DSSH 6301 - HW 08-09 Solutions

Problem 1

Does non-violent crime rates affect the crime rate of violent crimes?

Problem 2

The data source is table 5 of the 2013 crime offenses in the US statistics aggregated by the FBI. The source is found here:

http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2013/crime-in-the-u.s.-2013/tables

This data is used to track crime rates across the country.

Problem 3

Violent crime is the dependent variable. The presence of other low level or non-violent crimes should influence this. Knowing the relationship between non-violent and violent crimes can help cities improve the quality of life for residents.

Problem 4

The independent variables are the reported non-violent crimes. I would expect the higher the crime rates of the dependent variables the higher the crime rate of the dependent variable. There may be a relationship between the non-violent crime rates themselves.

Problem 5

The data comes in an excel spreadsheet. The variable/column names are good, but the formatting must be changed to a data format that R can easily interpret. Manually formatting the data allows you to export the data to a CSV file. The column names are updated to include the variable units and extraneous characters are removed from the state names and variable values using REGEX.

```r
data <- read.csv("state_crime_data_2013.csv")
head(data)
```

```
##        State ViolentCrimeRate.per.100000. BurglaryRate.per.100000.
## 1    ALABAMA                        430.8                    877.8
## 2     ALASKA                        640.4                    396.7
## 3    ARIZONA                        416.5                    732.4
## 4   ARKANSAS                        460.3                   1030.1
## 5 CALIFORNIA                        402.1                    605.4
## 6   COLORADO                        308.0                    476.1
##   LarcenyTheftRate.per.100000. MotorVehicleTheftRate.per.100000.
## 1                       2254.8                             218.7
```

```
## 2                        2258.0                              230.6
## 3                        2403.5                              263.2
## 4                        2380.6                              191.9
## 5                        1621.5                              431.2
## 6                        1944.5                              237.9
```

Problem 6

```r
burglary_mod <- lm(ViolentCrimeRate.per.100000. ~ BurglaryRate.per.100000., data=data)
larceny_mod <- lm(ViolentCrimeRate.per.100000. ~ LarcenyTheftRate.per.100000., data=data)
motor_veh_mod <- lm(ViolentCrimeRate.per.100000. ~ MotorVehicleTheftRate.per.100000.,
                    data=data)

summary(burglary_mod)
```

```
##
## Call:
## lm(formula = ViolentCrimeRate.per.100000. ~ BurglaryRate.per.100000.,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -227.58  -89.11  -30.11   58.12  956.29
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              191.4372    77.7838   2.461   0.0174 *
## BurglaryRate.per.100000.   0.2974     0.1246   2.386   0.0209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 173.7 on 49 degrees of freedom
## Multiple R-squared:  0.1041, Adjusted R-squared:  0.08584
## F-statistic: 5.695 on 1 and 49 DF,  p-value: 0.02092
```

```r
summary(larceny_mod)
```

```
##
## Call:
## lm(formula = ViolentCrimeRate.per.100000. ~ LarcenyTheftRate.per.100000.,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -239.68  -86.49  -15.66   86.80  412.76
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -179.73833   86.34503  -2.082   0.0426 *
## LarcenyTheftRate.per.100000.    0.28123    0.04328   6.497 3.99e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 134.5 on 49 degrees of freedom
## Multiple R-squared:  0.4628, Adjusted R-squared:  0.4518
## F-statistic: 42.21 on 1 and 49 DF,  p-value: 3.988e-08
```

```
summary(motor_veh_mod)
```

```
##
## Call:
## lm(formula = ViolentCrimeRate.per.100000. ~ MotorVehicleTheftRate.per.100000.,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -329.75  -75.45  -23.30   61.29  568.78
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        124.7134    48.7510   2.558   0.0137 *
## MotorVehicleTheftRate.per.100000.   1.2129     0.2213   5.480 1.47e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 144.5 on 49 degrees of freedom
## Multiple R-squared:   0.38,  Adjusted R-squared:  0.3673
## F-statistic: 30.03 on 1 and 49 DF,  p-value: 1.465e-06
```

Each dependent variable shows significance. If there is some relationship among the dependent variables, then one or two of these may lose significance in the multiple regression. For instance larceny and burglary are likely related and one would remove the significance of the other.

Problem 7

```
mod <- lm(ViolentCrimeRate.per.100000. ~ LarcenyTheftRate.per.100000. +
          MotorVehicleTheftRate.per.100000. + BurglaryRate.per.100000., data=data)
```

```
require(stargazer)
```

```
## Loading required package: stargazer
##
## Please cite as:
##
##  Hlavac, Marek (2014). stargazer: LaTeX code and ASCII text for well-formatted regression and summary
##  R package version 5.1. http://CRAN.R-project.org/package=stargazer
```

```r
vars <- c("Violent Crime Rate", "Larceny Theft Rate", "Motor Vehicle Theft Rate",
          "Burglary Rate")
vars <- paste(vars, "(per 100,000 people)")

stargazer(mod, align=TRUE, no.space=TRUE, dep.var.labels=vars[1],
          covariate.labels=vars[-1], omit.stat=c("LL","ser","f"), header=FALSE)
```

Table 1:

	Dependent variable:
	Violent Crime Rate (per 100,000 people)
Larceny Theft Rate (per 100,000 people)	0.208***
	(0.058)
Motor Vehicle Theft Rate (per 100,000 people)	0.619**
	(0.260)
Burglary Rate (per 100,000 people)	−0.045
	(0.108)
Constant	−134.720
	(87.033)
Observations	51
R^2	0.521
Adjusted R^2	0.490
Note:	*p<0.1; **p<0.05; ***p<0.01

Larceny and motor vehicle theft have a significant positive relationship with the violent crime rate. Burglary has lost the significance that it had in the bivariate regression.

Problem 8

```r
coef(mod)
```

```
##                        (Intercept)      LarcenyTheftRate.per.100000.
##                       -134.7195204                         0.2081372
## MotorVehicleTheftRate.per.100000.          BurglaryRate.per.100000.
##                          0.6190174                        -0.0451350
```

```r
coef(burglary_mod)
```

```
##              (Intercept) BurglaryRate.per.100000.
##              191.4371860                0.2974179
```

```r
coef(larceny_mod)
```

```
##                  (Intercept) LarcenyTheftRate.per.100000.
##                 -179.7383317                    0.2812339
```

```r
coef(motor_veh_mod)
```

```
##                        (Intercept) MotorVehicleTheftRate.per.100000.
##                         124.713351                          1.212891
```

The coefficients for each variable decreased from the bivariate regression, with burglary being close to 0. The variation in the dependent variable is in part being explained by each dependent variable in the multivariate regression, reducing the contribution of any one of the independent variables. This is an example of a chained causal pathway, with Burglary losing its significance through Larceny.

Problem 9

The results match the hypothesis, there is a positive relationship between the dependent variables and the dependent variable.

Problem 10

The model explains some of the variance in the dependent varaible. The adjusted R^2 and the R^2 are close since there are not many dependent variables.

Problem 11

I use a stepwise selection based on the AIC. The function step is an R builtin that automates the process of building different models and selecting based on AIC.

```r
step <- step(mod)
```

```
## Start:  AIC=500.13
## ViolentCrimeRate.per.100000. ~ LarcenyTheftRate.per.100000. +
##     MotorVehicleTheftRate.per.100000. + BurglaryRate.per.100000.
##
##                                     Df Sum of Sq     RSS    AIC
## - BurglaryRate.per.100000.           1      2962  794339 498.33
## <none>                                            791377 500.13
## - MotorVehicleTheftRate.per.100000.  1     95252  886629 503.93
## - LarcenyTheftRate.per.100000.       1    220539 1011916 510.67
##
## Step:  AIC=498.33
## ViolentCrimeRate.per.100000. ~ LarcenyTheftRate.per.100000. +
##     MotorVehicleTheftRate.per.100000.
##
##                                     Df Sum of Sq     RSS    AIC
## <none>                                            794339 498.33
## - MotorVehicleTheftRate.per.100000.  1     92515  886854 501.94
## - LarcenyTheftRate.per.100000.       1    229262 1023601 509.26
```

```r
step
```

```
##
## Call:
## lm(formula = ViolentCrimeRate.per.100000. ~ LarcenyTheftRate.per.100000. +
##     MotorVehicleTheftRate.per.100000., data = data)
##
## Coefficients:
##                       (Intercept)       LarcenyTheftRate.per.100000.
##                         -143.0414                             0.2001
## MotorVehicleTheftRate.per.100000.
##                            0.6050
```

This returns a model that drops the burglary term. I do agree with these results. The beta on the burglary independent variable was not significant and removing it provides the best AIC measure.

Problem 12

There is a significant, although perhaps not causative, relationship between violent and non-violent crimes. The root causes of violent crimes is a complex issue that must incorporate many factors that we did not look at here, such as cultural factors, etc. These factors may very well vary at a finer geographical level in the US then state, which was the aggregation level used here.

Problem 13

```r
summary(mod)
```

```
##
## Call:
## lm(formula = ViolentCrimeRate.per.100000. ~ LarcenyTheftRate.per.100000. +
##     MotorVehicleTheftRate.per.100000. + BurglaryRate.per.100000.,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -303.84  -76.21   -5.75   88.04  358.60
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -134.71952   87.03299  -1.548 0.128351
## LarcenyTheftRate.per.100000.        0.20814    0.05751   3.619 0.000721
## MotorVehicleTheftRate.per.100000.   0.61902    0.26026   2.378 0.021502
## BurglaryRate.per.100000.           -0.04513    0.10762  -0.419 0.676838
##
## (Intercept)
## LarcenyTheftRate.per.100000.      ***
## MotorVehicleTheftRate.per.100000. *
## BurglaryRate.per.100000.
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 129.8 on 47 degrees of freedom
## Multiple R-squared:  0.5206, Adjusted R-squared:   0.49
## F-statistic: 17.02 on 3 and 47 DF,  p-value: 1.278e-07
```

Part a

```
xmat <- cbind(1, data$LarcenyTheftRate.per.100000., data$MotorVehicleTheftRate.per.100000.,
              data$BurglaryRate.per.100000.)
betas <- solve(t(xmat)%*%xmat) %*% t(xmat) %*% data$ViolentCrimeRate.per.100000.
betas
```

```
##              [,1]
## [1,] -134.7195204
## [2,]    0.2081372
## [3,]    0.6190174
## [4,]   -0.0451350
```

Part b

```
n <- nrow(data)
k <- 3

x <- data$LarcenyTheftRate.per.100000.
y <- data$ViolentCrimeRate.per.100000.

mse <- sum((mod$fitted.values - y)^2) / (n-k-1)
se_b <- sqrt(solve(t(xmat) %*% xmat) * mse)
```

```
## Warning in sqrt(solve(t(xmat) %*% xmat) * mse): NaNs produced
```

```
se_b1 <- se_b[2, 2]
se_b1
```

```
## [1] 0.05751092
```

```
t_stat <- betas[2, 1] / se_b1
t_stat
```

```
## [1] 3.61909
```

```
p_val <- pt(t_stat, df = n-k-1, lower.tail=F)*2
p_val
```

```
## [1] 0.0007210921
```

Part c

```
tss <- sum((y - mean(y))^2)
sse <- sum((y - mod$fitted.values)^2)
r2 <- (tss-sse)/tss

dft <- n - 1
dfe <- n - k - 1

adj_r2 <- ((tss/dft) - (sse/dfe)) / (tss/dft)

r2
```

```
## [1] 0.5206375
```

```
adj_r2
```

```
## [1] 0.4900399
```

Part d

```
f <- (r2/k) / ((1-r2)/(n-k-1))
f
```

```
## [1] 17.01563
```

Problem 14

```
mod_quad <- lm(ViolentCrimeRate.per.100000. ~ LarcenyTheftRate.per.100000. +
               MotorVehicleTheftRate.per.100000. + BurglaryRate.per.100000. +
               I(BurglaryRate.per.100000.^2), data=data)
summary(mod_quad)
```

```
##
## Call:
## lm(formula = ViolentCrimeRate.per.100000. ~ LarcenyTheftRate.per.100000. +
##     MotorVehicleTheftRate.per.100000. + BurglaryRate.per.100000. +
##     I(BurglaryRate.per.100000.^2), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -314.19  -62.56  -14.46   92.32  352.98
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         9.846e+01  2.127e+02   0.463 0.645655
```

```
## LarcenyTheftRate.per.100000.       2.045e-01  5.732e-02   3.567 0.000856
## MotorVehicleTheftRate.per.100000.  7.058e-01  2.689e-01   2.624 0.011746
## BurglaryRate.per.100000.          -8.685e-01  6.943e-01  -1.251 0.217320
## I(BurglaryRate.per.100000.^2)      6.280e-04  5.233e-04   1.200 0.236208
##
## (Intercept)
## LarcenyTheftRate.per.100000.      ***
## MotorVehicleTheftRate.per.100000. *
## BurglaryRate.per.100000.
## I(BurglaryRate.per.100000.^2)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 129.2 on 46 degrees of freedom
## Multiple R-squared:  0.5352, Adjusted R-squared:  0.4948
## F-statistic: 13.24 on 4 and 46 DF,  p-value: 2.961e-07
```

```r
xbar <- mean(data$BurglaryRate.per.100000.)
y1 <- mod_quad$coefficients[4]*xbar + mod_quad$coefficients[5] * xbar^2
y2 <- mod_quad$coefficients[4]*(xbar+1) + mod_quad$coefficients[5]*(xbar+1)^2
y2 - y1
```

```
## BurglaryRate.per.100000.
##                -0.123233
```

The quadratic term is not significant in the model. There is a change in y of -0.123233 with a 1-unit increase in the quadratic term at from the mean when all other variables are held constant. We can also show this algebraically.

$$y1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \bar{x} + \beta_4 \bar{x}^2$$
$$y2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(\bar{x}+1) + \beta_4(\bar{x}+1)^2$$

$$y2 - y1 = \beta_3 - \beta_4 \bar{x}^2 + \beta_4(\bar{x}+1)^2 = \beta_3 - \beta_4 * \bar{x}^2 + \beta_4(\bar{x}^2 + 1 + 2\bar{x}) = \beta_3 + \beta_4(2\bar{x}+1)$$
$$\bar{x} = 592.8216, \beta_3 = -8.685e-01, \beta_4 = 6.280e-04$$
$$y2 - y1 = -0.8684574172 + 0.0006280105(2 * 592.8216 + 1) = -0.123233$$

Problem 15

```r
mod_inter <- lm(ViolentCrimeRate.per.100000. ~ LarcenyTheftRate.per.100000. +
                MotorVehicleTheftRate.per.100000. + BurglaryRate.per.100000. +
                LarcenyTheftRate.per.100000.*BurglaryRate.per.100000., data=data)
summary(mod_inter)
```

```
##
## Call:
## lm(formula = ViolentCrimeRate.per.100000. ~ LarcenyTheftRate.per.100000. +
##     MotorVehicleTheftRate.per.100000. + BurglaryRate.per.100000. +
```

```
##     LarcenyTheftRate.per.100000. * BurglaryRate.per.100000.,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -304.98  -76.43   -4.94   88.21  359.05
##
## Coefficients:
##                                                         Estimate
## (Intercept)                                           -1.238e+02
## LarcenyTheftRate.per.100000.                           2.025e-01
## MotorVehicleTheftRate.per.100000.                      6.215e-01
## BurglaryRate.per.100000.                              -6.602e-02
## LarcenyTheftRate.per.100000.:BurglaryRate.per.100000.  9.944e-06
##                                                       Std. Error t value
## (Intercept)                                            3.623e+02  -0.342
## LarcenyTheftRate.per.100000.                           1.899e-01   1.067
## MotorVehicleTheftRate.per.100000.                      2.753e-01   2.258
## BurglaryRate.per.100000.                               6.831e-01  -0.097
## LarcenyTheftRate.per.100000.:BurglaryRate.per.100000.  3.211e-04   0.031
##                                                       Pr(>|t|)
## (Intercept)                                             0.7341
## LarcenyTheftRate.per.100000.                            0.2916
## MotorVehicleTheftRate.per.100000.                       0.0287 *
## BurglaryRate.per.100000.                                0.9234
## LarcenyTheftRate.per.100000.:BurglaryRate.per.100000.   0.9754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 131.2 on 46 degrees of freedom
## Multiple R-squared:  0.5206, Adjusted R-squared:  0.479
## F-statistic: 12.49 on 4 and 46 DF,  p-value: 5.864e-07
```

```r
mean_burg <- mean(data$BurglaryRate.per.100000.)
mean_larc <- mean(data$LarcenyTheftRate.per.100000.)

y1 <- mod_inter$coefficients[2]*mean_larc + mod_inter$coefficients[5]*mean_larc*mean_burg
y2 <- mod_inter$coefficients[2]*(mean_larc+1) +
      mod_inter$coefficients[5]*(mean_larc+1)*mean_burg
y2 - y1
```

```
## LarcenyTheftRate.per.100000.
##                    0.2084355
```

The interaction term is not significant. There is an increase of 0.2084355 in y with a 1-unit increase in the Larceny term, holding the interacting term at the mean and all other independent variables constant.

Problem 16

```r
anova(mod, mod_inter)
```

```
## Analysis of Variance Table
##
## Model 1: ViolentCrimeRate.per.100000. ~ LarcenyTheftRate.per.100000. +
##     MotorVehicleTheftRate.per.100000. + BurglaryRate.per.100000.
## Model 2: ViolentCrimeRate.per.100000. ~ LarcenyTheftRate.per.100000. +
##     MotorVehicleTheftRate.per.100000. + BurglaryRate.per.100000. +
##     LarcenyTheftRate.per.100000. * BurglaryRate.per.100000.
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     47 791377
## 2     46 791361  1    16.495 0.001 0.9754
```

We fail to reject the null. The added interaction variable does not significantly improve the model.