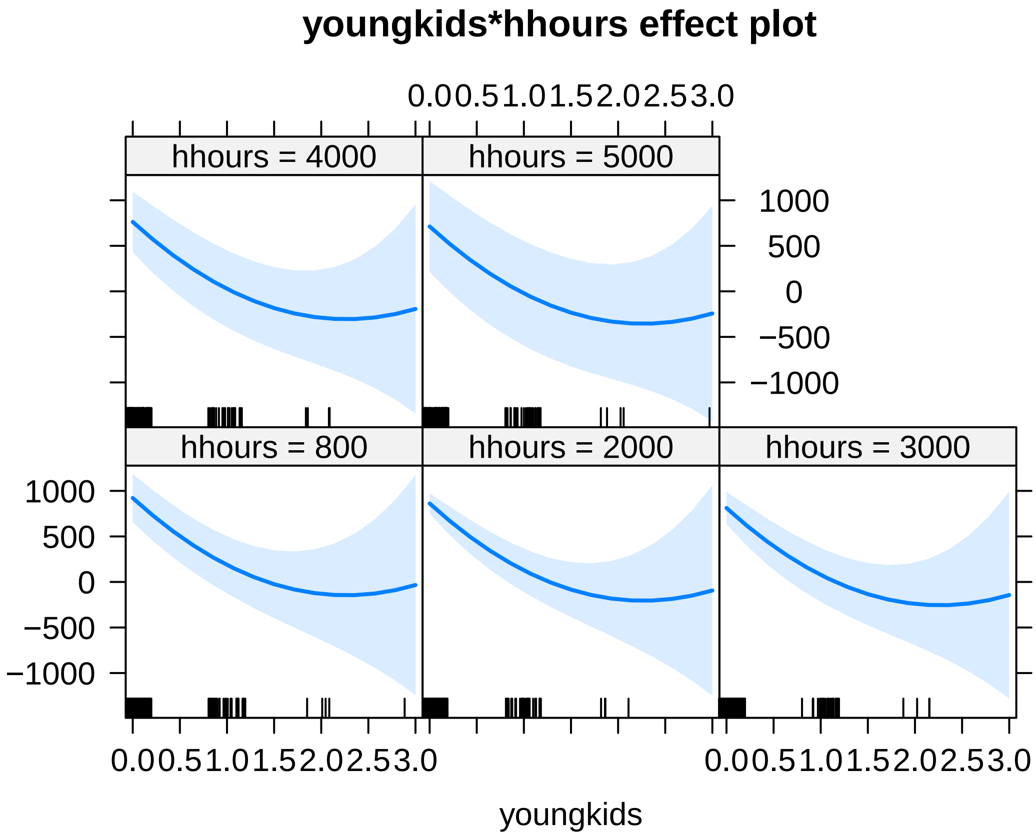
I want to see if $\beta_1 = \beta_2 = 0$

Using F-test (non-robust since it is homoscedastic):

```
## Linear hypothesis test
##
## Hypothesis:
## youngkids = 0
## I(youngkids^2) = 0
##
## Model 1: restricted model
## Model 2: hours ~ youngkids + I(youngkids^2) + I(havekids * hhours) + havekids +
##     age + experience + hhours + hwage
##
##   Res.Df      RSS Df Sum of Sq      F   Pr(>F)
## 1    181 92182607
## 2    179 78571813  2  13610794 15.504 6.17e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

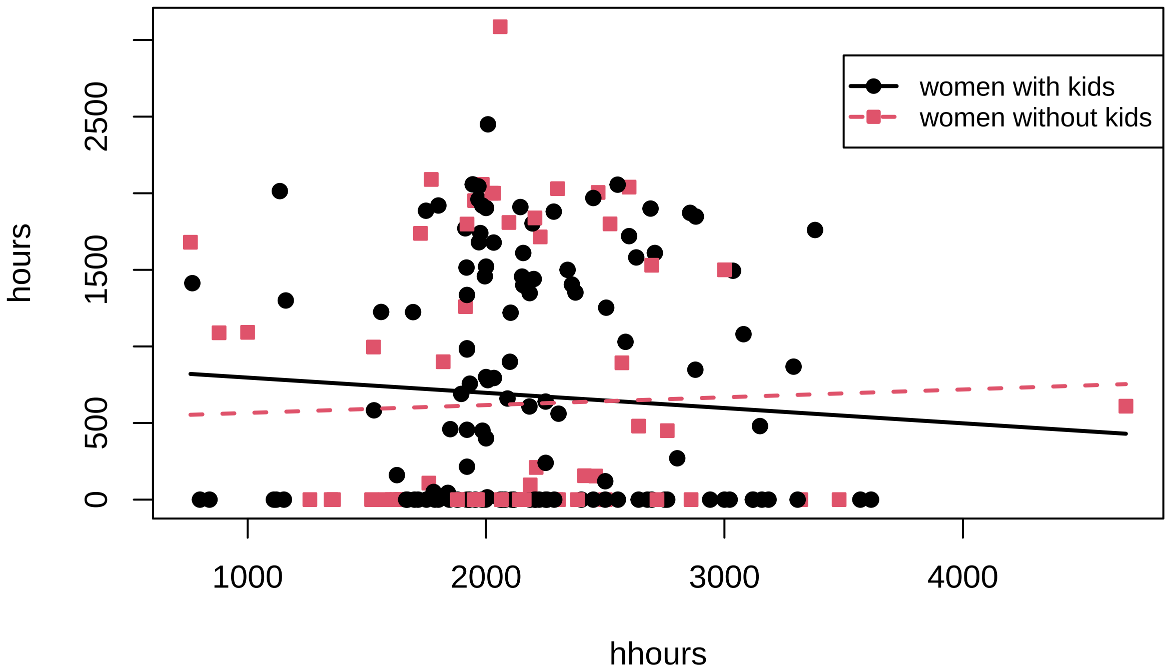
Since p-value is less than 0.05 we can reject the Null hypothesis and conclude that having children less than 6 does not have the same effect as having children between ages 6 and 18.

effect of youngkids and hhours



hhours and youngkids seem to have an inverse non linear relationship

WOMEN WITHOUT KIDS VS WOMEN WITH KIDS



Looking at the above graph, we can see that initially when the husband is working less hours (approximately less than 2500 hours), the wife with kids is working more hours but then we see after around 2500 hhours, the women with kids have their wroking hours reduce, implying that as the husband works more hours, the wife decides to work less hours if she had kids. On the other hand, we see an oopsite trend for women without kids,where as hhours increase so does the wife's working hours.

$$
\begin{aligned}
\widehat{hours} = & \underset{(526.1)^{***}}{2081} - \underset{(242.7)^{***}}{942.9}\ youngkids + \underset{(125.8)^{*}}{208.1}\ I(youngkids^2) - \underset{(0.1794)}{0.1504}\ I(havekids * hhours) \\
& + \underset{(402.5)}{380.9}\ havekids - \underset{(8.26)^{***}}{40.21}\ age + \underset{(6.323)^{***}}{45.25}\ experience \\
& + \underset{(0.1426)}{0.051}\ hhours - \underset{(11.19)}{16.77}\ hwage \\
& n = 188, \quad R^2 = 0.34477, SSR = 78571813
\end{aligned}
$$

*pv < 0.1; **pv < 0.05; ***pv < 0.01

Part B

4 different questions related to discrimination are:

1. Is discrimination different for people from Boston and Chicago?
2. Is discrimination different for male and female individuals with prior military experience or not?
3. Is the effect of experience on the probability of being called back the same for people with African-American sounding name who provided email address and Caucasian sounding name?
4. Is discrimination different for people who mention some volunteering experience VS those that didnot for people applying to EOE?

Now that I have my 4 questions, I can formulate my models for each question:

The initial model:

$$call = \beta_0 + \beta_1 ethnicity$$

this model only detects discrimination on average.

I will interpret each coefficient in the next section once i have formulated all 4 models for each question. I will also add experience and $exprience^2$ since the question asked me. This will explained in the next section as well.

Model 1

$$call = \beta_0 + \beta_1 ethnicity + \beta_2 city + \beta_3 (city \times ethnicity) + \beta_4 experience + \beta_5 exprience^2$$

For question 1, we want to see if discrimination is different for people from Boston and Chicago. So we need the interaction between city and ethnicity. Here I wanted to check if there is discrimination for people from Chicago

Model 2

$$\begin{aligned} call = \beta_0 + \beta_1 (ethnicity \times military) + \beta_2 (gender \times military) + \beta_3 (ethnicity \times \\ military \times gender) + \beta_4 (ethnicity \times gender) + \beta_5 ethnicity + \\ \beta_6 gender + \beta_7 military + \beta_8 experience + \beta_9 exprience^2 \end{aligned}$$

For question 2, I want to see if discrimination is different for male and female that have prior militay experience or not. Since I am checking for discrimination, I will need to have "ethnicity" in my model. Other than that, I need to interact "ethnicity" with "military" and "gender", so that if the person is African-American,is male and has military expeirence, the coefficient of that interaction will give me the probability of being called. I will also include interaction between “gender” and “military”. This will to differentiate for Caucasian male with military experience. I need the term “military” too so that if “gender” is 0 (female), “military” is 0 (no expereince) and ethnicity is Caucasian, the coefficienct of that term will give me the probability of being called for female Caucasian sounding name with military experience. Lastly, I need an interaction between ethnicity and military for African-American sounding name for females with no military background.

Model 3

The model is given in next section...

For question 3, I will need to add interaction between exprience and email since I am trying to check the specific relationship for the effect of experience on the probability of being called back the same for people who provided email address and who didnot.

Model 4

$$call = \beta_0 + \beta_1(ethnicity \times equal) + \beta_2(volunteer \times equal) + \beta_3(ethnicity \times volunteer) + \\ \beta_4 volunteer + \beta_5 ethnicity + \beta_6 equal + \beta_7 exprience + \beta_8 expereince^2$$

For question 4, I want to see if discrimination is different for male and female that have prior volunteer experience applying to an EOE. Since I am checking for discrimination, I will need to have “ethnicity” in my model. Other than that, I need to interact “ethnicity” with “equal” and “volunteer”, so that if the person is African-American has volunteer expeirence and applying to an EOE, the coefficient of that interaction will give me the probability of being called back. I will also include interaction between “volunteer” and “equal”. This will to differentiate for Caucasian vlunteer with volunteer experience or not.

Estimating the models defined above:

Model 1

$$\widehat{call} = \underset{(0.03945)}{0.05029} - \underset{(0.03026)^{**}}{0.06128}\ ethnicity - \underset{(0.03024)^{**}}{0.06545}\ city + \underset{(0.04079)^{*}}{0.0749}\ I(city * ethnicity)$$

$$+ \underset{(0.006684)^{*}}{0.01261}\ experience - \underset{(0.0002613)}{0.0003805}\ I(experience^2)$$

$$n = 623, \quad R^2 = 0.01854, SSR = 39.29009$$

*pv < 0.1; **pv < 0.05; ***pv < 0.01

Model 2

$$\widehat{call} = \underset{(0.03821)}{0.01521} - \underset{(0.02402)}{0.009755}\, ethnicity - \underset{(0.0391)}{0.02572}\, gender + \underset{(0.08251)}{0.03824}\, military$$

$$+ \underset{(0.007123)^{*}}{0.01207}\, experience - \underset{(0.0002737)}{0.0003429}\, I(experience^2) - \underset{(0.09844)}{0.08906}\, ethnicity : military$$

$$+ \underset{(0.1267)}{0.05662}\, gender : military + \underset{(0.05316)}{0.0001332}\, ethnicity : gender - \underset{(0.1542)}{0.01977}\, ethnicity : gender : military$$

$$n = 623, \quad R^2 = 0.01531, SSR = 39.41924$$

*pv < 0.1; **pv < 0.05; ***pv < 0.01

Model 3

$$\widehat{call} = \underset{(0.06482)}{0.05812} - \underset{(0.1001)}{0.09971}\, email + \underset{(0.09302)}{0.01803}\, ethnicity + \underset{(0.01315)}{0.002594}\, experience$$

$$- \underset{(0.0004985)}{0.0002367}\, I(experience^2) + \underset{(0.1348)}{0.01547}\, email : ethnicity + \underset{(0.02075)}{0.01946}\, email : experience$$

$$- \underset{(0.0008361)}{8.472e-05}\, email : I(experience^2) - \underset{(0.0186)}{0.005319}\, ethnicity : experience + \underset{(0.0006871)}{0.0002172}\, ethnicity : I(experience^2)$$

$$- \underset{(0.02782)}{0.01059}\, email : ethnicity : experience + \underset{(0.001117)}{0.0004156}\, email : ethnicity : I(experience^2)$$

$$n = 623, \quad R^2 = 0.04477, SSR = 38.23968$$

*pv < 0.1; **pv < 0.05; ***pv < 0.01

Model 4

$$\widehat{call} = \underset{(0.03831)}{0.02573} - \underset{(0.03469)}{0.004256}\, volunteer - \underset{(0.02932)}{0.03836}\, ethnicity - \underset{(0.03568)^{*}}{0.05934}\, equal$$

$$+ \underset{(0.04429)}{0.05216}\, I(ethnicity * equal) + \underset{(0.0458)}{0.04315}\, I(volunteer * equal) + \underset{(0.04214)}{0.006724}\, I(ethnicity * volunteer)$$

$$+ \underset{(0.006712)^{*}}{0.01291}\, experience - \underset{(0.0002618)}{0.0003715}\, I(experience^2)$$

$$n = 623, \quad R^2 = 0.01598, SSR = 39.39247$$

*pv < 0.1; **pv < 0.05; ***pv < 0.01

Now I will start answering the questions:

Question 1

Is discrimination different for people from Boston and Chicago?

I will use this model to answer:

$$call = \beta_0 + \beta_1 ethnicity + \beta_2 city + \beta_3(city \times ethnicity) + \beta_4 experience + \beta_5 exprience^2$$

Printing coefficient table with robust s.e. and test

```
##
## t test of coefficients:
##
##                         Estimate  Std. Error t value Pr(>|t|)
## (Intercept)           0.05029248  0.03992725  1.2596  0.20829
## ethnicity            -0.06128094  0.03476768 -1.7626  0.07847 .
## city                 -0.06544836  0.03456274 -1.8936  0.05874 .
## I(city * ethnicity)   0.07490281  0.04265341  1.7561  0.07957 .
## experience            0.01260612  0.00634459  1.9869  0.04737 *
## I(experience^2)      -0.00038045  0.00024993 -1.5222  0.12847
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First I will interpret the coefficients:

β_0: The probability of being called back for caucasians sounding names who live in boston is 5.03%.
ethnicity: The average probability of being called back for an individual with an African-American sounding name living in Boston is 6.13% lower than Caucasian sounding name in Boston.
city: The average probability of being called back for an individual with an Caucasian sounding name living in Chicago is 6.54% lower than in Boston.
$city \times ethnicity$: The difference between the average probability of being called back for the two ethnic groups for the two different city is 7.5%.

LPM's are heteroscedastic by construction, so all your tests and confidence intervals must be robust to heteroscedasticity.

Since we checking for discrimination for people from Boston and Chicago, if there was no discrimination, the probability of being called would be same for both ethnicity. So, inoreder to answer the question we need to check the significance of coefficient of ethnicity and $city \times ethnicity$. Notice that we donot need to check city since we are cheecking to see the difference between the two ethnicities but city is for the same etnicity.

Testing for African-American sounding names in Chicago VS Caucasian sounding names in Chicago: Using the robust F test:

```
## Linear hypothesis test
##
## Hypothesis:
## ethnicity  + I(city * ethnicity) = 0
##
## Model 1: restricted model
## Model 2: call ~ ethnicity + city + I(city * ethnicity) + experience +
##     I(experience^2)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    618
## 2    617  1 0.3022     0.5825
```

Since the p-value of the test is greater than 5%, we do not reject the Null hypothesis. Therefore, we can conclude that there is no significant difference in discrimination for for African-American sounding names in Chicago VS Caucasian sounding names in Chicago.

Testing for African-American sounding names in Boston VS Caucasian sounding names in Boston:

Using the robust F test:

```
## Linear hypothesis test
##
## Hypothesis:
## ethnicity = 0
##
## Model 1: restricted model
## Model 2: call ~ ethnicity + city + I(city * ethnicity) + experience +
##     I(experience^2)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df  Chisq Pr(>Chisq)
```

```
## 1    618
## 2    617  1 3.1067    0.07797 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value of the test is greater than 5%, we do not reject the Null hypothesis. Therefore, we can conclude that there is no significant difference in discrimination for African-American sounding names in Boston VS Caucasian sounding names in Boston.

Testing for African-American sounding names in Chicago VS African-American sounding names in Boston:

Using the robust F test:

```
## Linear hypothesis test
##
## Hypothesis:
## city  + I(city * ethnicity) = 0
##
## Model 1: restricted model
## Model 2: call ~ ethnicity + city + I(city * ethnicity) + experience +
##     I(experience^2)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    618
## 2    617  1 0.1305     0.7179
```

Since the p-value of the test is greater than 5%, we do not reject the Null hypothesis. Therefore, we can conclude that there is no significant difference in discrimination for African-American sounding names in Chicago VS African American sounding names in Boston.

Computing robust confidence intervals:

```
##                               2.5 %       97.5 %
## (Intercept)         -0.0279634892 0.1285484556
## ethnicity           -0.1294243436 0.0068624592
```

```
## city                 -0.1331900924 0.0022933757
## I(city * ethnicity) -0.0086963285 0.1585019559
## experience           0.0001709472 0.0250412921
## I(experience^2)     -0.0008703143 0.0001094071
```

Question 2

LPM's are heteroscedastic by construction, so all your tests and confidence intervals must be robust to heteroscedasticity.

Is discrimination different for male and female individuals with prior military experience having African-American sounding name?

I will use this model to answer:

$$call = \beta_0 + \beta_1(ethnicity \times military) + \beta_2(gender \times military) + \beta_4(ethnicity \times gender) + \beta_5 ethnicity + \\ \beta_6 gender + \beta_7 military + \beta_8 experience + \beta_9 exprience^2$$

Printing coefficient table with robust s.e. and test

```
##
## t test of coefficients:
##
##                              Estimate  Std. Error t value Pr(>|t|)
## (Intercept)                0.01520986  0.03556209  0.4277  0.66902
## ethnicity                 -0.00975463  0.02559138 -0.3812  0.70321
## gender                    -0.02572416  0.03825208 -0.6725  0.50152
## military                   0.03824185  0.10785410  0.3546  0.72303
## experience                 0.01206767  0.00708619  1.7030  0.08908 .
## I(experience^2)           -0.00034287  0.00026493 -1.2942  0.19609
## ethnicity:military        -0.08906349  0.10950321 -0.8133  0.41634
## gender:military            0.05662383  0.17010573  0.3329  0.73934
## ethnicity:gender           0.00013316  0.05024810  0.0027  0.99789
## ethnicity:gender:military -0.01977149  0.17201770 -0.1149  0.90853
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First I will interpret the coefficients:

intercept: The probability of being called back for female caucasian sounding names with no prior military experience is 1.5%.

ethnicity: The probability of being called back when applicants have female African-American sounding names with no military experience is 0.93 percentage less.

military: The probability of being called back when applicants have female Caucasian sounding names with military experience is 4.38 percentage higher.

gender: The probability of being called back when applicants have male Caucasian sounding names with no military experience is 2.44 percentage less.

$military \times ethnicity$: Having African-American sounding name changes the probability of being called by 9.7% percentage point less when the individual had military experience.

$military \times gender$: Being male changes the probability of being called by 4.33% percentage point less when the individual had military experience.

$ethnicity : gender$: Being male changes the probability of being called by 0.22% percentage point less when the individual had African-American sounding name.

There are a few conditions:

1. Male African American sounding names with military experience VS Male Caucasian sounding names with military experience
2. Female African American sounding names with military experience VS female Caucasian sounding names with military experience
3. Male African American sounding names with no military experience VS Male Caucasian sounding names with no military experience
4. Female African American sounding names with no military experience VS Female Caucasian sounding names with no military experience
5. Male African American sounding names with military experience VS Female African American sounding names with military experience
6. Male African American sounding names with no military experience VS Female African American sounding names with no military experience

Testing Condition 1:

Using the robust F test:

```
## Linear hypothesis test
##
## Hypothesis:
## ethnicity  + ethnicity:military  + ethnicity:gender  + ethnicity:gender:military = 0
##
## Model 1: restricted model
## Model 2: call ~ ethnicity + gender + military + ethnicity * military +
##     (gender * military) + (ethnicity * gender) + (ethnicity *
##     gender * military) + experience + I(experience^2)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    614
## 2    613  1 0.8979     0.3433
```

Since the p-value of the test if greater than 5%, we reject the Null hypothesis. Therefore, we can conclude that there is no significant difference in discrimination for Male African American sounding names with military experience VS Male Caucasian sounding names with military experience.

Testing Condition 2:

Using the robust F test:

```
## Linear hypothesis test
##
## Hypothesis:
## ethnicity  + ethnicity:military = 0
##
## Model 1: restricted model
## Model 2: call ~ ethnicity + gender + military + ethnicity * military +
##     (gender * military) + (ethnicity * gender) + (ethnicity *
##     gender * military) + experience + I(experience^2)
##
```

```
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    614
## 2    613  1 0.8624     0.3531
```

Since the p-value of the test if greater than 5%, we reject the Null hypothesis. Therefore, we can conclude that there is no significant difference in discrimination for female African American sounding names with military experience VS female Caucasian sounding names with military experience.

Testing Condition 3:

Using the robust F test:

```
## Linear hypothesis test
##
## Hypothesis:
## ethnicity  + ethnicity:gender = 0
##
## Model 1: restricted model
## Model 2: call ~ ethnicity + gender + military + ethnicity * military +
##     (gender * military) + (ethnicity * gender) + (ethnicity *
##     gender * military) + experience + I(experience^2)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    614
## 2    613  1 0.0498     0.8234
```

Since the p-value of the test if greater than 5%, we reject the Null hypothesis. Therefore, we can conclude that there is no significant difference in discrimination for Male African American sounding names with no military experience VS Male Caucasian sounding names with no military experience.

Testing Condition 4:

Using the robust F test:

```
## Linear hypothesis test
##
## Hypothesis:
## ethnicity = 0
##
## Model 1: restricted model
## Model 2: call ~ ethnicity + gender + military + ethnicity * military +
##     (gender * military) + (ethnicity * gender) + (ethnicity *
##     gender * military) + experience + I(experience^2)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    614
## 2    613  1 0.1453     0.7031
```

Since the p-value of the test if greater than 5%, we reject the Null hypothesis. Therefore, we can conclude that there is no significant difference in discrimination for female African American sounding names with no military experience VS female Caucasian sounding names with no military experience.

Testing Condition 5:

Using the robust F test:

```
## Linear hypothesis test
##
## Hypothesis:
## gender  + gender:military  + ethnicity:gender  + ethnicity:gender:military = 0
##
## Model 1: restricted model
## Model 2: call ~ ethnicity + gender + military + ethnicity * military +
##     (gender * military) + (ethnicity * gender) + (ethnicity *
##     gender * military) + experience + I(experience^2)
##
## Note: Coefficient covariance matrix supplied.
```

```
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    614
## 2    613  1 1.5094     0.2192
```

Since the p-value of the test if greater than 5%, we reject the Null hypothesis. Therefore, we can conclude that there is no significant difference in discrimination for male African American sounding names with military experience VS female African American sounding names with military experience.

Testing Condition 6:

Using the robust F test:

```
## Linear hypothesis test
##
## Hypothesis:
## gender  + ethnicity:gender = 0
##
## Model 1: restricted model
## Model 2: call ~ ethnicity + gender + military + ethnicity * military +
##     (gender * military) + (ethnicity * gender) + (ethnicity *
##     gender * military) + experience + I(experience^2)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    614
## 2    613  1 0.6173     0.4321
```

Since the p-value of the test if greater than 5%, we reject the Null hypothesis. Therefore, we can conclude that there is no significant difference in discrimination for male African American sounding names with no military experience VS female African American sounding names with no military experience.

Therefore, looking at all the cases, the answer to my question is: **There is no significant discrimination difference for male and female individuals with prior military experience having African-American sounding names.**

Computing robust confidence intervals:

```
##                                 2.5 %       97.5 %
## (Intercept)               -0.0544905608 0.0849102805
## ethnicity                 -0.0599128086 0.0404035509
## gender                    -0.1006968551 0.0492485360
## military                  -0.1731483082 0.2496320074
## experience                -0.0018210086 0.0259563388
## I(experience^2)           -0.0008621246 0.0001763879
## ethnicity:military        -0.3036858418 0.1255588656
## gender:military           -0.2767772693 0.3900249242
## ethnicity:gender          -0.0983513146 0.0986176361
## ethnicity:gender:military -0.3569199799 0.3173770073
```

Question 3

LPM's are heteroscedastic by construction, so all your tests and confidence intervals must be robust to heteroscedasticity.

Is the effect of experience on the probability of being called back the same for people provided email address and those that didnot?

Printing coefficient table with robust s.e. and test

```
##
## t test of coefficients:
##
##                             Estimate  Std. Error t value Pr(>|t|)
## (Intercept)               0.05812225  0.05022877  1.1572   0.2477
## email                    -0.09971026  0.09478713 -1.0519   0.2932
## ethnicity                 0.01803360  0.08072048  0.2234   0.8233
## experience                0.00259365  0.00974558  0.2661   0.7902
## I(experience^2)          -0.00023673  0.00034665 -0.6829   0.4949
## email:ethnicity           0.01546828  0.13109255  0.1180   0.9061
## email:experience          0.01946438  0.02172977  0.8957   0.3707
## email:I(experience^2)    -0.00008472  0.00097633 -0.0868   0.9309
```

```
## ethnicity:experience            -0.00531919  0.01526353 -0.3485   0.7276
## ethnicity:I(experience^2)        0.00021720  0.00051846  0.4189   0.6754
## email:ethnicity:experience      -0.01058898  0.03195068 -0.3314   0.7404
## email:ethnicity:I(experience^2)  0.00041562  0.00153074  0.2715   0.7861
```

First I will interpret the coefficients:

intercept: The the effect of experience on probability of being called back for people who didnot provided any email is 6.73%.

ethnicity ∗ *experience*: The the effect of experience on probability of being called back when applicants have African American sounding names with no email is 0.53 percentage less.

experience ∗ *email*: the effect of experience on the probability of being called is 1.55% percentage point more when the individual provided an email.

We need to test a few conditions:

1. African American sounding names who provided email VS Caucasian sounding names who provided email.

2. African American sounding names who didnot provide email VS Caucasian sounding names who didnot provide email

Testing Condition 1:

Using the robust F test:

```
## Linear hypothesis test
##
## Hypothesis:
## ethnicity  + email:ethnicity  + ethnicity:experience  + ethnicity:I(experience^2)  + email:ethnicity
##
## Model 1: restricted model
## Model 2: call ~ (email * ethnicity) * (experience + I(experience^2))
##
```

```
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    612
## 2    611  1 0.0541     0.8161
```

Since the p-value of the test is greater than 5%, we cannot reject the Null hypothesis. Therefore, we can conclude that the effect of experience on the probability of being called is not significant for African American sounding names who provided email VS Caucasian sounding names who provided email.

Testing Condition 2:

Using the robust F test:

```
## Linear hypothesis test
##
## Hypothesis:
## ethnicity  + ethnicity:experience  + ethnicity:I(experience^2) = 0
##
## Model 1: restricted model
## Model 2: call ~ (email * ethnicity) * (experience + I(experience^2))
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    612
## 2    611  1 0.0369     0.8476
```

Since the p-value of the test is greater than 5%, we cannot reject the Null hypothesis. Therefore, we can conclude that the effect of experience on the probability of being called is not significant for African American sounding names who didnot provide email VS Caucasian sounding names who didnot provide email.

Therefore, looking at all the cases, the answer to my question is: **There is no significant difference caused by the effect of experience on the probability of being called back for people who provided email address and those that didnot.**

Computing robust confidence intervals:

```
##                                        2.5 %       97.5 %
## (Intercept)                      -0.0403243277 0.1565688297
## email                            -0.2854896298 0.0860691098
## ethnicity                        -0.1401756232 0.1762428308
## experience                       -0.0165073360 0.0216946374
## I(experience^2)                  -0.0009161557 0.0004426975
## email:ethnicity                  -0.2414684010 0.2724049556
## email:experience                 -0.0231251868 0.0620539513
## email:I(experience^2)            -0.0019982986 0.0018288581
## ethnicity:experience             -0.0352351615 0.0245967821
## ethnicity:I(experience^2)        -0.0007989656 0.0012333764
## email:ethnicity:experience       -0.0732111558 0.0520331969
## email:ethnicity:I(experience^2) -0.0025845741 0.0034158083
```

Model 4

LPM's are heteroscedastic by construction, so all your tests and confidence intervals must be robust to heteroscedasticity.

Is discrimination different for people who mention some volunteering experience VS those that didnot for people applying to EOE?

I will use this model to answer:

$$
\begin{aligned}
call = \beta_0 + \beta_1(ethnicity \times equal) + \beta_2(volunteer \times equal) + \beta_3(ethnicity \times volunteer) + \\
\beta_4 volunteer + \beta_5 ethnicity + \beta_6 equal + \beta_7 exprience + \beta_8 expereince^2
\end{aligned}
$$

Printing coefficient table with robust s.e. and test

```
##
## t test of coefficients:
##
##                            Estimate  Std. Error t value Pr(>|t|)
## (Intercept)              0.02573218  0.03748248  0.6865  0.49265
## volunteer               -0.00425599  0.03871174 -0.1099  0.91249
```

```
## ethnicity               -0.03835830  0.03130862 -1.2252  0.22098
## equal                   -0.05933884  0.03292386 -1.8023  0.07199 .
## I(ethnicity * equal)     0.05215593  0.04346908  1.1998  0.23066
## I(volunteer * equal)     0.04315213  0.04736282  0.9111  0.36260
## I(ethnicity * volunteer)  0.00672417  0.04413334  0.1524  0.87895
## experience               0.01291067  0.00642064  2.0108  0.04478 *
## I(experience^2)         -0.00037149  0.00024976 -1.4874  0.13743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First I will interpret the coefficients:

intercept: The probability of being called back for caucasian sounding names with no prior volunteer experience is and not applying to a EOE is 2.57%.

ethnicity: The probability of being called back when applicants have African-American sounding names with no volunteer experience applying to a EOE is 3.84 percentage less.

volunteer: The probability of being called back when applicants Caucasian sounding names with volunteer experience is 0.43 percentage less when applying to a non-EOE.

equal: The probability of being called back when applicants have Caucasian sounding names with no volunteer experience is 5.93 percentage less when applying to a EOE.

$volunteer \times ethnicity$: Having African-American sounding name changes the probability of being called by 0.67% percentage point more when the individual had volunteer experience.

$equal \times volunteer$: Having volunteer experience changes the probability of being called by 4.31% percentage point more when the individual had applied to a EOE.

$ethnicity : equal$: Having African-American sounding name changes the probability of being called by 5.21% percentage point more when applying to a EOE.

There are a few conditions:

1. African American sounding names with Volunteer experience applying to EOE VS Caucasian sounding names with Volunteer experience applying to EOE.

2. African American sounding names with no Volunteer experience applying to EOE VS Caucasian sounding names with no Volunteer experience applying to EOE

3.African American sounding names with Volunteer experience applying to non-EOE VS Caucasian sounding names with Volunteer experience applying to non-EOE

4. African American sounding names with no Volunteer experience applying to non-EOE VS Caucasian sounding names with no Volunteer experience applying to non-EOE

Using the robust F test:

Testing Condition 1:

```
## Linear hypothesis test
##
## Hypothesis:
## ethnicity  + I(ethnicity * equal)  + I(ethnicity * volunteer) = 0
##
## Model 1: restricted model
## Model 2: call ~ volunteer + ethnicity + equal + I(ethnicity * equal) +
##     I(volunteer * equal) + I(ethnicity * volunteer) + experience +
##     I(experience^2)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    615
## 2    614  1 0.1826     0.6691
```

Since the p-value of the test if greater than 5%, we cannot reject the Null hypothesis. Therefore, we can conclude that there is no significant difference in discrimination for African American sounding names with Volunteer experience applying to EOE VS Caucasian sounding names with Volunteer experience applying to EOE.

Testing Condition 2:

```
## Linear hypothesis test
##
## Hypothesis:
## ethnicity  + I(ethnicity * equal) = 0
##
## Model 1: restricted model
## Model 2: call ~ volunteer + ethnicity + equal + I(ethnicity * equal) +
```

```
##     I(volunteer * equal) + I(ethnicity * volunteer) + experience +
##     I(experience^2)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    615
## 2    614  1 0.1497     0.6988
```

Since the p-value of the test is greater than 0.05, we cannot reject the Null hypothesis. Therefore, we can conclude that there is no significant difference African American sounding names with Volunteer experience applying to EOE VS African American sounding names with no Volunteer experience applying to EOE.

Testing Condition 3:

```
## Linear hypothesis test
##
## Hypothesis:
## ethnicity  + I(ethnicity * volunteer) = 0
##
## Model 1: restricted model
## Model 2: call ~ volunteer + ethnicity + equal + I(ethnicity * equal) +
##     I(volunteer * equal) + I(ethnicity * volunteer) + experience +
##     I(experience^2)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    615
## 2    614  1 0.7198     0.3962
```

Since the p-value of the test is greater than 5%, we cannot reject the Null hypothesis. Therefore, we can conclude that there is no significant difference in discrimination for African American sounding names with Volunteer experience applying to non-EOE VS Caucasian sounding names with Volunteer experience applying to non-EOE.

Testing Condition 4:

```
## Linear hypothesis test
##
## Hypothesis:
## ethnicity = 0
##
## Model 1: restricted model
## Model 2: call ~ volunteer + ethnicity + equal + I(ethnicity * equal) +
##     I(volunteer * equal) + I(ethnicity * volunteer) + experience +
##     I(experience^2)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df Chisq Pr(>Chisq)
## 1    615
## 2    614  1 1.501     0.2205
```

Since the p-value of the test if greater than 5%, we cannot reject the Null hypothesis. Therefore, we can conclude that there is no significant difference in discrimination for African American sounding names with no Volunteer experience applying to non-EOE VS Caucasian sounding names with no Volunteer experience applying to non-EOE.

Therefore, looking at all the cases, the answer to my question is: **There is no significant difference in discrimination for people who mention some volunteering experience VS those that didnot for people applying to EOE**

Computing robust confidence intervals:

```
##                                  2.5 %       97.5 %
## (Intercept)              -0.0477321434 0.0991964954
## volunteer                -0.0801296117 0.0716176360
## ethnicity                -0.0997220648 0.0230054743
## equal                    -0.1238684142 0.0051907428
## I(ethnicity * equal)     -0.0330418987 0.1373537562
## I(volunteer * equal)     -0.0496772813 0.1359815501
```

```
## I(ethnicity * volunteer) -0.0797755930 0.0932239344
## experience                 0.0003264569 0.0254948903
## I(experience^2)           -0.0008610056 0.0001180329
```

The Average Partial effect of experience on call is:

```
##  experience
## 0.005276949
```

Interpretation: The probability of being called for an an African American sounding name with an extra year of experience and has provided email on their resume is 0.53% higher, when their initial number of years of experience is equal to the sample average.

The confidence interval for this is:

```
## [1] -0.0001751514  0.0107290487
```

The Average Partial effect of email on call is:

```
##      email
## 0.03247654
```

The confidence interval for this is:

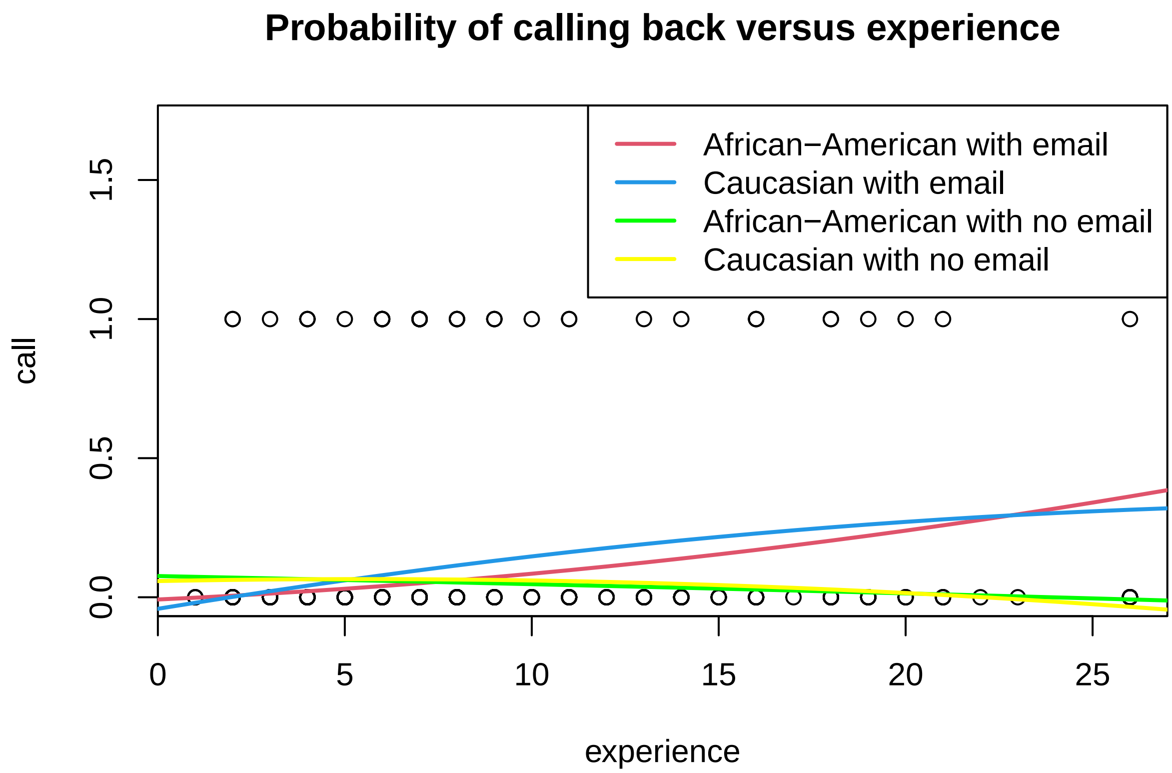
```
## [1] -0.00886122  0.07381430
```

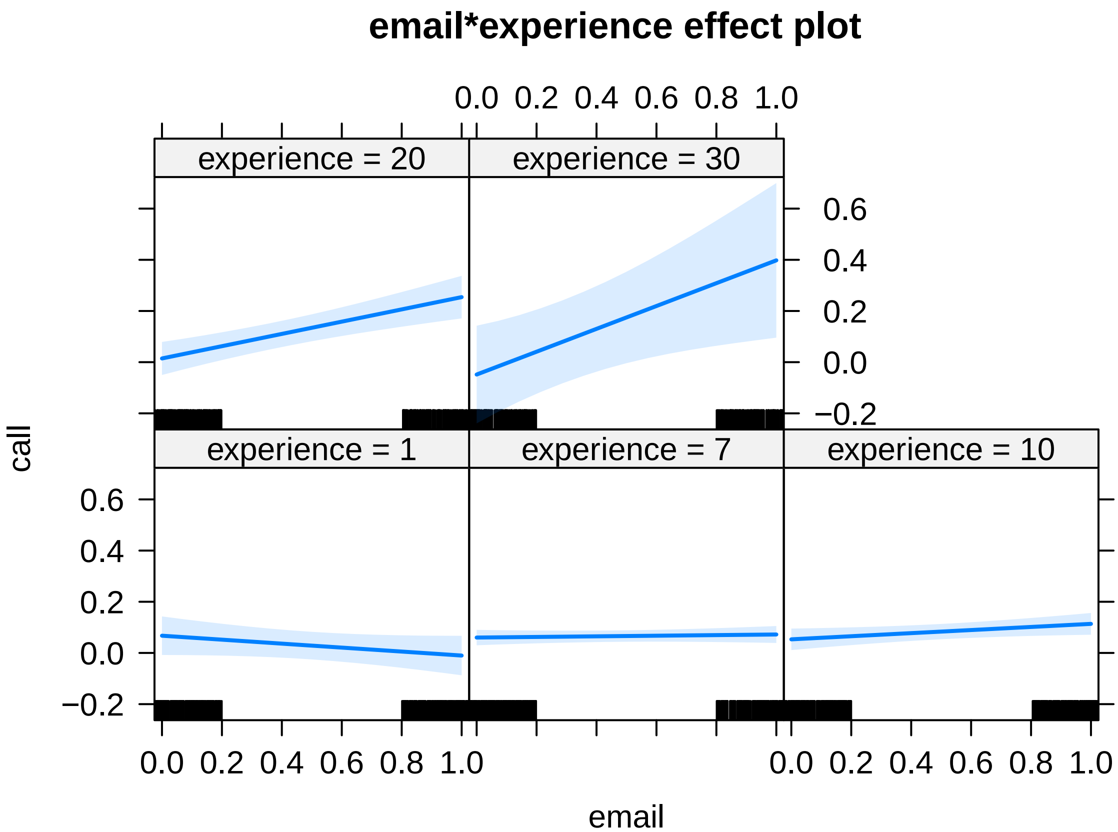
The Average Partial effect of ethnicity on call is:

```
##   ethnicity
## -0.01966382
```

The confidence interval for this is:

```
## [1] -0.06010273  0.02077509
```


Probability of calling back versus experience
African−American with email
Caucasian with email
African−American with no email
Caucasian with no email
call
0.0
0.5
1.0
1.5
0
5
10
15
20
25
experience



If the individual with African American sounding name provided an email, the probability of being called back is slightly less than an individual with Caucasian sounding name who provided an email. However, as level of experience increases, the probability of being called back for an individual with African American sounding name provided an email increases.

If the individual with African American sounding name didnot provide an email, the probability of being called back is almost indistinguisable from an individual with Caucasian sounding name who didnot an email.

Experience has a greater effect between African-American and Caucasian sounding names if they both provided an email.

Code

```
## Part A

library(stargazer)
library(lmtest)
library(sandwich)
library(car)
library(margins)


load("Data20745516.rda")

lm.ols <- lm(hours ~. , data = Hours)
Hours$residue <- lm.ols$residuals

library(ggplot2)
ggplot(data = Hours, aes(y = residue, x = hours)) + geom_point(col = 'blue') + geom_abline(slope = 0)

PartA<-na.omit(Hours)

# Creating Duumy Variable "haveKids"
PartA$havekids <- ifelse(Hours$youngkids > 0, 1, 0)
PartA$havekids <- ifelse((Hours$oldkids > 0), 1, PartA$havekids)

# The 3 models are:

model_1 <- lm(hours~youngkids+I(youngkids^2)+I(havekids*hhours)+havekids+
                  age+experience+hhours+hwage,data=PartA)
model_2 <- lm(log(hours)~youngkids+I(havekids*hhours)+experience+havekids+
                  log(age)+hhours+I(hhours*hwage^2)+log(hwage),
              data=PartA,subset=(hours>0))
model_3 <- lm(hours~youngkids+I(havekids*hhours)+age+experience+havekids+
```

```r
                  hhours+hwage,data=PartA)


# from Proj2 file
printEqu <- function(obj,digits=4, form=NULL, maxpl=5, adjrsq=FALSE,
                     robsd=NULL, stars=FALSE, dist=c("t","n"), label=NULL,
                     omit=NULL, ...)
{
  dist <- match.arg(dist)
  if (is.null(label))
  {
    typeEQ <- "equation*"
    label=""
  } else {
    typeEQ <- "equation"
    label <- paste("\\label{",label,"}", sep="")
  }
  if (!is.null(omit))
  {
    omit <- omit[omit != "(Intercept)"]
    chk <- lapply(omit, function(o) grep(o, names(obj$coef)))
    chk <- sapply(1:length(chk), function(i) length(chk[[i]])==0)
    omit <- omit[!chk]
    if (length(omit)==0)
      omit <- NULL
  }
  if (is.null(form))
  {
    cat("\\begin{",typeEQ, "}", label,"\n", sep="")
    cat("\\begin{split}\n")
    ncoef <- names(coef(obj))
    if (!is.null(omit))
    {
```

```
    omit <- do.call("c", lapply(omit, function(o) grep(o, ncoef)))
    omit <- -unique(omit)
    nomit <- length(omit)
    add <- paste("\\mbox{ (+ ", nomit, " omitted terms)}", sep="")
  } else {
    add <- ""
    omit <- 1:length(ncoef)
  }
  Intercept <- attr(obj$terms, "intercept")
  b <- formatC(abs(obj$coef), digits=digits, ...)[omit]
  if (!is.null(robsd))
    snum <- robsd[omit]
  else
    snum <- summary(obj)$coef[omit,2]
  s <- formatC(snum, digits=digits, ...)
  ncoef <- ncoef[omit]
  if (stars)
  {
    ttest <- coef(obj)[omit]/snum
    if (dist=="t")
      pv <- 2*pt(-abs(ttest), obj$df)
    else
      pv <- 2*pnorm(-abs(ttest))
    sym <- symnum(pv, cutpoints=c(0,.01,.05,.1,1),
                  symbols=c("^{***}","^{**}","^*"," "))
    symmess <- "\\\\& ^*\\text{pv}<0.1\\mbox{; }^{**}\\text{pv}<0.05\\mbox{; }^{***}\\text{pv}<0.01"
  } else {
    sym <- rep("", length(coef(obj)[omit]))
    symmess <- ""
  }
  ny <- rownames(attr(obj$terms, "factors"))[1]
  ny <- paste("\\widehat{",ny,"}",sep="")
```

```
cat(ny,"&=")
if (obj$coef[1] < 0)
  cat("\\underset{(",s[1],")",sym[[1]],"}{-",b[1],"}", sep="")
else
  cat("\\underset{(",s[1],")",sym[[1]],"}{",b[1],"}")
if (Intercept==0)
  cat("~",ncoef[1], sep="")
j <- 1
for (i in 2:length(b))
{
  if (j>maxpl)
  {
    j <- 1
    cat("\\\\&\\quad\n")
  }
  if ((obj$coef[omit])[i] < 0)
    cat("~-~")
  else
    cat("~+~")
  cat("\\underset{(",s[i],")",sym[[i]],"}{",b[i],"}~")
  cat(ncoef[i])
  j <- j+1
}
cat(add)
n <- length(obj$residuals)
cat("\\\\ &\\quad n=", n, ",~~R^2=", round(summary(obj)$r.squared,5))
cat(", SSR=", round(sum(obj$resid^2),5))
if (adjrsq)
  cat(", \\bar{R}^2=", round(summary(obj)$adj,5))
if (!is.null(robsd))
  cat("\\mbox{ (Robust S-E)}")
cat(symmess)
```

```r
    cat("\n\\end{split}\n")
    cat("\\end{", typeEQ, "}\n", sep="")
  } else {
    t <- terms(form)
    y <- rownames(attr(t, "factors"))[1]
    x <- colnames(attr(t, "factors"))
    cat("\\begin{",typeEQ, "}", label,"\n\\begin{split}\n",
        y,"&=\\beta_0", sep="")
    j  <-  1
    for (i in 1:length(x))
    {
      if (j>maxpl)
      {
        j <- 1
        cat("\\\\&\n")
      }
      cat("+\\beta_",i,x[i],sep="")
      j <- j+1
    }
    cat("+u\n\\end{split}\n\\end{",typeEQ,"}", sep="")
  }
}

printEqu(model_1, stars=TRUE, maxpl=3,adjrsq=TRUE, digits=5)
printEqu(model_2, stars=TRUE, maxpl=3,adjrsq=TRUE, digits=5)
printEqu(model_3, stars=TRUE, maxpl=3,adjrsq=TRUE, digits=5)

# We first estimate the models and test the homoscedasticity using the short
# White test:

yhat <- fitted(model_1)
yhat2 <- yhat^2
```

```
bptest(model_1, ~yhat+yhat2)

yhat <- fitted(model_2)
yhat2 <- yhat^2
bptest(model_2, ~yhat+yhat2)

yhat <- fitted(model_3)
yhat2 <- yhat^2
bptest(model_3, ~yhat+yhat2)

# Indirect t-test
coeftest(model_1, vcov.=vcovHC, df=Inf)

# J-test between Model 1 and 2
jtest(model_1, model_2,vcov=vcovHC, df=Inf, data=PartA)

# Checking if model properly specified
resettest(model_1, power=2:3)

bp <- bptest(model_1)
adj <- if(bp$p.value<.05) vcovHC else NULL
coeftest(model_1, vcov.=adj, df=Inf)

#Cook's Distance
plot(model_1,4)

# Removing outliers
fit <- lm(hours~youngkids+I(youngkids^2)+I(havekids*hhours)+havekids+
            age+experience+hhours+hwage,data=PartA,
          subset = !(rownames(PartA) %in% c(37,126,403,598,734,292)))
plot(fit,4)
```

```r
# impact of dropping these three observations:

stargazer(model_1, fit, type='text')

confint(fit)

# APE of youngkids on hours
b <- coef(fit)
m <- mean(PartA$youngkids)
APE <- b[2]+2*b[3]*m

APE
# 95% Confidence of the APE
v <- vcov(fit)
s <- sqrt(v[2,2]+4*m^2*v[3,3]+4*m*v[2,3])

crit <- qt(.975, fit$df)
c(APE-s*crit, APE+s*crit)

#  Does having children less than 6 have the same effect as having
#  children between ages 6 and 18?
linearHypothesis(fit,c("youngkids","I(youngkids^2)"))

## hhours vs youngkids
library(effects)
plot(Effect(c("youngkids","hhours"), fit))

## Women with kids VS Women without kids

fit <- lm(hours~youngkids+I(youngkids^2)+I(havekids*hhours)+havekids+
            age+experience+hhours+hwage,data=PartA,
          subset = !(rownames(PartA) %in% c(37,126,403,598,734,292)))
```

```
new<-PartA[!(rownames(PartA) %in% c(37,126,403,598,734,292)),]


r <- range(new$hhours)


hhours <- r[1]:r[2]
newd1 <- data.frame(hhours=hhours, youngkids= mean(new$youngkids),
                    havekids=1,  age= mean(new$age),
                    experience= mean(new$experience),hwage=mean(new$hwage))
newd2 <- data.frame(hhours=hhours, youngkids= mean(new$youngkids),
                    havekids=0,  age= mean(new$age),
                    experience= mean(new$experience),hwage=mean(new$hwage))


pr1 <- predict(fit, newdata=newd1)
pr2 <- predict(fit, newdata=newd2)
pred <- cbind(pr1,pr2)


col <- (new$havekids==1) +
  2*(new$havekids==0)
pch <- 21*(new$havekids==1) +
  22*(new$havekids==0)
plot(hours~hhours, col=col, bg=col, pch=pch, data=new)
matplot(hhours, pred, col=1:2, lty=1:2, lwd=2, type='l',
        main="Predicted working hours of women with and without kids in 1975",
        ylab="hours",xlab="husband's working hours", add=TRUE)
legend(x=3500, y=2900,legend=c("women with kids", "women without kids"),col=1:2,
       lty=1:2, lwd=2,pt.bg=1:4, pch=21:24, cex=.8)
grid()
```

PART B

```
PartB<-Names
```

```r
# Changinging categorical data to binary 1 and 0.

PartB$gender <- as.numeric(PartB$gender == "male")
PartB$call <- as.numeric(PartB$call == "yes")
PartB$ethnicity <- as.numeric(PartB$ethnicity == "afam")
PartB$email <- as.numeric(PartB$email == "yes")
PartB$military <- as.numeric(PartB$military == "yes")
PartB$volunteer <- as.numeric(PartB$volunteer == "yes")
PartB$equal <- as.numeric(PartB$equal == "yes")
PartB$city <- as.numeric(PartB$city == "chicago") # Chicago is 1 and boston is 0


# Estimating the models
fit_1 <- lm (call~ ethnicity + city + I(city * ethnicity) + experience +
               I(experience^2),data = PartB)

fit_2 <- lm (call ~ ethnicity+ gender + military + ethnicity * military
             + (gender * military) + (ethnicity * gender)+
               (ethnicity* gender *military) + experience +
               I(experience^2),data=PartB )

fit_3 <- lm(call ~(email*ethnicity)*(experience+I(experience^2)),data=PartB)
fit_4 <- lm(call ~ volunteer + ethnicity + equal+ I(ethnicity * equal) +
              I (volunteer * equal) + I (ethnicity* volunteer)+ experience +
              I(experience^2),data=PartB)

# QUESTION 1

# Printing coefficient table with robust s.e. and test

coeftest(fit_1, vcov.=vcovHC)
```

```
# Testing for African-American sounding names in Chicago VS Caucasian sounding names in Chicago:
# Using the robust F test:

linearHypothesis(fit_1, "ethnicity + I(city * ethnicity)=0", white.adjust=TRUE, test="Chisq")

# Testing for African-American sounding names in Boston VS Caucasian sounding names in Boston:

# Using the robust F test:

# ethnicity =0
linearHypothesis(fit_1, "ethnicity = 0", white.adjust=TRUE, test="Chisq")

# Testing for African-American sounding names in Chicago VS African-American sounding names in Boston:

# Using the robust F test:

# "city + I(city * ethnicity) = 0"
linearHypothesis(fit_1, "city + I(city * ethnicity) = 0",
                 white.adjust=TRUE, test="Chisq")

# Computing robust confidence intervals:

coefci(fit_1, level=.95, vcov.=vcovHC, df=Inf)

## QUESTION 2

# Testing Condition 1:

# Using the robust F test:

# "ethnicity + ethnicity:military+ethnicity:gender+ethnicity:gender:military= 0"
linearHypothesis(fit_2, "ethnicity + ethnicity:military+ethnicity:gender+
```

```
                 ethnicity:gender:military= 0",
                 white.adjust=TRUE,vcov.=vcovHC, test="Chisq")

# Testing Condition 2:

# Using the robust F test:

# "ethnicity+ethnicity:military =0"
linearHypothesis(fit_2, "ethnicity+ethnicity:military =0",
                 white.adjust=TRUE,vcov.=vcovHC, test="Chisq")

# Testing Condition 3:

# Using the robust F test:

# "ethnicity+ ethnicity:gender=0"
linearHypothesis(fit_2, "ethnicity+ ethnicity:gender=0",
                 white.adjust=TRUE,vcov.=vcovHC, test="Chisq")

# Testing Condition 4:

# Using the robust F test:

# "ethnicity=0"
linearHypothesis(fit_2, "ethnicity=0",
                 white.adjust=TRUE,vcov.=vcovHC, test="Chisq")

# Testing Condition 5:

# Using the robust F test:
# "gender +gender:military+ethnicity:gender+ethnicity:gender:military=0"
linearHypothesis(fit_2, "gender+gender:military+ethnicity:gender+
```

```
                 ethnicity:gender:military=0",
                 white.adjust=TRUE,vcov.=vcovHC, test="Chisq")

# Testing Condition 6:

# Using the robust F test:

# "gender + ethnicity:gender= 0"
linearHypothesis(fit_2, "gender + ethnicity:gender= 0",
                 white.adjust=TRUE,vcov.=vcovHC, test="Chisq")

# Computing robust confidence intervals:

coefci(fit_2, level=.95, vcov.=vcovHC, df=Inf)

# QUESTION 3

# Printing coefficient table with robust s.e. and test

coeftest(fit_3, vcov.=vcovHC)

# Testing Condition 1:

# Using the robust F test:

# "ethnicity+email:ethnicity+ethnicity:experience+ethnicity:experience^2+
#  email:ethnicity:experience+email:ethnicity:I(experience^2)= 0"

linearHypothesis(fit_3, "ethnicity+email:ethnicity+
                 ethnicity:experience+ethnicity:I(experience^2)+
                 email:ethnicity:experience+
                 email:ethnicity:I(experience^2)= 0",
```

```r
                 white.adjust=TRUE, test="Chisq")

# Testing Condition 2:

# Using the robust F test:

# "ethnicity+ethnicity:experience+ethnicity:I(experience^2)=0"

linearHypothesis(fit_3, "ethnicity+ethnicity:experience+
                 ethnicity:I(experience^2)=0",white.adjust=TRUE, test="Chisq")

# Computing robust confidence intervals:

coefci(fit_3, level=.95, vcov.=vcovHC, df=Inf)


## QUESTION 4

# Printing coefficient table with robust s.e. and test

coeftest(fit_4, vcov.=vcovHC)

# Testing Condition 1:

linearHypothesis(fit_4, "ethnicity+I(ethnicity * equal)+I(ethnicity * volunteer)= 0", white.adjust=TRUE

# Testing Condition 2:

linearHypothesis(fit_4, "ethnicity + I(ethnicity * equal) = 0", white.adjust=TRUE, test="Chisq")

# Testing Condition 3:
```

```r
linearHypothesis(fit_4, "ethnicity+I(ethnicity * volunteer)=0", white.adjust=TRUE, test="Chisq")

# Testing Condition 4:

linearHypothesis(fit_4, "ethnicity = 0", white.adjust=TRUE, test="Chisq")

# Computing robust confidence intervals:

coefci(fit_4, level=.95, vcov.=vcovHC, df=Inf)



library(margins)
a<-margins_summary(fit_3, vcov=vcovHC(fit_3))

# The Average Partial effect of experience on call is:

a$AME[3]

# The confidence interval for this is:

c(a$lower[3],a$upper[3])

# The Average Partial effect of email on call is:

a$AME[1]

# The confidence interval for this is:

c(a$lower[1],a$upper[1])

# The Average Partial effect of ethnicity on call is:
```

```r
a$AME[2]

# The confidence interval for this is:

c(a$lower[2],a$upper[2])

# Graph of different groups
newd <- data.frame(email=1, experience=seq(0,30),ethnicity=1)
pr <- predict(fit_3, newdata=newd)
newd <- data.frame(email=1, experience=seq(0,30),ethnicity=0)
pr2 <- predict(fit_3, newdata=newd)
newd <- data.frame(email=0, experience=seq(0,30),ethnicity=1)
pr3 <- predict(fit_3, newdata=newd)
newd <- data.frame(email=0, experience=seq(0,30),ethnicity=0)
pr4 <- predict(fit_3, newdata=newd)
plot(call~experience, data=PartB,
     main="Probability of calling back versus experience",ylab="call",
     xlab="experience",ylim=c(0,1.7))
lines(newd[,2], pr, col=2, lwd=2)
lines(newd[,2], pr2, col=4, lwd=2)
lines(newd[,2], pr3, col="green", lwd=2)
lines(newd[,2], pr4, col="yellow", lwd=2)
legend("topright",c("African-American with email","Caucasian with email",
                    "African-American with no email","Caucasian with no email"),
       col=c(2,4,"green","yellow"), lty=2,lwd=2)

plot(Effect(c("email","experience"), fit_3))
```