36-401 Modern Regression HW #3 Solutions

DUE: 9/22/2017

Problem 1 [10 points]

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_i &= \frac{1}{n}\sum_{i=1}^{n}(\widehat{\beta}_0 + \widehat{\beta}_1 X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\overline{Y} - \widehat{\beta}_1\overline{X} + \widehat{\beta}_1 X_i) \\
&= \overline{Y} - \widehat{\beta}_1\overline{X} + \frac{\widehat{\beta}_1}{n}\sum_{i=1}^{n} X_i \\
&= \overline{Y}.
\end{aligned}$$

Problem 2 [40 points total]

(a) (20 pts.)

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n-2}\sum_{i=1}^{n} e_i^2\right] &= \frac{\sigma^2}{n-2}\mathbb{E}\left[\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2}\right] \\
&= \frac{\sigma^2}{n-2}\cdot(n-2) \qquad \text{since } \frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \sim \chi^2_{n-2} \\
&= \sigma^2.
\end{aligned}$$

(b) (20 pts.)

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} e_i^2\right] - \sigma^2 &= \frac{n-2}{n}\mathbb{E}\left[\frac{1}{n-2}\sum_{i=1}^{n} e_i^2\right] - \sigma^2 \\
&= \frac{n-2}{n}\sigma^2 - \sigma^2 \qquad \text{from part (a)} \\
&= -\frac{2}{n}\sigma^2.
\end{aligned}$$

As n becomes large the bias approaches 0.

Problem 3 [50 points total]

(a) (2 pts.)

```r
dat <- read.csv("bea-2006.csv")
dim(dat)
```

```
## [1] 366   7
```

```r
head(dat)
```

```
##                             MSA pcgmp    pop finance prof.tech     ict
## 1                  Abilene, TX 24490 158700 0.09750        NA 0.01621
## 2                    Akron, OH 32890 699300 0.12940   0.05440      NA
## 3                   Albany, GA 24270 163000 0.08217        NA 0.00708
## 4 Albany-Schenectady-Troy, NY 36840 850300 0.15780   0.09399 0.04511
## 5              Albuquerque, NM 37660 816000 0.15990   0.09978 0.20500
## 6               Alexandria, LA 25490 152200 0.09152   0.03790 0.01134
##   management
## 1         NA
## 2   0.054310
## 3         NA
## 4         NA
## 5   0.006509
## 6   0.015210
```

The data file has a column for the name of the city, and one column for each of the six statistics.

(b) (2 pts.)

```r
summary(dat[,2:7])
```

```
##      pcgmp            pop              finance          prof.tech
##  Min.   :14920   Min.   :   54980   Min.   :0.03845   Min.   :0.01474
##  1st Qu.:26532   1st Qu.:  135625   1st Qu.:0.10403   1st Qu.:0.02932
##  Median :31615   Median :  231500   Median :0.14140   Median :0.04212
##  Mean   :32923   Mean   :  680898   Mean   :0.15082   Mean   :0.04905
##  3rd Qu.:38212   3rd Qu.:  530875   3rd Qu.:0.18122   3rd Qu.:0.05932
##  Max.   :77860   Max.   :18850000   Max.   :0.38480   Max.   :0.19080
##                                     NA's   :12        NA's   :112
##       ict            management
##  Min.   :0.00349   Min.   :0.00042
##  1st Qu.:0.01215   1st Qu.:0.00294
##  Median :0.02218   Median :0.00651
##  Mean   :0.03910   Mean   :0.00908
##  3rd Qu.:0.04072   3rd Qu.:0.01191
##  Max.   :0.58600   Max.   :0.05431
##  NA's   :76        NA's   :157
```

(c) (6 pts.)

```r
par(mfrow=c(1,2))
hist(dat$pop, breaks = 100, main = "", xlab = "Population", ylab = "Frequency")
text(max(dat$pop), 20, adj = 0.8, labels = "New York")
points(max(dat$pop),0, col = "red", pch = 19)
hist(log(dat$pop), breaks = 100, main = "", xlab = "log(Population)", ylab = "Frequency")
```

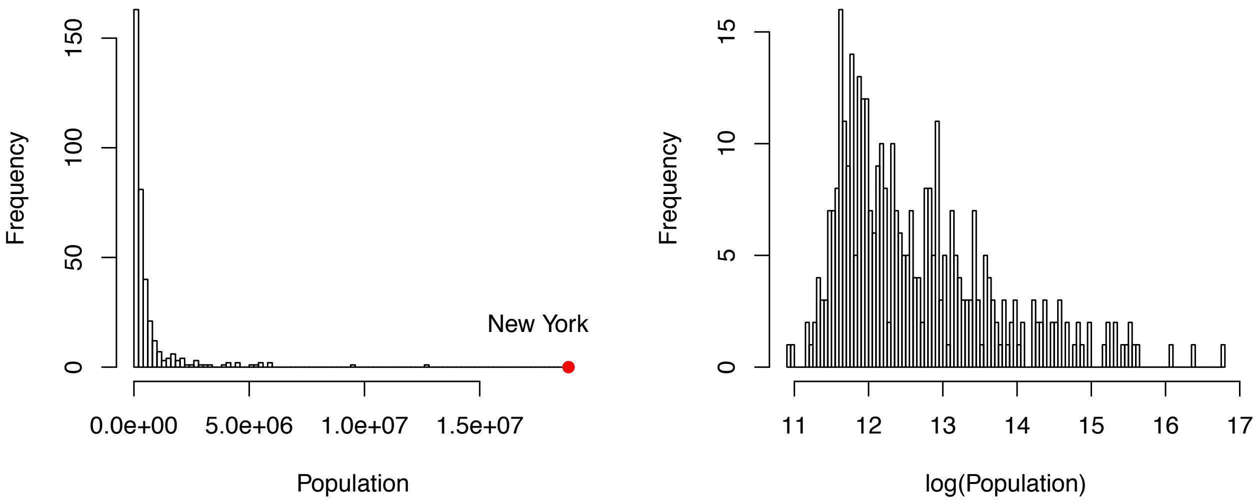

Figure 1: Histogram of Populations for 366 U.S. Metropolitan Areas in 2006 (**Left**: Raw scale. **Right**: Log scale)

As seen in Figure 1, the distribution of city (metro area) populations has a highly positive skewness, with the New York-Northern New Jersey-Long Island area having the highest documented population. Plotting the distribution on the log scale (right panel of Figure 1) allows for a more informative inspection.

```r
boxplot(dat$pcgmp, boxwex = 0.7, main = "", ylab = "Per Capita GMP")
text(max(dat$pcgmp), labels = "Bridgeport-Stamford-Norwalk, CT", cex = 0.7, adj = -0.075)
points(max(dat$pcgmp), pch = 19, col = "red")
```

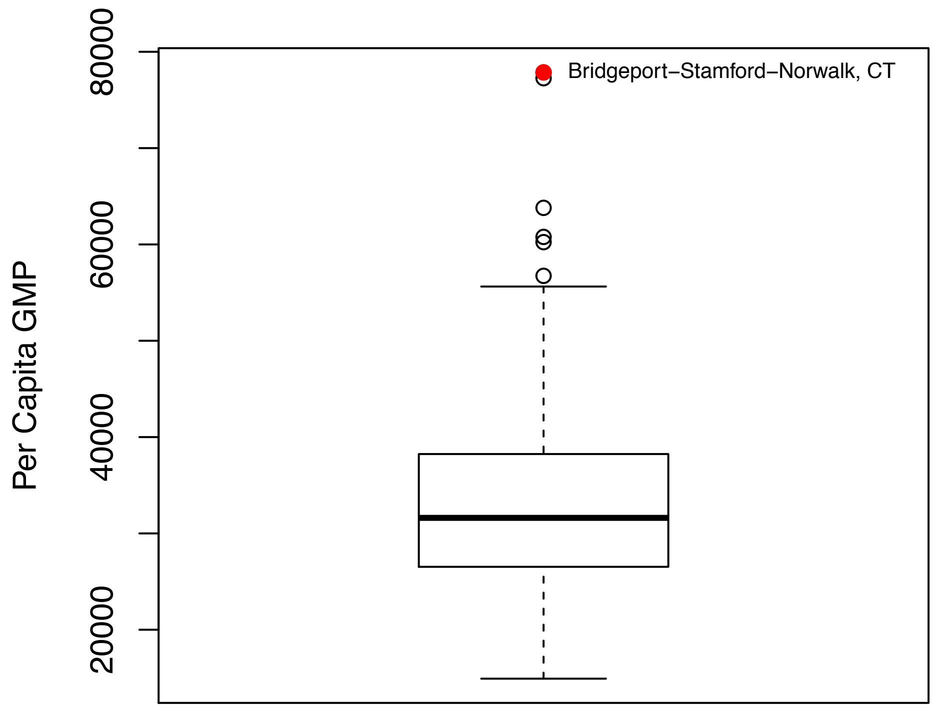

Figure 2: Box plot of Per Capita GMP for 366 U.S. Metropolitan Areas in 2006

Similar to population, the distribution of per-capita GMP has a positive skewness. The mean per-capita GMP over all 366 cities is approximately $\$33,000$ per person-year, while Bridgeport-Stamford-Norwalk (shown in red), CT boasts a per-capita GMP of $\approx \$78,000$ per person-year, approximately 5 standard deviations above the mean.

(d) (6 pts.)

```
par(mfrow = c(1,2))
with(dat, plot(pop, pcgmp, xlab = "Population", ylab = "Per Capita GMP"))
with(dat, plot(log(pop), pcgmp, xlab = "log(Population)", ylab = "Per Capita GMP"))
```

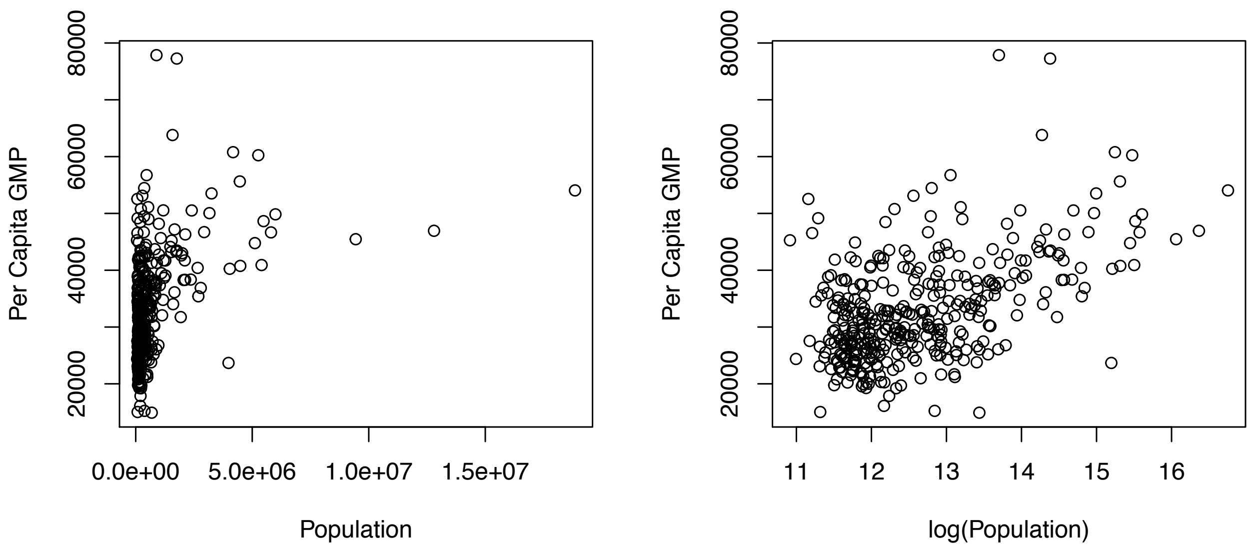

Figure 3: Scatterplot of Per Capita GMP vs. Population for 366 U.S. Metropolitan Areas in 2006 (**Left**: Raw scale; **Right**: Log scale).

Per-capita GMP and population have a positive association. This trend is most apparent on the log scale (right panel).

(e) (5 pts.)

```
n <- nrow(dat)
b1 <- with(dat, (n-1)/n * cov(pop,pcgmp) / ((n-1) / n * var(pop)))
b0 <- with(dat, mean(pcgmp) - b1 * mean(pop))
print(list(b0,b1))
```

```
## [[1]]
## [1] 31277.57
##
## [[2]]
## [1] 0.002416201
```

The estimated linear regression parameters are

$$\widehat{\beta}_0 = 31277.57 \quad \text{and} \quad \widehat{\beta}_1 = 0.002416201.$$

(f) (3 pts.)

```r
model <- lm(pcgmp ~ pop, data = dat)
model$coefficients
```

```
##  (Intercept)          pop
## 3.127757e+04 2.416201e-03
```

The intercept and slope parameters given by `lm` are

$$\widehat{\beta}_0 = 31277.57 \quad \text{and} \quad \widehat{\beta}_1 = 0.002416201,$$

which, as expected, matches what we computed by hand in part (d).

(g) (4 pts.)

```r
with(dat, plot(pop, pcgmp, xlab = "Population", ylab = "Per Capita GMP",
               cex = 0.5, log = "x"))
abline(model, lwd = 2, col = "red")
```

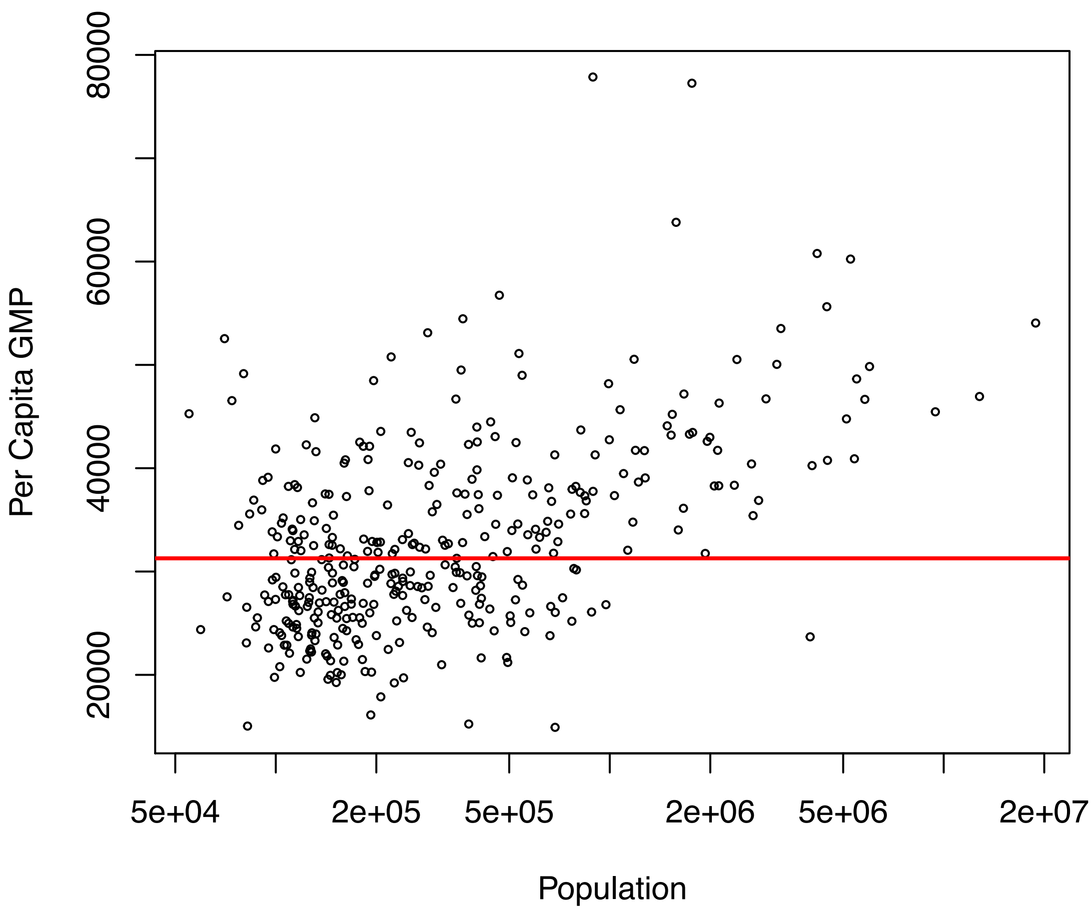

Figure 4: Scatterplot of Per Capita GMP vs. Population for 366 U.S. Metropolitan Areas in 2006, with linear regression shown in red.

Figure 4 shows per-capita GMP vs. Population with the least squares regression line plotted in red. Here we have plotted the x-axis on the log-scale so we can better examine the regression fit, but it is important to note that the regression is still on `Population`, *not* `log(Population)`. The assumptions of the simple linear regression do not hold here. In particular, the linear model significantly understimates the the per-capita GMP for cities with relatively large populations. The fit is somewhat better for small populations.

(h) (4 pts.)

```
index <- which(dat$MSA == "Pittsburgh, PA")
dat[index,]
```

```
##                MSA pcgmp     pop finance prof.tech     ict management
## 262 Pittsburgh, PA 38350 2361000  0.2018    0.0777 0.03434    0.02946
```

In 2006, the population of Pittsburgh, PA was approximately $2,361,000$, with a per-capita GMP of $\$38,350$ per person-year.

```
fitted(model)[index]
```

```
##      262
## 36982.22
```

```
residuals(model)[index]
```

```
##      262
## 1367.775
```

The simple linear model predicts a per-capita GMP of $\sim \$37,000$ per person-year for Pittsburgh, yielding a residual of $\sim \$1,370$ per person-year.

(i) (2 pts.)

```
mean(residuals(model)^2)
```

```
## [1] 70697145
```

The empirical MSE of the simple linear regression is approximately 7.07×10^7.

(j) (1 pts.)

```
residuals(model)[index] ^ 2
```

```
##     262
## 1870810
```

Pittsburgh's squared residual is 1.87×10^6, which is relatively small when compared to the MSE. Notice that we need to square the residual in order to make it directly comparable to the MSE.

(k) (3 pts.)

```r
with(dat, plot(pop, residuals(model), xlab = "Population", ylab = "Residuals",
               cex = 0.5, log = "x"))
abline(0,0, lty = 2)
```

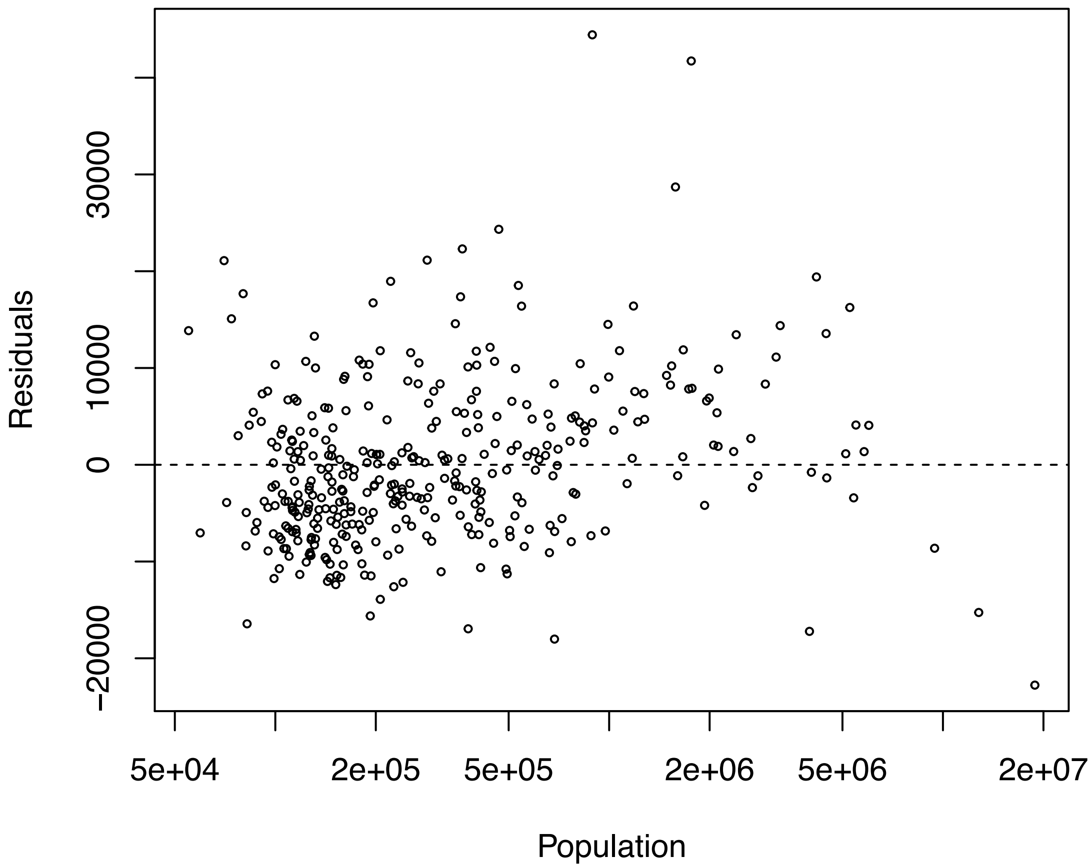

Figure 5: Linear regression residuals of Per Capita GMP on Population.

If the assumptions of the linear regression held, the residuals would have a symmetric and homoskedastic scatter about 0. Figure 5 is not compatible with the standard linear regression assumptions. In particular, most of the residuals are negative at low populations and most of the residuals are positive at populations above $\sim 1,000,000$. Furthermore, there are several highly positive outliers which are not well-explained by the homoskedastic linear model.

(l) (3 pts.)

```
plot(dat$pop, residuals(model)^2, xlab = "Population", ylab = "Squared residuals", pch = 19,
     cex = 0.95, log = "x", col = addTrans("red", 50), font.lab = 2, xaxt = "n", yaxt = "n")
```

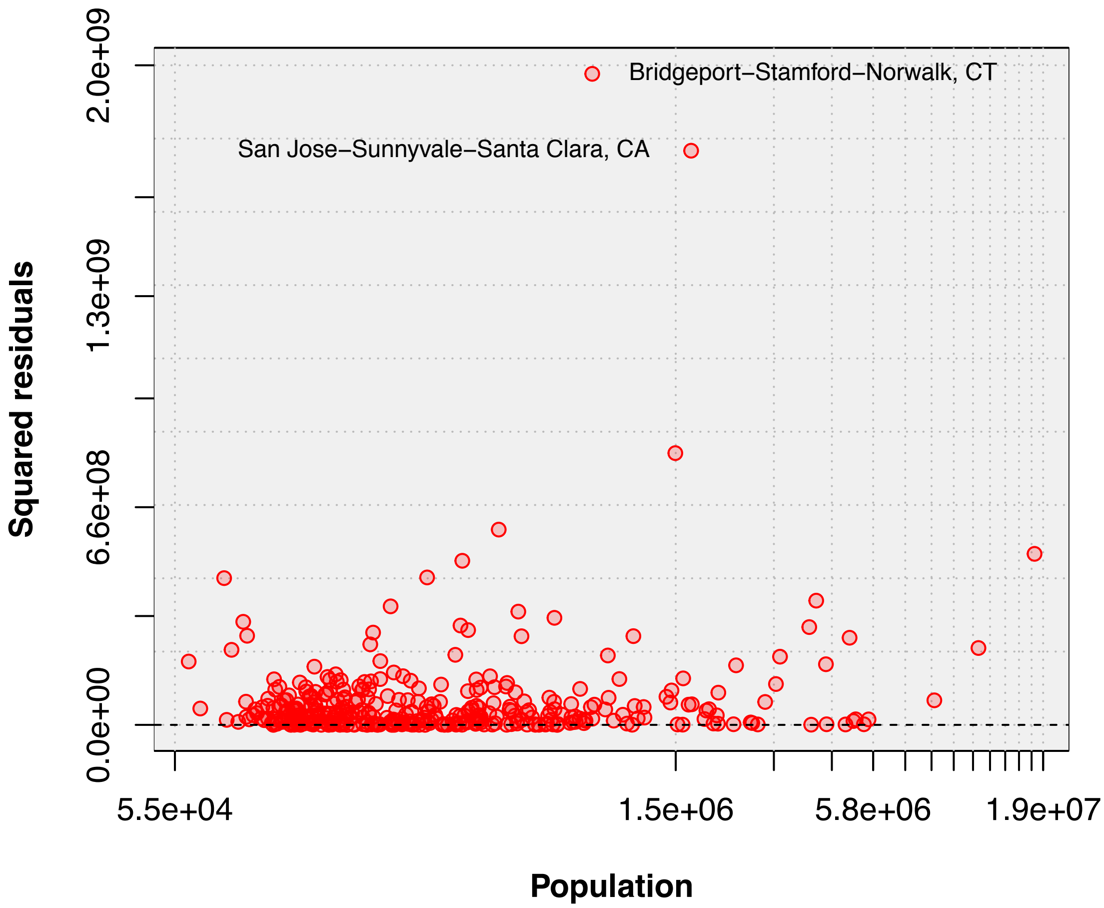

Figure 6: Linear regression squared residuals.

If the homoskedastic assumption of the linear model held, then the squared residuals would have an approximately constant amplitude over all values of population. This assumption is violated by two or three residuals (`Bridgeport-Stamford-Norwalk, CT` and `San Jose-Sunnyvale-Santa Clara, CA` being the worst), but not too egregiously elsewhere.

(m) (3 pts.)

Based on this data set, the total value of all goods and services produced for sale in a city in 2006 (per person) has a highly significant positive correlation with the population of the city. In particular, on the average, a one person increase in population is associated with a \$0.002416201 per person-year increase in per-capita GMP.

(n) (3 pts.)

```r
predict(model, newdata = data.frame(pop = dat$pop[262] + 1e5))
```

```
##        1
## 37223.84
```

The model predicts a per-capita GMP of approximately $\$37,200$ per person-year for a city with 100,000 more people than Pittsburgh, PA.

(o) (3 pts.)

```r
model$coefficients[2] * 1e5
```

```
##      pop
## 241.6201
```

If, by a policy intervention, we added 100,000 people to Pittsburgh's population, the model predicts that the per-capita GMP would increase by approximately $\widehat{\beta}_1 \cdot 100,000 \approx \$240$ per person-year. Note, however, that such a prediction assumes there is a causal relationship between population and per-capita GMP.