An In-depth Examination of the Racial Income Gap among Subgroups

Thomas Redding, Yichen Shen

Introduction

While many people believe we are entering a "post-racial" era, many sociologists continue to insist that their research continues to show racially based biases. The census figure still shows significant income gap between different ethnicity groups. However to conclude that racism still pervades American society just based on a simple mean difference is overly simplistic.

There are many previous studies addressing racial discrimination in the employment, as discrimination in employment is potentially a big source of the income gap between different subgroups of population. Gottschalk (1997) indicate that in 1980 and 1990 black men in the United States were suffering a 12 to 15 percent loss in earnings due to labor market discrimination. This discrimination is found in many aspects of employment, including the hiring process (M Bertrand2004) and merit-based reward systems (Castilla 2008).

However, the regional and sex differences of discrimination against ethnicity subgroups are far less studied in past literatures. On the other hand, more recent and comprehensive data sets allow us to draw more recent conclusions that are comparable to previous studies.

This study serves the purpose of filling this gap of literatures as well as trying to do a more up-to-date analysis about income gap and discrimination. In this study we, after controlling for a myriad of other relevant variables, want to determine whether race could

interact with someone's region, sex, or parental income in predicting a child's future income. In this study, we decided that, although both measures are probably affected by a child's race, we would adjust for standardized test score and highest grade completed (in addition to other variables), because we wanted to focus on the effects a person's race has *after* entering the labor force.

The answer to this question could significantly influence our understanding of which group of blacks are most affected by racism and, therefore, point to subgroups that may be discriminated against more than blacks as a group are, thereby giving a more precise direction to future analysis and awareness of racial issues.

Methods

Our data was collected by the U.S. Bureau of Labor Statistics between the years of 1997 and 2011. The age of all the cases varied between 14 and 17 on the original collection date and therefore between 28 and 31 during the last sampling period. The cases were selected to be representative of the US population as a whole, but with an oversampling of Hispanic children and non-Hispanic black children. For the purposes of this study, we will assume that this will not significantly bias in our results.

Although the data collected by the study contained literally thousands of variables, we restricted ourselves to what we considered the most relevant due to both model simplicity and also because most of the variables were missing some values and so, using too many variables would have significantly shrunk our sample size. Our independent variable was the case's income in 2011.

The explanatory variables we considered were the case's year of birth, sex, income in 1997, whether s/he lived in a urban or rural environment, percentile on the ASVAB

Math-Verbal Test (which we take as a proxy for intelligence)[1], highest grade completed (in years) as of 2011, an ethnicity factor variable with levels Hispanic, Black, Mixed Race, and Other (Non-Black, Non-Hispanic), a factor variable indicating which region the case was born in, and a factor variable indicating the familial structure (e.g. both biological parents in household, single biological mom in household, married biological mom in household, etc.). Additionally, we included some variables about the case's mother (variables about the father were excluded to the numerous missing values): her age at her first birth (in years), her age at the case's birth (in years), the highest grade she completed (in years), and her self-reported parenting style (uninvolved, permissive, authoritarian, or authoritative).

Because examining family structure was not the primary purpose of this study and because this vastly simplified the model, we condensed the 10 levels into 3: whether both biological parents were present in the household, whether one biological parent was present, or whether neither biological parent was present.

Results

Adjusting the Data Set

We were missing values for several variables, notably the income values for 2011 and 1997. Additionally, due to the collection method of capping the incomes variables at the 98th percentile, we restricted the income range to below 200,000. After these corrections, we retained. Also, to eliminate the problem of a large number of reported $0 incomes, we restricted the range to positive incomes. We presumed that those reporting incomes of $0 received some kind of economic assistance; therefore, $0 was most likely an inaccurate representation of their true income. Therefore, overall, we looked only at cases

[1] this percentile was multiplied by 1000 to obtain an integer from 0 to 100,000

with positive incomes less than 200,000 in both 1997 and also 2011. Thus, any conclusions we draw must be restricted to the annual income range of $1 to $199,999. After removing those cases, we were left with a sample size of 4,460 out of an original 8,984. Less problematically, though still relevantly, we were missing roughly 1,390 values in the other explanatory variables, leaving us with a final 3,070 cases

Variable Overview

Exploratory plots of the data suggested that the income variables were right skewed. We attempted a logarithmic transformation; however, we found this overcorrected the right skewness. We found a square-root transformation of both variables largely eliminated the skewness. Additionally, the square-root transformation produced a linear relationship between income in 1997 and income in 2011 when a least squares linear regression model was applied.

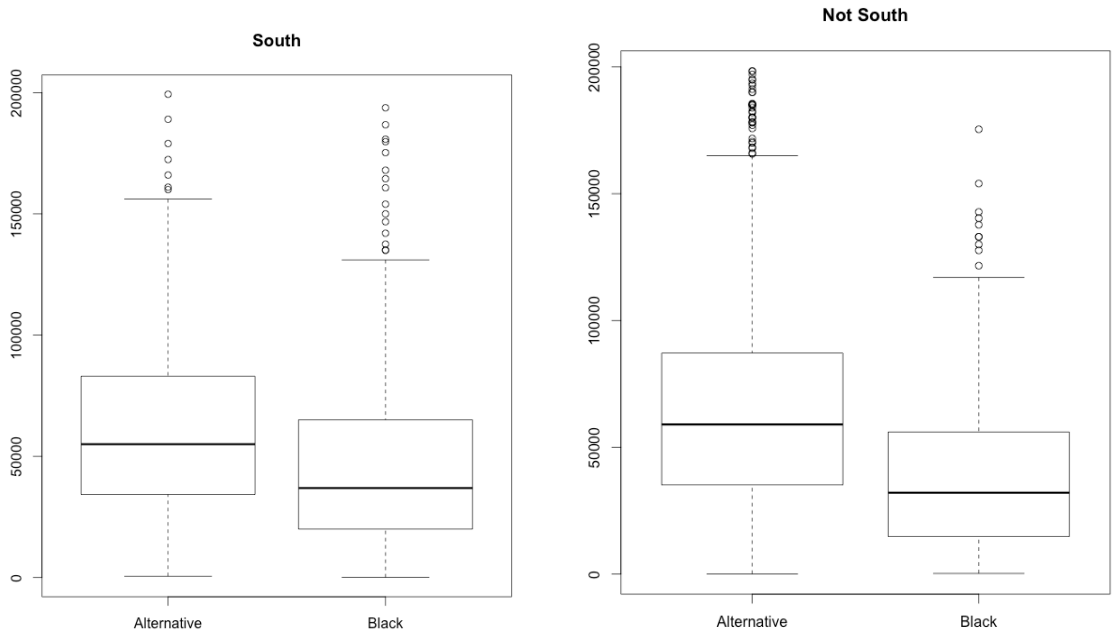


Figure 1: Income distributions by race and region

The other variables are summarized by the tables below:

Table 1 Summaries for 2011 income, 1997 income, year of birth, age of mother at first birth, age of mother at case's birth, household size as a child, mother's highest grade completed, mother's parenting style, standardized test score, and highest grade completed

Variable	Min	1st Qu	Median	Mean	3rd Qu	Max
Income11	10	$30000	$52750	$59626	$82000	$199330
Income97	5	$21025	$40500	$45097	$62000	$162501
BirthYear	1980	1981	1982	1982	1983	1984
MomAgeFirst	1	19	22	22.9	26	40
MomAgeCase	12	22	25	25.51	29	51
HouseholdSize	2	4	4	4.418	5	11
MotherEducation	1	12	12	12.81	14	95
MotherParentingStyle	1	2	3	2.855	4	4
Intelligence	0	22374	47096	48235	73419	100000
HighestGrade	6	12	14	14	16	95

Table 2 Summaries for region of residence, ethnicity, childhood householdType, sex, and whether the case lived in an urban or rural area

Region		Ethnicity		HouseholdType		Sex		Urban	
NorthCentral	816	Black	693	BothBio	1702	M	1490	Urban	2243
NorthEast	482	Hispanic	581	OneBio	1351	F	1580	Rural	827
South	1099	MixedRace	25	Other	17	-	-	-	-
West	673	Alternative	1771	-	-	-	-	-	-

Assumptions

Our most important assumption is that the people who either were not interviewed or refused to report specific variables were randomly distributed. We also need to assume independence of the cases, because they were not completely randomly selected from the population, but were instead selected to represent the US population in 1997 as a whole, with a deliberate oversampling of Hispanics and non-Hispanic blacks. Furthermore, we assume that the relationships between the square root of the two income variables and the

intelligence variable are linear. We also assume that variance in the square root of income in 2011 is constant across all observations. Finally, we assume that in the original survey done in 1997, all of the cases were living with at least one legal guardian who was responsible for the vast majority of the households income. This assumption is probably fairly accurate, given that all cases were between the ages of 14 and 17 at the time.

First, we fit a regression including all possible explanatory variables except race and region. We then eliminated the variables with statistically insignificant coefficients resulting in a simplified model.[2] We used a p-value of 0.01 throughout our analysis to determine statistical significance.

Next, we added ethnicity, region, and an interaction term between the two. We found that only the black category was significant in both the ethnicity term and the interaction terms, so we combined all other ethnicities into a base ethnicity. Next, we found that the Northeast's and Wests' coefficients were not statistically significant, nor were their interaction terms with ethnicity. Therefore, we combined Northeast, West, and NorthCentral (base factor) to create a Region factor variable with values "South" and "not South."

Next, we added sex-race and income-race interaction terms and found that they were statistically significant. Ultimately, then, our final model was:

[2] This model included the following explanatory variables: sqrt(Income97), BirthYear, MomAgeCase, HouseholdSize, MotherEducation, Intelligence, HighestGrade, HouseholdType

$$E\left[\sqrt{Income11}\right] = \beta_0 + \beta_1\sqrt{Income97} + \beta_2 blac\square + \beta_3 BirthYear + \beta_4 Intelligence$$
$$+ \beta_5 HighestGrade + \beta_6 HouseholdType + \beta_7 Region + \beta_8 Sex$$
$$+ \beta_9 black\sqrt{Income97} + \beta_{10} black \cdot Region + \beta_{11} black \cdot Sex$$

Table 3 Estimates of the coefficients for the regression model

	Estimate	Std. Error	t value	P-value
(Intercept)	7207	1932	3.73	0.000194
sqrt(Income97)	0.0988	0.0239	4.13	0.0000373
black: Black	-26.7	13.9	-1.92	0.0552
BirthYear	-3.57	9.74	-3.67	0.000250
Intelligence	0.000258	0.0000613	4.21	0.0000258
HighestGrade	5.77	0.616	9.37	< 2-16
HouseholdType: OneBio	-12.0	3.09	-3.90	0.0000993
HouseholdType: Other	-32.7	18.3	-1.787	0.0741
Region: South	-5.22	3.45	-1.51	0.131
Sex	2.84	3.09	0.918	0.359
sqrt(Income97)*black	0.131	0.0443	2.96	0.00309
black*Region	18.8	6.92	2.72	0.00660
black*Sex	-20.0	6.52	-3.06	0.00222

Although we found no influential outliers by any measure, we removed the six most deviant points and found no change in significance levels. The most likely reason that three of these points were abnormal was because the mothers reported an education of over 30 years – no other mothers did so. We ultimately retained this model, because the model with these deviants points had a very low p-value under the Breusch-Pagan Non-Constant Variance Test, indicating that there was a possibility of non-constant variance, where as the model without these points showed much less evidence of doing so according to the Breusch-Pagan Non-Constant Variance Test and visual analysis (see below).

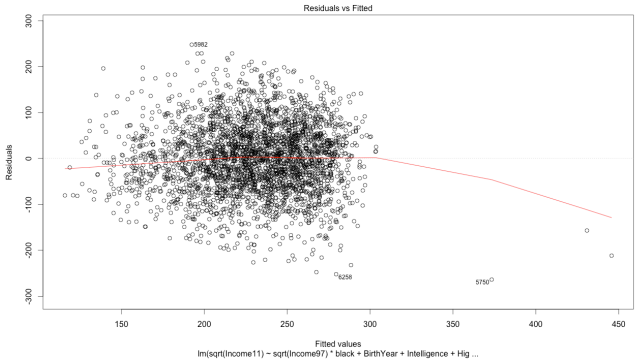


Figure 2 Residuals plot may have non-constant Variance

Because of the square-root transformation of income in 2011, it is very difficult for us to give intuitive interpretations of any of the coefficients. However, it may be helpful to analyze the expected incomes in some carefully selected situations.

For instance, the expected racial gap between two people born in 1982 who both grew up with both biological parents, who earned an income of $40,500, both scored at the 50th percentile on the standardized test, and both had 14 years of education is[3]

Table 4 Expected racial income gap for subgroups after controlling for various factors including a parental income of $40,500

	Not South	South
Male	$9,451	$725
Female	$18,158	$9,867

Given the same households described above but with parental incomes of $20,000, we find the following expected income gaps

Table 5 Expected racial income gap for subgroups after controlling for various factors including a parental income of $20,000

	Not South	South
Male	$12,541	$4,250
Female	$20,743	$12,887

[3] Positive values indicate that the white case is expected to earn more than the black case

However, when comparing the confidence intervals for white expected income and black expected income, the top of the black 95% confidence intervals were always smaller than the bottom of the white 95% confidence intervals except for males in the South. The table below shows the intervals used when calculating Table 4, but the results for Table 5 are similar.

Table 6 Confidence intervals for various subgroups after controlling for all other factors

	Not South	South
Male, Black	($44164, $54959)	($50768, $60721)
Male, White	($56375, $61410)	($53123, $59689)
Female, Black	($37301, $47177)	($43692, $52192)
Female, White	($57653, $62908)	($54416, $61111)

It should be clear that the racial income gap after controlling for these variables appears smaller in the South and greater for lower incomes.

Unfortunately, due to the large amount of variability that remains unexplained by any of the variables included in our model, prediction intervals have very little practical meaning. For example, for a black male born in the South in 1982 who all grew up with both biological parents, who earned an income of $40,500, all scored at the 50th percentile on the standardized test, and all had 14 years of education, their predicted income intervals with 95% confidence ranges from $7,907 to $146,544.

Discussion

After controlling for parental income, intelligence, highest grade completed, household type, and sex, it appears that blacks in the South are expected to earn more than their black, non-Southern counterparts. Providing some evidence to counter the belief that

racism is worse in the South. Likewise, after controlling for intelligence, highest grade completed, household type, sex, and region, race appears to be less important as parental income increases, suggesting that, if one attributes income gaps between races to racism after controlling for various relevant factors, racism may be a significantly worse problem among the poor and working class.

The interaction term between race and sex implies that the difference in expected earnings between nonblack males and black females is greater than the combined gaps between nonblack males and nonblack females and the gap between nonblack females and black females after controlling for intelligence, highest grade completed, household type, and parental income, and region.

We found it quite surprising that the sex and race terms were statistically insignificant. While further investigation is needed to determine why we came to this particular conclusion, we believe that the two most likely explanations are that our exclusion of cases missing data may not have been randomly distributed or that much of the income gap between races is due to unemployment differences rather than wage differences among the employed (given that we considered only those with positive incomes), which could cast doubt on the entire study.

It is also crucial to remember that we are analyzing income gaps *after* controlling for standardized test scores and highest grade level completed, both of which may be affected by discrimination. Thus, any conclusions drawn from our analysis of race and its interactions should keep in mind that we focused exclusively on income gap developments after the cases entered the workforce.

Future studies could improve upon our conclusions by extending more effort to obtain complete rather than large data sets to remove this potential bias. Moreover, this could possibly allow for the averaging of several years of income to remove annual variations within individual cases. Most importantly though, our findings suggest that future analyses of economic inequality due to race consider region, sex, and income level, as these all could be very significant factors in how much power race holds.

Appendix A

PubID	Sex	BirthYear	MomAgeFirst	MomAgeCase	Income97	HouseholdSize	Urban	MotherEducation
1	2	1981	19	26	-3	6	1	8
2	1	1982	19	19	0	4	1	15
3	2	1983	26	26	63000	2	1	12
4	2	1981	20	33	11700	2	1	12
5	1	1982	34	34	-3	4	1	12
6	2	1982	21	25	-3	5	1	12
7	1	1983	21	26	-3	5	1	12
8	2	1981	36	36	-3	5	1	12
9	1	1982	36	37	-3	5	1	12
10	1	1984	36	38	-3	5	1	12

PubID	MotherParentingStyle	Income11	Intelligence	HighestGrade	Ethnicity Factor	Region	HouseholdType
1	4	50000	45070	16	Alternative	Northeast	BothBio
2	4	81000	58483	14	Hispanic	Northeast	OneBio
3	2	150250	27978	14	Hispanic	Northeast	OneBio
4	4	-3	37012	13	Hispanic	Northeast	OneBio
5	3	130000	-4	12	Hispanic	Northeast	BothBio
6	2	55000	22001	14	Hispanic	Northeast	OneBio
7	4	14766	3585	11	Hispanic	Northeast	OneBio
8	2	66750	-4	18	Alternative	Northeast	BothBio
9	2	110000	-4	18	Alternative	Northeast	BothBio
10	4	-5	-4	-5	Alternative	Northeast	BothBio

Appendix B

All data used in this study came from the following source:

"NLS Investigator." *Investigator*. Bureau of Labor Statistics, n.d. Web. 8 June 2014. <https://www.nlsinfo.org/investigator/pages/search.jsp>.

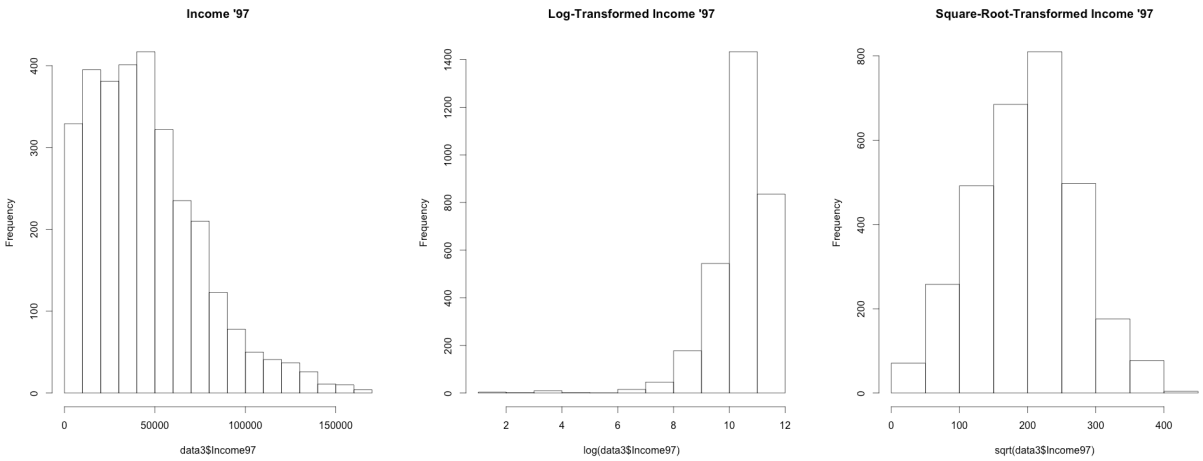
Cited Literature

Bertrand, Marianne, and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. No. w9873. National Bureau of Economic Research, 2003.

Castilla, Emilio J. "Gender, race, and meritocracy in Organizational careers." Academy of Management Proceedings. Vol. 2005. No. 1. Academy of Management, 2005.

Gottschalk, Peter. "Inequality, income growth, and mobility: The basic facts." Journal of Economic Perspectives 11 (1997): 21-40.

Appendix C



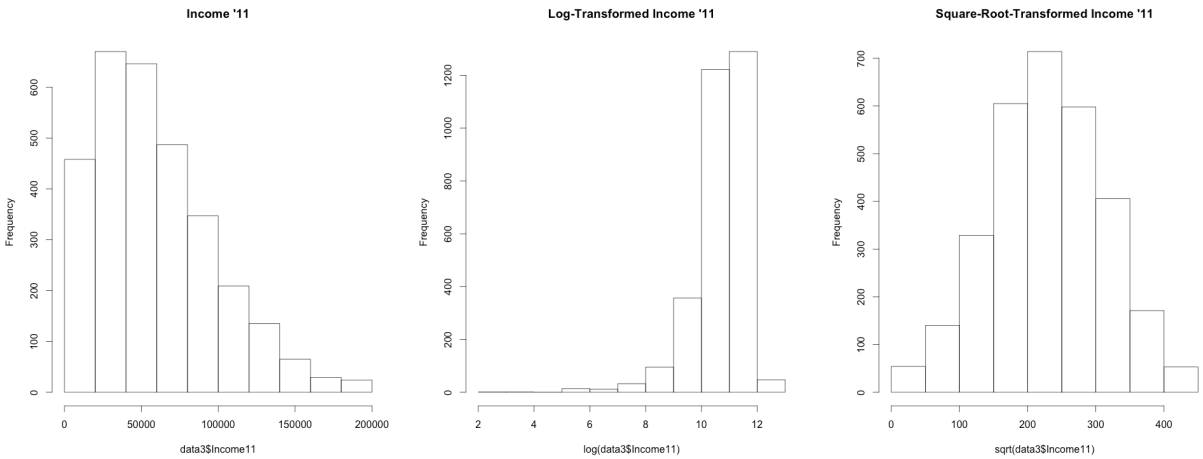


Figure 3 Distribution of incomes in 1997 and 2011: no transformation (left), log transformation (middle), and square root transformation (right)

Appendix D

```
# throw out all cases without income data, with incomes below or equal to zero, and/or
with incomes greater or equal to 200,000
# this restricts our conclusions to households with positive incomes below 200,000 and
adds a source of error (the deletion of nonreporting households)
index <- which(data$Income97>0 & data$Income97<200000 & data$Income11>0 &
data$Income11<200000)
data2 <- data[index,]

boxplot(Income11~EthnicityFactor, data=data2)
library(lattice)
xyplot(Income11 ~ Income97 | Sex, data = data2, panel = function(x,y){panel.xyplot(x,y);
panel.lmline(x,y)})
xyplot(Income11 ~ Income97 | EthnicityFactor, data = data2, panel =
function(x,y){panel.xyplot(x,y); panel.lmline(x,y)})
xyplot(Income11 ~ Intelligence | EthnicityFactor, data = data2, panel =
function(x,y){panel.xyplot(x,y); panel.lmline(x,y)})


# check income distributions (expected to be skewed)
```

```
hist(data2$Income97)
hist(log(data2$Income97))
hist(sqrt(data2$Income97))
hist(data2$Income11)
hist(log(data2$Income11))
hist(sqrt(data2$Income11))

# calculate income-only model
model1 <- lm(sqrt(Income11)~sqrt(Income97), data=data2)
summary(model1)

# eliminate cases which did not answer HouseholdType question (about 1390 such cases)
data2$HouseholdType2 <- relevel(data2$HouseholdType, ref = "BothBio")
index <- which(data2$HouseholdType2!="-" & data2$Urban!=2 & data2$Intelligence>=0 &
data2$MomAgeFirst>=0 & data2$MomAgeCase>=0 & data2$MotherParentingStyle>=0 &
data2$MotherEducation>=0 & data2$HighestGrade>=0)
data3 <- data2[index,]

# make sure income distribution is still skewed
hist(data3$Income97, main="Income '97")
hist(log(data3$Income97), main="Log-Transformed Income '97")
hist(sqrt(data3$Income97), main="Square-Root-Transformed Income '97")
hist(data3$Income11, main="Income '11")
hist(log(data3$Income11), main="Log-Transformed Income '11")
hist(sqrt(data3$Income11), main="Square-Root-Transformed Income '11")

boxplot(data3$Income11~data3$black, subset=which(data3$Region2=="South"),
main="South", maxY=100000)
boxplot(data3$Income11~data3$black, subset=which(data3$Region2=="NorthCentral"),
main="Not South")

# model with all explanatory variables included
model2 <-
lm(sqrt(Income11)~sqrt(Income97)+Sex+BirthYear+MomAgeFirst+MomAgeCase+Househ
oldSize+Urban+MotherEducation+MotherParentingStyle+Intelligence+HighestGrade+Hous
eholdType2, data=data3)
summary(model2)

model3 <-
lm(sqrt(Income11)~sqrt(Income97)+BirthYear+MomAgeCase+HouseholdSize+MotherEdu
cation+Intelligence+HighestGrade+HouseholdType2, data=data3)
summary(model3)

model4 <-
lm(sqrt(Income11)~sqrt(Income97)+BirthYear++Intelligence+HighestGrade+HouseholdT
ype2, data=data3)
```

```
summary(model4)
anova(model3,model4)

# p-value = 0.0173 > 0.01, so we kept the simpler model (model4)

# add Ethnicity & Region Interaction
model5 <-
lm(sqrt(Income11)~sqrt(Income97)+BirthYear+Intelligence+HighestGrade+HouseholdTy
pe2+EthnicityFactor*Region, data=data3)
summary(model5)
anova(model4,model5)

# combine all ethnicities except black
library(reshape)
data3$black <- combine_factor(data3$EthnicityFactor, c(0,1,0,0))
model6 <-
lm(sqrt(Income11)~sqrt(Income97)+BirthYear+Intelligence+HighestGrade+HouseholdTy
pe2+black*Region, data=data3)
summary(model6)

# combine Northeast, Midwest, and West
data3$Region2 <- combine_factor(data3$Region, c(0,0,1,0))
model7 <-
lm(sqrt(Income11)~sqrt(Income97)+BirthYear+Intelligence+HighestGrade+HouseholdTy
pe2+black*Region2, data=data3)
summary(model7)

anova(model7,model6)

# Keep the simpler model (model7)
library(car)
vif(model7)

model8 <-
lm(sqrt(Income11)~sqrt(Income97)+BirthYear+Intelligence+HighestGrade+HouseholdTy
pe2+black*Region2+black*Sex, data=data3)
summary(model8)

model9 <-
lm(sqrt(Income11)~sqrt(Income97)*black+BirthYear+Intelligence+HighestGrade+Househ
oldType2+black*Region2+black*Sex, data=data3)
summary(model9)

anova(model7,model9)

# p-value = 0.0001747<0.01, so we keep the more complex model (model9)
```

```
plot(model9)
plot(model9, which=1)
ncvTest(model9)
index <- which(cooks.distance(model11)<0.04 & fitted(model11)<350)
model9b <-
lm(sqrt(Income11)~sqrt(Income97)*black+BirthYear+Intelligence+HighestGrade+Household
oldType2+black*Region2+black*Sex, data=data3, subset=index)
summary(model9b)
summary(model9)
plot(model9b)
plot(model9b, which=1)
ncvTest(model9b)

# we removed 6 potential outliers to improve the constant-variance assumption; no
significance levels changed

vif(model9b)

# two values above 10, but this is probably due to interaction terms, since we didn't see
this in model7

summary(model9b)

predict(model9b,data.frame(Income97=40500, black="Black", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio", Region2="South",
Sex=1),interval="confidence")
predict(model9b,data.frame(Income97=40500, black="Alternative", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio", Region2="South",
Sex=1),interval="confidence")
predict(model9b,data.frame(Income97=40500, black="Black", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio",
Region2="NorthCentral", Sex=1),interval="confidence")
predict(model9b,data.frame(Income97=40500, black="Alternative", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio",
Region2="NorthCentral", Sex=1),interval="confidence")
predict(model9b,data.frame(Income97=40500, black="Black", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio", Region2="South",
Sex=2),interval="confidence")
predict(model9b,data.frame(Income97=40500, black="Alternative", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio", Region2="South",
Sex=2),interval="confidence")
predict(model9b,data.frame(Income97=40500, black="Black", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio",
Region2="NorthCentral", Sex=2),interval="confidence")
```

```
predict(model9b,data.frame(Income97=40500, black="Alternative", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio",
Region2="NorthCentral", Sex=2),interval="confidence")

predict(model9b,data.frame(Income97=20000, black="Black", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio", Region2="South",
Sex=1),interval="confidence")
predict(model9b,data.frame(Income97=20000, black="Alternative", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio", Region2="South",
Sex=1),interval="confidence")
predict(model9b,data.frame(Income97=20000, black="Black", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio",
Region2="NorthCentral", Sex=1),interval="confidence")
predict(model9b,data.frame(Income97=20000, black="Alhistternative", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio",
Region2="NorthCentral", Sex=1),interval="confidence")
predict(model9b,data.frame(Income97=20000, black="Black", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio", Region2="South",
Sex=2),interval="confidence")
predict(model9b,data.frame(Income97=20000, black="Alternative", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio", Region2="South",
Sex=2),interval="confidence")
predict(model9b,data.frame(Income97=20000, black="Black", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio",
Region2="NorthCentral", Sex=2),interval="confidence")
predict(model9b,data.frame(Income97=20000, black="Alternative", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio",
Region2="NorthCentral", Sex=2),interval="confidence")

predict(model9b,data.frame(Income97=40500, black="Black", BirthYear=1982,
Intelligence=50000, HighestGrade=14, HouseholdType2="BothBio", Region2="South",
Sex=1),interval="prediction")

head(data,10)
```