# ECOM20001: Econometrics 1

## Assignment 2: Suggested Solutions

---

1. Summary statistics and standard deviations reported below. A typical observation is a US state in a year with robbery, assault and burglary rates of 104.7, 259.1, and 693.5 crimes/events per 100,000 people, with 10.4% of the population being black, earning $46,343 USD per year, an age of 36.7 years old, and where 50.7% of the population is female. For regressions, the key variable to rescale is income, and we will rescale it using a variable called income_scale=income/10000 such that income_scale is in terms of $10,000.

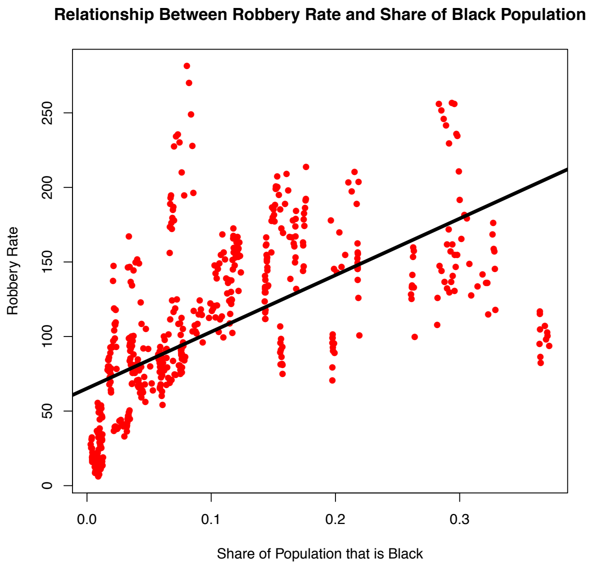
Means, mins, max

```
        state           year       robbery_rate      assault_rate     burglary_rate         black         
 Alabama   : 11   Min.   :2000   Min.   :  6.148   Min.   : 42.58   Min.   : 292.3   Min.   :0.003095  
 Alaska    : 11   1st Qu.:2002   1st Qu.: 67.121   1st Qu.:158.94   1st Qu.: 506.6   1st Qu.:0.025444  
 Arizona   : 11   Median :2005   Median : 98.917   Median :223.39   Median : 650.8   Median :0.073862  
 Arkansas  : 11   Mean   :2005   Mean   :104.697   Mean   :259.09   Mean   : 693.5   Mean   :0.104037  
 California: 11   3rd Qu.:2008   3rd Qu.:147.598   3rd Qu.:346.51   3rd Qu.: 909.6   3rd Qu.:0.155621  
 Colorado  : 11   Max.   :2010   Max.   :281.584   Max.   :626.46   Max.   :1244.6   Max.   :0.372139  
 (Other)   :484                                                                                         
     income           age            female      
 Min.   :29359   Min.   :30.63   Min.   :0.4792  
 1st Qu.:40986   1st Qu.:36.03   1st Qu.:0.5029  
 Median :45748   Median :36.81   Median :0.5085  
 Mean   :46343   Mean   :36.71   Mean   :0.5072  
 3rd Qu.:51236   3rd Qu.:37.58   3rd Qu.:0.5129  
 Max.   :68059   Max.   :40.59   Max.   :0.5197  
```

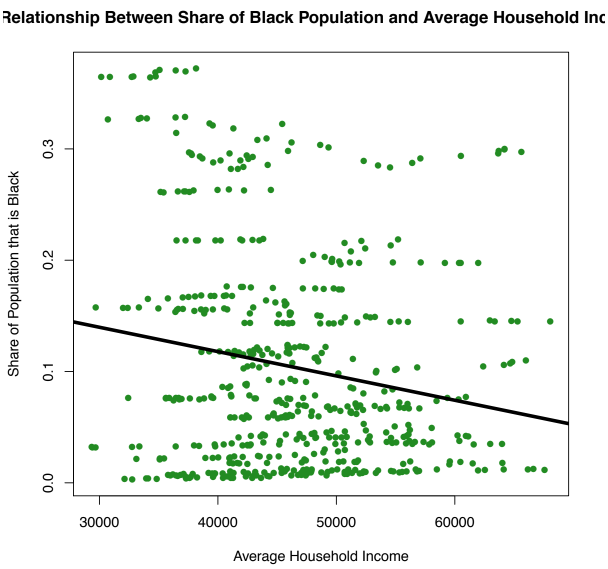
Standard Deviations

```
> sd(mydata$robbery_rate)
[1] 58.38495
> sd(mydata$assault_rate)
[1] 128.4194
> sd(mydata$robbery_rate)
[1] 58.38495
> sd(mydata$assault_rate)
[1] 128.4194
> sd(mydata$burglary_rate)
[1] 234.6821
> sd(mydata$black)
[1] 0.09538564
> sd(mydata$income)
[1] 7820.835
> sd(mydata$age)
[1] 1.533901
> sd(mydata$female)
[1] 0.007407594
```

2. Scatter plots with estimated single linear regression lines presented below. regression lines.








3. The regression of interest is robbery_rate = B0 + B1 x black + u. From the graphs, we see that there is a positive **(+)** relationship between robbery_rate and income and a negative relationship **(-)** between black vs. income. Therefore we should expect the correlation between black and u in the regression to be which creates **(+)** **x** **(-)** = **(-)** downward (negative) omitted variable bias in B1. Given we expect B1 to be positive based on the historical experience of racial discrepancies in the US, the bias will cause B1 to be too small in magnitude.

4. See the as2.R code for the construction of the year dummy variables using the as.numeric() command in R. If you tried to run a regression of robbery_rate on a constant and d2000, d2001, d2002, d2003, d2004, d2005, d2006, d2007, d2008, d2009, d2010, you would run into a perfect collinearity problem because the sum of the dummy variables would always equal 1 in the dataset, which is exactly what the constant regressor on B0 is. R drops the d2010 dummy variable to avoid the dummy variable trap.

5. The regression table created by stargazer() is outputted on the next page.

6. Answers to parts A-E as follows:

   A. Comparing Reg (1) and (2) results, we see that the coefficient on rises from 379.72 to 401.54, meaning its magnitude in Reg (1) was too small due to omitted variable bias due to not controlling for income, exactly as we predicted from questions 2 and 3 above.

   B. Comparing the results across Reg (2) to Reg (5), we see that the magnitude does fall between Reg (2) and (5), with a pronounced fall to 349.144 in Reg (4) when we control for the share of the population that is female. In our richest specification, once year dummies are controlled for, we find that the "final" estimate rises to 382.614 and is statistically significantly different from 0 at the 5% level of significance as the p-value for the test is less than 0.01.

   C. The base group for the year dummy variables is the excluded category, which is the year 2000 as the d2000 dummy variable is not included in the regression. From the table, the coefficients on d2009 and d2010 are statistically significantly different from 0 at the 5% level. Interpreting these coefficients, holding all other regressors fixed, they imply that relative to the year 2000, there are 20.052 and 29.742 fewer robberies per 100,000 people, implying an overall drop in the robbery rate on average across US states over time in the sample.

   D. Returning the statistically significant slope coefficient estimate on black in column (5) of 382.614, the two interpretations for the associated change in robbery_rate are as follows:

      - If black changed by 1 unit, this would only be possible if the entire population changed from being all non-black to all black (e.g,. a 100 percentage point increase in black). In this case, there is a predicted increase in robbery_rate of 382.614 robberies per year, holding all other regressors fixed.

- If black changed by 1 standard deviation, which from question 1 the standard deviation of black is 0.095 (e.g,. a 9.5 percentage point increase in black), then, holding all other regressors fixed, the predicted increase in robbery_rate is 382.614 x 0.095 = 36.35 robberies per year, holding all other regressors fixed.
- The latter 9.5 percentage point change in black is clearly more plausible and relevant since a standard deviation is by definition a "standard change" in a variable in the data, whereas the notion of a state going from 0% to 100% black is completely unrealistic.

E. The sample mean from question 1 for black is 104.7 and the predicted more relevant change from 6D is an increase in robbery_rate of 36.35 robberies per year. That is, the predicted change in robbery_rate from a 9.5 percentage point one-standard deviation change in black is 100* 36.35/104.7=34.7% of the sample mean (holding all other regressors fixed). This is a very large-magnitude change, highlighting just how large racial disparities as they relate to US robberies.

**Regression Output for Question 5**

| | Dependent variable: | | | | |
|---|---|---|---|---|---|
| | Robbery Rate | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| Share of Population that is Black | 379.721*** | 401.542*** | 397.865*** | 349.144*** | 382.614*** |
| | (25.556) | (22.660) | (22.995) | (33.344) | (35.377) |
| Average Household Income (ten thousands) | | 14.849*** | 14.859*** | 15.651*** | 19.335*** |
| | | (2.277) | (2.249) | (2.231) | (2.456) |
| Average Age | | | -3.115** | -5.432*** | -2.505 |
| | | | (1.253) | (1.730) | (2.003) |
| Share of Population that is Female | | | | 1,032.686** | 524.144 |
| | | | | (497.615) | (546.971) |
| 2001 | | | | | 1.546 |
| | | | | | (9.142) |
| 2002 | | | | | -0.994 |
| | | | | | (8.955) |
| 2003 | | | | | -4.891 |
| | | | | | (8.918) |
| 2004 | | | | | -11.105 |
| | | | | | (8.878) |
| 2005 | | | | | -11.650 |
| | | | | | (9.252) |
| 2006 | | | | | -6.812 |
| | | | | | (9.853) |
| 2007 | | | | | -12.263 |
| | | | | | (10.109) |
| 2008 | | | | | -13.625 |
| | | | | | (10.207) |
| 2009 | | | | | -20.052** |
| | | | | | (9.879) |
| 2010 | | | | | -29.742*** |
| | | | | | (9.907) |
| Constant | 65.192*** | -5.892 | 108.768** | -328.553 | -188.652 |
| | (2.949) | (10.878) | (47.834) | (216.067) | (226.578) |
| Observations | 550 | 550 | 550 | 550 | 550 |
| Adjusted R2 | 0.384 | 0.421 | 0.427 | 0.433 | 0.439 |

Note: *p<0.1; **p<0.05; ***p<0.01

7. Regression output provided in the table below:

**Regression Output for Question 7**

| | Dependent variable: | | |
| --- | --- | --- | --- |
| | Robbery Rate (1) | Assault Rate (2) | Burglary Rate (3) |
| Share of Population that is Black | 382.614*** | 732.961*** | 1,046.685*** |
| | (35.377) | (101.319) | (111.214) |
| Average Household Income (ten thousands) | 19.335*** | -23.450*** | -133.380*** |
| | (2.456) | (7.591) | (10.867) |
| Average Age | -2.505 | -3.795 | -31.026*** |
| | (2.003) | (4.764) | (7.614) |
| Share of Population that is Female | 524.144 | -4,152.593*** | -2,010.886 |
| | (546.971) | (1,199.535) | (1,872.000) |
| 2001 | 1.546 | -2.732 | 17.381 |
| | (9.142) | (23.889) | (35.984) |
| 2002 | -0.994 | -6.905 | 31.583 |
| | (8.955) | (23.895) | (36.992) |
| 2003 | -4.891 | -13.173 | 44.117 |
| | (8.918) | (23.138) | (36.962) |
| 2004 | -11.105 | -13.887 | 54.971 |
| | (8.878) | (23.712) | (38.646) |
| 2005 | -11.650 | -6.427 | 71.625* |
| | (9.252) | (24.605) | (39.016) |
| 2006 | -6.812 | -1.703 | 107.701*** |
| | (9.853) | (25.064) | (39.026) |
| 2007 | -12.263 | 2.780 | 125.695*** |
| | (10.109) | (25.778) | (40.061) |
| 2008 | -13.625 | -3.801 | 137.411*** |
| | (10.207) | (25.663) | (41.186) |
| 2009 | -20.052** | -13.925 | 123.280*** |
| | (9.879) | (25.521) | (41.539) |
| 2010 | -29.742*** | -19.462 | 115.200*** |
| | (9.907) | (25.511) | (41.179) |
| Constant | -188.652 | 2,544.273*** | 3,286.170*** |
| | (226.578) | (492.758) | (726.924) |
| Observations | 550 | 550 | 550 |
| Adjusted R2 | 0.439 | 0.235 | 0.402 |

Note: *p<0.1; **p<0.05; ***p<0.01

8. Answers to parts A-D as follows:

   A. Recalling that the standard deviation of black is 0.095, a one standard deviation increase in black is associated with a 732.961 x 0.095=69.63 in the annual assault_rate (per 100,000 people) and a 1046.685 x 0.095=99.44 increase in the burglary_rate, holding all other regressors fixed. From the table, both of the coefficient estimates are statistically significantly different from 0 at the 5% level.

B. The respective adjusted R-Squared's for the robbery_rate, assault_rate, and burglary_rate regressions are 0.439, 0.235, and 0.402 implying that the regressors are best able to predict robbery_rate relative to their ability to predict assault_rate or burglary_rate.

C. The respective 95% confidence intervals for the predicted annual state-level changes in robbery_rate, assault_rate, and burglary_rate for a one-standard deviation change in black of 0.095, holding all other regressors fixed, are[1]:

- robbery_rate
  - 95% CI: [(382.614-35.377 x 1.96) x 0.095,(382.614+35.377 x 1.96) x 0.095]=**[29.76, 42.94]** robberies per 100,000 people
- assault_rate
  - 95% CI: [(732.961-101.319 x 1.96) x 0.095,(732.961+101.319 x 1.96) x 0.095]=**[50.77, 88.50]** assaults per 100,000 people
- burglary_rate
  - 95% CI: [(1046.685-111.214 x 1.96) x 0.095,(1046.685+111.214 x 1.96) x 0.095]=**[78.73,120.14]** burglaries per 100,000 people

D. With n=550 observations, and k=14 regressors in Reg (1)-(3), the overall regression F-statistics which impose q=k restrictions on the model are distributed F(q,n-k-1)=F(14,550-14-1)=F(14,535) with df1=14 and df2=535 degrees of freedom. The restrictions come from the null H0: Bj=0 for regression coefficient j, for j=1,…,k against the alternative that H1: at least one Bj !=0 for j=1,…,k (where != means "not equals"). From the as2.R code, the regression F-statistics and corresponding p-values for the test of the null are as follows:

- robbery_rate
  - F=35.107, p<0.01
- assault_rate
  - F=7.200, p<0.01
- burglary_rate
  - F=45.631, p<0.01

[1] In the as2.R code, I produce the confidence intervals to extreme precision based on the regression output in the code from Reg (1)-(3), and the intervals are [29.88,43.11], [50.97, 88.86] and [79.05, 120.63].

For each model Reg (1)-(3) with a p-value less than 0.01, we reject the null H0 at the 1% level of significance, which in words means that we reject the null that each of the models are, statistically, not at all useful for explaining variation in the respective forms of crime, namely robbery_rate, assault_rate, and burglary_rate. See the code as2.R for the F-test output; I produce an example from the code for the Reg (1) robbery_rate regression here for quick reference:

**Example Overall Regression F-statistic Code and Output from as2.R for Question 8**

```
> ## Overall regression F-statistic for the robbery rate regression
> linearHypothesis(reg1,c("black=0","income_scale=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(reg1, "HC1"))
Linear hypothesis test

Hypothesis:
black = 0
income_scale = 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

Model 1: restricted model
Model 2: robbery_rate ~ black + income_scale + age + female + d2001 +
    d2002 + d2003 + d2004 + d2005 + d2006 + d2007 + d2008 + d2009 +
    d2010

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1    549
2    535 14 35.107 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

9. As per the question on the assignment, full marks for the R code will be given it is as clear as the code in as2.R (or better!).