STAT-S631

Exam 1

John Koo

Statement

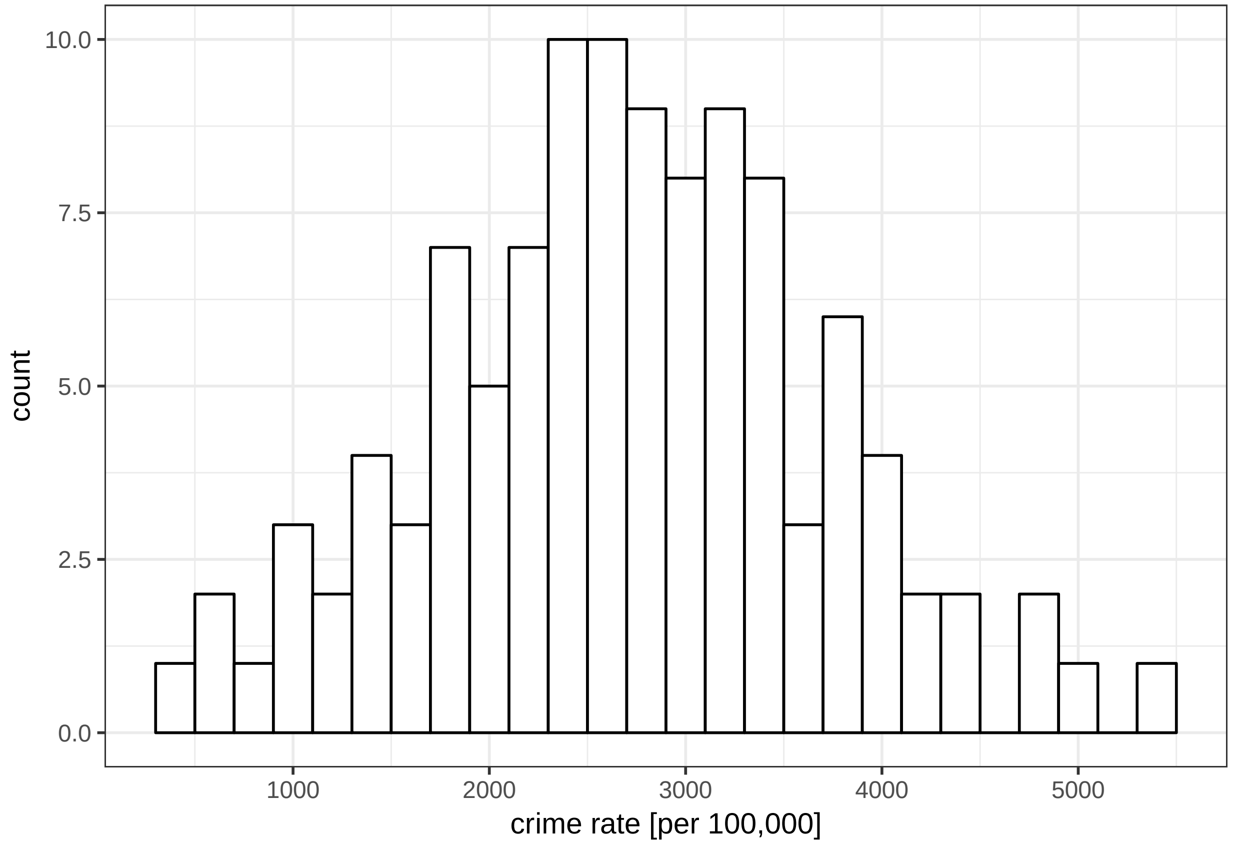
On my honor, I have not had any form of communicationabout this exam with any other individual (including other students, teaching assistants, instructors, etc.).

Signed: *John Koo*

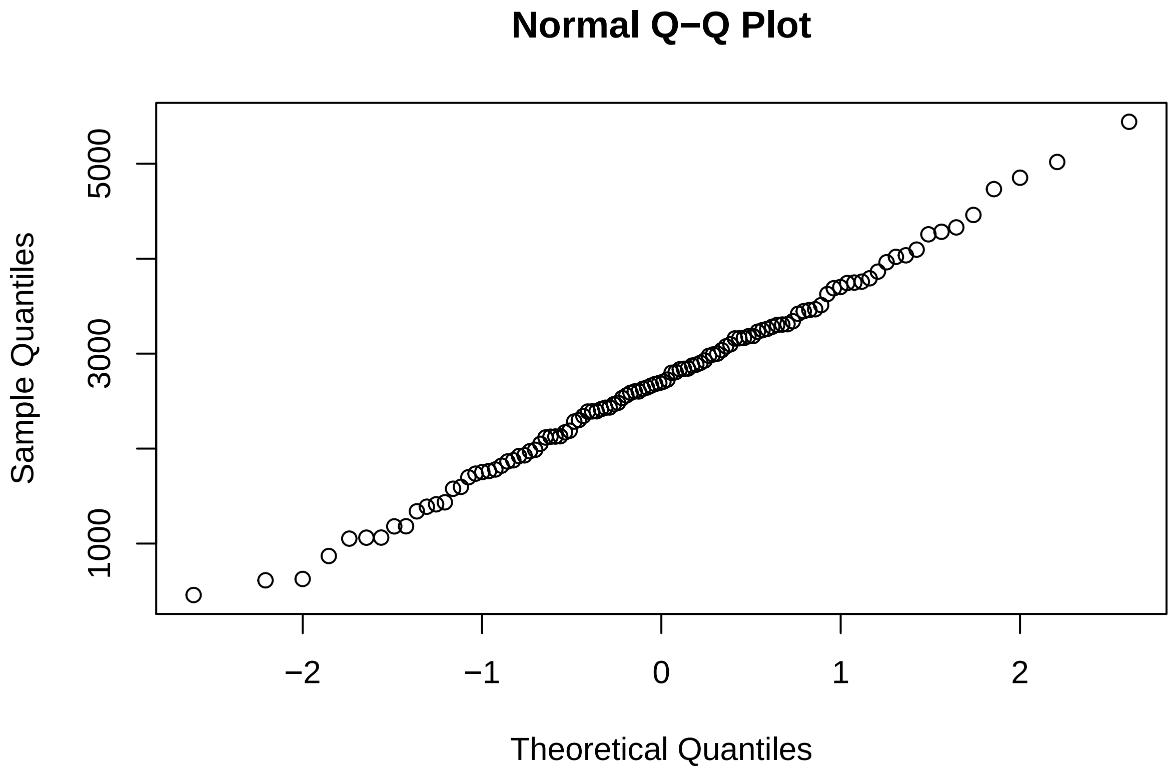
```r
# packages
import::from(magrittr, `%>%`)
dp <- loadNamespace('dplyr')
library(ggplot2)
import::from(GGally, ggpairs)
import::from(gridExtra, grid.arrange)

# plotting stuff
theme_set(theme_bw())

# read data
city.df <- read.table('~/dev/stats-hw/stat-s631/takehome1.txt')
```

Question 1

```r
N <- nrow(city.df)
prop.crime.above.3200 <- nrow(dp$filter(city.df, crime > 3200)) / N

ggplot(city.df) +
  geom_histogram(aes(x = crime),
                 fill = 'white',
                 colour = 'black',
                 binwidth = 200) +
  labs(x = 'crime rate [per 100,000]')
```



```
qqnorm(city.df$crime)
```



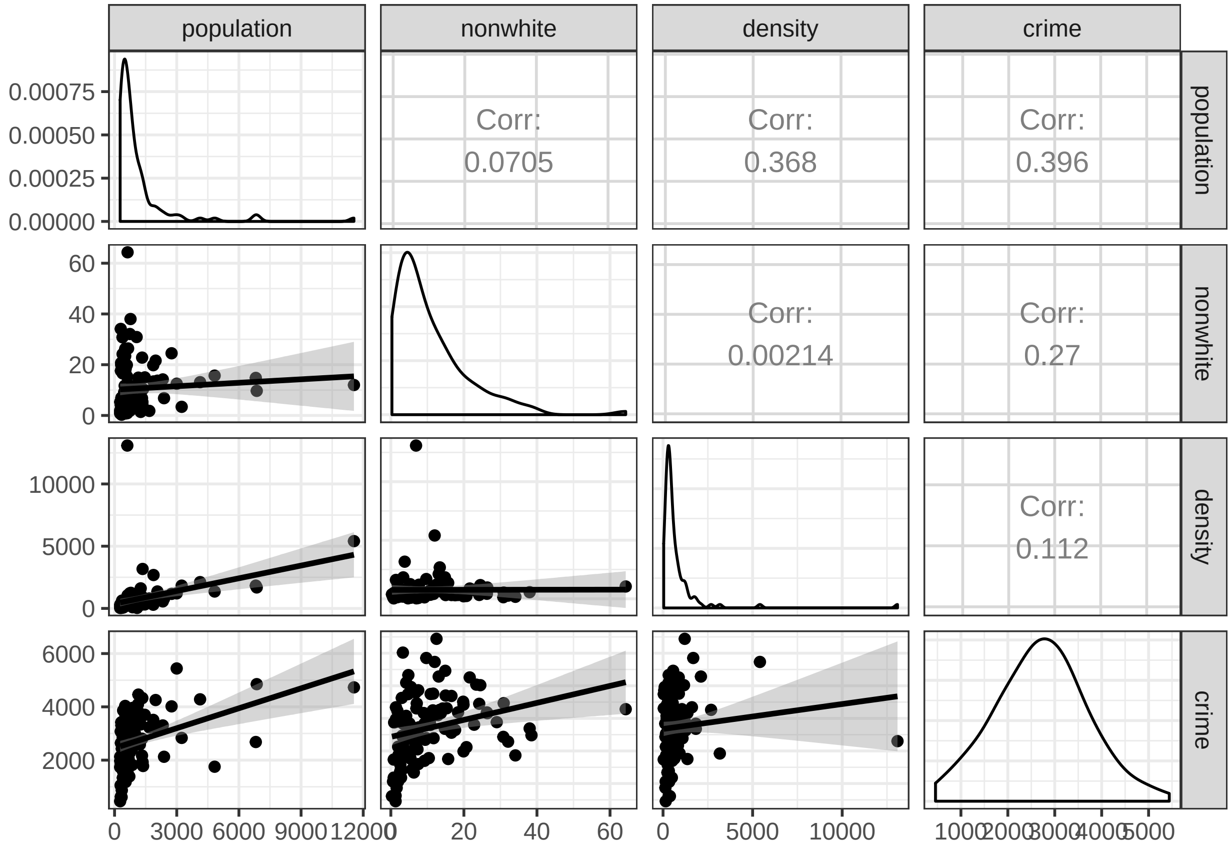
From the data, the estimated probability that a randomly selected city will have a crime rate higher than

3,200 per 100,000 people is 0.3.

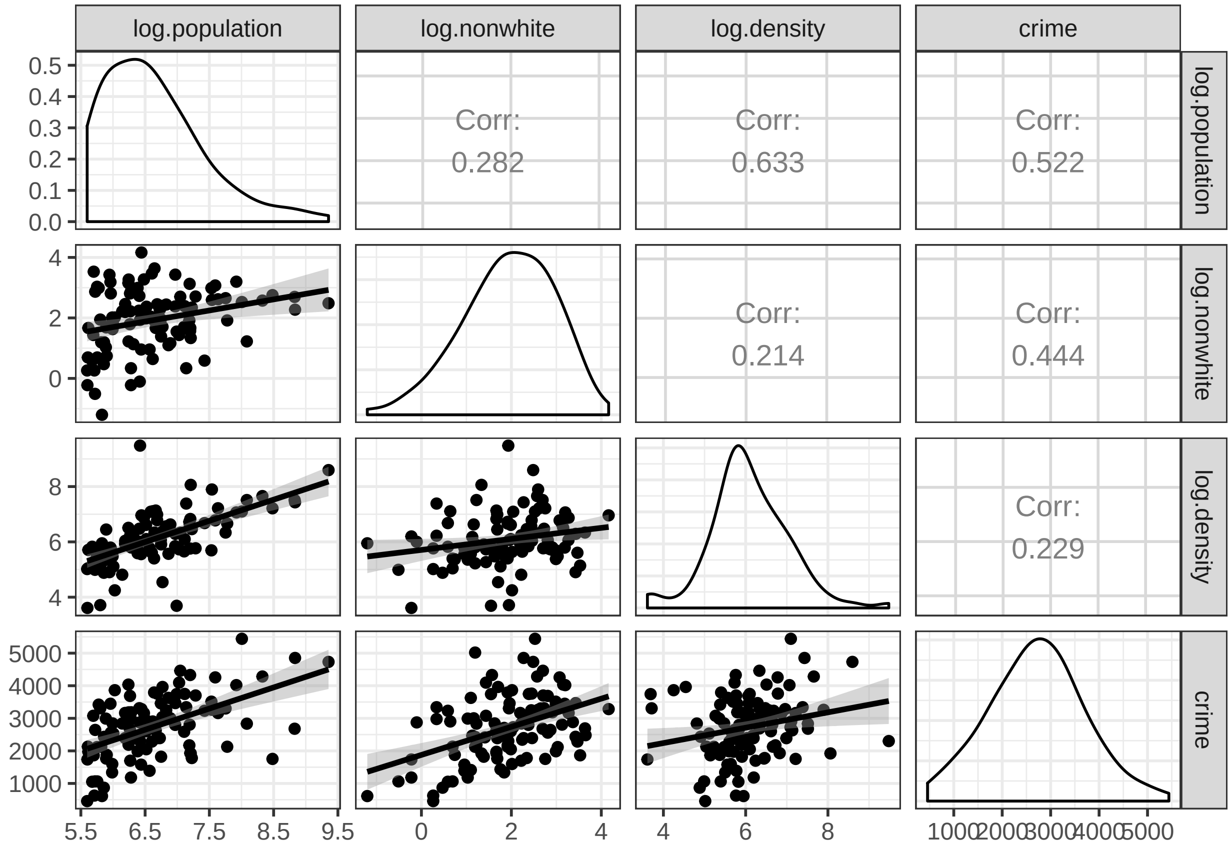
The histogram of `crime` seems to suggest that the data are normally distributed. The histogram is a symmetric, unimodal bell curve. A QQ-plot confirms this.

Question 2

```r
ggpairs(city.df, lower = list(continuous = 'smooth'))
```



```r
city.df %>%
  dp$transmute(log.population = log(population),
               log.nonwhite = log(nonwhite),
               log.density = log(density),
               crime) %>%
  ggpairs(lower = list(continuous = 'smooth'))
```



From the pairwise scatterplots, if `crime` is the response variable, a log transformation is appropriate for all three of the covariates.

```
model.2 <- lm(crime ~ log(population), data = city.df)
summary(model.2)
```

```
Call:
lm(formula = crime ~ log(population), data = city.df)

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
```

Part a

From the scatterplot of `crime` vs `population`, there are no obvious violations of our assumptions. The trend looks linear and there is an even number of points to either side of the OLS line.

Part b

We can interpret $\hat{\beta}_1$, the coefficient for `log(population)`, as follows: For a unit increase in `log(population)`, we expect that, on average, the response variable, `crime`, to increase by 648.9 units.

The underlying hypothesis test is $H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$. We know that $\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t_{n-p-1}$. Under H_0, $\beta_1 = 0$, so $\frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{n-p-1}$. Then our p-value for this hypothesis test is $P(|t| > \frac{\hat{\beta}_1}{se(\hat{\beta}_1)})$. In this case, this value is 2.55×10^{-8}. In other words, assuming H_0 is true, the probability that $\hat{\beta}_1$ is this far or further from 0 (our H_0) is extremely small. So for most reasonably small values of α (e.g., 0.05 or 0.01), we can say that we reject H_0. So we can say that `log(population)` explains some of the variability in `crime`.

Question 3

As mentioned previously in question (2), it appears that a log transformation for all of the covariates is appropriate.

```r
model.3 <- lm(crime ~ log(population) + log(nonwhite) + log(density),
              data = city.df)
summary(model.3)
```

```

Call:
lm(formula = crime ~ log(population) + log(nonwhite) + log(density),
    data = city.df)

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
  (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
```

Part b

Here $\hat{\beta}_1$ corresponds to the coefficient on `log(population)`.

We can interpret $\hat{\beta}_1$, the coefficient for `log(population)`, as follows: For a unit increase in `log(population)`, given that everything else is held constant, we expect that, on average, the response variable, `crime`, to increase by 670.279 units.

The underlying hypothesis test is $H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$. We know that $\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t_{n-p-1}$. Under H_0, $\beta_1 = 0$, so $\frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{n-p-1}$. Then our p-value for this hypothesis test is $P(|t| > \frac{\hat{\beta}_1}{se(\hat{\beta}_1)})$. In this case, this value is 9.72×10^{-7}. In other words, assuming H_0 is true, the probability that $\hat{\beta}_1$ is this far or further from 0 (our H_0) is extremely small. So for most reasonably small values of α (e.g., 0.05 or 0.01), we can say that we reject H_0. So we can say that `log(population)` explains some of the variability in `crime`, even taking into account the other regressors.

Part c

The $\hat{\beta}_1$ values for the models in questions (2) and (3) are fairly similar. This occurs when the regressors are not too correlated. In the pairwise scatterplot of all of our variables (from question (2)), we can see that there is some correlation, especially between `log(population)` and `log(density)` (which makes sense logically), but for the most part it didn't affect `log(population)`'s contribution.

Even though the $\hat{\beta}_1$s are reasonably similar, the one in `model.3` is larger. This may be because `log(density)`, which correlates fairly well with `log(population)`, has a negative estimated coefficient. So given a value of `log(density)` (and `log(nonwhite)`), we get some negative contribution from `log(density)`, which has to be made up by `log(population)`.

Part d

$R^2_{model.2} = 0.2725$ while $R^2_{model.3} = 0.4123$. So `log(population)` explains only ~27% of the variation in `crime` while `log(population)`, `log(nonwhite)`, and `log(density)` together explain ~41% of the variability in `crime`. Comparing the two models, we can say `log(nonwhite)` and `log(density)` contributed a fair amount to the explanation of `crime`, on top of `log(population)`.

Question 4

Part a

We saw that at least one of the two additional predictors contributed to the explanation of `crime` and didn't correlate very highly with `log(population)`. So we would prefer a multiple linear regression model over a simple one (i.e., throw away `model.2`).

Based on the fact that `log(population)` and `log(density)` are fairly correlated (with a logical argument that this would be so, namely that $density = population/area$), we might want to construct another model excluding `log(density)`:

```
model.4 <- lm(crime ~ log(population) + log(nonwhite), data = city.df)
summary(model.4)
```

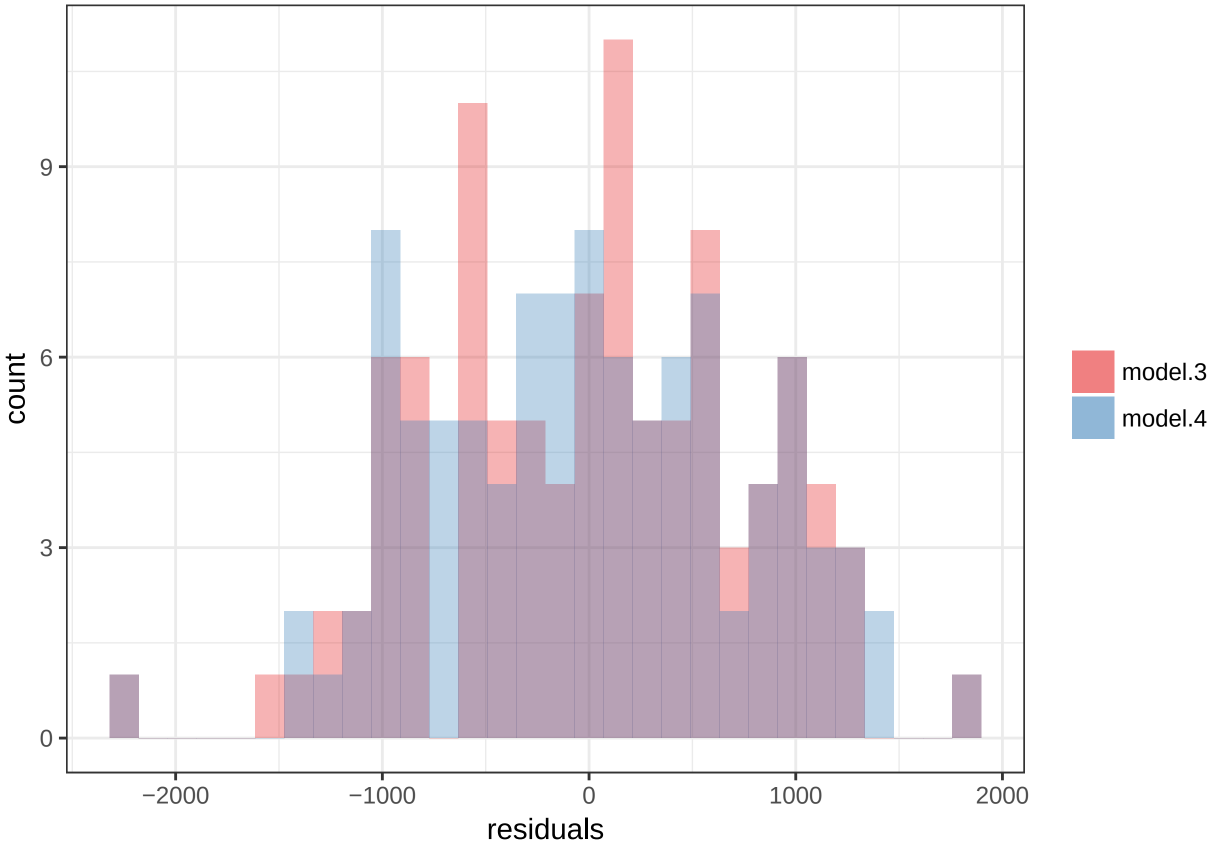
```

Call:
lm(formula = crime ~ log(population) + log(nonwhite), data = city.df)

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
  (10 observations deleted due to missingness)
Multiple R-squared:  0.3907,    Adjusted R-squared:  0.3781 
F-statistic:  31.1 on 2 and 97 DF,  p-value: 3.669e-11
```

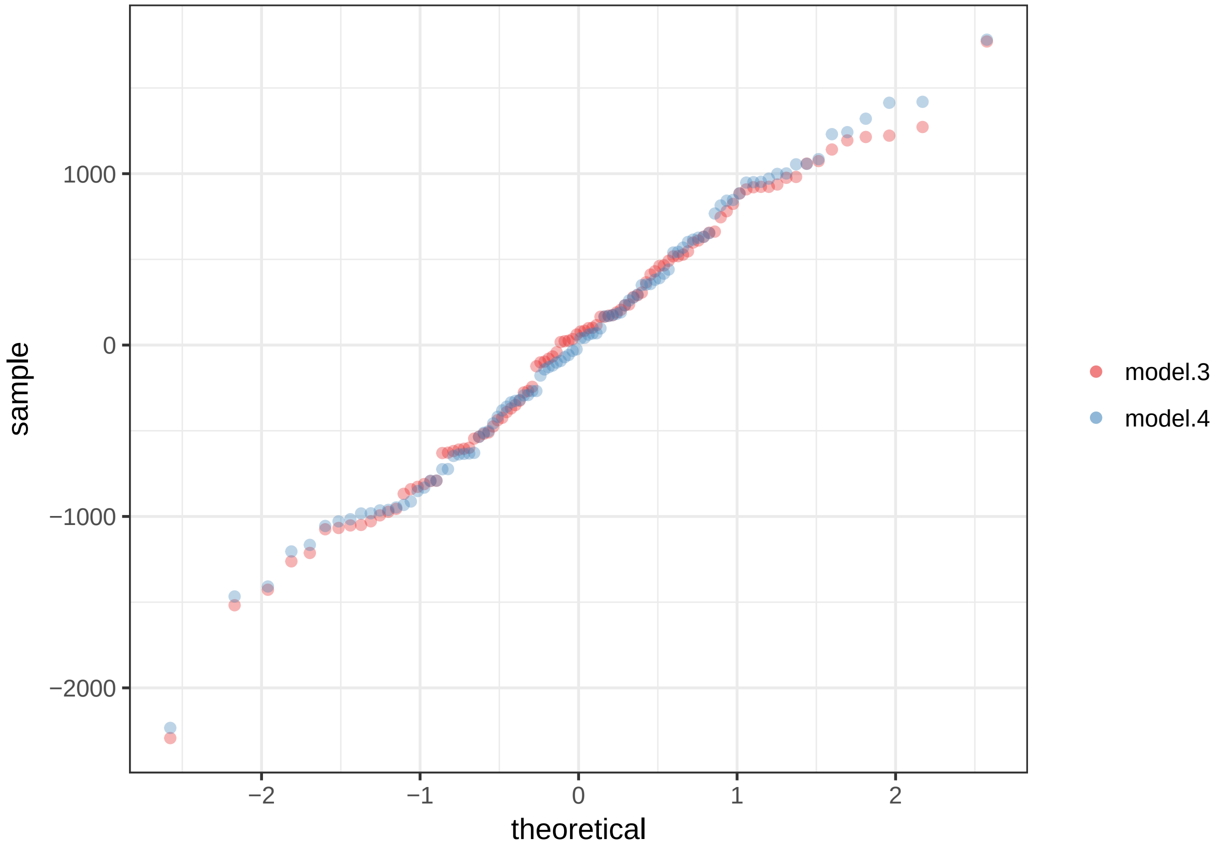
Here for a reasonable α such as 0.05 or 0.01, we would reject the null hypotheses for each of the βs. However, R^2 decreases somewhat, which makes sense—density contains information that population does not, namely, the area of the city (and in general, removing predictors always reduces R^2). We could also perhaps try omitting `log(population)` and keeping `log(density)`, but the scatterplots from before showed that `log(population)` correlates more highly with `crime` than `log(density)` does.

Another thing we could try is seeing how well the models fit our assumptions. We already know that there is some correlation among two of the regressors in `model.3`, but we can also check the residuals for both models:

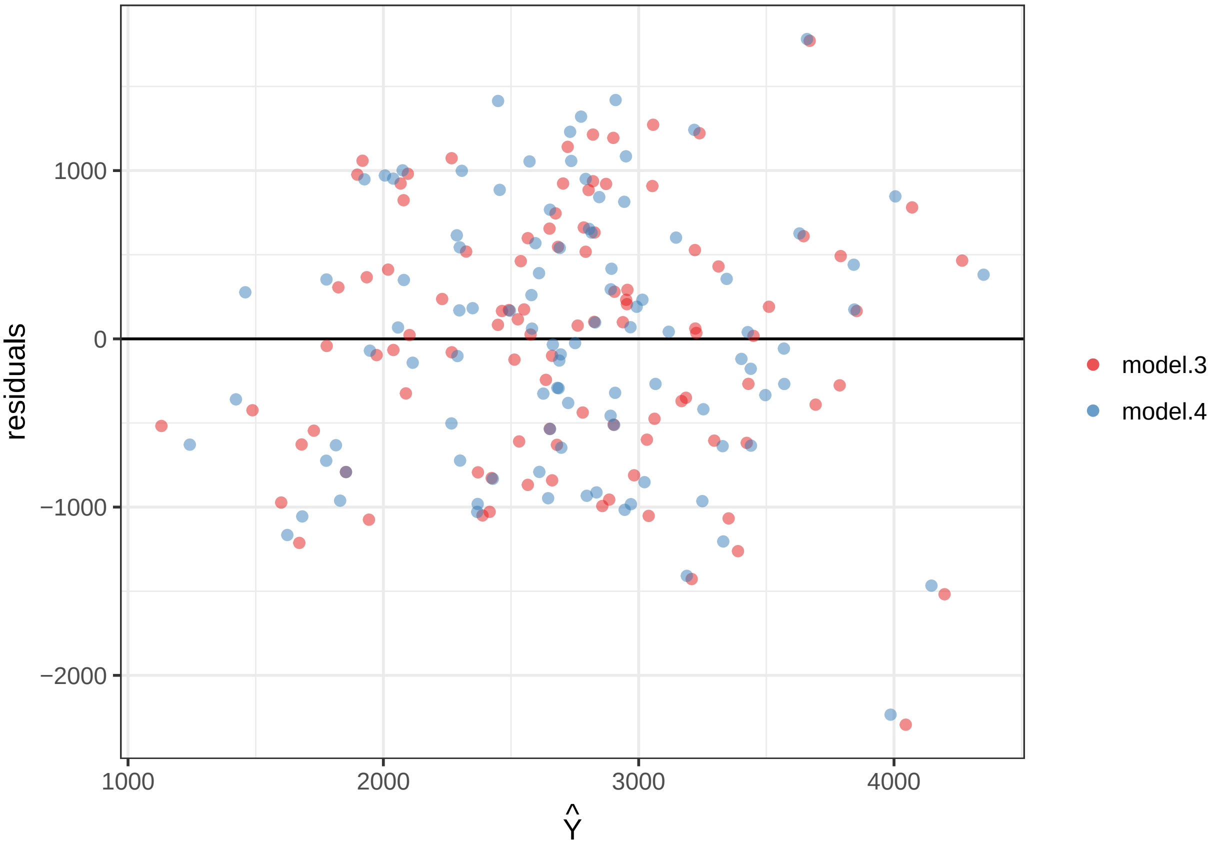
```r
city.df %<>%
  # add in the predictions from models 3 and 4
  dp$mutate(y3 = predict(model.3, newdata = .),
            y4 = predict(model.4, newdata = .)) %>%
  # ... and their residuals
  dp$mutate(r3 = crime - y3,
            r4 = crime - y4)

# plot a histogram of the residuals
ggplot(city.df) +
  geom_histogram(aes(x = r3, fill = 'model.3'),
                 alpha = 1 / 3) +
  geom_histogram(aes(x = r4, fill = 'model.4'),
                 alpha = 1 / 3) +
  labs(x = 'residuals', fill = NULL) +
  scale_fill_brewer(palette = 'Set1')
```



```r
# qq-plots of the residuals
ggplot(city.df) +
  stat_qq(aes(sample = r3, colour = 'model.3'),
          alpha = 1 / 3) +
  stat_qq(aes(sample = r4, colour = 'model.4'),
          alpha = 1 / 3) +
  labs(colour = NULL) +
  scale_colour_brewer(palette = 'Set1')
```



```
# residuals vs y-hat
ggplot(city.df) +
  geom_hline(yintercept = 0) +
  geom_point(aes(x = y3, y = r3, colour = 'model.3'), alpha = .5) +
  geom_point(aes(x = y4, y = r4, colour = 'model.4'), alpha = .5) +
  scale_colour_brewer(palette = 'Set1') +
  labs(x = expression(hat(Y)), y = 'residuals', colour = NULL)
```



From the three above visualizations, we have no reason to believe that either set of residuals are not normal or that there is some correlation between either pair of $\hat{Y}$ and residuals. The only remaining issue is the fact that there is a moderate correlation between two of the predictors, and whether our final model should include both as regressors.

Here we will use `model.3` since it has a higher R^2, we would fail to reject H_0 for β_3 for a not too unreasonable $\alpha = 0.1$, and since a city's population density adds information compared to just the city's total population.

Part b

```
confint(model.3, 'log(population)', level = .98)
```

```
                      1 %     99 %
log(population) 367.3513 973.206
```

We are 98% confident that the interval contains the true mean. In other words, assuming that `model.3` specifies the correct underlying structure of the population, 98% of our models from repeated samples of this population will result in confidence intervals that capture the true mean, β_1.

Part c

Here we will just use the sample means for the other regressors.

```
predict(model.3,
        newdata = data.frame(population = 1150,
                             nonwhite = mean(city.df$nonwhite, na.rm = TRUE),
                             density = mean(city.df$density, na.rm = TRUE)),
        interval = 'confidence',
        level = .99)
```

```
       fit      lwr      upr
1 3065.608 2822.315 3308.902
```

Our 99% confidence interval for an otherwise average city with a population of 1.15×10^6 is 3,065.6 to 3,308.9 per 100,000 people.