Practicum Four

Josh Virene

2023-11-27

2 Measuring and Interpreting Wage Discrimination by Race

The purpose of this exercise is to have you implement and then interpret three alternative procedures for measuring wage discrimination by race. The econometric methodology employed is based on that developed by Blinder [1973] and Oaxaca [1973b]. See the notes posted on Canvas for the details. You will also conduct a Chow test for parameter equality. Choose either the 1978 or the 1985 data set in CPS78 and CPS85, respectively, and use this dataset for all portions of this exercise.

a. Begin by estimating the effects of wage discrimination by race using simple dummy variable procedures. Specifically, using OLS, estimate parameters in the equation:

$$LNWAGE = \alpha + \alpha_f * FE + \alpha_u * UNION + \alpha_n * NONWH + \alpha_h * HISP + \beta_1 * ED + \beta_2 * EX + \beta_3 * EXSQ + \epsilon$$

Where FE is gender dummy(female if FE=1), UNION is union status, NONWH is nonwhite and not Hispanic (NONWH=1), HISP is Hispanic(HISP = 1), ED is school years, EX is experience and EXSQ is the square of EX. Interpret the estimated coefficients a_N , and a_H . Test the null hypothesis that, gender, union status education, and experience being held fixed, racial status has no effect on log wages. Then test and interpret the null hypothesis that the coefficients on NONWH and HISP are equal to each otherbut are not necessarily equal to zero.

Table 1: Summary Statistics for Model One

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6246849	0.1203640	5.1899637	0.0000003
FE	-0.2317750	0.0386147	-6.0022549	0.0000000
UNION	0.2116322	0.0504893	4.1916240	0.0000325
NONWH	-0.1252402	0.0576528	-2.1723193	0.0302775
HISP	-0.0807345	0.0877940	-0.9195896	0.3582088
ED	0.0888187	0.0079602	11.1578618	0.0000000
EX	0.0344669	0.0053782	6.4086390	0.0000000
EXSQ	-0.0005270	0.0001181	-4.4630997	0.0000099

Table 2: Linear Hypothesis Test Examining the Joint Statistical Significance of Race Coefficients

Res.Df	Df	F	Pr(>F)
528	NA	NA	NA
526	2	2.973128	0.0520033

Res.Df	Df	F	Pr(>F)

Table 3: Linear Hypothesis Test to determine whether coefficients are equal to eachother

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
527	100.3450	NA	NA	NA	NA
526	100.3076	1	0.0374627	0.1964497	0.6577842

Part one: Interpret the alpha coefficients

a_n: This is a coefficient on the dummy variable NONWH, which, when NONWH = 1, means the individual is non-White and non-Hispanic. The regression estimated a value on this coefficient of -0.1252402. The interpretation of this is that holding all other variables in the model equal, someone who is non-White and non-Hispanic will have a wage that is 12.5% lower than someone who is either White or Hispanic.

a_h: This is a coefficient on the dummy variable HISP, which when HISP = 1, means the individual is Hispanic. The regression estimated a value on this coefficient of -0.0807345. The interpretation of this is that holding all other variables in the model equal, someone who is Hispanic will have a wage that is 8.07% lower than someone who is non-Hispanic.

Part two: Test the null hypothesis that, gender, union status education, and experience being held fixed, racial status has no effect on log wages. The variables a_n and a_h are the variables whose coefficients will determine the effect of racial status on log wages.

Testing the coefficient statistical significance individually:

The p value for the a_n coefficient is statistically significant using the 95% threshold, p = 0.0303 (which is < 0.05), meaning that for this coefficient, we *reject the null hypothesis* that gender, union status education, and experience being held fixed, racial status as captured by a_n has no effect on log wages.

The p value for the a_h coefficient is not statistically significant using the 95% threshold, p = 0.3582 (which is not < 0.05), meaning that for this coefficient, we *fail to reject the null hypothesis* that gender, union status education, and experience being held fixed, racial status as captured by a_h has no effect on log wages.

Joint statistical significance tests:

Because there are different conclusions for these two coefficients, the overall null hypothesis that racial status has no effect on log wages is still uncertain, thus, a joint hypothesis test is needed. From this hypothesis test, a p-value of p = 0.052 is reached, and thus *fail to reject the null hypothesis* that gender, union status education, and experience being held fixed, racial status as captured by a_h and a_n jointly has no effect on log wages.

Part three: Examining that the coefficients on NONWH and HISP are equal to each other, a separate linear hypothesis test (see “Linear Hypothesis Test to determine whether coefficients are equal to eachother”) revealed a p value of p = 0.6578, which indicates we *fail to reject the null hypothesis* that the coefficients are equal. From the final hypothesis test given in part two (see “Linear Hypothesis Test Examining the Joint Statistical Significance of Race Coefficients”), we *fail to reject the null hypothesis* that a_h and a_n are jointly equal to zero. A note on this hypothesis test, we fail to reject at the 95% significance level, but we reject at the 90% significance level which is potentially important.

b. Next sort your data set into three subsamples, according to whether the individual is nonwhite and not Hispanic (NONWH = 1), Hispanic (HISP = 1), or other (primarily whites, hereafter denoted OTHER). How many observations are there in each of the three subsamples? Calculate sample means for the variables LNWAGE, FE, UNION, ED, EX, and EXSQ separately for each of the three subsamples. Then for each of these variables, calculate differences in the sample means between the OTHER and NONWH samples and between OTHER and HISP. What is the mean difference in LNWAGE in the OTHER and NONWH samples? What is the mean difference in LNWAGE between OTHER and HISP? On average, do NONWH and HISP workers have less or more schooling than OTHER workers? What are the potential experience differences among these three sub-samples? Are there substantial gender (FE) differences by race?

Table 4: Mean Values for the Non-Hispanic and Non-White Subset of Data

	MeanValue
LNWAGE	1.9663298
FE	0.4179104
UNION	0.2686567
ED	12.6417910
EX	18.7910448
EXSQ	500.2238806

Table 5: Mean Values for the Hispanic Subset of Data

	MeanValue
LNWAGE	1.8185185
FE	0.4814815
UNION	0.1851852
ED	11.5185185
EX	17.0000000
EXSQ	497.2222222

Table 6: Mean Values for the “Other” Subset of Data

	MeanValue
LNWAGE	2.0880882
FE	0.4636364
UNION	0.1659091
ED	13.1681818
EX	17.7250000
EXSQ	464.4522727

Table 7: Difference in mean values for each variable between the “Other” and Hispanic Dataframes

	MeanValue
LNWAGE	0.2695697
FE	-0.0178451
UNION	-0.0192761
ED	1.6496633
EX	0.7250000

	MeanValue
EXSQ	-32.7699495

Table 8: Difference in mean values for each variable between the "Other" and non-White, non-Hispanic Dataframes

	MeanValue
LNWAGE	0.1217583
FE	0.0457259
UNION	-0.1027476
ED	0.5263908
EX	-1.0660448
EXSQ	-35.7716079

Part one: How many observations are there in each sample:

There are 440 observatoins for the "other" subset of the data, there are 27 observations for the Hispanic subset of the data, and there are 67 observations for the non-Hispanic and non-White subset of the data.

Part two: Calculate sample means for the variables LNWAGE, FE, UNION, ED, EX, and EXSQ separately for each of the three subsamples. These are reported in the tables above.

Part three: Calculating the mean difference between OTHER and NONWHITE, OTHER and HISPANIC for each variable These calculated results are reported in the tables above.

What is the mean difference in LNWAGE between OTHER and NONWH? **The mean difference is 0.1217583** What is the mean difference in LNWAGE between OTHER and HISP? **The mean difference is 0.269569**

On average, do NONWH and HISP workers have less or more schooling than OTHER workers? The difference in mean schooling for other versus Hispanic workers is 1.6496633, and the difference in mean schooling for other versus non-white workers is 0.526390. These numbers show that the NONWH and HISP workers *have less schooling* than OTHER workers.

What are the potential experience differences among these three sub-samples? Examining the difference in experience between the OTHER workers and those that are Hispanic, there is a 0.725 difference, so the OTHER workers have slightly more experience (less than a year). For the workers that are non-white or hispanic, the difference between OTHER worker experience and NONWH is -1.0660448, which shows that the NONWH workers have more experience (approximately a year's worth more) than the OTHER workers.

Are there substantial gender (FE) differences by race? The difference among these racial demographics in terms of gender are *nearly negligible* at -0.0178451 between OTHER and HISP and 0.0457259 between other and NONWH.

c. Using the sorted data by race from part (b), estimate parameters in separate NONWH. HISP, and OTHER equations, each of the form:

$$LNWAGE = \alpha + \alpha_f * FE + \alpha_u * UNION + \beta_1 * ED + \beta_2 * EX + \beta_3 * EXSQ + \epsilon$$

Now compare magnitudes of the estimated coefficients in the three equations, calculating the standard errors of differences in estimated coefficients between two subsamples using the formula in the instruction notes. In particular, is the effect of schooling (ED) on LNWAGE different between OTHER and NONWH? Between OTHER and HISP? Interpret these results. Does the gender-related (FE) wage differential vary significantly by race?

Table 9: Summary Statistics for Model Two: The non-White, non-Hispanic demographic

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4271196	0.3198104	1.335540	0.1866618
FE	-0.1449883	0.1137073	-1.275101	0.2071087
UNION	0.2816406	0.1247246	2.258101	0.0275282
ED	0.0824925	0.0215278	3.831906	0.0003031
EX	0.0458119	0.0156722	2.923131	0.0048561
EXSQ	-0.0007588	0.0003798	-1.998057	0.0501742

Table 10: Summary Statistics for Model Three: The Hispanic demographic

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6763156	0.3978822	1.6997882	0.1039365
FE	-0.4193190	0.2017993	-2.0779011	0.0501724
UNION	0.2846697	0.2474100	1.1505990	0.2628310
ED	0.0824417	0.0270854	3.0437701	0.0061713
EX	0.0316863	0.0232221	1.3644915	0.1868575
EXSQ	-0.0003960	0.0004359	-0.9083844	0.3739825

Table 11: Summary Statistics for Model Four: The “Other” (White) demographic

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5877189	0.1375061	4.274129	0.0000236
FE	-0.2341419	0.0427290	-5.479701	0.0000001
UNION	0.1881871	0.0576041	3.266906	0.0011736
ED	0.0926237	0.0091817	10.087844	0.0000000
EX	0.0343376	0.0061253	5.605842	0.0000000
EXSQ	-0.0005396	0.0001346	-4.009046	0.0000718

Table 12: Standard errors of difference between the White and non-Hispanic, non-White demographics

	x
(Intercept)	0.4573165
FE	0.1564363
UNION	0.1823286
ED	0.0307095
EX	0.0217975
EXSQ	0.0005144

Table 13: Standard errors of difference between the White and Hispanic demographics

	x
(Intercept)	0.5353884
FE	0.2445283
UNION	0.3050141
ED	0.0362671
EX	0.0293474
EXSQ	0.0005705

Formula (used in calculating the standard error of differences in estimated coefficients):

$$Var(\beta^* - \beta_*) = Var(\beta^*) + Var(\beta_*)$$

Interpretation: Refer to the table above for all the standard error of difference value estimates.

Evaluating the effect of *schooling* for the white versus nonwhite population, the coefficient estimates show that the effect is different for the two demographics; the marginal effect on percentage increase of wage on education for the white demographic is 1.02% higher than that of the non-White, non-Hispanic demographic (*Calculation:* 9.26237% - 8.24925%). Evaluating the effect of schooling for the white versus Hispanic demographics, the marginal effect on percentage increase of wage on education is also 1.02% higher for the white demographic than for the Hispanic demographic (*Calculation:* 0.926237% - 0.824417% x 100).

Evaluating the gender related wage differential, quantified by the FE coefficient, the following statements explain how these vary across race. Across all three demographics, there is a percentage decrease in wage associated with an individual being female in this dataset relative to if they were male. Comparing the white demographic to the non-white demographic, the wage decrease associated with being female and white is lower than that of being female and non-white by 8.9% (*Calculation:* -0.1449883 + 0.2341419 x 100). Comparing the white demographic to the Hispanic demographic, the wage decrease associated with being female and Hispanic is lower than that of being female and white by 18.5% (*Calculation:* 0.4193190 - 0.2341419 x 100).

These marginal effects illustrate that the gender related wage differential does vary significantly by race; furthermore, they illustrate that wage discrimination is even higher for those that fall into multiple categories of marginalized demographics, highlighting the importance of recognizing intersectional identities and how the discrimination that these individuals face is different compared to that of individuals who fall into a single marginalized identity.

d. Using a reasonable level of significance, perform a Chow test of the null hypothesis that all intercept and slope coefficients are simultaneously equal in the NONWH, HISP, and OTHER equations. To do this, first re-estimate the equation in part (a), omitting the NONWH and HISP variables, and retrieve the sum of squared residuals from this equation. Using the Chow test procedure (be careful in computing degrees of freedom), compare this sum of squared residuals with the sum from the three separate regressions in part (c). Interpret your results. Outline what additional information would be required in order to test the null hypothesis that parameters from the HISP sample are equal to those from the OTHER sample. Calculate and interpret results from such a test.

Table 14: Summary Statistics for Model Five

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5802548	0.1179364	4.920067	1.20e-06
FE	-0.2308973	0.0387206	-5.963168	0.00e+00
UNION	0.2021730	0.0504693	4.005860	7.07e-05
ED	0.0907418	0.0078944	11.494397	0.00e+00
EX	0.0344467	0.0053743	6.409498	0.00e+00
EXSQ	-0.0005243	0.0001181	-4.438159	1.11e-05

Table 15: Model Six: Evaluating whether coefficients are equal across models

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5877189	0.1376866	4.2685270	0.0000240
FE	-0.2341419	0.0427850	-5.4725194	0.0000001
HISP	0.0885967	0.4112638	0.2154254	0.8295322
UNION	0.1881871	0.0576797	3.2626242	0.0011869
ED	0.0926237	0.0091938	10.0746236	0.0000000
EX	0.0343376	0.0061334	5.5984955	0.0000000
EXSQ	-0.0005396	0.0001348	-4.0037917	0.0000728
FE:HISP	-0.1851771	0.2011522	-0.9205823	0.3577564
HISP:UNION	0.0964826	0.2477804	0.3893876	0.6971716
HISP:ED	-0.0101820	0.0279369	-0.3644631	0.7156815
HISP:EX	-0.0026513	0.0234348	-0.1131336	0.9099746
HISP:EXSQ	0.0001436	0.0004455	0.3223781	0.7473143

Sum of Square Residuals Comparison: Comparing the sum of square residuals in the model that includes the data from all demographics ($SS_{res} = 101.306$) to the sum of the sum of square residuals from each model split by demographic group ($SS_{res} = 99.27263$), the original- restricted model has a higher sum of square residuals than the unrestricted model.

Chow Procedure, F statistic:

Formula:

$$F = ((RSS_{unres} - (\Sigma RSS_{res})/(k_{unres} - k_{res}))/((\Sigma RSS_{res})/(n - 2 * k_{res}))$$

$$F = ((101.3061 - 99.27263)/(12))/((99.27263)/(522)) = 0.8910362$$

Interpretation: The F statistic: F = 0.8910362, which indicates that the specification under the unrestricted models do not constitute an improvement in fit over that of restricted model. In terms of the Chow procedure, the interpretation of this is that these models have similar performance in terms of their explanatory power for the factors that drive discrepancies in log wage across observations by racial demographic in the data.

Testing the null hypothesis that parameters from the HISP sample are equal to those from the OTHER sample No additional information is needed for either approach, as there are observations in the dataset for all variables in each demographic in question- white and Hispanic. The approach to accomplish this is by running a dummy variable regression, with interaction between a dummy and all dependent variables for HISP where HISP = 1 if the individual is Hispanic, else = 0. Examining model six, the p-values show that none of the estimates are statistically significant at any reasonable level of significance.

One consideration however, revisiting part c. the summary statistic tables contradict this, as the coefficient estimates between model 3 - Hispanic, and model 4- white, are different. To support this, an F stat comparing the model with an interaction term to the model without the interaction term, F = 0.645, which shows that including the dummy variables on the endowment terms does not improve the explanatory power of the model, and therefore, we fail to reject the null hypothesis that the marginal effects for the Hispanic demographic are equal to those of the white demographic.

e. Now implement the Blinder-Oaxaca procedures to measure wage discrimination by race. In particular, using sample mean data from part (b) and the Eq.(1) in the instruction notes , where OTHER is the advantaged group and NONWH is the disadvantaged group, decompose the difference in mean LNWAGE between the OTHER and NONWH groups into that portion due to differences in endowments (weight the differences in sample means of the regressors by the estimated coefficients from the OTHER equation) and the residual part due to discrimination by race. (Note that according to Eq.(1), the discrimination portion can be calculated indirectly as the above residual or directly as the difference in the estimated coefficients delta b weighted by mean values of the NONWH regressors.) What proportion of mean differences in LNWAGE between OTHER and NONWH is due to differences in endowments? Which endowment differences appear to be particularly important? What proportion of the mean differences in LNWAGE can be attributed to effects of discrimination by race? How do these proportions compare with the dummy variable procedure in part (a) of this exercise?

Table 16: Weights of each endownment driving wage discrepency between the White and non-White demographic

	FE	UNION	ED	EX	EXSQ
FE	-0.0107064	-0.0193358	0.0487562	-0.0366054	0.0193021

From the equation used under the Blinder-Oaxaca procedures to measure wage discrimination by race, the result measuring the percentage difference between the other (White) demographic and the non-White and non-Hispanic demographic is 13.8%, meaning that the observations in this dataset in the White demographic have a wage that is 13.8% *higher* than those that are non-White and non-Hispanic. Decomposing this into the wage effect driven by endowment and by race, the effect driven by endowment is 1.2%, The effect driven by racial discrimination then, is 98.8%. This illustrates that most of the wage discrepancy between the white and non-white demographic is driven by racial discrimination, not differences in the endowments of the variables used in the regression models.

Examining the table reported above, the effect that is particularly important, with a weight of -0.036, is experience. When examining all values in the table, experience is the variable that accounts for the largest share in terms of lower wages for the non-White, non-Hispanic demographic.

Drawing comparison to the dummy variable procedure that was run in part one, the effect of wage discrimination under this measure is very similar to that of the dummy variable procedure. The Blinder-Oaxaca procedure calculated that someone who is non-White and non-Hispanic will have a wage that is 13.8% lower on account of their race relative to someone who is White, whereas the dummy variable model measured this effect as someone who is non-White and non-Hispanic will have a wage that is 12.1% lower compared to someone who is white.

f. Now measure wage discrimination for Hispanics. Specifically, repeat procedures in part (e), but replace NONWH with HISP, thereby contrasting the HISP and OTHER groups. Compare your results with those obtained in part (e) in terms of relative endowments and the proportions of differences in LNWAGE attributable to differences in endowments and to discrimination. What can you say concerning the relative wage discrimination experienced by members of NONWH as compared to HISP? How do these proportions compare with the dummy variable procedure in part (a) of this exercise?

Table 17: Weights of each endownment driving wage discrepency between the White and Hispanic demographic

	FE	UNION	ED	EX	EXSQ
FE	-0.0107064	-0.0193358	0.0487562	-0.0366054	0.0193021

From the equation used under the Blinder-Oaxaca procedures to measure wage discrimination by race, the result measuring the percentage difference between the other (white) demographic and the Hispanic demographic is 17.3%, meaning that the observations in this dataset in the White demographic have a wage that is 17.3% *higher* than those that are Hispanic.

Decomposing this into the wage effect driven by endowment and by race, the effect driven by endowment is 72.68%, The effect driven by racial discrimination then, is 27.31%. This illustrates that a result different from that of the previous in part e, comparing the Hispanic and white demographics, most of the wage discrepancy between the white and non-white demographic is driven by differences in endowments, there is still difference in wage driven by racial discrimination, though a much lower proportion of total wage discrimination compared to that of the non-white, non-Hispanic demographic.

Examining the table reported above, the effect that is particularly important, with a weight of -0.036 is experience. When examining all values in the table, experience is the variable that accounts for the largest share in terms of lower wages for the Hispanic demographic.

Drawing comparison to the dummy variable procedure that was run in part one, the effect of wage discrimination under this measure is much higher than that of the dummy variable procedure. The Blinder-Oaxaca calculated that someone who is Hispanic will have a wage that is 17.3% lower on account of their race relative to someone who is White, whereas the dummy variable model, and measured this effect as someone who is Hispanic will have a wage that is 8.07% lower compared to someone who is white. This comparison illustrates that the two models are less consistent in their prediction of wage discrimination for the Hispanic demographic, whereas the two measures were very consistent when predicting wage discrimination for the non-white, non-Hispanic demographics.

g. The measures of discrimination by race computed in parts (e) and (f) of this exercise involve weighting differences in endowments by the estimated parameters of the advantaged (OTHER) group, as in Eq.(1) in the instruction notes. An alternative procedure, discussed in the instruction notes and summarized in Eq.(2), involves weighting differences in mean endowments by estimated parameters of the disadvantaged group. Using the estimated parameters from the NONWH equation from part (c) and OTHER-NONWH mean differences in endowments, substitute into Eq.(2) and compute this alternative measure of wage discrimination. What proportion of the mean difference in LNWAGE between OTHER and NONWH is now estimated as being due to differences in mean endowments? What portion represents wage discrimination by race? Compare these estimates to results obtained in part (e). Finally, use parameters from the estimated HISP equation, mean differences between OTHER and HISP in LNWAGE and endowments, and Eq.(2) to compute an alternative estimate of the effects of wage discrimination by race experienced by HISP individuals. Compare these new results to those obtained in part (f).

Table 18: Weights of each endownment driving wage discrepency between the White and non-White demographic (under the alternative equation)

	FE	UNION	ED	EX	EXSQ
FE	-0.0066297	-0.0289379	0.0434233	-0.0488375	0.0271433

Table 19: Weights of each endownment driving wage discrepency between the White and Hispanic demographic (under the alternative equation

	FE	UNION	ED	EX	EXSQ
FE	-0.0066297	-0.0289379	0.0434233	-0.0488375	0.0271433

What proportion of the mean difference in LNWAGE between OTHER and NONWH is now estimated as being due to differences in mean endowments?

The proportion of the mean difference in LNWAGE between OTHER and NONWH due to the difference in mean endowments is -11.3%. The interpretation of this is that under this alternative approach, endowments are associated with a higher wage for non-White, non-Hispanic people relative to White people (i.e., discrimination driven by endowments goes in the other direction). The proportion that represents wage discrimination by race is 111%. The interpretation is that all of the discrimination in wage for non–White, non-Hispanic people is driven by racial discrimination. Comparing these to the estimates given in part (e), (the effect driven by endowment is 1.1%, The effect driven by racial discrimination then, is 98.8%) the effect driven by endowment flipped signs, so non-White, non-Hispanic people realize a positive wage increase due to endowments, while the effect of race on wage is virtually the same, as in both models, nearly all wage discrepancy is driven by racial discrimination.

Decomposing this into the wage effect driven by endowment and by race under this alternative equation, the effect driven by endowment is 64.52%, The effect driven by racial discrimination then, is 35.47%. (In part f, percentages were: endowment = 72.68%, race = 27.31%). This illustrates that a result that is very similar to that of part f. In both equations, most of the wage discrepancy between the white and non-white demographic is driven by differences in endowments, there is still difference in wage driven by racial discrimination, though still a much lower proportion of total wage discrimination compared to that of the non-white, non-Hispanic demographic.

h. On the basis of these alternative measures of wage discrimination by race, what do you conclude concerning the relative importance of racial discrimination in the United States? Defend your conclusions.

To give a broad overview of the results, racial discrimination is present across multiple demographics and gender identities within this dataset.

The dummy variable model (reported in table 1) showed that there is a percentage wage decrease associated with being a part of the Hispanic demographic, the non-White, non-Hispanic demographic, and being female. It is also important to note that from this regression equation, the dummy variable on NONWH was statistically significant at the 95% threshold, though, the dummy variable on HISP was not statistically significant at any threshold, suggesting that wage discrimination is not as apparent for the Hispanic demographic as it is for the nonwhite, nonhispanic demographic. This is further supported later on under the Blinder-Oaxaca procedures.

Although wage discrimination associated with gender is not in the scope of this analysis, part c illustrated that intersectional identities (i.e., being female and Hispanic, or being female and non-White, non-Hispanic) was associated with larger wage decreases relative to those associated with being only one of these identities and not the other.

Using the Blinder-Oaxaca procedures, this analysis is able to parse out percentage decreases in wage that are driven by endowments (as specified in the variables female, union membership, education, and experience). From this portion of the analysis, it is clear that the wage discrimination against non-White, non-Hispanic people is driven almost entirely by racial discrimination (part e reported 99.8% of discrimination attributed to race, and part g reported 111% attributed to race). Conversely, wage discrimination against Hispanic people is more evenly split between that attributed to endowments and that attributed to race (in part f: endowment = 72.68%, race = 27.31% and in part g endowment = 64.52%, race = 35.48%).

Of course, this sample of 534 observations is not entirely representative across the United States population, wage discrimination varies by sector, and many other factors, so the external validity of this study may be called into question. Still, this analysis draws some important conclusions that are likely present currently.

3 Heteroskedasticity in the Statistical Earnings Function

The purpose of this exercise is to have you assess whether disturbances in an estimated statistical earnings function are homoskedastic, to compare traditional and robust estimates of standard errors of coefficients when heteroskedasticity may be present, and to examine the sensitivity of estimated coefficients to alternative stochastic specifications involving heteroskedasticity. Choose either the 1978 or the 1985 data set in CPS78 and CPS85, respectively, and use that data set for all portions of this exercise.

a. Begin by estimating a traditional statistical earnings function. More specifically, employing OLS, estimate parameters in the equation:

   $$LNWAGE = \alpha + \alpha_f * FE + \alpha_u * UNION + \alpha_n * NONWH + \alpha_h * HISP + \beta_1 * ED + \beta_2 * EX + \beta_3 * EXSQ + \epsilon$$

   Compute both the traditional and the heteroskedasticity-consistent standard errors. Are the heteroskedasticity-consistent standard error estimates always larger than the (inconsistent) OLS estimates? Is this what you expected? Why or why not?

Table 20: Traditional Standard Errors

	x
(Intercept)	0.1203640
FE	0.0386147
UNION	0.0504893
NONWH	0.0576528
HISP	0.0877940
ED	0.0079602
EX	0.0053782
EXSQ	0.0001181

Table 21: Heteroscedasticity Robust Standard Errors

	x
(Intercept)	0.1234638
FE	0.0391861
UNION	0.0458050
NONWH	0.0535959
HISP	0.0861886
ED	0.0082780
EX	0.0060174
EXSQ	0.0001305

Comparing the traditional and the heteroscedasticity-robust standard errors, the tables above show that the heteroscedasticity-robust standard errors are not always larger, the standard error for NONWH is larger for the inconsistent OLS estimate. This is not what I would have expected; I would have expected that the heteroscedasticity-robust standard errors would always be larger because we allow error terms to vary by values of independent variables.

b. Even though OLS estimated parameters in part (a) are consistent if heteroskedasticity is present, they are not efficient. To obtain efficient estimates, a generalized least squares (GLS) procedure is required. Since the form of heteroskedasticity is unknown, the FGLS procedure should be used here. Estimate the model using FGLS procedure. Compare your FGLS and OLS estimated parameters and standard errors. Any surprises? Why or why not?

Table 22: Summary Statistics for Ordinary Least Squares Regression Model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6246849	0.1203640	5.1899637	0.0000003
FE	-0.2317750	0.0386147	-6.0022549	0.0000000
UNION	0.2116322	0.0504893	4.1916240	0.0000325
NONWH	-0.1252402	0.0576528	-2.1723193	0.0302775
HISP	-0.0807345	0.0877940	-0.9195896	0.3582088
ED	0.0888187	0.0079602	11.1578618	0.0000000
EX	0.0344669	0.0053782	6.4086390	0.0000000
EXSQ	-0.0005270	0.0001181	-4.4630997	0.0000099

Table 23: Coefficient Estimates for Feasible Generalized Least Squares Regression Model:

	x
(Intercept)	0.6603784
FE	-0.2335230
UNION	0.2472426
NONWH	-0.1205256
HISP	-0.0789893
ED	0.0845969
EX	0.0370789
EXSQ	-0.0006059

Table 24: Standard Errors for Feasible Generalized Least Squares Regression Model:

	x
(Intercept)	0.1124680
FE	0.0373182
UNION	0.0451700
NONWH	0.0511313
HISP	0.0806091

	x
ED	0.0074886
EX	0.0057319
EXSQ	0.0001258

Comparing the FLGS parameters to the original OLS parameters there, is not a very large difference in the coefficient estimates. This is not very surprising; we know that heteroscedasticity does not bias coefficient estimates, so a model that accounts for heteroscedasticity should not change the coefficients very much relative to a model that does not do this.

Looking at the standard errors, we see that again, there is not a large difference in the standard errors between these two models. This is more surprising because standard error is calculated differently in the presence of heteroscedasticity, so I would think that the heteroscedasticity-robust standard errors, would all be larger than those in the model that is not heteroscedasticity-robust, though, comparing the two models, it is clear that this is not the case.

c. In typical econometric theory textbooks a number of tests are presented for testing the null hypothesis of homoskedasticity against an alternative hypothesis consisting of either a specific or some unspecified form of heteroskedasticity. One very simple test is that proposed by Halbert J. White [1980]; As in part (b), retrieve the residuals from the part (a) regression, and square them. White's procedure consists of running an auxiliary regression in which the squared OLS residual is the dependent variable and the regressors consist of the original set of regressors, plus the cross-products and squares of all the regressors in the original OLS equation. In our context this implies running a regression of the residuals squared on a constant, ED, EX, EXSQ, FE, UNION, NONWH, HISP, and 17 cross-products regressors, constructed as FE · UNION, FE · NONWH, FE · HISP, FE · ED, FE · EX, FE · EXSQ, UNION · NONWH, UNION · HISP, UNION · ED · UNION · EX, UNION · EXSQ, NONWH · ED, NONWH · EX, NONWH · EXSQ, HISP · ED, HISP · EX, HISP · EXSQ, and the two squared terms ED · ED, and EXSQ · EXSQ (note that squares of the dummy variables such as FE are identical to FE, and so they are not included as additional regressors). Run this auxiliary regression, and retrieve the R 2 measure. White has shown that if the original disturbances are homokurtic (that is, if the expected value of e^4_i is a constant), then under the null hypothesis, N (the sample size) times the R 2 from this auxiliary regression is distributed asymptotically as a chi-square random variable with 27 degrees of freedom (the total number of zero slope coefficients in the auxiliary regression under the null hypothesis). Compute this chi-square test statistic for homoskedasticity, and compare it to the 5% critical value. Are your results consistent with the null hypothesis of homoskedasticity? If not, make the appropriate adjustments and reestimate the equation in part (a) by GLS using a weighted least squares procedure. Does adjusting for heteroskedasticity affect the parameter estimates significantly? The estimated standard errors? The t-statistics of significance? Is this what you expected? Why?

Table 25: Summary Statistics for White's Procedure Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0192922	0.1478699	-0.1304671	0.8962484
FE	0.0845114	0.2083713	0.4055809	0.6852210
UNION	0.0057898	0.2896236	0.0199908	0.9840585
NONWH	-0.0904353	0.3118605	-0.2899863	0.7719448
HISP	-0.1406946	0.3412744	-0.4122624	0.6803207
ED	0.0147095	0.0099720	1.4750839	0.1408082
EX	0.0011080	0.0070125	0.1580032	0.8745169
EXSQ	0.0000793	0.0001570	0.5052048	0.6136337

	Estimate	Std. Error	t value	Pr(>|t|)
FE:UNION	0.0009960	0.0923110	0.0107893	0.9913958
FE:NONWH	0.0885735	0.1085206	0.8161910	0.4147727
FE:HISP	-0.0629449	0.1679495	-0.3747846	0.7079768
FE:ED	-0.0001725	0.0138791	-0.0124256	0.9900910
FE:EX	-0.0108847	0.0093480	-1.1643921	0.2448106
FE:EXSQ	0.0001334	0.0002047	0.6516804	0.5149015
UNION:NONWH	-0.0004474	0.1222195	-0.0036604	0.9970808
UNION:HISP	0.1524079	0.2196275	0.6939381	0.4880376
UNION:ED	0.0000403	0.0185054	0.0021773	0.9982636
UNION:EX	-0.0054690	0.0137354	-0.3981716	0.6906707
UNION:EXSQ	0.0000331	0.0002888	0.1145167	0.9088734
NONWH:ED	0.0034547	0.0209229	0.1651180	0.8689167
NONWH:EX	-0.0019469	0.0151463	-0.1285403	0.8977722
NONWH:EXSQ	0.0000715	0.0003654	0.1958080	0.8448386
HISP:ED	0.0049805	0.0231346	0.2152822	0.8296336
HISP:EX	0.0061818	0.0195417	0.3163378	0.7518758
HISP:EXSQ	-0.0000135	0.0003737	-0.0361932	0.9711425

The R2 of the model ran using White's procedure was R^2 = 0.033. Multiplying this by the sample size gives: 18.0452, which is less than the chi-square test statistic for homoscedasticity of 40.11327 at the 5% significance level. These results are *consistent* with the null hypothesis of homoscedasticity and thus we do not need to re-estimate the equation.

Does adjusting for heteroscedasticity affect the parameter estimates significantly? No, this adjustment does not affect the parameter estimates, which is what we would expect by the fact that coefficients are not biased in the presence of heteroscedasticity, so any procedure to account for this should not change the coefficient estimates in any way.

Examining the standard errors, the adjustment whereby we compute heteroscedasticity-robust standard errors will change them compared to those that are computed in a traditional manner (i.e., with the lm() function). This is because they are White's standard errors, the formula to calculate these is different, and they are generally (though not always) larger than those computed with lm().

Adjusting for heteroscedasticity will also affect the t-statistics of significance because part of the t-stat formula includes the standard error. Recall, the formula is $t = b_k - h_0/(se(b_k))$

```r
library(knitr)
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, fig.width = 4,
    fig.height = 4, tidy = TRUE)

# load packages:
library(car)
library(dplyr)
# load dataset:
library(readxl)
cps85 <- read_excel("cps85.xlsx")

# running the regression:
model_1 <- lm(LNWAGE ~ FE + UNION + NONWH + HISP + ED + EX + EXSQ, data = cps85)
sum_1 <- summary(model_1)
kable(coefficients(sum_1), caption = "Summary Statistics for Model One")

# ** linear hypothesis that both coefficients are equal to zero (i.e., racial
# status has no effect on log wages)
test1 <- linearHypothesis(model_1, c("NONWH=0", "HISP=0"), white.adjust = "hc1")
test_x <- linearHypothesis(model_1, c("NONWH=0", "HISP=0"))
kable(test1, caption = "Linear Hypothesis Test Examining the Joint Statistical Significance of Race Coe

# running a linear hypothesis test on the two coefficients on racial status
test2 <- linearHypothesis(model_1, "NONWH=HISP")
kable(test2, caption = "Linear Hypothesis Test to determine whether coefficients are equal to eachother
# for non-white, non-hispanic demographic:
columns <- c("LNWAGE", "FE", "UNION", "ED", "EX", "EXSQ")
mean_function <- function(cps85, columns) {
    subset_data <- cps85[cps85[["NONWH"]] == 1, columns, drop = FALSE]
    means <- colMeans(subset_data, na.rm = TRUE)
    result_df <- data.frame(MeanValue = means)
    return(result_df)
}
# call the function:
mean_values_cps85nonWH <- mean_function(cps85, columns)
# use kable to display the mean values:
kable(mean_values_cps85nonWH, caption = "Mean Values for the Non-Hispanic and Non-White Subset of Data"


# for Hispanic demographic:
columns <- c("LNWAGE", "FE", "UNION", "ED", "EX", "EXSQ")
mean_function <- function(cps85, columns) {
    subset_data <- cps85[cps85[["HISP"]] == 1, columns, drop = FALSE]
    means <- colMeans(subset_data, na.rm = TRUE)
    result_df <- data.frame(MeanValue = means)
    return(result_df)
}
# Call the function
mean_values_cps85hisp <- mean_function(cps85, columns)
# use kable to display the mean values:
kable(mean_values_cps85hisp, caption = "Mean Values for the Hispanic Subset of Data")

# for the White demographic:
columns <- c("LNWAGE", "FE", "UNION", "ED", "EX", "EXSQ")
mean_function <- function(cps85, columns) {
    subset_data <- cps85[cps85[["HISP"]] == 0 & cps85[["NONWH"]] == 0, columns, drop = FALSE]
    means <- colMeans(subset_data, na.rm = TRUE)
    result_df <- data.frame(MeanValue = means)
    return(result_df)
}
# Call the function
mean_values_cps85other <- mean_function(cps85, columns)
# use kable to display the mean values:
kable(mean_values_cps85other, caption = "Mean Values for the \"Other\" Subset of Data")

# since I wrote these as dataframes, no need to calculate the difference for
# each by hand,
DifferenceOTH_HISP <- mean_values_cps85other - mean_values_cps85hisp
kable(DifferenceOTH_HISP, caption = "Difference in mean values for each variable between the \"Other\"
DifferenceOTH_NonWH <- mean_values_cps85other - mean_values_cps85nonWH
kable(DifferenceOTH_NonWH, caption = "Difference in mean values for each variable between the \"Other\"

# running regressions for each of the data subsets:

# NONWH
model_2 <- lm(LNWAGE ~ FE + UNION + ED + EX + EXSQ, data = subset(cps85, NONWH ==
    1))
sum_2 <- summary(model_2)
kable(coefficients(sum_2), caption = "Summary Statistics for Model Two: The non-White, non-Hispanic dem

# HISPANIC
model_3 <- lm(LNWAGE ~ FE + UNION + ED + EX + EXSQ, data = subset(cps85, HISP ==
    1))
sum_3 <- summary(model_3)
kable(coefficients(sum_3), caption = "Summary Statistics for Model Three: The Hispanic demographic")

# OTHER
model_4 <- lm(LNWAGE ~ FE + UNION + ED + EX + EXSQ, data = subset(cps85, HISP ==
    0 & NONWH == 0))
sum_4 <- summary(model_4)
kable(coefficients(sum_4), caption = "Summary Statistics for Model Four: The \"Other\" (White) demograp

# calculating the calculating the standard errors of differences in estimated
# coefficients between two subsamples:

# OTHER versus NONWH

sd_oth <- sum_4$coefficients[, 2]
sd_nonwh <- sum_2$coefficients[, 2]
SEdiff_OTNW <- sd_oth + sd_nonwh
kable(SEdiff_OTNW, caption = "Standard errors of difference between the White and non-Hispanic, non-Whi

# OTHER versus Hispanic
sd_hisp <- sum_3$coefficients[, 2]
SEdiff_OTHISP <- sd_oth + sd_hisp
kable(SEdiff_OTHISP, caption = "Standard errors of difference between the White and Hispanic demographi



# regression that omits the NONWH and HISP variables
model_5 <- lm(LNWAGE ~ FE + UNION + ED + EX + EXSQ, data = cps85)
sum_5 <- summary(model_5)
kable(coefficients(sum_5), caption = "Summary Statistics for Model Five")

# Extracting the sum of square residuals from each model:
model_2ssRES <- sum(model_2$residuals^2)
model_3ssRES <- sum(model_3$residuals^2)
model_4ssRES <- sum(model_4$residuals^2)
models2to4_ssRES <- sum(model_2ssRES, model_3ssRES, model_4ssRES)
model5_ssRES <- sum(model_5$residuals^2)

# Code to calculate the Chow Test F-statistic:
```

```
n <- nrow(cps85)
# degrees of freedom numberator = df unrestricted - restricted
df_ur <- 3 * length(model_2$coefficients)
df_r <- length(model_5$coefficients)
Fstat_1 = ((model5_ssRES - models2to4_ssRES)/(df_ur - df_r))/(models2to4_ssRES/(n -
    2 * (df_r)))

# Code for testing the null hypothesis that parameters from the HISP sample are
# equal to those from the OTHER sample

# model 1- unrestricted:
model_6 <- lm(LNWAGE ~ (FE * HISP) + (UNION * HISP) + (ED * HISP) + (EX * HISP) +
    (EXSQ * HISP), data = subset(cps85, NONWH == 0))
sum_6 <- summary(model_6)
kable(coefficients(sum_6), caption = "Model Six: Evaluating whether coefficients are equal across models

model_res <- lm(LNWAGE ~ FE + UNION + ED + EX + EXSQ, data = subset(cps85, NONWH ==
    0))
sum_res <- summary(model_res)
newdf <- subset(cps85, NONWH == 0)

# F stat to evaluate significance
F_stat_2 <- ((0.3214 - 0.3157)/6)/((1 - 0.3214)/(467 - 6))


# gathering all of the values needed for calculation within this formula:

# difference in means for each variable, stored in: DifferenceOTH_NonWH
# difference in beta coefficients
coefficient_diff1 <- coefficients(model_4) - coefficients(model_2)

# difference in means:
diff_y <- DifferenceOTH_NonWH[1, 1]
diff_x <- sum((DifferenceOTH_NonWH[2, 1] * model_4$coef[2]), (DifferenceOTH_NonWH[3,
    1] * model_4$coef[3]), DifferenceOTH_NonWH[4, 1] * model_4$coef[4], DifferenceOTH_NonWH[5,
    1] * model_4$coef[5], DifferenceOTH_NonWH[6, 1] * model_4$coef[6])

# difference due to endowment
endowment1 <- diff_x/diff_y
# difference due to race
race1 <- 1 - diff_x/diff_y

# code to calculate create a table that shows the weight of each endowment:
endowment_table <- data.frame(FE = (DifferenceOTH_NonWH[2, 1] * model_4$coef[2]),
    UNION = (DifferenceOTH_NonWH[3, 1] * model_4$coef[3]), ED = (DifferenceOTH_NonWH[4,
        1] * model_4$coef[4]), EX = (DifferenceOTH_NonWH[5, 1] * model_4$coef[5]),
    EXSQ = (DifferenceOTH_NonWH[6, 1] * model_4$coef[6]))

kable(endowment_table, caption = "Weights of each endownment driving wage discrepency between the White


# gathering all of the values needed for calculation within this formula:
```

```
# difference in means for each variable, stored in: DifferenceOTH_HISP
# difference in beta coefficients
coefficient_diff2 <- coefficients(model_4) - coefficients(model_3)

# difference in means:
diff_y1 <- DifferenceOTH_HISP[1, 1]
diff_x1 <- sum((DifferenceOTH_HISP[2, 1] * model_4$coef[2]), (DifferenceOTH_HISP[3,
    1] * model_4$coef[3]), DifferenceOTH_HISP[4, 1] * model_4$coef[4], DifferenceOTH_HISP[5,
    1] * model_4$coef[5], DifferenceOTH_HISP[6, 1] * model_4$coef[6])

# difference due to endowment
endowment2 <- diff_x1/diff_y1
# difference due to race
race2 <- 1 - diff_x1/diff_y1

# code to calculate create a table that shows the weight of each endowment:
endowment_table2 <- data.frame(FE = (DifferenceOTH_NonWH[2, 1] * model_4$coef[2]),
    UNION = (DifferenceOTH_NonWH[3, 1] * model_4$coef[3]), ED = (DifferenceOTH_NonWH[4,
        1] * model_4$coef[4]), EX = (DifferenceOTH_NonWH[5, 1] * model_4$coef[5]),
    EXSQ = (DifferenceOTH_NonWH[6, 1] * model_4$coef[6]))

kable(endowment_table, caption = "Weights of each endownment driving wage discrepency between the White



# gathering all of the values needed for calculation within this formula:

# difference in means for each variable, stored in: DifferenceOTH_NonWH
# difference in beta coefficients
coefficient_diff1 <- coefficients(model_4) - coefficients(model_2)

# difference in means:
diff_y <- DifferenceOTH_NonWH[1, 1]
diff_x <- sum((DifferenceOTH_NonWH[2, 1] * model_2$coef[2]), (DifferenceOTH_NonWH[3,
    1] * model_2$coef[3]), DifferenceOTH_NonWH[4, 1] * model_2$coef[4], DifferenceOTH_NonWH[5,
    1] * model_2$coef[5], DifferenceOTH_NonWH[6, 1] * model_2$coef[6])

# difference due to endowment
endowment3 <- diff_x/diff_y
# difference due to race
race3 <- 1 - diff_x/diff_y

# code to calculate create a table that shows the weight of each endowment:
endowment_table <- data.frame(FE = (DifferenceOTH_NonWH[2, 1] * model_2$coef[2]),
    UNION = (DifferenceOTH_NonWH[3, 1] * model_2$coef[3]), ED = (DifferenceOTH_NonWH[4,
        1] * model_2$coef[4]), EX = (DifferenceOTH_NonWH[5, 1] * model_2$coef[5]),
    EXSQ = (DifferenceOTH_NonWH[6, 1] * model_2$coef[6]))

kable(endowment_table, caption = "Weights of each endownment driving wage discrepency between the White

# difference in means for each variable, stored in: DifferenceOTH_HISP
# difference in beta coefficients
coefficient_diff2 <- coefficients(model_4) - coefficients(model_3)
```

```r
# difference in means:
diff_y1 <- DifferenceOTH_HISP[1, 1]
diff_x1 <- sum((DifferenceOTH_HISP[2, 1] * model_3$coef[2]), (DifferenceOTH_HISP[3,
    1] * model_3$coef[3]), DifferenceOTH_HISP[4, 1] * model_3$coef[4], DifferenceOTH_HISP[5,
    1] * model_3$coef[5], DifferenceOTH_HISP[6, 1] * model_3$coef[6])

# difference due to endowment
endowment4 <- diff_x1/diff_y1
# difference due to race

race4 <- 1 - diff_x1/diff_y1

# code to calculate create a table that shows the weight of each endowment:
endowment_table2 <- data.frame(FE = (DifferenceOTH_NonWH[2, 1] * model_3$coef[2]),
    UNION = (DifferenceOTH_NonWH[3, 1] * model_3$coef[3]), ED = (DifferenceOTH_NonWH[4,
        1] * model_3$coef[4]), EX = (DifferenceOTH_NonWH[5, 1] * model_3$coef[5]),
    EXSQ = (DifferenceOTH_NonWH[6, 1] * model_3$coef[6]))

kable(endowment_table, caption = "Weights of each endownment driving wage discrepency between the White




# packages to install: install.packages('fixest', repos =
# 'http://cran.us.r-project.org')
library(fixest)

# code to compute the traditional standard errors:
model_11 <- lm(LNWAGE ~ FE + UNION + NONWH + HISP + ED + EX + EXSQ, data = cps85)
sum_11 <- summary(model_11)
se_11 <- sum_11$coefficients[, 2]
kable(se_11, caption = "Traditional Standard Errors")

# code to compute the heteroscedasticity-consistent standard errors:
model_12 <- feols(LNWAGE ~ FE + UNION + NONWH + HISP + ED + EX + EXSQ, data = cps85,
    "hetero")
sum_12 <- summary(model_12)
se_12 <- sum_12$se
kable(se_12, caption = "Heteroscedasticity Robust Standard Errors")

# step one: regress y on x, this is done already- ** should we use the feols
# model that is robust, or the inconsistent model?
model_12 <- feols(LNWAGE ~ FE + UNION + NONWH + HISP + ED + EX + EXSQ, data = cps85,
    "hetero")
mod_uhat <- model_12$residuals
mod_squhat <- log((mod_uhat)^2)
mod_fitted <- model_12$fitted.values

# step two: regress ln(uhat^2) on x saving the fitted values
model_13 <- lm(mod_squhat ~ FE + UNION + NONWH + HISP + ED + EX + EXSQ, data = cps85)
g_fitted <- model_13$fitted.values

# step three: calculate h_hat, these are the weights that we use for the
# regression of y on x.
cps85$h_hat <- exp(g_fitted)

# last, we can run the regression, with 1/h_hat as the weights:
model_14 <- feols(LNWAGE ~ FE + UNION + NONWH + HISP + ED + EX + EXSQ, data = cps85,
    "hetero", weights = (1/cps85$h_hat))
sum_14 <- summary(model_14)
kable(coefficients(sum_1), caption = "Summary Statistics for Ordinary Least Squares Regression Model:")

kable(coefficients(sum_14), caption = "Coefficient Estimates for Feasible Generalized Least Squares Reg
kable(sum_14$se, caption = "Standard Errors for Feasible Generalized Least Squares Regression Model:")


# square residuals
cps85$sqres <- (model_1$residuals^2)
model_15 <- lm(sqres ~ FE + UNION + NONWH + HISP + ED + EX + EXSQ + (FE * UNION) +
    (FE * NONWH) + (FE * NONWH) + (FE * HISP) + (FE * ED) + (FE * EX) + (FE * EXSQ) +
    (UNION * NONWH) + (UNION * HISP) + (UNION * ED) + (UNION * EX) + (UNION * EXSQ) +
    (NONWH * ED) + (NONWH * EX) + (NONWH * EXSQ) + (HISP * ED) + (HISP * EX) + (HISP *
    EXSQ) + (ED^2) + (EX^2), data = cps85)
sum_15 <- summary(model_15)
kable(coefficients(sum_15), caption = "Summary Statistics for White's Procedure Model")

# compute sample size * R^2
test_statistic <- nrow(cps85) * sum_15$r.squared


# computing the chi-square test statistic: qnorm(p = 0.05/2, mean = 0, sd = 1)
# qnorm(p = 0.05/2, mean = 0, sd = 1, lower.tail = FALSE) # get chi-squared 5%
# critical value chi <- qchisq(p=0.05, df=27, lower.tail=FALSE)
```