Chapter 12: Multicollinearity, Model Selection

Context

Intro to data

Our data set comes from Ericksen et al. (1989), who were expert witnesses in a US Supreme Court case about corrections to the U.S. Census undercount. (The descriptive text here quotes from that article.) This is important because Census numbers are used in many ways throughout the government, including setting the amount of funding that is passed from the federal government to those locations. Several cities and states sued the U.S. Census Bureau, to correct the census results. One of the arguments made was that census undercounts were primarily concentrated in regions with high prevalence of minorities.

We have measurements on 66 different regions in the U.S.: 16 large cities, the remaining parts of the states in which these cities are located, and the other U.S. states. For each of these regions, we have:

- `name`: The name of the region.
- `minority`: Percentage black or Hispanic.
- `crime`: Rate of serious crimes per 1000 population.
- `poverty`: Percentage poor.
- `language`: Percentage having difficulty speaking or writing English.
- `highschool`: Percentage age 25 or older who had not finished highschool.
- `housing`: Percentage of housing in small, multiunit buildings.
- `geographic_unit`: "city" or "state".
- `conventional`: Percentage of households counted by conventional personal enumeration (the method in primary use up until about 1950, and used less in 1980).
- `undercount`: Preliminary estimate of percentage undercount, based on a second survey.

```
head(census, 3)
```

```
##      name minority crime poverty language highschool housing geographic_unit conventional undercount
## 1 Alabama     26.1    49    18.9      0.2       43.5     7.6           state            0      -0.04
## 2  Alaska      5.7    62    10.7      1.7       17.5    23.6           state          100       3.35
## 3 Arizona     18.9    81    13.2      3.2       27.6     8.1           state           18       2.48
```

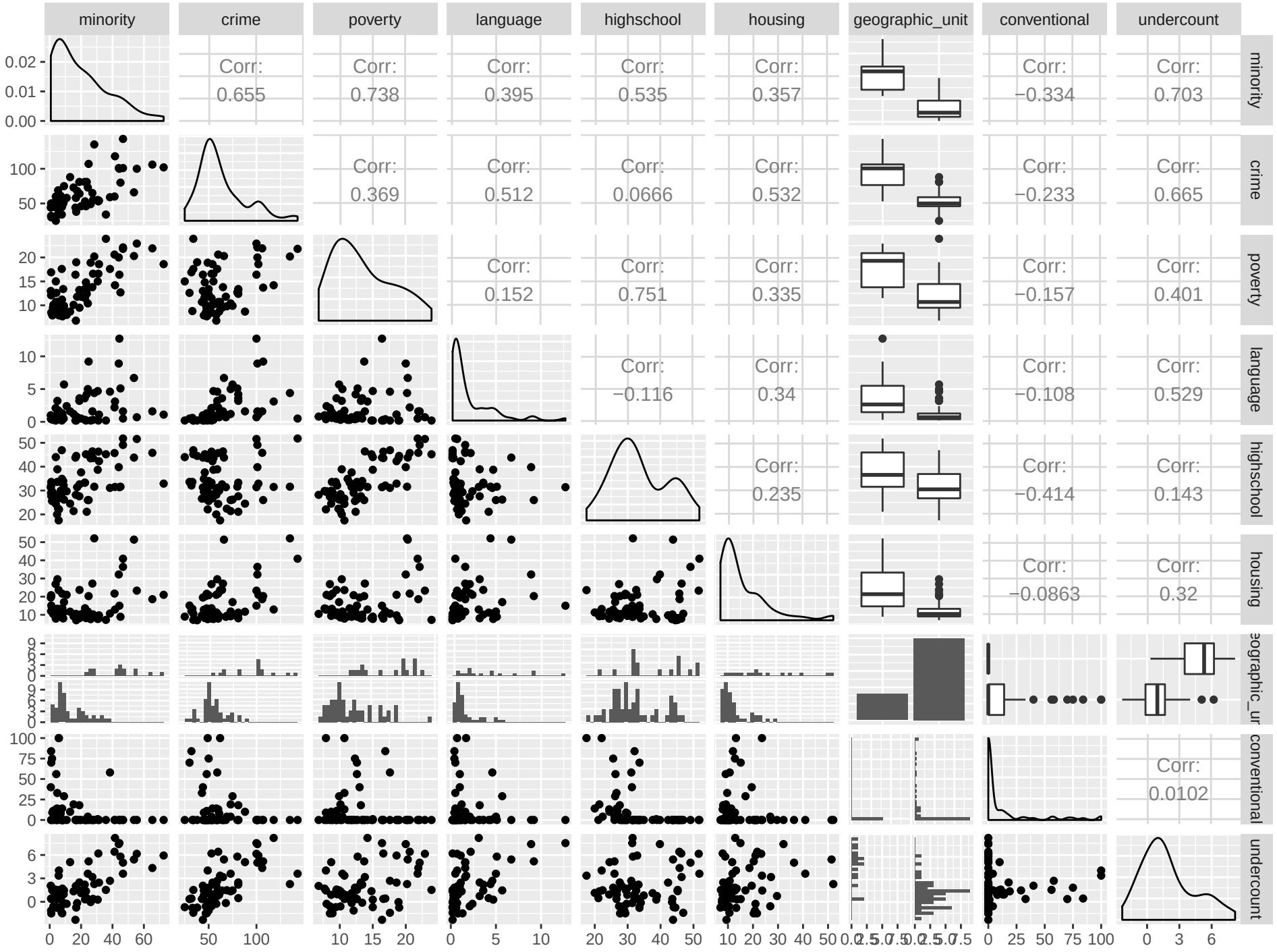
References:

- Fox, J. and Weisberg, S. (2011) An R Companion to Applied Regression, Second Edition, Sage.
- Ericksen, E. P., Kadane, J. B. and Tukey, J. W. (1989) Adjusting the 1980 Census of Population and Housing. *Journal of the American Statistical Association* 84, 927–944.

Let's consider a model for `undercount` (response) based on all of the other explanatory variables in the data set (other than the name of the location).

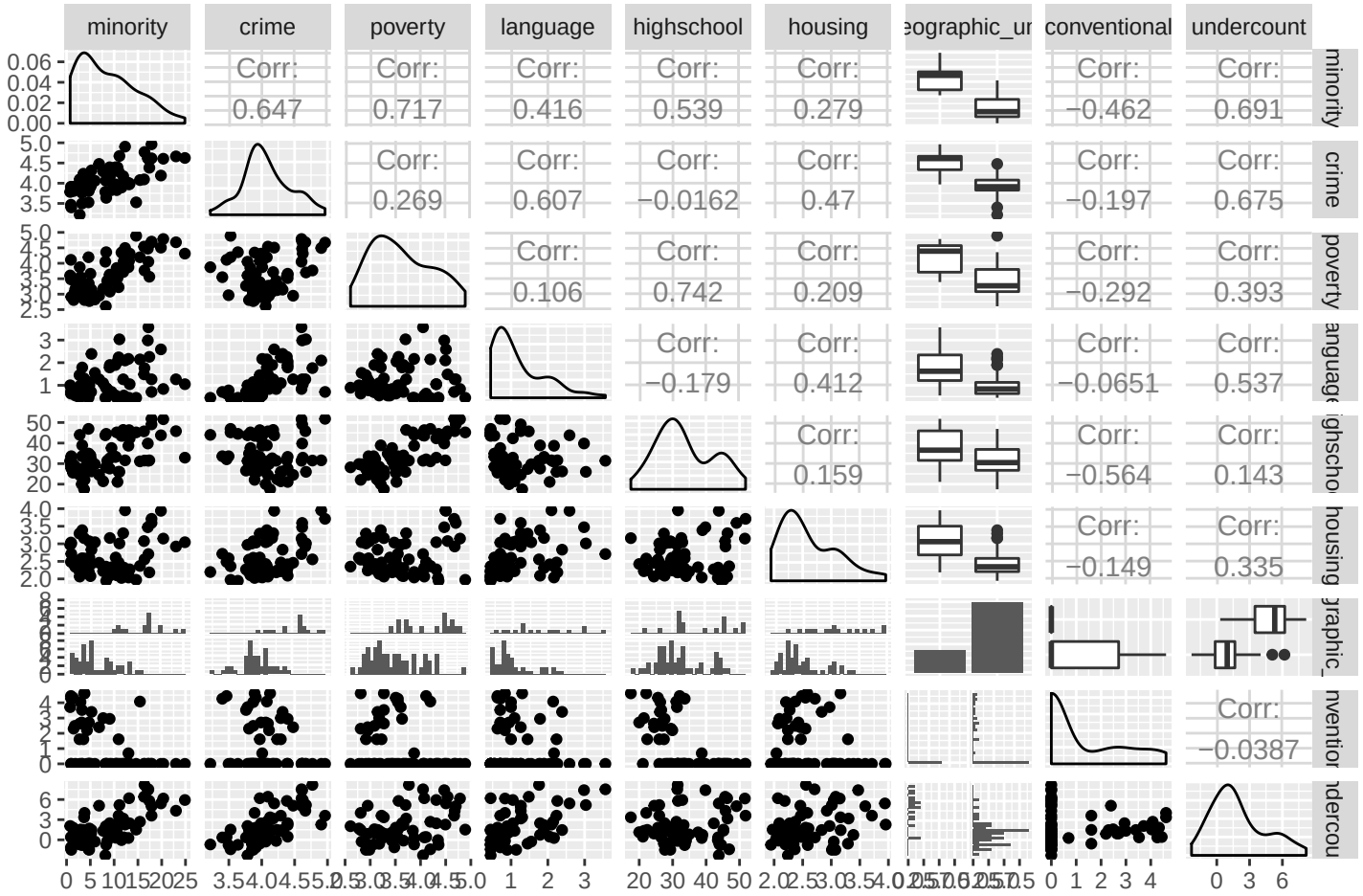
Pairs Plots

```
library(GGally)
ggpairs(census %>% select(-name))
```


minority
crime
poverty
language
highschool
housing
geographic_unit
conventional
undercount
Corr: 0.655
Corr: 0.738
Corr: 0.395
Corr: 0.535
Corr: 0.357
Corr: −0.334
Corr: 0.703
Corr: 0.369
Corr: 0.512
Corr: 0.0666
Corr: 0.532
Corr: −0.233
Corr: 0.665
Corr: 0.152
Corr: 0.751
Corr: 0.335
Corr: −0.157
Corr: 0.401
Corr: −0.116
Corr: 0.34
Corr: −0.108
Corr: 0.529
Corr: 0.235
Corr: −0.414
Corr: 0.143
Corr: −0.0863
Corr: 0.32
Corr: 0.0102

```
census_transformed <- census %>%
  transmute(
    minority = minority^0.75,
    crime = log(crime),
    poverty = sqrt(poverty),
    language = sqrt(language),
    highschool = highschool,
    housing = log(housing),
    geographic_unit = geographic_unit,
    conventional = log(conventional + 1),
    undercount = undercount
  )
```

```
ggpairs(census_transformed)
```

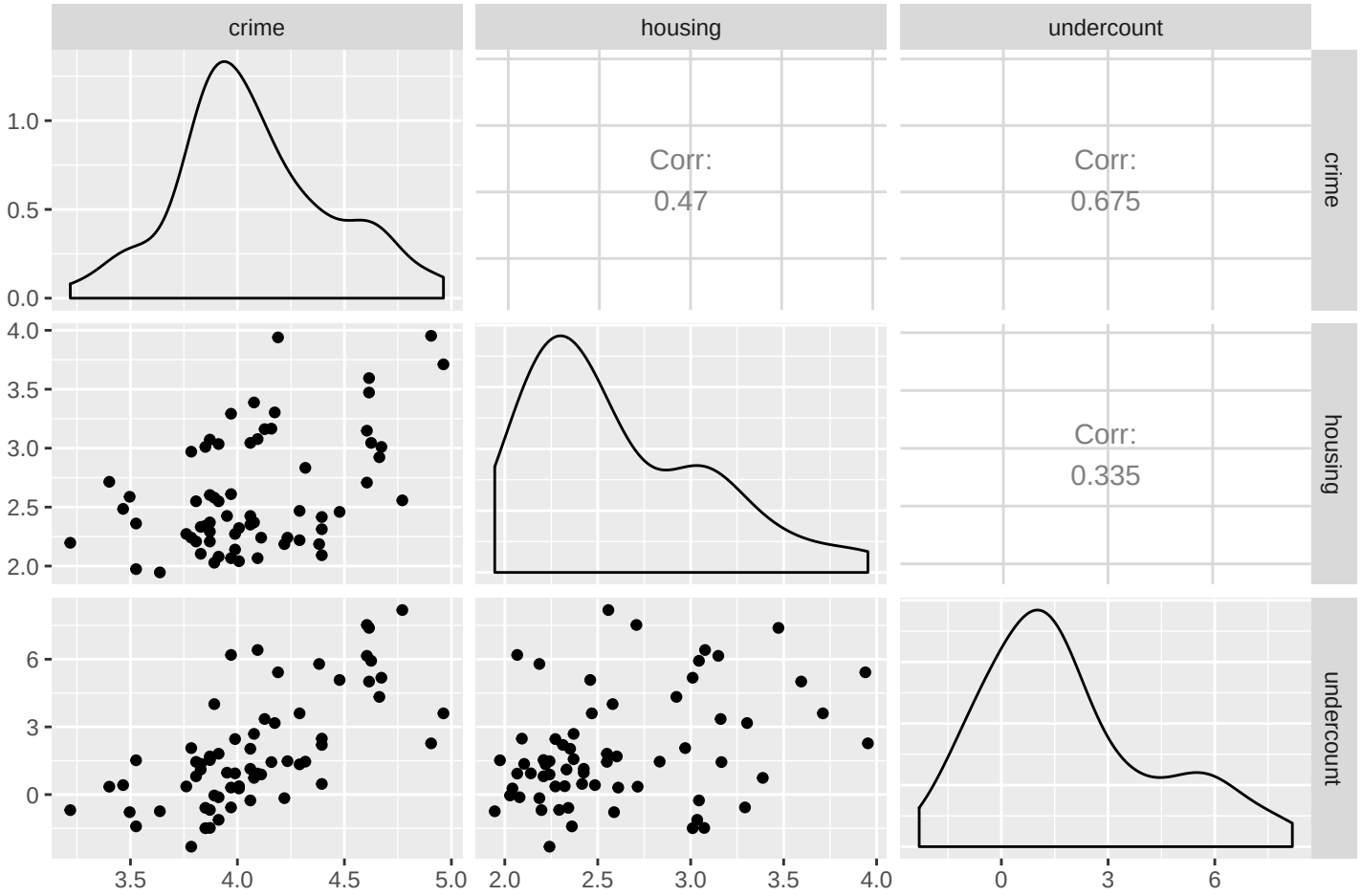


Challenge: Multicollinearity

- **Multicollinearity:** Some of the explanatory variables are linearly associated with each other. Effects:
  - Apparent association between a given explanatory variable and the response can change if we add or remove other explanatory variables from the model.
  - Standard errors of coefficient estimates are large.
    - We are not certain of the coefficient value: which of the correlated explanatory variables is actually responsible for changes in the response?
    - This can lead to small t statistics/large p-values even if a variable is associated with the response.

Illustration, focusing on just the crime and housing explanatory variables

```
ggpairs(census_transformed %>% select(crime, housing, undercount))
```



```r
lm_highschool <- lm(undercount ~ crime, data = census_transformed)
summary(lm_highschool)
```

```
##
## Call:
## lm(formula = undercount ~ crime, data = census_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4388 -1.2250 -0.0878  1.1312  4.7525
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16.6997     2.5562  -6.533 1.22e-08 ***
## crime         4.5682     0.6246   7.313 5.23e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.838 on 64 degrees of freedom
## Multiple R-squared:  0.4553, Adjusted R-squared:  0.4467
## F-statistic: 53.49 on 1 and 64 DF,  p-value: 5.233e-10
```

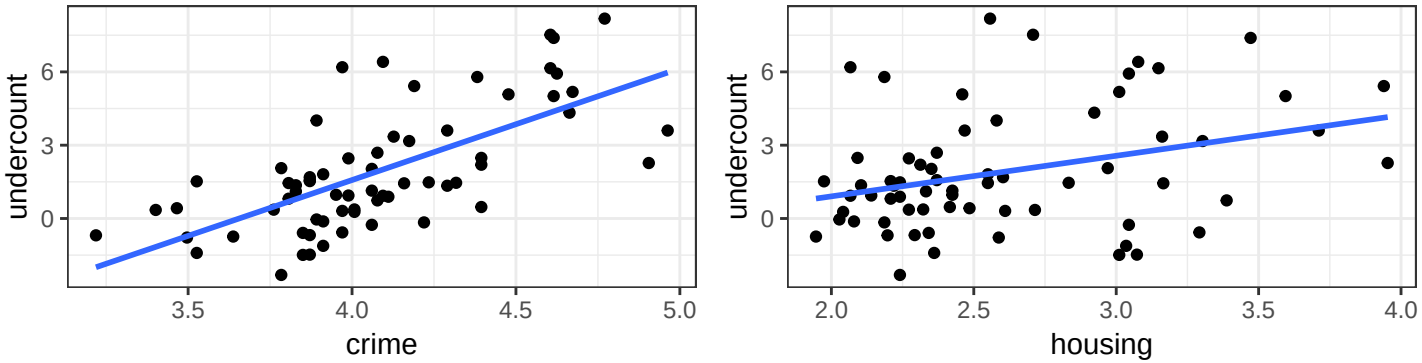
```r
lm_poverty <- lm(undercount ~ housing, data = census_transformed)
summary(lm_poverty)
```

```
##
## Call:
## lm(formula = undercount ~ housing, data = census_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1672 -1.3946 -0.2819  1.1527  6.3489
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.4163     1.5506  -1.558  0.12409
## housing       1.6609     0.5834   2.847  0.00592 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.346 on 64 degrees of freedom
## Multiple R-squared:  0.1124, Adjusted R-squared:  0.09854
## F-statistic: 8.106 on 1 and 64 DF,  p-value: 0.005925
```

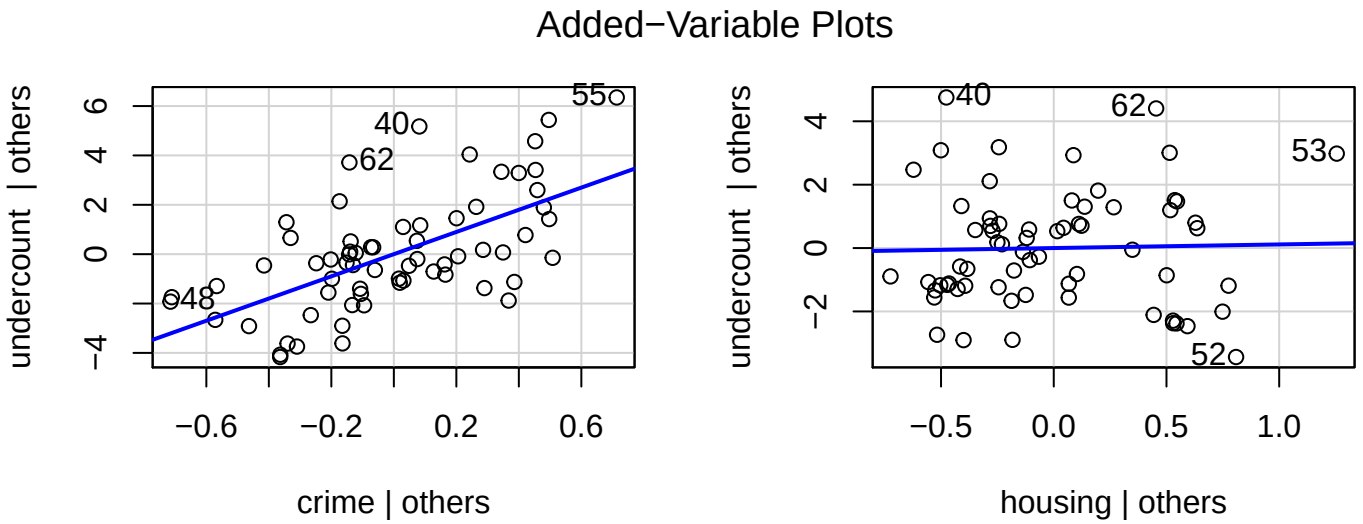
```r
lm_highschool_poverty <- lm(undercount ~ crime + housing, data = census_transformed)
summary(lm_highschool_poverty)
```

```
##
## Call:
## lm(formula = undercount ~ crime + housing, data = census_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5308 -1.2343 -0.0999  1.0951  4.8067
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16.6986     2.5754  -6.484 1.57e-08 ***
## crime         4.4950     0.7132   6.303 3.22e-08 ***
## housing       0.1139     0.5218   0.218    0.828
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.852 on 63 degrees of freedom
## Multiple R-squared:  0.4557, Adjusted R-squared:  0.4384
## F-statistic: 26.37 on 2 and 63 DF,  p-value: 4.785e-09
```

```r
p1 <- ggplot(data = census_transformed, mapping = aes(x = crime, y = undercount)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()

p2 <- ggplot(data = census_transformed, mapping = aes(x = housing, y = undercount)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()

grid.arrange(p1, p2, ncol = 2)
```




```r
library(car)
avPlots(lm_highschool_poverty)
```



A model with all the explanatory variables

```
bigfit <- lm(undercount ~ minority + crime + poverty + language + highschool + housing + geographic_unit + conventional,
    data = census_transformed)
summary(bigfit)
```

```
##
## Call:
## lm(formula = undercount ~ minority + crime + poverty + language +
##     highschool + housing + geographic_unit + conventional, data = census_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8986 -0.9230  0.1316  0.7453  4.4242
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2.64397    4.40182  -0.601 0.550451
## minority              0.21050    0.07229   2.912 0.005118 **
## crime                 1.38213    0.92024   1.502 0.138636
## poverty              -1.13818    0.62294  -1.827 0.072922 .
## language              0.58031    0.35387   1.640 0.106538
## highschool            0.07221    0.04925   1.466 0.148128
## housing              -0.42226    0.51322  -0.823 0.414070
## geographic_unitstate -1.93606    0.79686  -2.430 0.018289 *
## conventional          0.66436    0.16203   4.100 0.000132 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.449 on 57 degrees of freedom
## Multiple R-squared:  0.6982, Adjusted R-squared:  0.6559
## F-statistic: 16.48 on 8 and 57 DF,  p-value: 2.495e-12
```

It seems like we probably don't need all of those variables in the model.

Variance Inflation Factors

- "How much larger is the variance of $\hat{\beta}_j$ than it would be if all explanatory variables were uncorrelated?"

```
vif(bigfit)
```

```
##        minority           crime         poverty        language      highschool         housing geographic_unit     conventional
##        6.067582        3.489176        4.273742        2.037437        5.414260        2.027117        3.663772        2.068474
```

- A VIF of 4 means our confidence interval for that coefficient is twice as wide as it would be if all explanatory variables were uncorrelated.

- As a rough guide, a VIF > 5 indicates we may want to do something to address multicollinearity.
  - In this class, that means drop explanatory variables that don't help the model but do inflate variances
  - Other strategies include principle components analysis and penalized regression; see stat 340/other classes!
- $VIF = \frac{1}{1-R_X^2}$, where R_X^2 is the R^2 of the regression of the given explanatory variable on all other explanatory variables.

Which explanatory variables should we include in our model?

Note: this topic will be explored in much more depth in Statistics 340. We introduce some first ideas here.

Basic Motivations

- Need **enough**:
  - Include all the important variables we are interested in/necessary to answer our scientific questions
  - Include any important potential confounding variables
- But **not too many**:
  - Try to avoid including highly correlated explanatory variables, unless we really need them.
- Acknowledge **there is no perfect model**:
  - We will probably identify multiple sets of explanatory variables that are about as good as each other.
  - We should present results from all of these models (or a representative group of them).
- Our goal is **not** to find models that support our favorite theories/get us statistically significant results/etc.!
- **Our goal is to see what the data can and can not tell us about what we're studying**

All Subsets Regression

- Fit every possible model based on all the different subsets of explanatory variables
  - Model with only minority
  - . . . all other models with one explanatory variable. . .
  - Model with minority and crime
  - . . . all other models with 2, 3, 4, 5, 6, or 7 explanatory variables. . .
  - Model with all 8 explanatory variables
- Pick "best" models to discuss

```
library(leaps)
candidate_models <- regsubsets(
  undercount ~ minority + crime + poverty + language + highschool + housing + geographic_unit + conventional,
  data = census_transformed,
  nbest = 1)
summary(candidate_models)
```

```
## Subset selection object
## Call: regsubsets.formula(undercount ~ minority + crime + poverty +
##     language + highschool + housing + geographic_unit + conventional,
##     data = census_transformed, nbest = 1)
## 8 Variables  (and intercept)
##                       Forced in Forced out
## minority                  FALSE      FALSE
## crime                     FALSE      FALSE
## poverty                   FALSE      FALSE
## language                  FALSE      FALSE
## highschool                FALSE      FALSE
## housing                   FALSE      FALSE
## geographic_unitstate      FALSE      FALSE
## conventional              FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          minority crime poverty language highschool housing geographic_unitstate conventional
## 1  ( 1 ) "*"      " "   " "     " "      " "        " "     " "                   " "
## 2  ( 1 ) "*"      " "   " "     " "      " "        " "     " "                   "*"
## 3  ( 1 ) "*"      "*"   " "     " "      " "        " "     " "                   "*"
## 4  ( 1 ) "*"      " "   "*"     " "      " "        " "     "*"                   "*"
## 5  ( 1 ) "*"      " "   "*"     "*"      " "        " "     "*"                   "*"
## 6  ( 1 ) "*"      "*"   "*"     "*"      " "        " "     "*"                   "*"
## 7  ( 1 ) "*"      "*"   "*"     "*"      "*"        " "     "*"                   "*"
## 8  ( 1 ) "*"      "*"   "*"     "*"      "*"        "*"     "*"                   "*"
```

- For a particular number of explanatory variables, "best" models have smallest sum of squared residuals
- How should we select the number of explanatory variables to use?

Not by looking for smallest sum of squared residuals

- Adding more explanatory variables will *always* reduce sum of squared residuals, but may not give a better model

```
summary(candidate_models)$rss
```

```
## [1] 207.5440 167.7937 142.1790 131.3334 126.4677 124.4305 121.1672 119.7451
```

Not by looking for largest R^2

- Adding more explanatory variables will *always* increase R^2, but may not give a better model

```
summary(candidate_models)$rsq
```

```
## [1] 0.4769295 0.5771118 0.6416680 0.6690021 0.6812651 0.6863994 0.6946239 0.6982080
```

One Option: BIC

- We want a criterion that says "Make the Residual Sum of Squares small, but don't include too many explanatory variables"
- Bayesian Information Criterion: Minimize

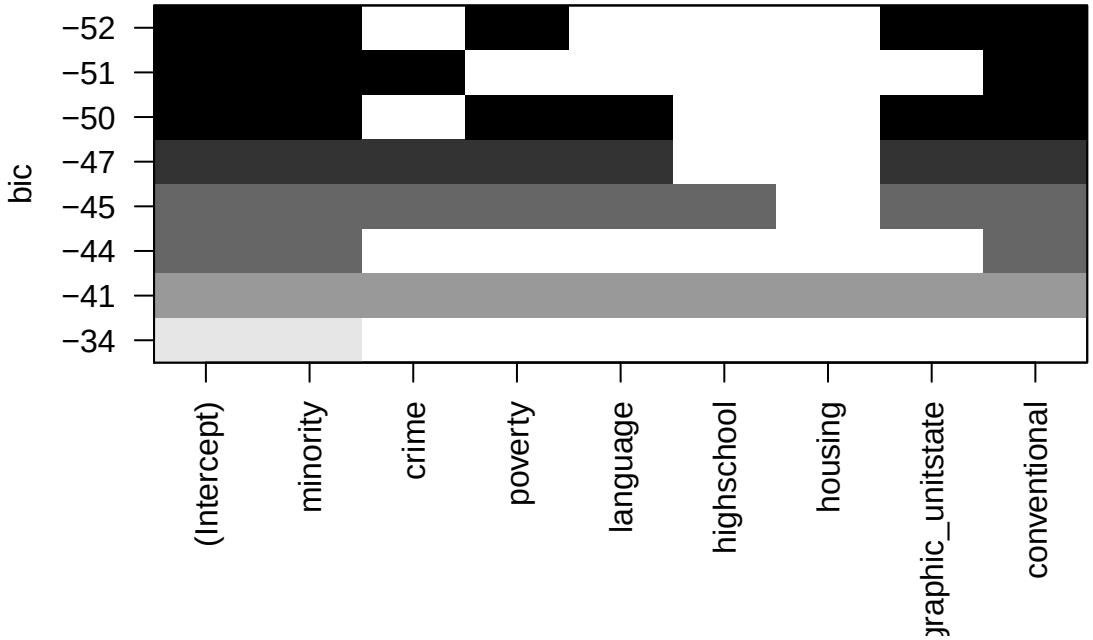
$$BIC = n \times \log\left(\frac{\text{Sum of Squared Residuals}}{n}\right) + \log(n) \times (p+1)$$

- The first term is small if the Sum of Squared Residuals is small
- The second term is small if the number of explanatory variables, p, is small

```
summary(candidate_models)$bic
```

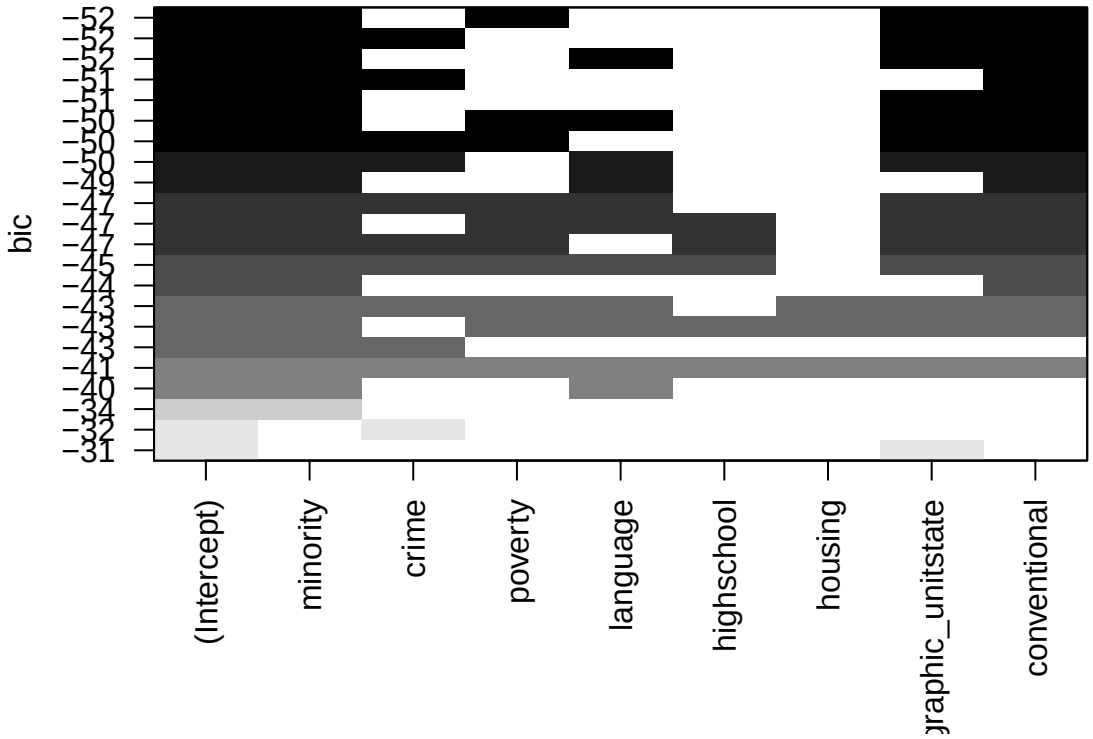
```
## [1] -34.39127 -44.23377 -50.97688 -52.02418 -50.32618 -47.20833 -44.77270 -41.36226
```

```
plot(candidate_models)
```



- Generally, models with BIC within about 2 of the smallest BIC are equivalent in terms of utility.
  - We should report on **all 3** of the models with lowest BIC here.
- We could also look at more than 1 "best model" of each size

```
candidate_models3 <- regsubsets(
  undercount ~ minority + crime + poverty + language + highschool + housing + geographic_unit + conventional,
  data = census_transformed,
  nbest = 3)
plot(candidate_models3)
```




```
summary(candidate_models3)$bic %>%
  sort()
```

```
##  [1] -52.02418 -51.76637 -51.55289 -50.97688 -50.84480 -50.32618 -49.96010 -49.70790 -48.66521 -47.20833 -47.02759 -46.56884 -44.77270 -44.2
## [16] -42.99031 -42.55693 -41.36226 -40.48006 -34.39127 -31.71105 -30.78063
```

- In an actual analysis, I would examine and report on **all 6** of the models with lowest BIC. (Not doing that here for the sake of time/space/our attention spans.)

Finishing off the analysis

We have 3 candidate models - which findings, if any, are robust to the covariates included?

```
fit1 <- lm(
  undercount ~ minority + poverty + geographic_unit + conventional,
  data = census_transformed)
summary(fit1)
```

```
##
## Call:
## lm(formula = undercount ~ minority + poverty + geographic_unit +
##     conventional, data = census_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3948 -0.9026 -0.0018  0.7943  4.4731
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.91732    1.56374   2.505  0.01493 *
## minority               0.30672    0.05601   5.476 8.70e-07 ***
## poverty               -0.99704    0.43851  -2.274  0.02652 *
## geographic_unitstate -2.16770    0.62994  -3.441  0.00105 **
## conventional           0.59760    0.12911   4.628 1.97e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.467 on 61 degrees of freedom
## Multiple R-squared:  0.669,  Adjusted R-squared:  0.6473
## F-statistic: 30.82 on 4 and 61 DF,  p-value: 4.844e-14
```

```
confint(fit1)
```

```
##                           2.5 %     97.5 %
## (Intercept)            0.7904220  7.0442092
## minority               0.1947092  0.4187259
## poverty               -1.8739021 -0.1201855
## geographic_unitstate -3.4273525 -0.9080521
## conventional           0.3394186  0.8557725
```

```r
fit2 <- lm(
  undercount ~ minority + crime + conventional,
  data = census_transformed)
summary(fit2)
```

```
##
## Call:
## lm(formula = undercount ~ minority + crime + conventional, data = census_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9498 -1.0441  0.0266  0.7823  3.8835
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -10.02613    2.55182  -3.929 0.000217 ***
## minority       0.24897    0.04498   5.536 6.67e-07 ***
## crime          2.28238    0.68291   3.342 0.001413 **
## conventional   0.48512    0.13428   3.613 0.000608 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.514 on 62 degrees of freedom
## Multiple R-squared:  0.6417, Adjusted R-squared:  0.6243
## F-statistic: 37.01 on 3 and 62 DF,  p-value: 7.827e-14
```

```r
confint(fit2)
```

```
##                    2.5 %     97.5 %
## (Intercept)  -15.1271453 -4.9251165
## minority       0.1590693  0.3388784
## crime          0.9172543  3.6475025
## conventional   0.2167037  0.7535325
```

```r
fit3 <- lm(
  undercount ~ minority + poverty + language + geographic_unit + conventional,
  data = census_transformed)
summary(fit3)
```

```
##
## Call:
## lm(formula = undercount ~ minority + poverty + language + geographic_unit +
##     conventional, data = census_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3482 -0.7504  0.0424  0.8456  4.7206
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.56877    1.78374   1.440  0.15504
## minority              0.27865    0.05842   4.770 1.22e-05 ***
## poverty              -0.76388    0.46022  -1.660  0.10217
## language              0.46634    0.30693   1.519  0.13393
## geographic_unitstate -1.85647    0.65609  -2.830  0.00633 **
## conventional          0.55517    0.13077   4.246 7.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.452 on 60 degrees of freedom
## Multiple R-squared:  0.6813, Adjusted R-squared:  0.6547
## F-statistic: 25.65 on 5 and 60 DF,  p-value: 9.578e-14
```

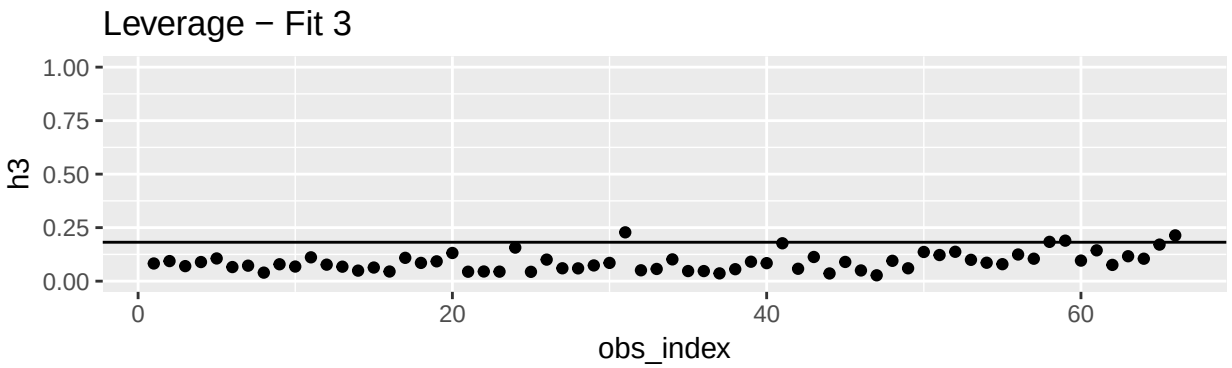
```r
confint(fit3)
```

```
##                           2.5 %     97.5 %
## (Intercept)          -0.9992425  6.1367813
## minority              0.1617919  0.3955101
## poverty              -1.6844633  0.1567016
## language             -0.1476174  1.0803030
## geographic_unitstate -3.1688493 -0.5440877
## conventional          0.2936051  0.8167445
```

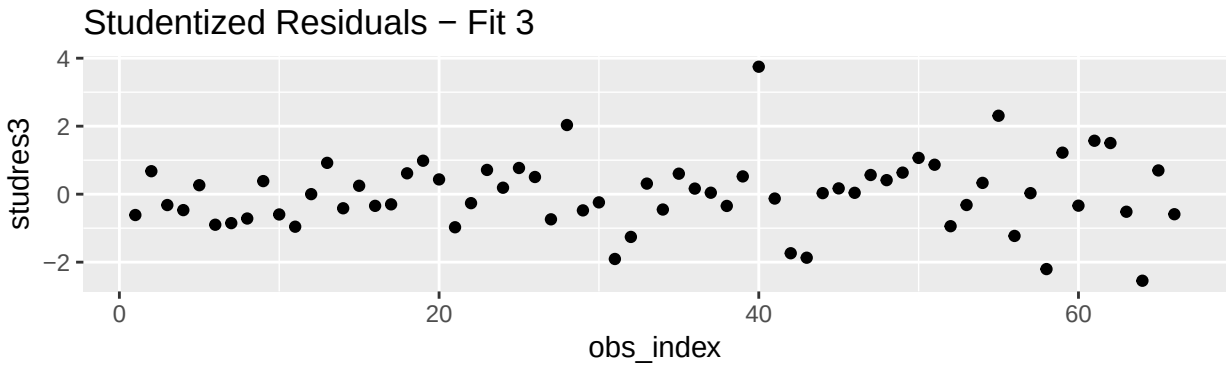
Do we have any outliers or high leverage observations?

- I will look at just the most complex model here. The story would probably be the same for the other models, but in an analysis in preparation for publication I would examine these results for all models.

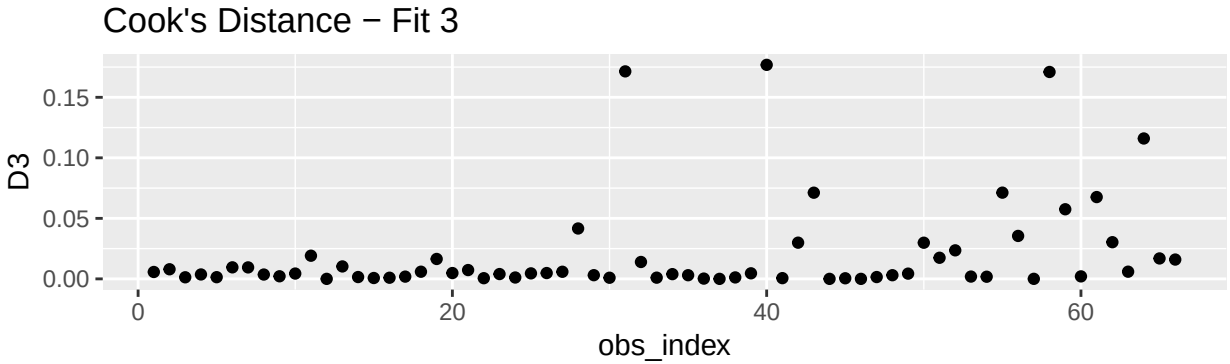
```
census_transformed <- census_transformed %>%
  mutate(
    obs_index = row_number(),
    h3 = hatvalues(fit3),
    studres3 = rstudent(fit3),
    D3 = cooks.distance(fit3)
  )

ggplot(data = census_transformed, mapping = aes(x = obs_index, y = h3)) +
  geom_point() +
  geom_hline(yintercept = 2 * 6 / nrow(census_transformed)) +
  ylim(0, 1) +
  ggtitle("Leverage - Fit 3")
```




```
ggplot(data = census_transformed, mapping = aes(x = obs_index, y = studres3)) +
  geom_point() +
  ggtitle("Studentized Residuals - Fit 3")
```




```
ggplot(data = census_transformed, mapping = aes(x = obs_index, y = D3)) +
  geom_point() +
  ggtitle("Cook's Distance - Fit 3")
```



Yes. Are our findings robust to whether or not these observations are included?

```
suspicious_obs <- c(31, 40, 58, 66)
census_transformed_nosuspicious <- census_transformed[-suspicious_obs, ]

fit3_nosuspicious <- lm(
  undercount ~ minority + poverty + language + geographic_unit + conventional,
  data = census_transformed_nosuspicious)

summary(fit3_nosuspicious)
```

```
##
## Call:
## lm(formula = undercount ~ minority + poverty + language + geographic_unit +
##     conventional, data = census_transformed_nosuspicious)
##
## Residuals:
```

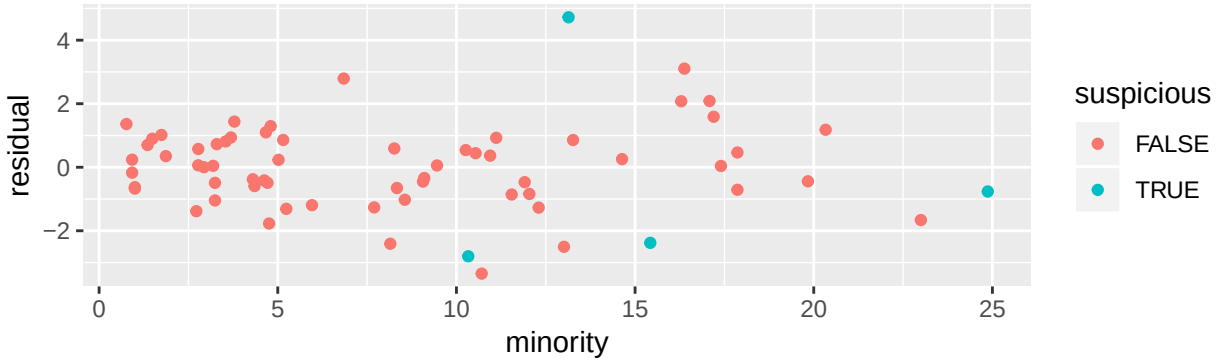
```
##     Min      1Q  Median      3Q     Max
## -3.7371 -0.6246 -0.0500  0.7937  2.7357
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.82448    1.74793   2.188 0.032858 *
## minority               0.28256    0.05733   4.929 7.73e-06 ***
## poverty               -0.99338    0.41873  -2.372 0.021135 *
## language               0.41730    0.29447   1.417 0.161987
## geographic_unitstate -2.47668    0.65065  -3.806 0.000352 ***
## conventional           0.64648    0.11973   5.400 1.41e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.248 on 56 degrees of freedom
## Multiple R-squared:  0.7567, Adjusted R-squared:  0.735
## F-statistic: 34.84 on 5 and 56 DF,  p-value: 5.153e-16
```

```r
confint(fit3_nosuspicious)
```

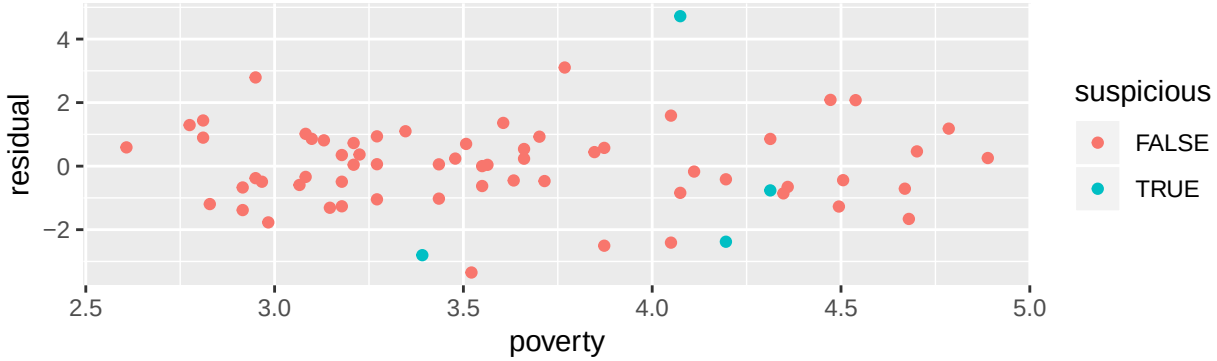
```
##                           2.5 %     97.5 %
## (Intercept)            0.3229632  7.3259979
## minority               0.1677125  0.3974114
## poverty               -1.8322023 -0.1545590
## language              -0.1725940  1.0071887
## geographic_unitstate -3.7800868 -1.1732740
## conventional           0.4066323  0.8863249
```

How about residuals diagnostics for linearity/normality/etc?

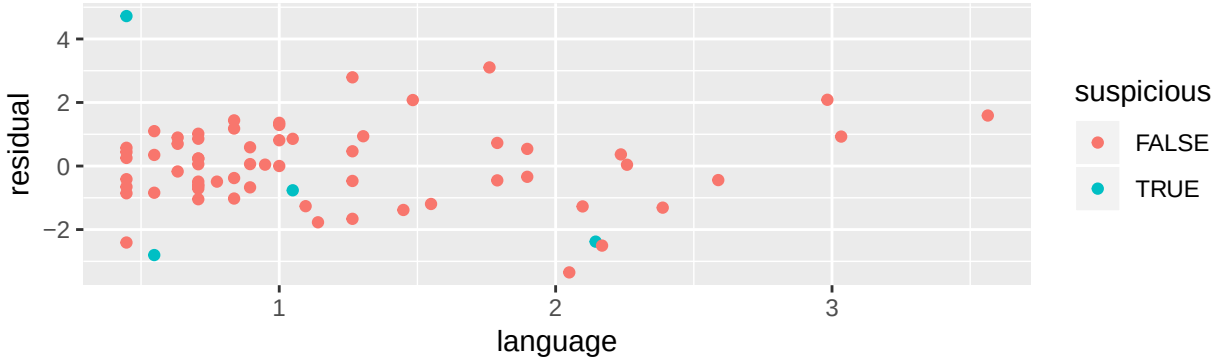
```r
census_transformed <- census_transformed %>%
  mutate(
    residual = residuals(fit3),
    suspicious = row_number() %in% suspicious_obs
  )

ggplot(data = census_transformed, mapping = aes(x = minority, y = residual, color = suspicious)) +
  geom_point()
```




```
ggplot(data = census_transformed, mapping = aes(x = poverty, y = residual, color = suspicious)) +
  geom_point()
```

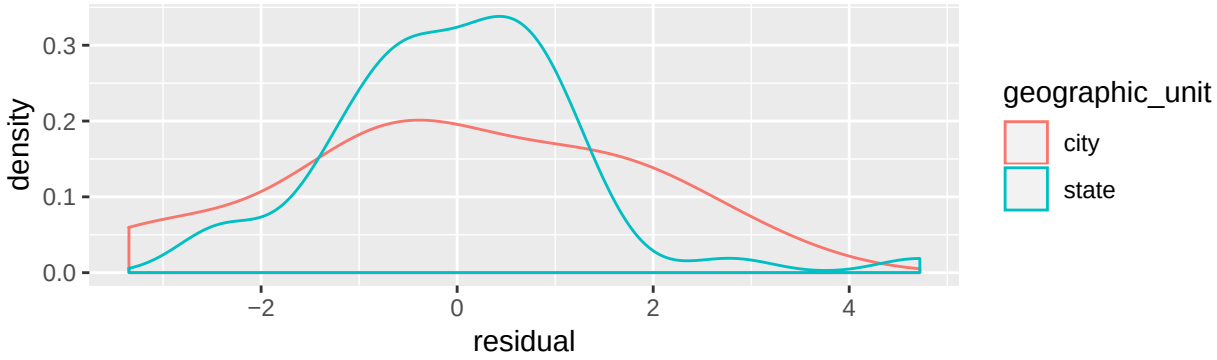



```
ggplot(data = census_transformed, mapping = aes(x = language, y = residual, color = suspicious)) +
  geom_point()
```




```
ggplot(data = census_transformed, mapping = aes(x = residual, color = geographic_unit, color = suspicious)) +
  geom_density()
```

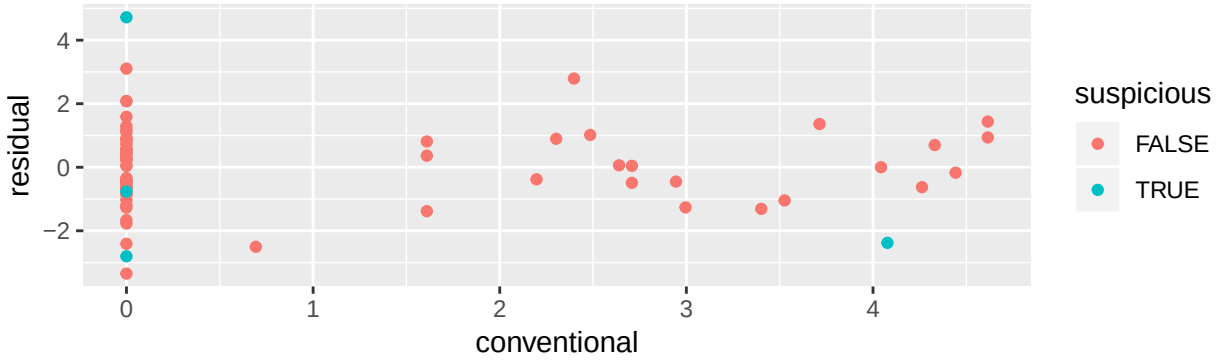
```
## Warning: Duplicated aesthetics after name standardisation: colour
```




```
census_transformed %>%
  group_by(geographic_unit) %>%
  summarize(sd(residual))
```

```
## # A tibble: 2 x 2
##   geographic_unit `sd(residual)`
##   <fct>                   <dbl>
## 1 city                     1.79
## 2 state                    1.26
```

```
ggplot(data = census_transformed, mapping = aes(x = conventional, y = residual, color = suspicious)) +
  geom_point()
```



What are our conclusions?

- We found consistent and strong evidence that higher concentrations of minorities and higher rates of use of conventional census methods were both associated with larger census undercounts, after controlling for the other covariates selected for inclusion in the model.
- We also found consistent and strong evidence that census undercounts were larger in large cities than in the other parts of those states or other states taken as a whole, after accounting for the associations of other covariates with census undercounts.
- These findings were robust to the covariates included in our model, and held whether or not four outlying high leverage observations were included.
- We saw less consistent associations between census undercounts and the poverty rate, crime rate, and rate at which residents had difficulty speaking or writing in English, after controlling for the associations found above. These variables were not always included in models selected by BIC, and the strength of evidence of an association with the census undercount depended on which other covariates were included and whether or not high leverage observations were included.
- This is an observational study and our findings are limited in scope to the particular regions included in the analysis, and to the procedures used 1980 Census.

In the authors' words...

- Here's what they said about their variable selection procedure

mainders of SMSA it was 1.1%, and for the remainder of the country it was .6%. We concluded from these results that central cities should be separated from suburbs, and we opted for a dummy variable indicating central-city location, rather than the census definition of urban, as one of our eight predictor variables.

5.2 Choosing a Best Set of Predictors

To select a best set of predictor variables, we regressed Series 2-9 on all subsets of two, three, and four variables. We considered all equations in which *each* regression coefficient was at least twice its standard error, and selected the one that minimized σ^2, the unexplained variance. Best results were obtained with the three-variable subset consisting of the reported crime rate and the minority and conventional percentages. The margin of choice was small, and regressions substituting the language difficulty or central-city variables for the reported crime rate also fit the data well. Omitting any of the three selected variables to produce a two-variable regression led to substantially increased values of σ^2. Finally, when these calculations were made for dependent variables based on the other 11 PEP series, the optimal choice of predictor variables was the same or similar. It was the same for Series 2-8, 2-20, 3-8, 3-9, 3-20, 5-8, and 5-9. For Series 10-8, 14-8, 14-9, and 14-20, the optimal set included only two predictors, the reported crime rate and the percent conventional.

5.3 The Results From the Composite Method

The composite method is based on a hierarchical model we explained in Ericksen and Kadane (1985, 1987). Cressie (1988a,b) explored an interesting variant of our model with a different assumption about variances in the second stage.

Our estimate of the undercount rate is a certain matrix-weighted average of a regression estimate and the initial sample estimates. (See Table 8.) The regression estimate is

$$\hat{Y} = -3.0 + .059 \text{ min} + .026 \text{ conv} + .055 \text{ crime}, \quad (2)$$

where Y is the Series 2-9 undercount rate, *min* is the proportion Black or Hispanic, *conv* is the proportion counted

- Here is the first paragraph of the conclusions:

Our major substantive finding is that the largest undercounts of the 1980 census occurred in central cities with large minority populations. Above-average undercounts occurred in many western states, especially where the Census Bureau relied on the conventional method. Undercounts were very low in the northeast and north-central regions outside of large cities as well as in states of the "upper south," such as West Virginia and Kentucky. This conclusion does not depend on selecting a particular PEP series or a set of independent variables. Indeed, even if we ignore the PEP data and rely on synthetic estimation, adjustment shifts population to central cities from states and state remainders with fewer than the average shares of minorities.

- Here is a statement of the scope of conclusions, second paragraph of conclusions:

an adjustment is most needed. Our findings do not permit definitive conclusions for suburban areas, for central cities other than the 16 included in our data set, or for other rural or urban parts of individual states. To compute estimates for such areas, we would prefer not to extrapolate from the regression equations presented in this article. Instead, we would prefer to regroup the CPS sample areas, going from the 16 central cities and 50 states and state remainders to the more homogeneous areas described in Section 6. We would then follow our composite procedure with these redefined units, perhaps with different predictor variables.