ECOM20001 ASSIGNMENT 2 COVER PAGE

Name	Student ID Number
Gia Han Ly	1074109
Rui Zheng	1084537
Alexandre Bormann	1072524

Question 1:

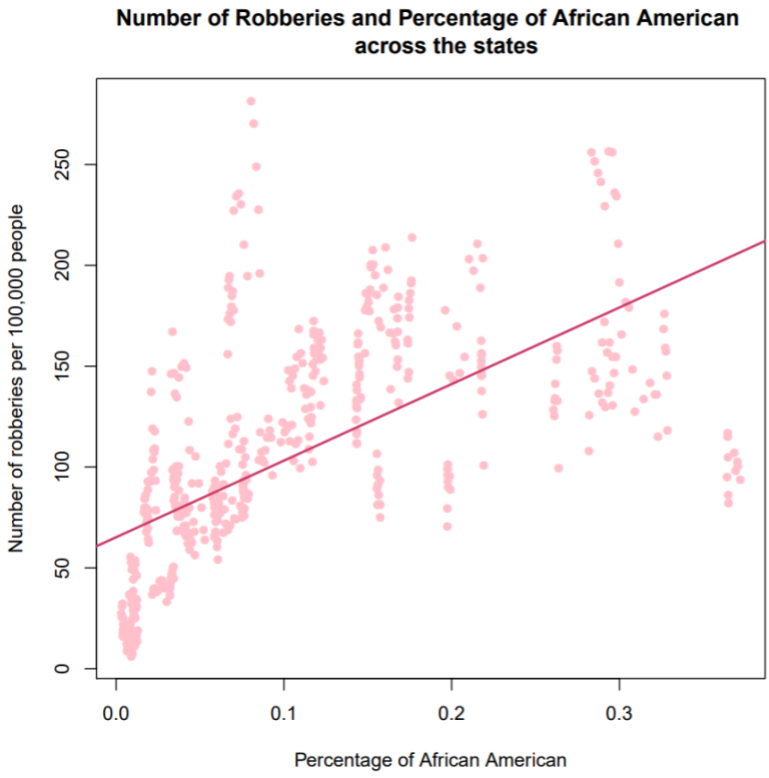
	Variables						
Summary statistics	robbery_rate	assault_rate	burglary_rate	black	income	age	female
Mean	104.697	259.090	693.500	0.104	46343	36.71	0.507
Standard Deviation	58.385	128.419	234.682	0.095	7820.835	1.534	0.007
Minimum	6.148	42.580	292.300	0.003	29359	30.63	0.479
Maximum	281.584	626.460	1244.600	0.372	68059	40.59	0.520

On average, most states in USA has about 105 cases of robbery, 259 cases of assault and lastly about 694 burglary cases for every 100,000 people in the population. About 10.4% of the US population is African American. The average income per household across all states is on $46343.00. The average age of individuals is about 37 years old. Lastly, there is about 50.7% of the US population that is female.

The variable rescaled is average household income for a better interpretation. We decide to rescale income by thousands, dividing all by 1000. The average household income being rescaled is 46.343 in thousands.

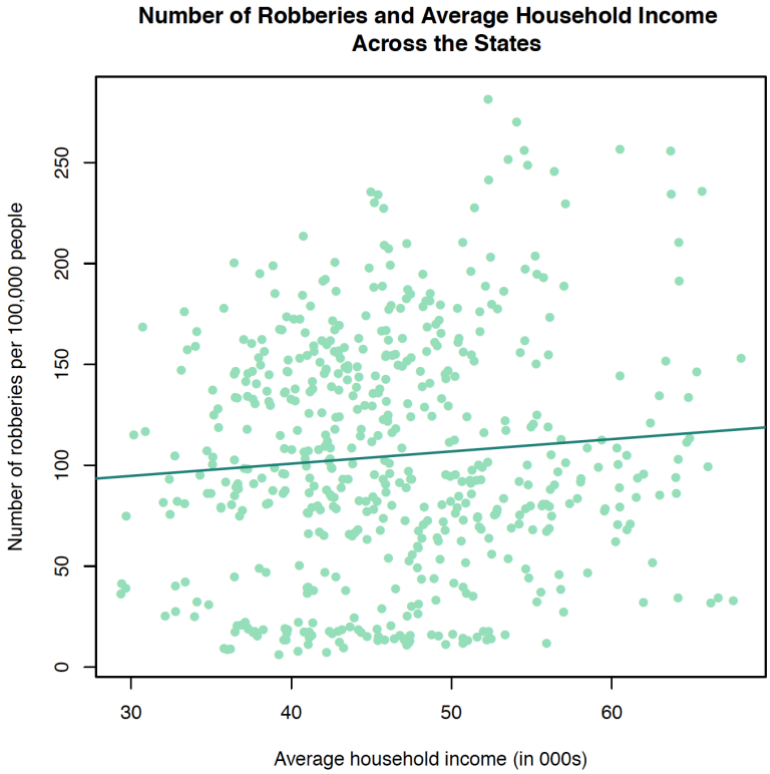
Question 2:

(a) Robbery_rate vs. Black



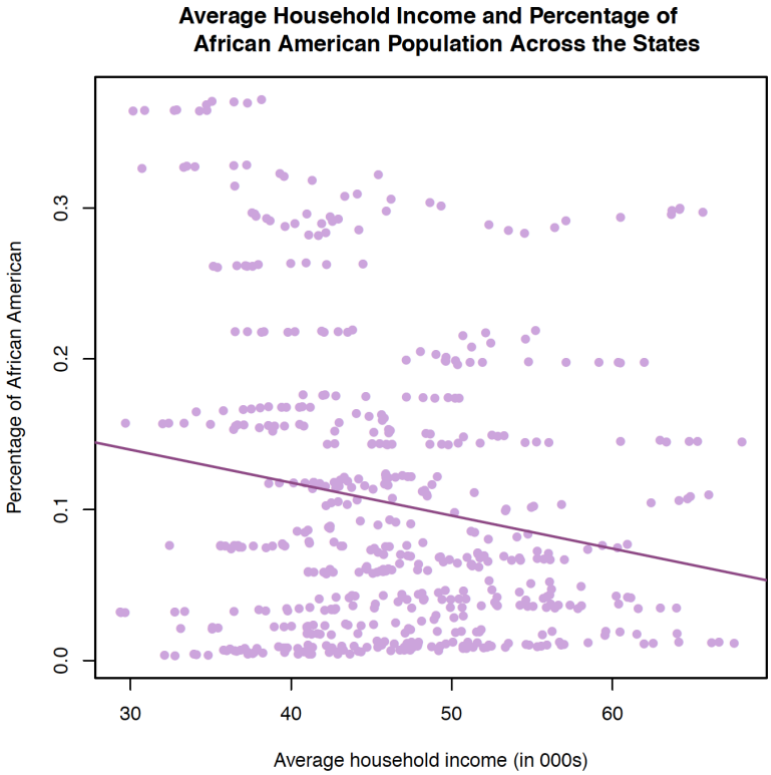
$$\widehat{robbery\_rate} = \underset{(2.94)}{65.19} + \underset{(25.55)}{379.72}\, Black;\ R^2 = 0.383,\ SER = 45.83$$

(b) Robbery_rate vs. Income_rescale



$$\widehat{robbery\_rate} = \underset{(2.949)}{76.56} + \underset{(0.32)}{0.61}\, Income\_rescale;\ R^2 = 0.0048,\ SER = 58.24$$

(c) Black vs. Income_rescale



$$\widehat{black} = \underset{(0.0295)}{0.2053} - \underset{(0.0005)}{0.0022}\, Income\_rescale;\ R^2 = 0.03,\ SER = 0.094$$

Question 3:
Robbery_rate on black with income omitted:

$$\widehat{robbery\_rate} = \underset{(2.94)}{65.19} + \underset{(25.55)}{379.72}\, Black;\ R^2 = 0.383,\ SER = 45.83$$

In diagram two (robbery_rate vs. income_rescale), it appears that average household income is positively related to robbery rate (per 100,000 people). Furthermore, income is negatively related to the African-American population in the states as evident in diagram 3 (black vs. income_rescale). Therefore, based on the signs of the estimated regression line, the OLS regression coefficient estimate for black will present a downward bias.

Based on historical experience in the US on matters of race, it is believed that crime rates are higher amongst areas with a higher African-American population. However, areas with more densely populated African-Americans have lower household income, and a higher household income leads to increased robbery rates. Therefore, in the regression of robbery_rate on black, the slope coefficient would be too small in magnitude.

Question 4:
If regression between robbery_rate and d2000, d2001, d2002, d2003, d2004, d2005, d2006, d2007, d2008, d2009, d2010 was run, R would return the regression result but with “1 not defined because of singularities” warning.

Dummy variable trap
Each of the year dummy variables falls into one and only category. All of the dummy variables are binary variables. There is perfect multicollinearity, the sum of the dummy variables is equal to the intercept (constant regressor). R removed the variable that caused the multicollinearity (d2010) in order to run the regression.

Question 5: **Regression table with different controlled variable**

	Dependent variable:				
	Number of Robberies Per 100,000 People				
	(1)	(2)	(3)	(4)	(5)
Percentage of Population that is African American	379.72***	401.54***	397.86***	349.14***	382.61***
	(25.56)	(22.66)	(23.00)	(33.34)	(35.38)
Average Household Income (in 000s)		1.48***	1.49***	1.57***	1.93***
		(0.23)	(0.22)	(0.22)	(0.25)
Average Individual Age (Years)			−3.11**	−5.43***	−2.51
			(1.25)	(1.73)	(2.00)
Percentage of Population that is Female				1,032.69**	524.14
				(497.61)	(546.97)
d2001					1.55
					(9.14)
d2002					−0.99
					(8.95)
d2003					−4.89
					(8.92)
d2004					−11.11
					(8.88)
d2005					−11.65
					(9.25)
d2006					−6.81
					(9.85)
d2007					−12.26
					(10.11)
d2008					−13.63
					(10.21)
d2009					−20.05**
					(9.88)
d2010					−29.74***
					(9.91)
Constant	65.19***	−5.89	108.77**	−328.55	−188.65
	(2.95)	(10.88)	(47.83)	(216.07)	(226.58)
Observations	550	550	550	550	550
R2	0.38	0.42	0.43	0.44	0.45
Adjusted R2	0.38	0.42	0.43	0.43	0.44
Residual Std. Error	45.83 (df = 548)	44.42 (df = 547)	44.21 (df = 546)	43.96 (df = 545)	43.73 (df = 535)
F Statistic	342.84*** (df = 1; 548)	200.62*** (df = 2; 547)	137.19*** (df = 3; 546)	105.86*** (df = 4; 545)	31.70*** (df = 14; 535)
Note:					*p<0.1; **p<0.05; ***p<0.01

Question 6:

A. In (1) and (2), the coefficient on black changes according to question 2 and 3. There is evidence of omitted variable bias as average household income has a negative bias relative to coefficient of black.

B. The coefficient on black seems to settle down when the average individual age is added to the regression model. The number of robberies only change by 3.68/100,000 when average age is factored in. When the percentage of the population that is female is added however, the change in the coefficient on black becomes larger again. In Reg (5) the regression coefficient on black is significant at the 5% and significance value, as the p-value is smaller than 0.05. This is indicated by the stars in the table above.

C. Base group is d2000 as it is the dummy variable that is not included.
Robbery rate is significantly lower in 2009 and 2010 across US states under 5% level of significance as indicated by the stars in the table above.

CI for d2009
[-20.05197 - 1.96(9.87856), -20.05197 + 1.96(9.87856)]
[-39.052, -0.6899]
The coefficient of d2009 is significantly different from 0 at 95% confidence
CI for d2010
[-29.74176 - 1.96(9.90698), -29.74176 + 1.96(9.90698)]
[-49.159, -10.324]
The coefficient of d2010 is significantly different from 0 at 95% confidence

From the coefficient, robbery rate in 2009 is lower than other years by 20.05 cases per 100,000 and in 2010 it is lower by 29.74 cases per 100,000 people

D. According to the coefficient estimate of Reg (5) the number of robberies per 100,000 people increases by 382.61 if the percentage of the population that is African American increases by 1 unit. On the other hand, when the standard deviation (0.095) of the percentage of the population that is African American increases by one, the robberies per 100,000 people increases by 36.25. The interpretation of changing one standard deviation is more relevant because the standard deviation of black of 0.095 is so far away from one unit. Increasing the percentage of the population that is African American by 1 unit is almost 10% of the average percentage of the population that is African American.

E. When comparing a change of one standard deviation in the percentage of the population that is African American to the mean number of robberies in the dataset (104.70 per 100 000 people), increasing the percentage of the population that is African American by one standard deviation of 0.095 leads to a 34.72% increase in robberies compared to the mean, which can be considered a large magnitude change in the robbery rate.

Question 7: **Regression table with (1)robbery_rate, (2)assault_rate, (3)burglary_rate**

	Dependent variable:		
	Robbery Rate Per 100,000 People	Assault Rate Per 100,000 People	Burglary Rate Per 100,000 People
	(1)	(2)	(3)
Percentage of population that is black	382.61***	732.96***	1,046.69***
	(35.38)	(101.32)	(111.21)
Average household income (in 000s)	1.93***	-2.34***	-13.34***
	(0.25)	(0.76)	(1.09)
Average individual age	-2.51	-3.80	-31.03***
	(2.00)	(4.76)	(7.61)
Percentage of population that is female	524.14	-4,152.59***	-2,010.89
	(546.97)	(1,199.54)	(1,872.00)
d2001	1.55	-2.73	17.38
	(9.14)	(23.89)	(35.98)
d2002	-0.99	-6.91	31.58
	(8.95)	(23.89)	(36.99)
d2003	-4.89	-13.17	44.12
	(8.92)	(23.14)	(36.96)
d2004	-11.11	-13.89	54.97
	(8.88)	(23.71)	(38.65)
d2005	-11.65	-6.43	71.62*
	(9.25)	(24.61)	(39.02)
d2006	-6.81	-1.70	107.70***
	(9.85)	(25.06)	(39.03)
d2007	-12.26	2.78	125.69***
	(10.11)	(25.78)	(40.06)
d2008	-13.63	-3.80	137.41***
	(10.21)	(25.66)	(41.19)
d2009	-20.05**	-13.92	123.28***
	(9.88)	(25.52)	(41.54)
d2010	-29.74***	-19.46	115.20***
	(9.91)	(25.51)	(41.18)
Constant	-188.65	2,544.27***	3,286.17***
	(226.58)	(492.76)	(726.92)
Observations	550	550	550
R2	0.45	0.25	0.42
Adjusted R2	0.44	0.23	0.40
Residual Std. Error (df = 535)	43.73	112.34	181.42
F Statistic (df = 14; 535)	31.70***	13.02***	27.40***
Note:			*p<0.1; **p<0.05; ***p<0.01

Question 8:

A. Regression coefficient of black variable under reg(2) is 732.96, this suggests that the assault rate per 100,000 people increases by 732.96 for every one-unit (100%) increase in the African American population. Furthermore, the coefficient estimate on black has a p-value<0.01, suggesting statistical significance at 5% significance level.

   Coefficient estimate of black under reg(3) shows that one-unit increase in African American increases burglary rate by 1,046.49 per 100,000 people. P-value for the regression coefficient of black has a p-value<0.01, therefore there is sufficient evidence to reject the null at 5% significance level.

B. The crime that is best predicted is the robbery rate. The adjusted r^2 value for robbery rate is 0.44 compared to an r^2 value of 0.40 for the burglary rate and 0.23 for the assault rate. As the r^2 value is indicative for the model fit, the crime that is best predicted is the robbery rate followed by the burglary rate and assault rate.

C. **95% CI for changes in robbery_rate with a 1SD increase in black**
= [(382.61 - 1.96(35.38)) x 0.095, (382.61+1.96(35.38)) x 0.095]
= [29.760, 42.936]

95% CI for changes in assault_rate with a 1SD increase in black
= [(732.96 - 1.96 (101.32)) x 0.095, (732.96 + 1.96 (101.32)) x 0.095]
= [50.765, 88.497]

95% CI for changes in burglary rate with a 1SD increase in black
= [(1046.69 - 1.96(111.21)) x 0.095, (1046.69 +1.96(111.21)) x 0.095]
= [78.728, 120.143]

D. $H_0: \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \ldots, \beta_{14} = 0$
$H_A: the\ null\ is\ false$

reg(1) robbery_rate
F- stats: 35.107
Degree of freedom: q = 14, n - k -1 = 535
p-value < 2.2e-16
p-value < 0.005, suggesting statistical significant at 5% confidence level
There is sufficient evident at 5% significance level that at least one of the coefficient differs from 0. This suggest that the model has some significant in explaining the variation in robbery rate

reg(2) assault_rate
F- stats: 7.204
Degree of freedom: q = 14, n - k -1 = 535
p-value < 8.536e-14
p-value < 0.005, suggesting statistical significant at 5% confidence level
There is sufficient evident at 5% significance level that at least one of the coefficient differs from 0. This suggest that the model has some significant in explaining the variation in assault rate

reg(3) burglary_rate
F- stats: 45.631
Degree of freedom: q = 14, n - k -1 = 535
p-value < 2.2e-16
p-value < 0.005, suggesting statistical significant at 5% confidence level
There is sufficient evident at 5% significance level that at least one of the coefficient differs from 0. This suggest that the model has some significant in explaining the variation in burglary rate

Appendix

```
#assignment 2
#*****************************************************************************
# Load Applied Econometrics package
library(AER)
library(stargazer)

# Load dataset
as2_crime=read.csv(file="as2_crime.csv")

#Question 1
#Summary statistic for all variables
summary(as2_crime)
sd(as2_crime$state)
sd(as2_crime$year)
sd(as2_crime$robbery_rate)
sd(as2_crime$assault_rate)
sd(as2_crime$burglary_rate)
sd(as2_crime$black)
sd(as2_crime$income)
sd(as2_crime$age)
sd(as2_crime$female)
#rescale variables
as2_crime$income_rescale=as2_crime$income/1000

#Question2
#scatterplot
#robbery_rate vs black
#SLR result between robbery rate and black assuming heteroskedasticity
reg1=lm(robbery_rate~black,data=as2_crime)
summary(reg1)
coeftest(reg1, vcov = vcovHC(reg1,"HC1" ))
summary(reg1)$adj.r.squared
#scatterplot of relationship between robbery rate and black
pdf("as2_robbery_black.pdf")
plot(as2_crime$black,as2_crime$robbery_rate,
    main = "Number of Robberies and Percentages of
    African American Population Across the States",
    xlab = "Percentage of African American",
    ylab = "Number of robberies per 100,000 people",
    col="pink",
    pch=16)
abline(reg1, col="#ace5ee", lwd=2)
dev.off()

#robbery_rate vs income_rescale
#SLR result between robbery rate and income(rescaled) assuming heteroskedasticity
reg2=lm(robbery_rate~income_rescale,data=as2_crime)
summary(reg2)
coeftest(reg2, vcov = vcovHC(reg1,"HC1" ))
summary(reg2)$adj.r.squared
#scatterplot of relationship between robbery rate and income(rescaled)
pdf("as2_robbery_incomerescale.pdf")
plot(as2_crime$income_rescale,as2_crime$robbery_rate,
    main = "Number of Robberies and Average Household Income
    Across the States",
    xlab = "Average household income (in 000s)",
    ylab = "Number of robberies per 100,000 people",
```

```
    col="#93dfb8",
    pch=16)
abline(reg2, col="#158078", lwd=2)
dev.off()

#black vs income_rescale
#SLR result between black and income(rescaled) assuming heteroskedasticity
reg3=lm(black~income_rescale,data=as2_crime)
summary(reg3)
coeftest(reg3, vcov = vcovHC(reg1,"HC1" ))
summary(reg3)$adj.r.squared

#scatterplot of relationship between black and income(rescaled)
pdf("as2_black_income.pdf")
plot(as2_crime$income_rescale,as2_crime$black,
    main = "Average Household Income and Percentage of
    African American Population Across the States",
    xlab = "Average household income (in 000s)",
    ylab = "Percentage of African American",
    col="#cda4de",
    pch=16)
abline(reg3, col="#8e4585", lwd=2)
dev.off()


#Question 4
#construct dummy variables for all years in the dataset 2000-2010
as2_crime$d2000=as.numeric(as2_crime$year==2000)
as2_crime$d2001=as.numeric(as2_crime$year==2001)
as2_crime$d2002=as.numeric(as2_crime$year==2002)
as2_crime$d2003=as.numeric(as2_crime$year==2003)
as2_crime$d2004=as.numeric(as2_crime$year==2004)
as2_crime$d2005=as.numeric(as2_crime$year==2005)
as2_crime$d2006=as.numeric(as2_crime$year==2006)
as2_crime$d2007=as.numeric(as2_crime$year==2007)
as2_crime$d2008=as.numeric(as2_crime$year==2008)
as2_crime$d2009=as.numeric(as2_crime$year==2009)
as2_crime$d2010=as.numeric(as2_crime$year==2010)

#Robbery_rate regression controlling for year
reg_dummy=lm(robbery_rate~d2000+d2001+d2002+d2003+d2004+d2005+d2006+d2007+d2008+d200
9+d2010,data=as2_crime)
summary(reg_dummy)
coeftest(reg_dummy, vcov = vcovHC(reg_dummy,"HC1" ))
summary(reg_dummy)$adj.r.squared

#Question 5
# Single linear regression of robbery_rate on black
reg4=lm(robbery_rate~black,data=as2_crime)
cov1=vcovHC(reg4, type = "HC1")
se1=sqrt(diag(cov1))

#Controlling for income
reg5=lm(robbery_rate~black+income_rescale,data=as2_crime)
cov2=vcovHC(reg5, type = "HC1")
se2=sqrt(diag(cov2))

#Controlling for income, age
reg6=lm(robbery_rate~black+income_rescale+age,data=as2_crime)
```

```
cov3=vcovHC(reg6, type = "HC1")
se3=sqrt(diag(cov3))

#Controlling for income, age, female
reg7=lm(robbery_rate~black+income_rescale+age+female,data=as2_crime)
cov4=vcovHC(reg7, type = "HC1")
se4=sqrt(diag(cov4))

#Controlling for income, age, female, years
reg8=lm(robbery_rate~black+income_rescale+age+female+d2001+d2002+d2003+d2004+d2005+d2006
+d2007+d2008+d2009+d2010,data=as2_crime)
cov5=vcovHC(reg8, type = "HC1")
se5=sqrt(diag(cov5))

#regression output table for the 5 regressions reg4, reg5, reg6, reg7, reg8
stargazer(reg4, reg5, reg6, reg7, reg8, type="text",
        se=list(se1, se2, se3, se4, se5),
        digits=2,
        dep.var.labels=c("Number of Robberies Per 100,000 People"),
        covariate.labels=
          c("Percentage of Population that is African American",
            "Average Household Income (in 000s)",
            "Average Individual Age (Years)",
            "Percentage of Population that is Female",
            "d2001",
            "d2002",
            "d2003",
            "d2004",
            "d2005",
            "d2006",
            "d2007",
            "d2008",
            "d2009",
            "d2010",
            "Constant"),
        out="Q5_output.txt")
```

#Question 7

```
#Robbery_rate regression controlling on income, age, female, year
reg9=lm(robbery_rate~black+income_rescale+age+female+d2001+d2002+d2003+d2004+d2005+d2006
+d2007+d2008+d2009+d2010,data=as2_crime)
cov6=vcovHC(reg9, type = "HC1")
se6=sqrt(diag(cov6))

#Assault_rate regression controlling on income, age, female, year
reg10=lm(assault_rate~black+income_rescale+age+female+d2001+d2002+d2003+d2004+d2005+d200
6+d2007+d2008+d2009+d2010,data=as2_crime)
cov7=vcovHC(reg10, type = "HC1")
se7=sqrt(diag(cov7))

#Burglary_rate regression controlling on income, age, female, year
reg11=lm(burglary_rate~black+income_rescale+age+female+d2001+d2002+d2003+d2004+d2005+d20
06+d2007+d2008+d2009+d2010,data=as2_crime)
cov8=vcovHC(reg11, type = "HC1")
se8=sqrt(diag(cov8))

#Constructing regression table for robbery_rate, assault_rate, burglary_rate
stargazer(reg9, reg10, reg11, type="text",
        se=list(se6, se7, se8),
```

```
        digits=2,
        dep.var.labels=c("Robbery Rate Per 100,000 People","Assault Rate Per 100,000
People","Burglary Rate Per 100,000 People"),
        covariate.labels=
          c("Percentage of population that is black",
            "Average household income (in 000s)",
            "Average individual age",
            "Percentage of population that is female",
            "d2001",
            "d2002",
            "d2003",
            "d2004",
            "d2005",
            "d2006",
            "d2007",
            "d2008",
            "d2009",
            "d2010",
            "Constant"),
        out="Q7_output.txt")

#Question 8
#f-test accounted for heteroskedasticity (reg9)
linearHypothesis(reg9,c("black=0", "income_rescale=0", "age=0", "female=0",
                        "d2001=0", "d2002=0", "d2003=0", "d2004=0",
                        "d2005=0", "d2006=0", "d2007=0", "d2008=0" ,
                        "d2009=0", "d2010=0"), vcov = vcovHC(reg9, "HC1"))
#f-test accounted for heteroskedasticity (reg10)
linearHypothesis(reg10,c("black=0", "income_rescale=0", "age=0", "female=0",
                        "d2001=0", "d2002=0", "d2003=0", "d2004=0",
                        "d2005=0", "d2006=0", "d2007=0", "d2008=0" ,
                        "d2009=0", "d2010=0"), vcov = vcovHC(reg10, "HC1"))
#f-test accounted for heteroskedasticity (reg11)
linearHypothesis(reg11,c("black=0", "income_rescale=0", "age=0", "female=0",
                        "d2001=0", "d2002=0", "d2003=0", "d2004=0",
                        "d2005=0", "d2006=0", "d2007=0", "d2008=0" ,
                        "d2009=0", "d2010=0"), vcov = vcovHC(reg11, "HC1"))
```