S631 Takehome 1

Erik Parker

October 9, 2017

1. What is the probability that a randomly selected city has a crime rate higher than 3200 (per 100000 people)? Is the crime rate normally distributed?

> 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.)
>
> Signed: Erik Parker

```
rm(list = ls())
library(alr4)
library(ggplot2)

cities <- read.table("takehome1.txt", header = TRUE)

str(cities)
```

```
## 'data.frame':    110 obs. of  4 variables:
##  $ population: int  675 713 NA 534 1261 1330 331 1981 315 305 ...
##  $ nonwhite  : num  7.3 2.6 3.3 0.8 1.4 22.8 7 21.6 20.7 0.6 ...
##  $ density   : int  746 322 NA 491 1612 770 41 877 240 147 ...
##  $ crime     : int  2602 1388 5018 1182 3341 2805 3306 4256 2117 1063 ...
```

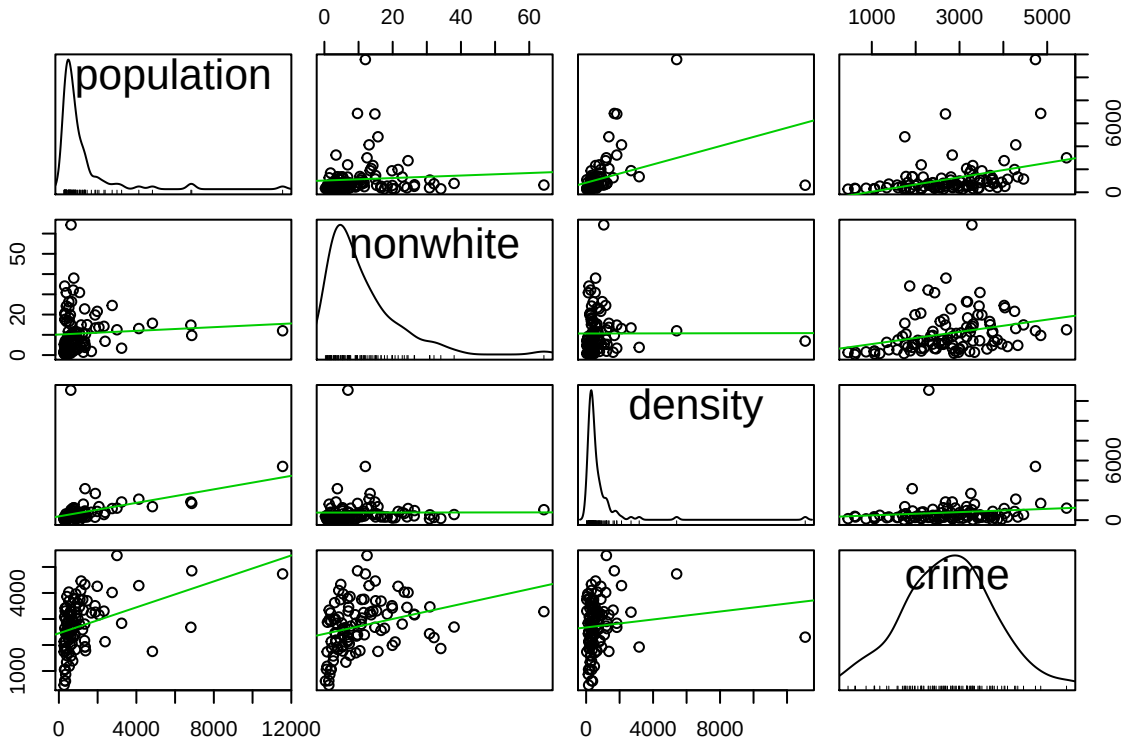
```
summary(cities)
```

```
##    population         nonwhite        density            crime
##  Min.   :  270.0   Min.   : 0.30   Min.   :   37.0   Min.   : 458
##  1st Qu.:  398.8   1st Qu.: 3.40   1st Qu.:  266.5   1st Qu.:2067
##  Median :  664.0   Median : 7.20   Median :  412.0   Median :2698
##  Mean   : 1136.0   Mean   :10.80   Mean   :  765.7   Mean   :2714
##  3rd Qu.: 1167.8   3rd Qu.:14.88   3rd Qu.:  773.2   3rd Qu.:3305
##  Max.   :11551.0   Max.   :64.30   Max.   :13087.0   Max.   :5441
##  NA's   :10                        NA's   :10
```

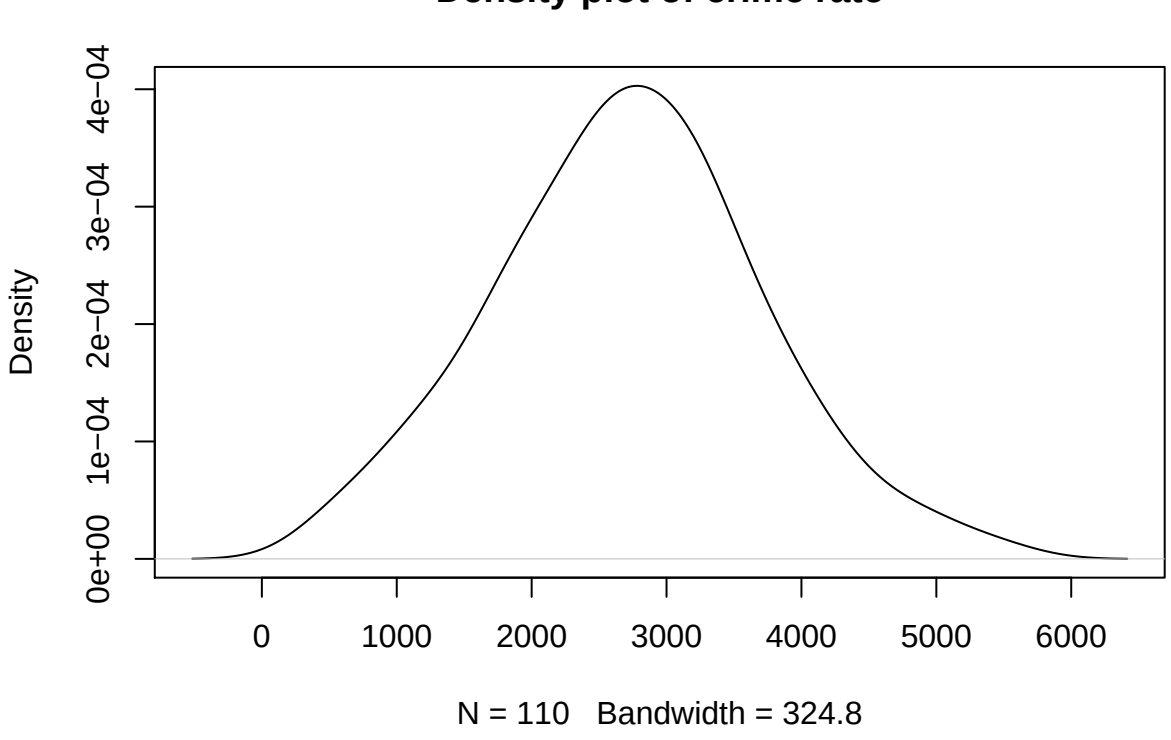
```
prob32 <- sum(cities$crime > 3200)/length(cities$crime)
prob32
```

```
## [1] 0.3
```

```
scatterplotMatrix(cities, smooth = FALSE)
```




```
plot(density(cities$crime), main = "Density plot of crime rate")
```



Here we can see that the probability of randomly selecting a city from this dataset that ahs a crime rate higher than 3200 is 0.3, or 30%. Furthermore, from the density plot we can also see that the crime rate is roughly normally distributed.

2. Regress *crime* on *population*.

```
m1 <- lm(crime ~ log(population), data = cities)

# log transformation of population seems necessary as population is very
# right skewed.
```

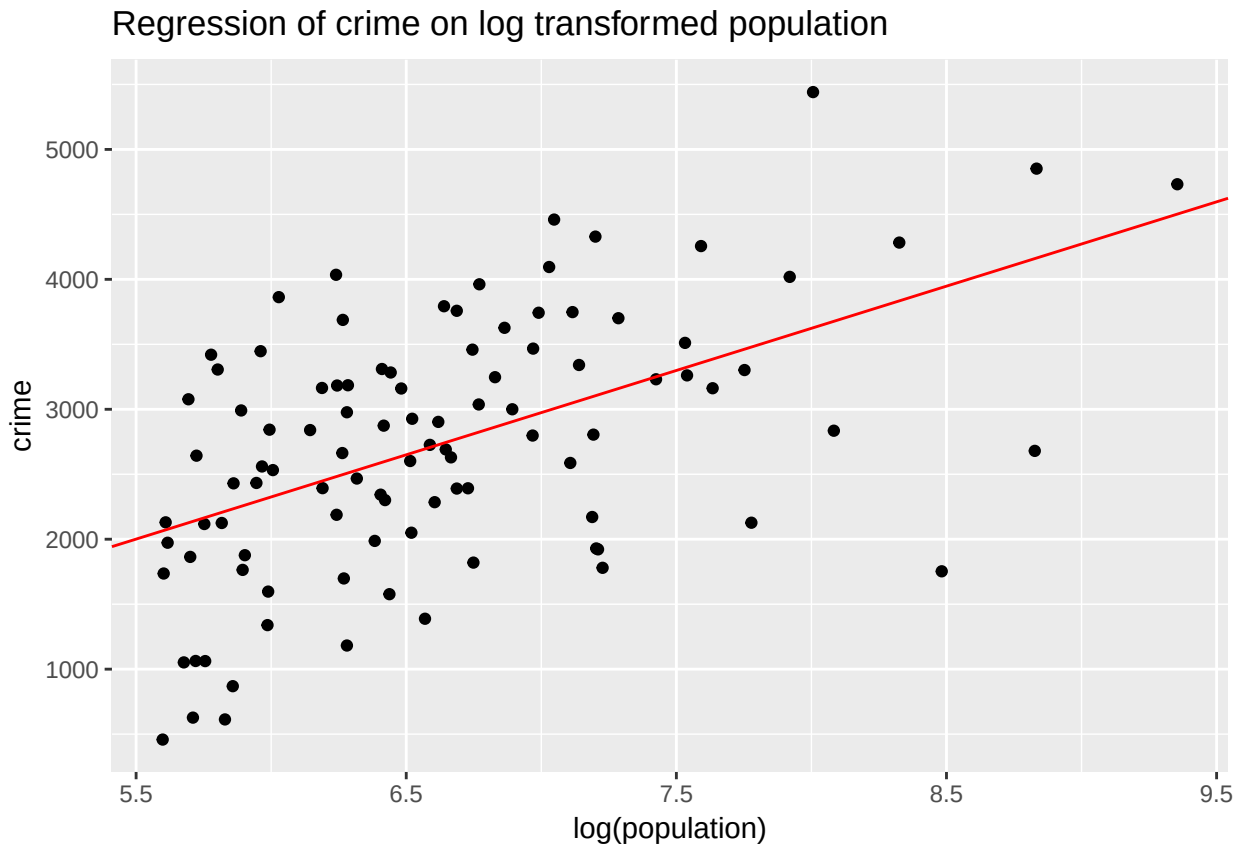
a. Describe visually the relationship between the response and regressor. Are there any clear violations of the linear model assumptions?

```
summary(m1)
```

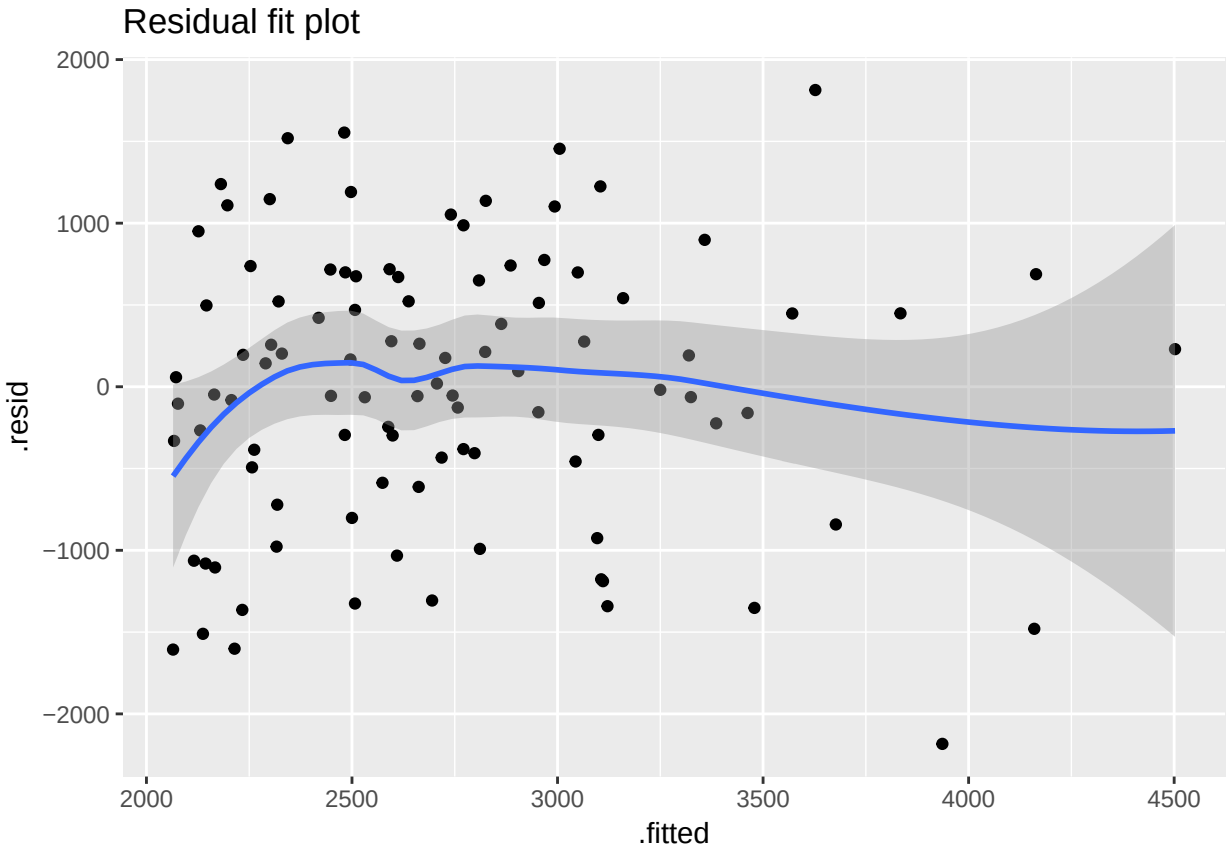
```
##
## Call:
## lm(formula = crime ~ log(population), data = cities)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2183.20  -465.95    38.71   655.15  1813.91
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1568.0      714.9  -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
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.2725, Adjusted R-squared:  0.2651
## F-statistic:  36.7 on 1 and 98 DF,  p-value: 2.553e-08
```

```
int1 <- coef(summary(m1))[1, 1]
beta1 <- coef(summary(m1))[2, 1]

ggplot(cities, aes(x = log(population), y = crime)) + geom_point() + geom_abline(intercept = int1,
    slope = beta1, col = "red") + labs(title = "Regression of crime on log transformed population")
```




```
ggplot(m1, aes(.fitted, .resid)) + geom_point() + geom_smooth() + labs(title = "Residual fit plot")
```



Based on these plots, I don't see any clear violations of the linear model assumptions. From the first plot of log(population) vs crime rate, we can see that there is a pretty clear increasing, linear, relationship between the regressor and response. There is no obvious curviture to the data, there are no obvious, egregious outliers present, and despite the higher concentration of points at lower population levels, there is still a pretty good spread of points along the X axis. The second plot showing the residuals vs the fitted values for the regression also seems to have no real issues. It closely resembles a null-plot as we see no real pattern to the data, there is no cone shape to the data indicating that the variance increases or decreases over different fitted values, and this data can be well summarized by a straight line.

b. Interpret the estimate $\hat{\beta}_1$, explain the underlying hypothesis test assumed in the R output, and interpret the results of the test.

```
summary(m1)
```

```
##
## Call:
## lm(formula = crime ~ log(population), data = cities)
##
## 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
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.2725, Adjusted R-squared:  0.2651
## F-statistic:  36.7 on 1 and 98 DF,  p-value: 2.553e-08
```

```
exp(1) * 1000
```

```
## [1] 2718.282
```

The estimate $\hat{\beta}_1$ of 648.9 obtained from this first model tells us that for every unit increase in log(population) (which translates to an increase in population by roughly 2718 people), we see an increase in the crime rate (per 100,000 people) of a city by 648.9.

The hypothesis test from this output, for β_1 is a two-tailed test testing the null hypothesis that $\beta_1 = 0$ against the alternative that $\beta_1 \neq 0$. From the summary of our model, we can see that this test returned a very low p-value of 2.55e-08, meaning that we can safely reject the null hypothesis that $\beta_1 = 0$ and support the alternative that β_1 is not equal to zero.

3. Regress *crime* on all three predictors. Use *log* transformation on the regressors if appropriate.

```
m2 <- lm(crime ~ log(population) + log(density) + log(nonwhite), data = cities)

# here I log transformed all the predictors as they were all strongly
# left-skewed, just like was seen with population in the previous question.

summary(m2)
```

```
##
## Call:
## lm(formula = crime ~ log(population) + log(density) + log(nonwhite),
##     data = cities)
##
## 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(density)     -195.29     103.91  -1.879   0.0632 .
## log(nonwhite)     349.74      78.07   4.480 2.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 765.7 on 96 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.4123, Adjusted R-squared:  0.394
## F-statistic: 22.45 on 3 and 96 DF,  p-value: 4.257e-11
```

a. Interpret the estimate $\hat{\beta}_1$, explain the underlying hypothesis test assumed in the *R* output for β_1 and interpret the corresponding output results.

In this multiple regression model, $\hat{\beta}_1$ corresponds to the log transformed population regressor, just like in the previous single regression. However, the interpretation of this estimate is now quite different. In the single regression case we could say that the estimated coefficient was simply the influence that log(population) had on crime rate for every unit change, but now this value of $\hat{\beta}_1$ corresponds to the *additional* contribution of log(population) on crime rate once all the contributions of the other regressors (log(density) and log(nonwhite)) have already been accounted for. So, the $\hat{\beta}_1$ value of 670.28 means that for every unit change in log(population) (so like before, an addition of roughly 2718 people), there is an increase in crime rate by 670.28 per 100,000, *only when* all other regressors are already present in the model and are held constant.

Additionally, the meaning and interpretation of the hypothesis test for β_1 has also changed. Now the null hypothesis that is being tested is that $\beta_1 = 0 \quad with \quad \beta_0, \beta_2, \beta_3 \quad arbitrary$. This essentially means that we are testing the influence of the regressor corresponding to β_1 only, while not caring about the other regressors. So this test is looking at the effect of adding log(population) last to a multiple regression mean function that already contains all of the other regressors. The results of this hypothesis test on β_1 show us a p-value of near 0, so we can safely reject the null hypothesis that $\beta_1 = 0$ when added last to a model containing the other regressors, meaning that log(population) has a significant explanatory influence on crime rate even when added last to our model.

b. Compare the effect that *population* has on *crime* between the single and multiple regression model and explain why the effect changes.

```
summary(lm(crime ~ log(population) + log(density), data = cities))
```

```
##
## Call:
## lm(formula = crime ~ log(population) + log(density), data = cities)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2234.9  -596.2    81.5   621.1  1805.2
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1392.3      719.5  -1.935   0.0559 .
## log(population)    781.4      137.4   5.686 1.37e-07 ***
## log(density)      -172.9      113.5  -1.523   0.1310
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 837.6 on 97 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.2895, Adjusted R-squared:  0.2748
## F-statistic: 19.76 on 2 and 97 DF,  p-value: 6.332e-08
```

As discussed above, the effect that *population* (specifically the log of population) has on *crime* is positive and significant in both models. The main difference between the two, is that the influence of *population* is actually greater in the multiple regression compared to the single one when we consider the value for $\hat{\beta}_1$. This is a bit confusing, as when adding variables to a multiple regression we might most often expect to see a *decrease* in influence as different variables can often explain the same variation and so when they are added last to the model there is already

another variable present which explains the same variation they do, so their effect is less than when they are in a single regression model alone. Here though, we see that the additional influence of log(population) on crime rate when the contributions of the other regressors are accounted for than when we consider the influence of log(population) on crime rate by itself. Why is this? It could be a meaningless difference, as the significance from the hypothesis test given by the p-value for β_1 of log(population) is actually slightly lower in the multiple regression, and the standard error is higher, meaning that the difference between the coefficients between the two models could be just due to "noise" and not any real effect. Alternatively, we can see that the regressor log(density) has a marginally non-significant and negative slope estimate when added to the multiple regression. So it could be that the inclusion of this regressor leads to a compensatory increase in the slope of log(population) when added last to our model. This idea is suppoted by the multiple regression shown above between just the log of population and density. There we see that the slope of log(population) increases a great deal over it's single regression and full multiple regression model values, suggesting that the slope is increasing in these instances to counteract the influence of log(density) which is even less significant when added last to a model just containg log(population).

So overall, it seems fairly likely that the effect of *population* on *crime* is increasing between the first and second model due most to the presence of *density*.

Compare R^2 from parts 2 and 3. What can you conclude?

The values of R^2 from parts 2 and 3 respectively are 0.2725 and 0.4123. This tells us that in the single regression model, 27.25% of the variation found in the response (crime rate) was being explained, while 41.23% of this variation was explained by the regressors in the multiple regression model. This means that the addition of the other two regressors helped explain 13.98% more variation in the response, and so their addition was useful, though we don't know yet the additional contribution of each of the new regressors.

4. What would be the most adequate linear model used to explain changes in crime? Use that model to answer the following questions.

Will choose the best model by looking at R^2 values of the different possible models, and comparing them to find the model that explains the most variation in the response.

```
summary(lm(crime ~ log(nonwhite), data = cities))$r.squared
```

```
## [1] 0.1971594
```

```
summary(lm(crime ~ log(population), data = cities))$r.squared
```

```
## [1] 0.272482
```

```
summary(lm(crime ~ log(density), data = cities))$r.squared
```

```
## [1] 0.05263081
```

```
summary(lm(crime ~ log(nonwhite) + log(population) + log(density), data = cities))$r.squared
```

```
## [1] 0.4123184
```

```
summary(lm(crime ~ log(nonwhite) + log(population), data = cities))$r.squared
```

```
## [1] 0.3906948
```

a. Why is this your chosen model?

When comparing the R^2 values between the different models above we can see that none of the single regressions explain more than 30% of the variation seen in our response. When we look at the two best multiple regression models we find that even though in the last question we saw that log(density) was only marginally significant when included in the full model, dropping that regressor leads to a slightly lower R^2 value, meaning that *density* is still useful in explaining some of the variation seen in the response (roughly 2%). So, even though it is only marginally important, I will include it in my chosen model which will be the full multiple regression model of form: *lm(crime ~ log(nonwhite) + log(population) + log(density), data = cities)*.

b. Obtain and interpret a 98% confidence interval for β_1

```
confint(m2, level = 0.98)
```

```
##                         1 %      99 %
## (Intercept)     -2753.2403 366.04220
## log(population)   367.3513 973.20595
## log(density)     -441.1289  50.54051
## log(nonwhite)     165.0292 534.44811
```

So, from this we can say that we are 98% confident that the true value for β_1 falls within the interval of 367.35 and 973.21.

c. With 99% confidence, what would the crime rate for a city with a population of 1.15 million people be?

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

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

```
# Because we are only conditioning on population here, we need to use a
# model that contains population as the only predictor.
```

So using our single linear regression model, we can say with 99% confidence that the crime rate for a city with a population of 1.15 million would be between 775.89 and 5234.33 per 100,000 people.