Take-home exam1

Dandong Tu

2017/10/12

On my honor, I have not had any form of communication about this exam with any other individual(including other students, teaching assistants, instructors, etc.) -Dandong Tu

1.

```
library(alr4)
```

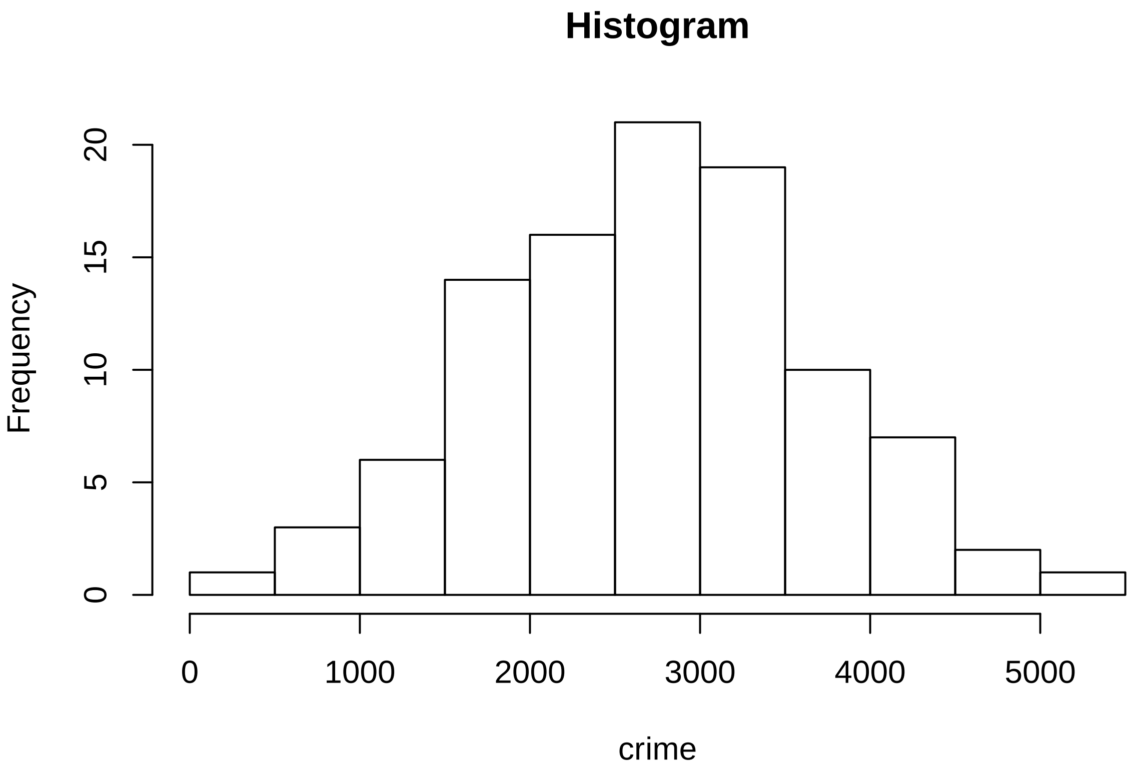
```
## Loading required package: car

## Loading required package: effects

##
## Attaching package: 'effects'

## The following object is masked from 'package:car':
##
##     Prestige
```

```
mydata = read.table("/Users/dandongtu/Desktop/S631 HW/exam1/takehome1.txt", header = TRUE)
mydata = mydata[complete.cases(mydata), ]
sum(mydata['crime'] > 3200) / nrow(mydata)
```

```
## [1] 0.32
```

The probability that a randomly selected city has a crime rate higher than 3,200 is 0.32.

```
hist(mydata$crime, xlab = 'crime', main = 'Histogram')
```

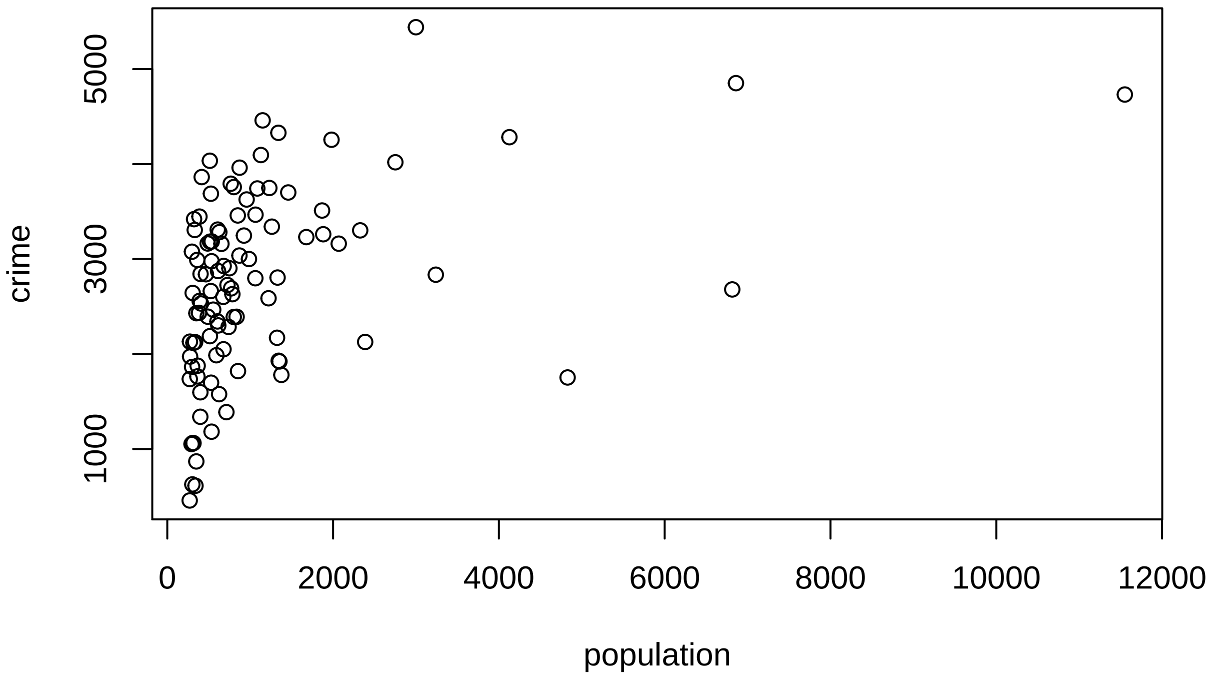


Based on the hitogram, we observe that the distribution of **crime** is close to a normal distribution and has no obvious skewness,therefore, we can conclude that **crime** is normally distributed.

2

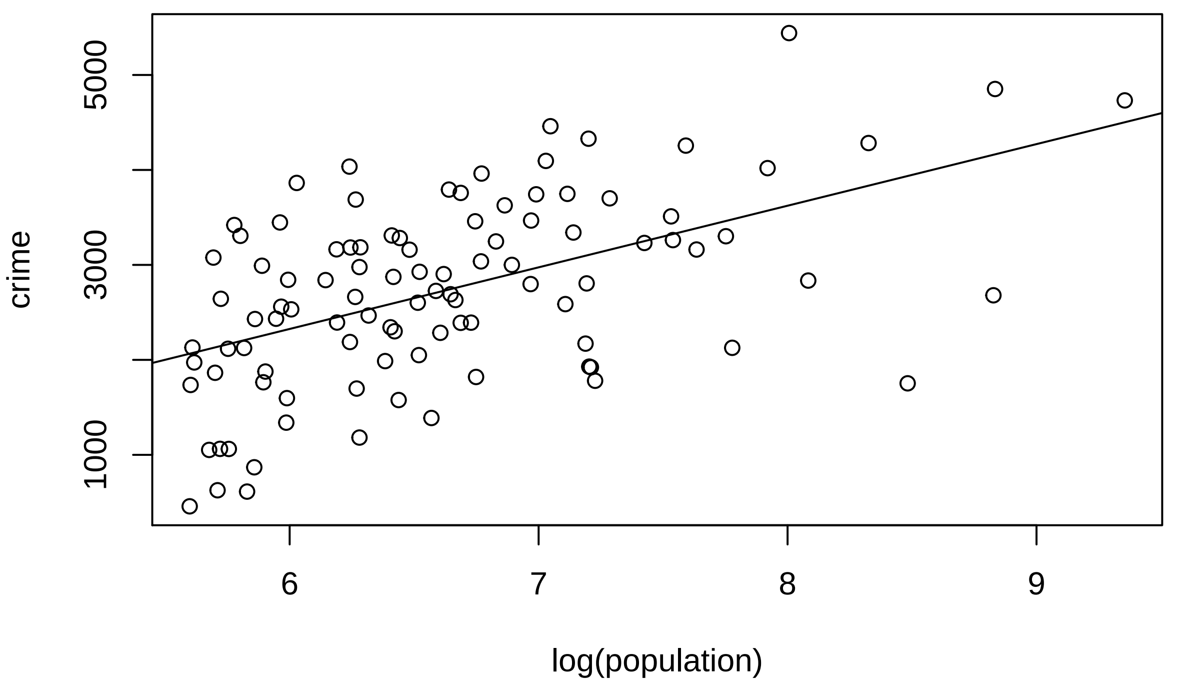
2a

```
plot(mydata$crime ~ mydata$population, xlab = 'population', ylab = 'crime')
```



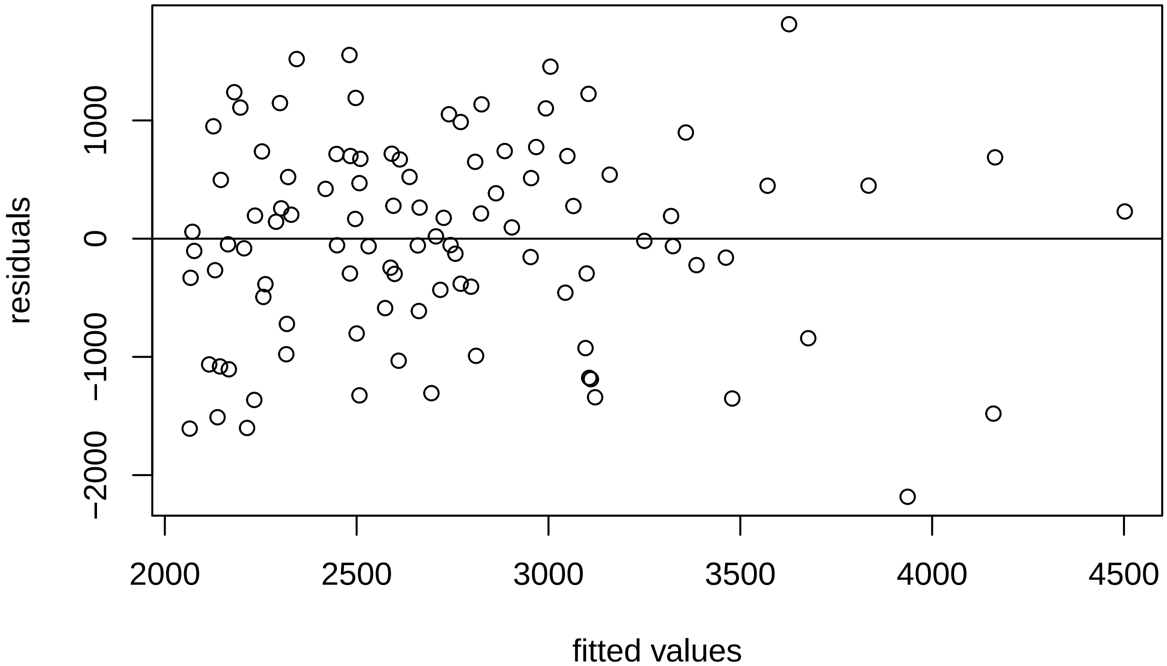
In general, as **population** increases, **crime** also increases. However, most of this increses are in the **crime** happens for smaller values of **population** and the increasing rate decrease as **population** increses. This means that the relationship is not showing a linear trend but a power function or exponential trend. So a straight-line mean function does not seems appropriate.

```
plot(mydata$crime ~ log(mydata$population), xlab = 'log(population)', ylab = 'crime')
abline(lm(crime ~ log(population), data=mydata))
```



The plot above is the distribution using log transformation on **population**. Here, a linear regression model seems appropriate as the mean function appears to be linear, and the variance across the plot is at least plausible, if not completely certain. As one might expect, there may be a few outliers with unusuallly high or low **crime**

```r
m1=lm(crime ~ log(population), data=mydata)
plot(predict(m1),resid(m1), ylab="residuals", xlab="fitted values")
abline(0, 0)
```



Based on the graph of residuals versus fitted values, we observe that the residuals do not follow any distinct pattern. Thus we can conclude that it is not a potential violation.

2b

```r
summary(m1)
```

```
##
## Call:
## lm(formula = crime ~ log(population), data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2183.20  -465.95    38.71   655.15  1813.91
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1568.0      714.9  -2.193   0.0307 *
## log(population)    648.9      107.1   6.058 2.55e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 843.2 on 98 degrees of freedom
## Multiple R-squared:  0.2725, Adjusted R-squared:  0.2651
## F-statistic:  36.7 on 1 and 98 DF,  p-value: 2.553e-08
```

From the linear model summary, $\hat{\beta}_1 = 648.9$ which indicates that If we increase the log(**population**) by 1 unit, the value of **crime** will increase on average in 648.9 units.

Hypothesis test

$H_0 : \beta_1 = 0$

$H_\alpha : \beta_1 \neq 0$

The test statistic,t,provided in the output is obtained from

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1|X)} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1|X)}$$

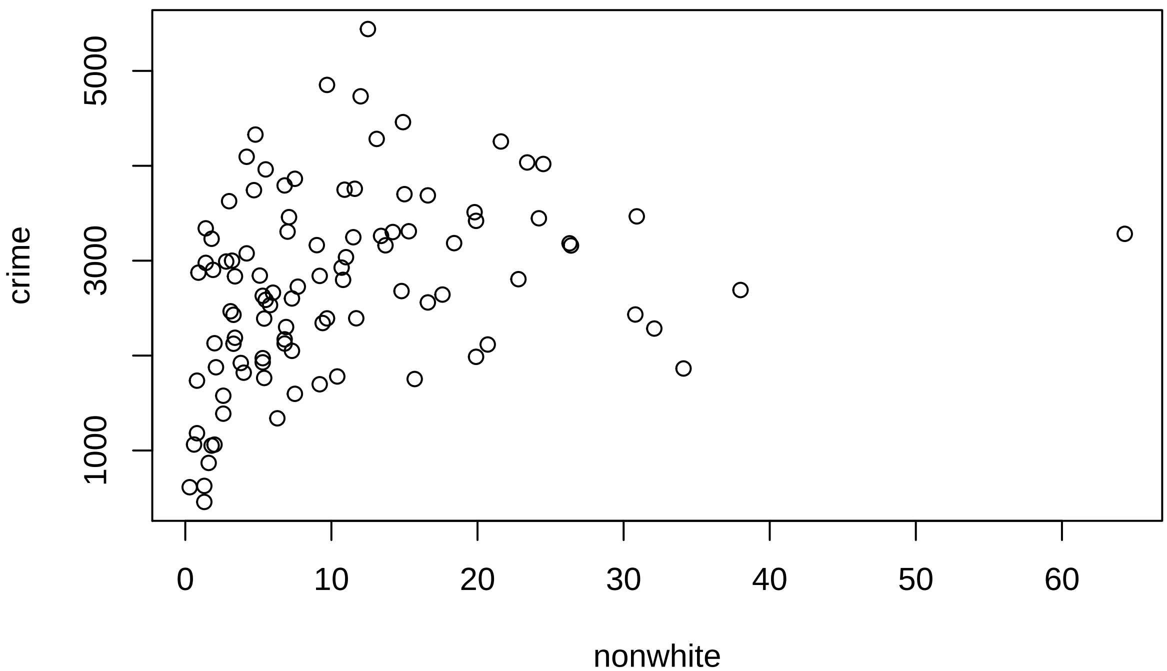
and it follows a **t** distribution($t - value = 6.06$) with $n - (p+1) = 98$ degrees of freedom. The corresponding $p - value = 2.55 \times 10^{-08}$ for this two-tailed test is also provided in the output. Since the p-value is very small, we reject the hypothesis that the coefficient of $\log(population)$ is zero. It means that the log(**population**) has a significant influence on **crime**.

Moreover,$R^2 = 0.2725$ means that, if our model assumptions hold, about 27.25% of the variation in **crime** is explained by log(**population**)

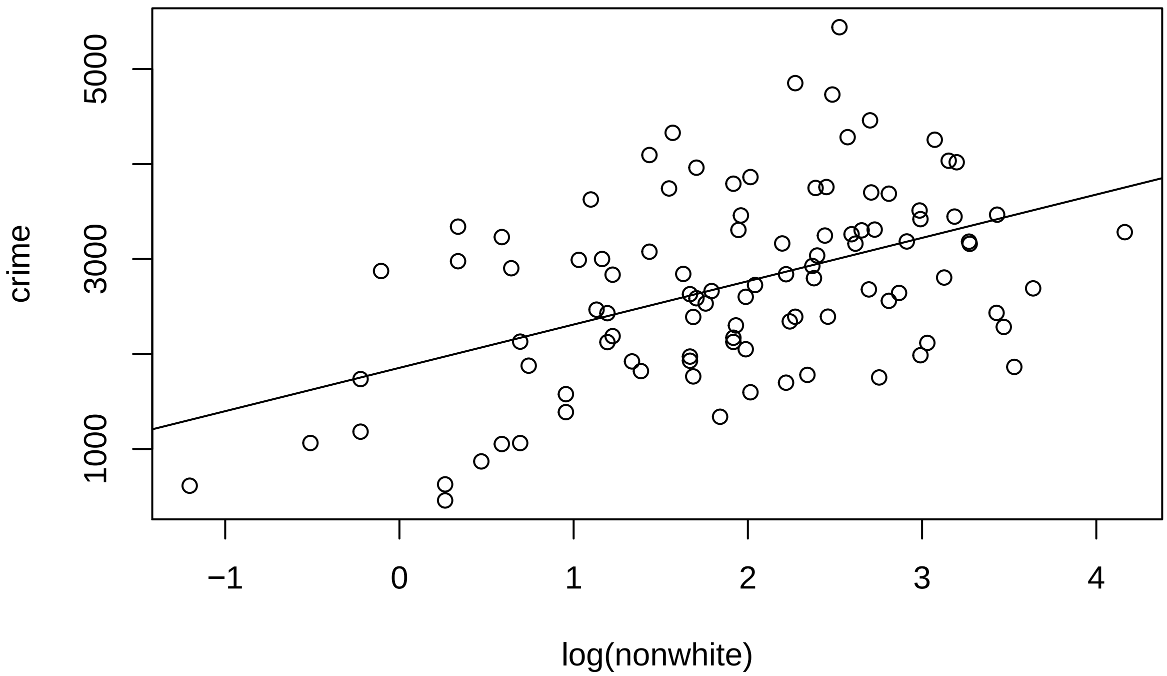
3

3a

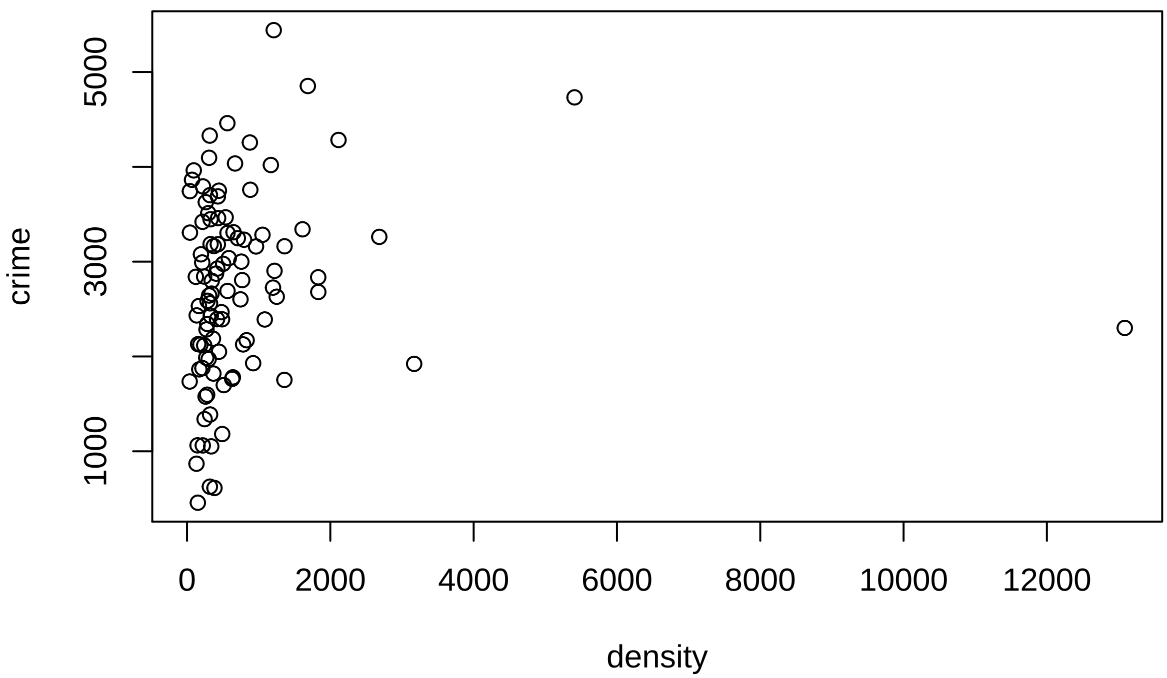
```r
plot(mydata$crime ~ mydata$nonwhite, xlab = 'nonwhite', ylab = 'crime')
```



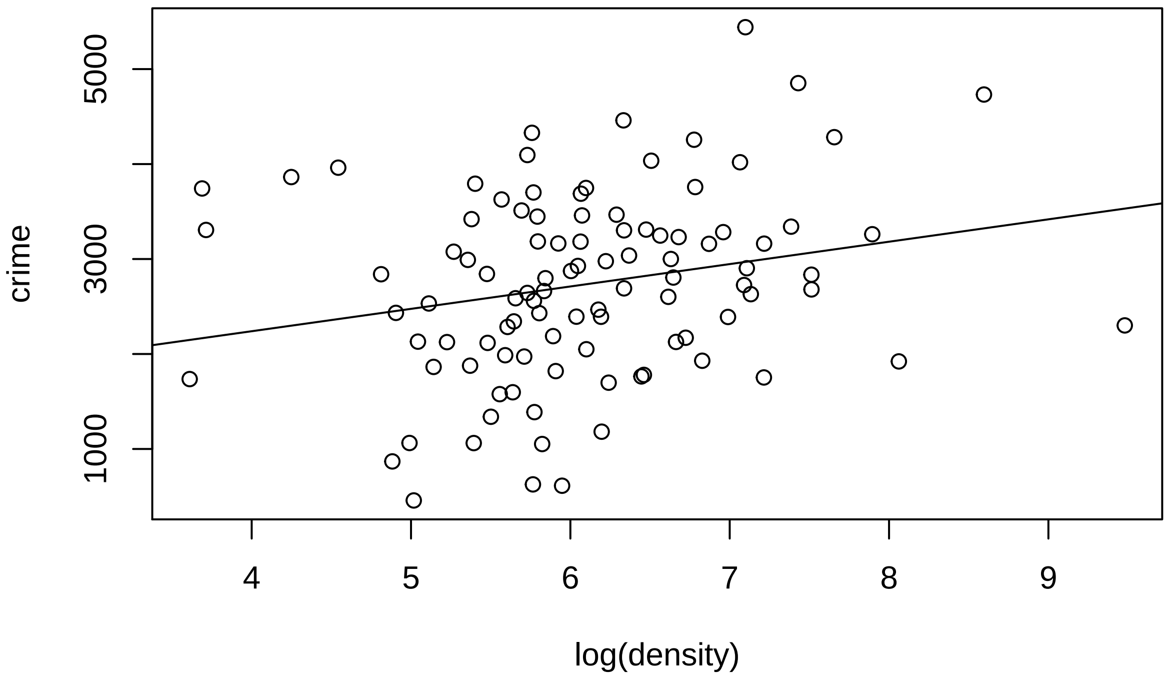

```r
plot(mydata$crime ~ log(mydata$nonwhite),xlab= 'log(nonwhite)', ylab = 'crime')
abline(lm(crime ~ log(nonwhite), data=mydata))
```




```r
plot(mydata$crime ~ mydata$density,xlab='density', ylab = 'crime')
```




```r
plot(mydata$crime ~ log(mydata$density), xlab = 'log(density)', ylab = 'crime')
abline(lm(crime ~ log(density), data=mydata))
```



Similar as part(2), we found that a linear regression model seems appropriate when using log transformation on **nonwhite** and **density**

```
m2=lm(crime ~ log(population)+log(nonwhite)+log(density), data=mydata)
summary(m2)
```

```
##
## Call:
## lm(formula = crime ~ log(population) + log(nonwhite) + log(density),
##     data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2293.1  -559.2    70.1   532.5  1771.2
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1193.60     659.24  -1.811   0.0733 .
## log(population)   670.28     128.04   5.235 9.72e-07 ***
## log(nonwhite)     349.74      78.07   4.480 2.06e-05 ***
## log(density)     -195.29     103.91  -1.879   0.0632 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 765.7 on 96 degrees of freedom
## Multiple R-squared:  0.4123, Adjusted R-squared:  0.394
## F-statistic: 22.45 on 3 and 96 DF,  p-value: 4.257e-11
```

From the linear model summary,new $\hat{\beta}_1 = 670.3$ which indicates that If we increase the log(**population**) by 1 unit, keeping the log(**nonwhite**) and log(**density**) fixed, the value of **crime** will increase on average in 670.3 units.

Hypothesis test

$H_0 : \beta_1 = 0$ for any given β_i

$H_\alpha : \beta_1 \neq 0$ for any given β_i

The test statistic,t,provided in the output is obtained from

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1|X)} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1|X)}$$

and it follows a **t** distribution with $n-(p+1) = 96$ degrees of freedom. The corresponding $p-value = 9.72 \times 10^{-07}$ for this two-tailed test is also provided in the output. Since the p-value is very small, we reject the hypothesis that the coefficient of log(**population**) is zero. It means that the log(**population**) has a significant influence on **crime**

Moreover,$R^2 = 0.4123$ means that, if our model assuptions hold, about 41.23% of the variation in **crime** is explained jointly by log(**population**),log(**nonwhite**) and log(**density**)

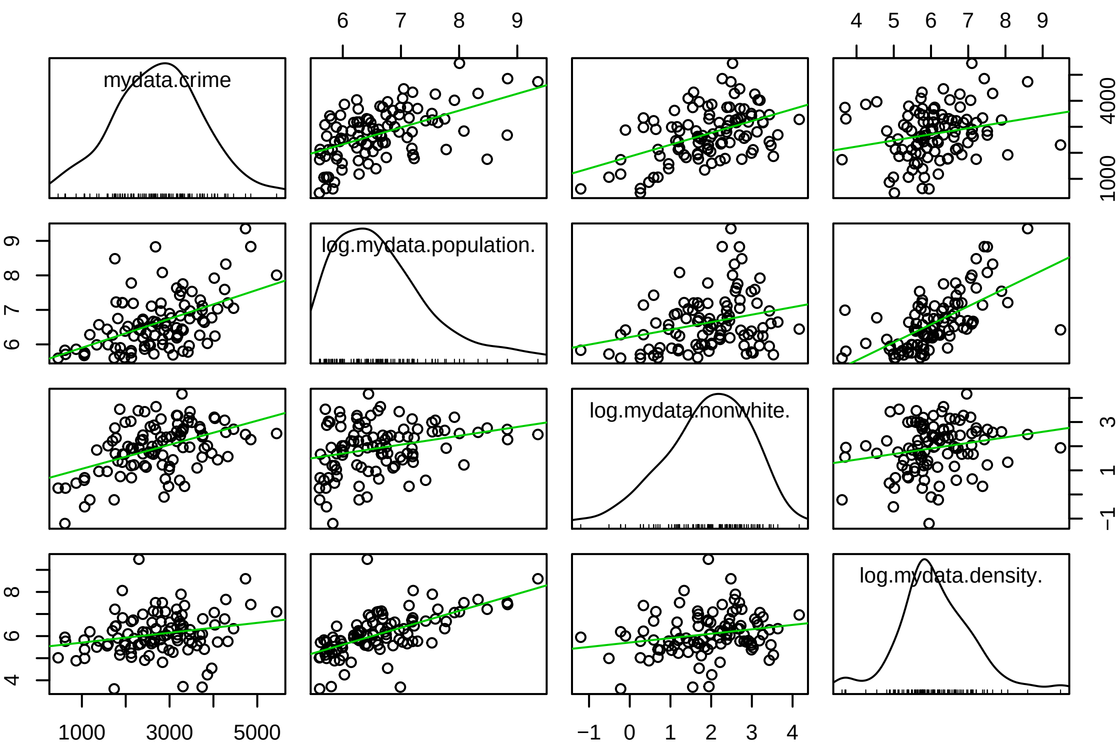
3b

Simple linear regression(part2): $\hat{\beta}_1 = 648.9$.

Multiple regression(part3):$\hat{\beta}_1 = 670.3$.

We observe that the $\hat{\beta}_1$ in part(3) is larger compared with part(2).

```
scatterplotMatrix(~mydata$crime+log(mydata$population)+log(mydata$nonwhite)+log(mydata$density),
                  smoother =F)
```




```
cor(log(mydata))
```

```
##            population  nonwhite   density     crime
## population  1.0000000 0.2819964 0.6328604 0.4761941
## nonwhite    0.2819964 1.0000000 0.2141434 0.5435710
## density     0.6328604 0.2141434 1.0000000 0.2316185
## crime       0.4761941 0.5435710 0.2316185 1.0000000
```

By obtaining scatter plot matrix and the correlation matrix, we conclue that log(**nonwhite**) and log(**density**) have obvious relationships with correlation coefficient 0.281996 and 0.6329 relatively. Thus, as log(**density**) and log(**nonwhite**) are also in this linear model, the effect of log(**population**) on **crime** has changed.

3c

Simple linear regression(part2): $R^2 = 0.2725$.

Multiple regression(part3): $R^2 = 0.4123$.

In simple linear regression, 27.25% of the variation in **crime** is explained by log(**population**). In terms of multiple regression, 41.23% of the variation in **crime** is explained jointly by log(**population**), log(**nonwhite**) and log(**density**).

As more predictors are fitted in multiple regression,the multiple regression in part 3 performs better than the simple linear regression in part 2.

4

4a

Based on the summay of our model **m2**,we found that the **p-value** for regressor log(**density**) is 0.063(with t-value=(-1.88),df=96), which is larger than 0.05 and indicates insignificance in model **m2**. Therefore, we construct a new model contains only the log(**population**) and log(**nonwhite**)

```
m3=lm(crime ~ log(population)+log(nonwhite), data=mydata)
summary(m3)
```

```
##
## Call:
## lm(formula = crime ~ log(population) + log(nonwhite), data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2233.64  -629.75     7.43   576.79  1781.74
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1395.7      658.9  -2.118   0.0367 *
## log(population)    523.3      102.7   5.096 1.71e-06 ***
## log(nonwhite)      342.7       79.0   4.338 3.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 775.6 on 97 degrees of freedom
## Multiple R-squared:  0.3907, Adjusted R-squared:  0.3781
## F-statistic:  31.1 on 2 and 97 DF,  p-value: 3.669e-11
```

From the summary, we observed the $p-value$ for log(**population**) and log(**nonwhite**) are 1.71×10^{-06} and 3.52×10^{-05} relatively. These small **p-value**s indicates both regressior are significant in our model. Meanwhile, we have $R^2 = 0.3907$, indicates that 39.07% of variation in **crime** is explained by log(**population**) and log(**nonwthie**).

In general, model **m3** seems the most adequate linear model to explain changes in **crime**

4b

```
confint(m3,level=0.98)
```

```
##                         1 %     99 %
## (Intercept)     -2954.0935 162.8020
## log(population)   280.3683 766.1831
## log(nonwhite)     155.8346 529.5472
```

From the result, It shows that if the assumptions of the model hold, we are 98% confident that the coefficient of log(**population**) is in the interval (280.368,766.183)

4c

```
newdata = data.frame(population=1150)
predict(m1, newdata, interval="predict", level=0.99)
```

```
##        fit      lwr      upr
## 1 3005.106 775.8865 5234.325
```

As we can see the fitted value, $\hat{y_*}$, with the value of 3005.106. It means that when a city's population is 1.15 million, the predicted **crime** will be 3005.106 with 99% confidence interval $(775.8865, 5234.325)$. Also we observe that the confidence interval is wide, it may because it takes into account the uncertainty of predicting a new response.