Regression: Model diagnostics

Author: Nicholas G Reich

This material is part of the **statsTeachR** *project*

Outline

- Model validation vs. diagnostics
- Outlier classification and influence
- Model selection

Model validation vs. diagnostics

Validation: Is my model generalizable?

- ideally: need external test sample
- internal test sample can offer (limited) help in evaluating overfitting

Diagnostics: Does my model fit the data well?

- very hard to automate this process if something isn't quite right
- requires inspection of residuals
- doesn't necessarily require detailed understanding of the model

types of outliers

Some terminology

- *Outliers* are points that lie away from the cloud of points.
- Outliers that lie horizontally away from the center of the cloud are called *high leverage* points.
- High leverage points that actually influence the slope of the regression line are called *influential* points.
- In order to determine if a point is influential, visualize the regression line with and without the point. Does the slope of the line change considerably? If so, then the point is influential. If not, then it's not an influential point.

Influence

Intuitively, "influence" is a combination of outlying-ness and leverage. More specifically, we can measure the "deletion influence" of each observation: quantify how much $\hat{\boldsymbol{\beta}}$ changes if an observation is left out.

- Mathematically: $|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}|$
- Cook's distance is a value we can calculate for each observation in our dataset that measures this deletion influence. (It uses some nice tricks of linear algebra without having to refit the regression iteratively without each point.)

Poverty data

This is a dataset on US states, containing the % of residents living below the poverty line, as well as some demographic characteristics of states:

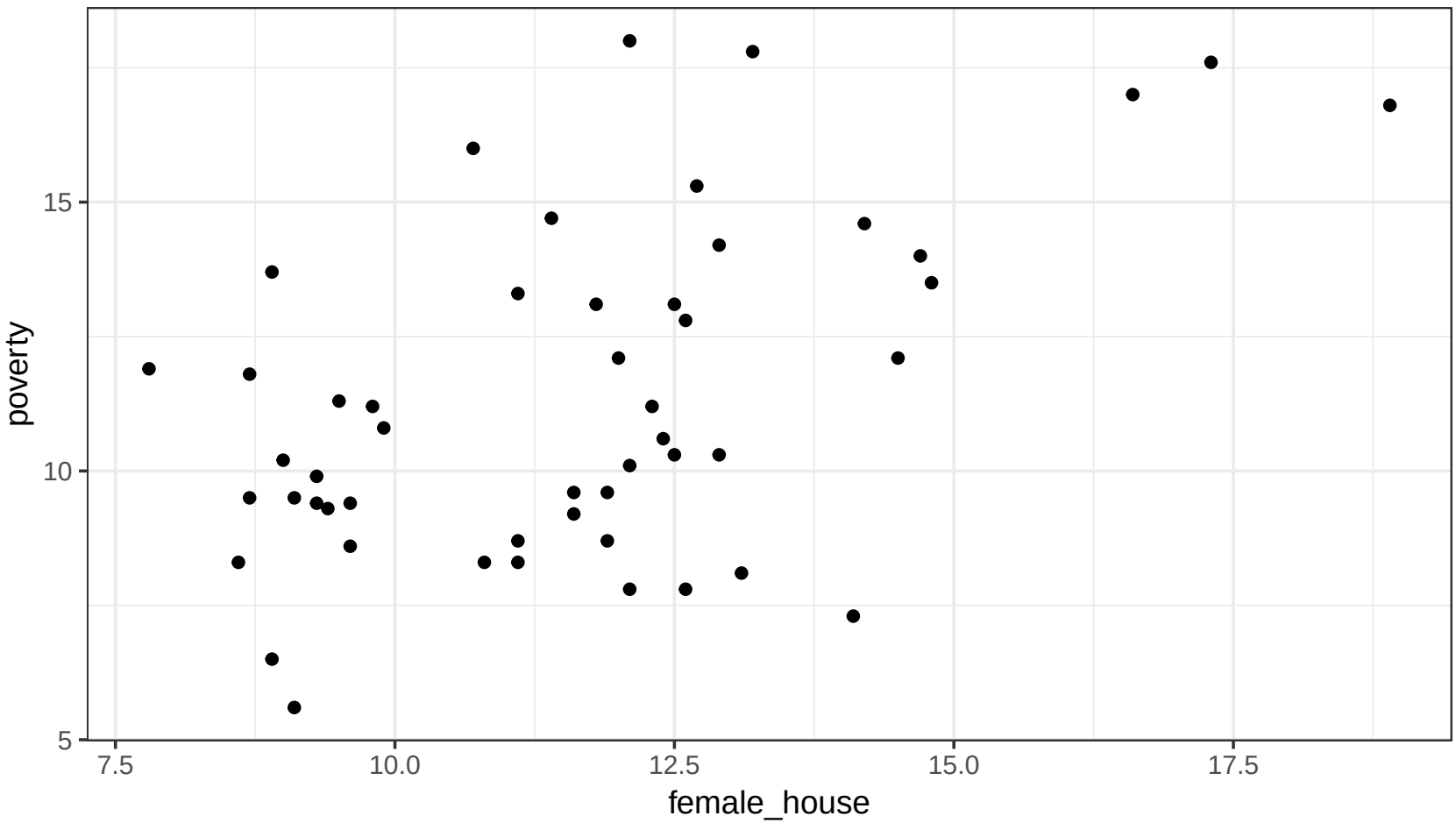
- % living in a metro region
- % white
- % high-school graduates
- % with female head-of-household

```
head(poverty)
```

```
##         state metro_res white hs_grad poverty female_house
## 1     Alabama      55.4  71.3    79.9    14.6         14.2
## 2      Alaska      65.6  70.8    90.6     8.3         10.8
## 3     Arizona      88.2  87.7    83.8    13.3         11.1
## 4    Arkansas      52.5  81.0    80.9    18.0         12.1
## 5  California      94.4  77.5    81.1    12.8         12.6
## 6    Colorado      84.5  90.2    88.7     9.4          9.6
```

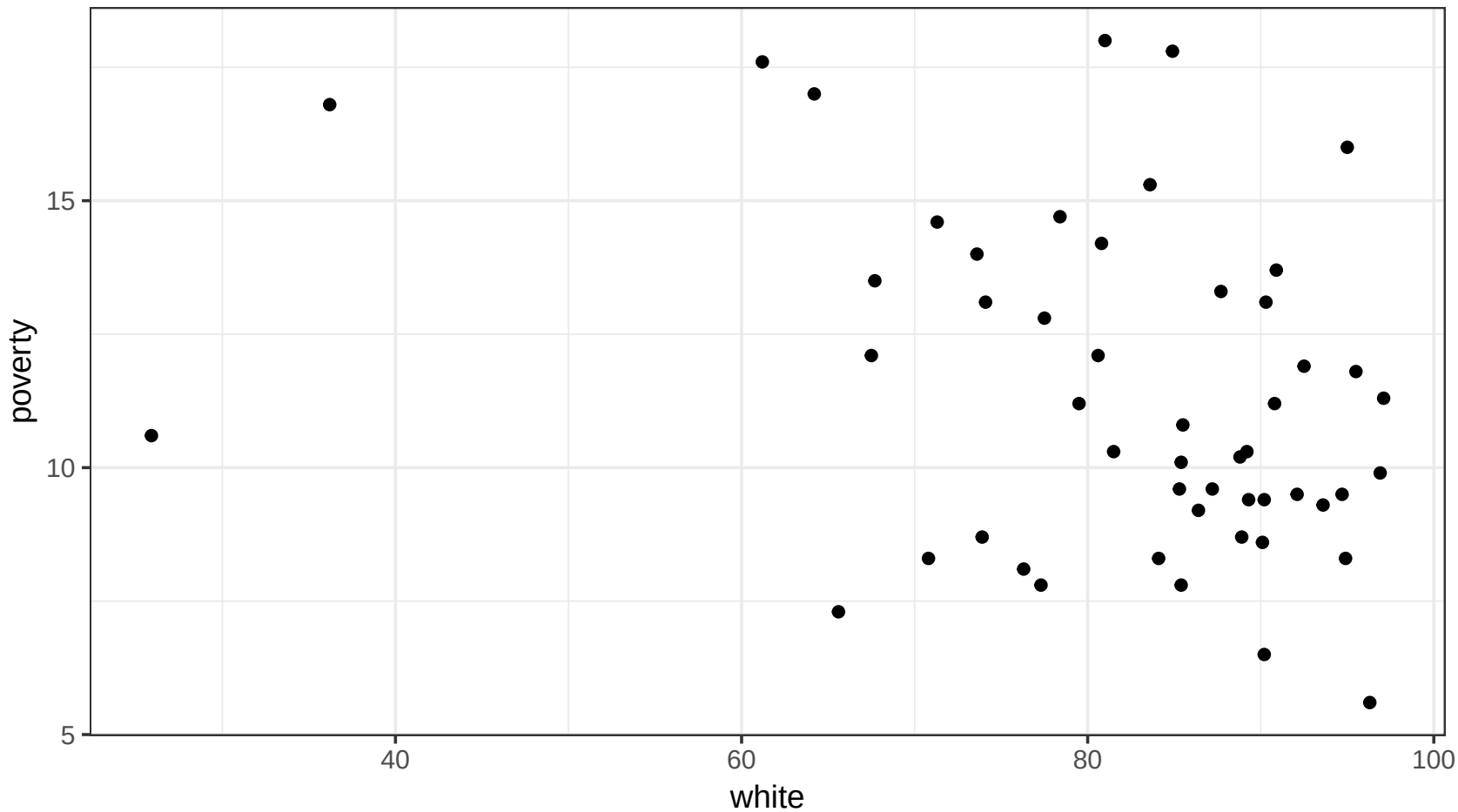
Poverty data: poverty vs % of female HH head

```
ggplot(poverty, aes(x=female_house, y=poverty)) + geom_point()
```



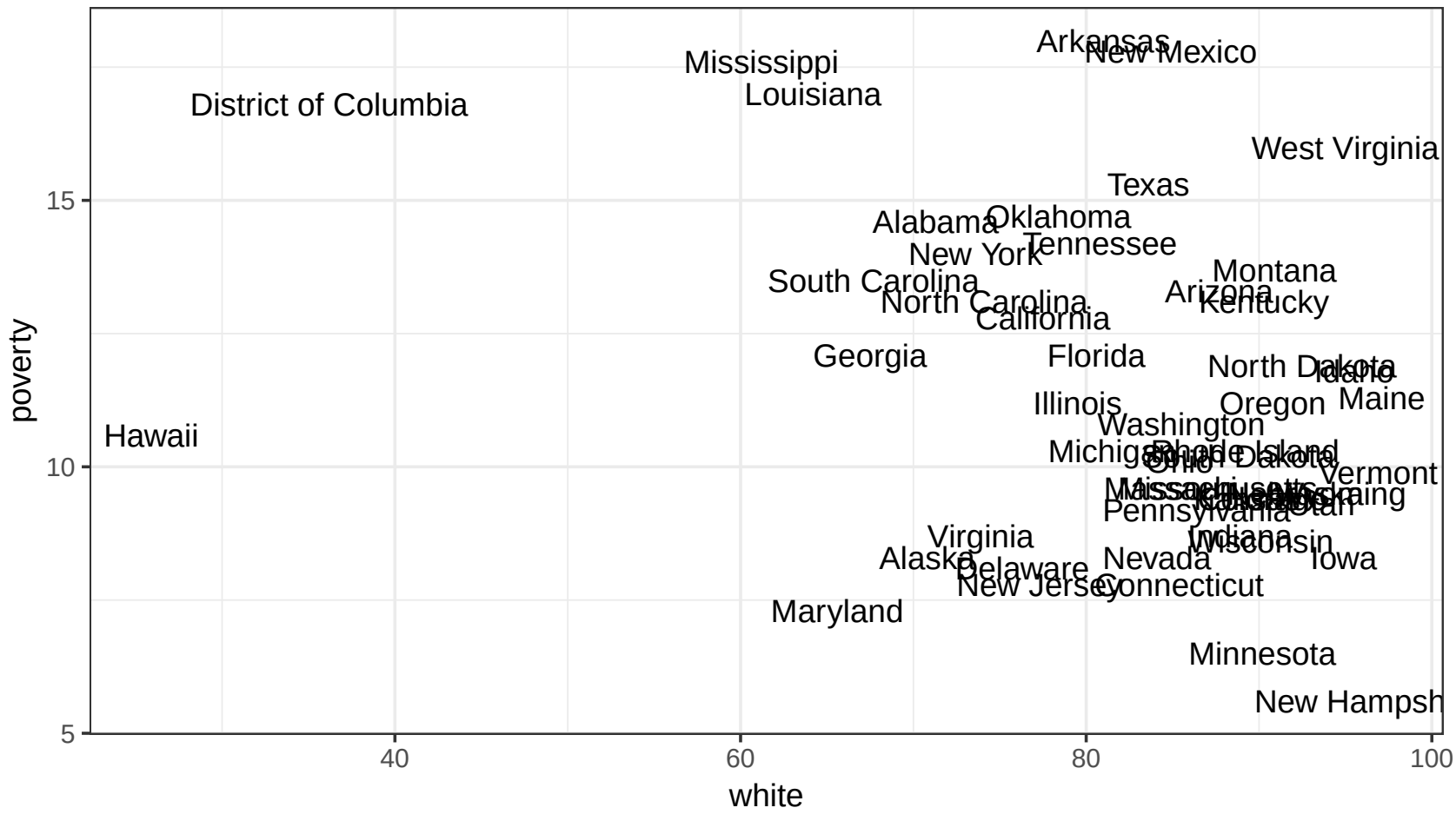
Poverty data: poverty vs % white

```
ggplot(poverty, aes(x=white, y=poverty)) + geom_point()
```



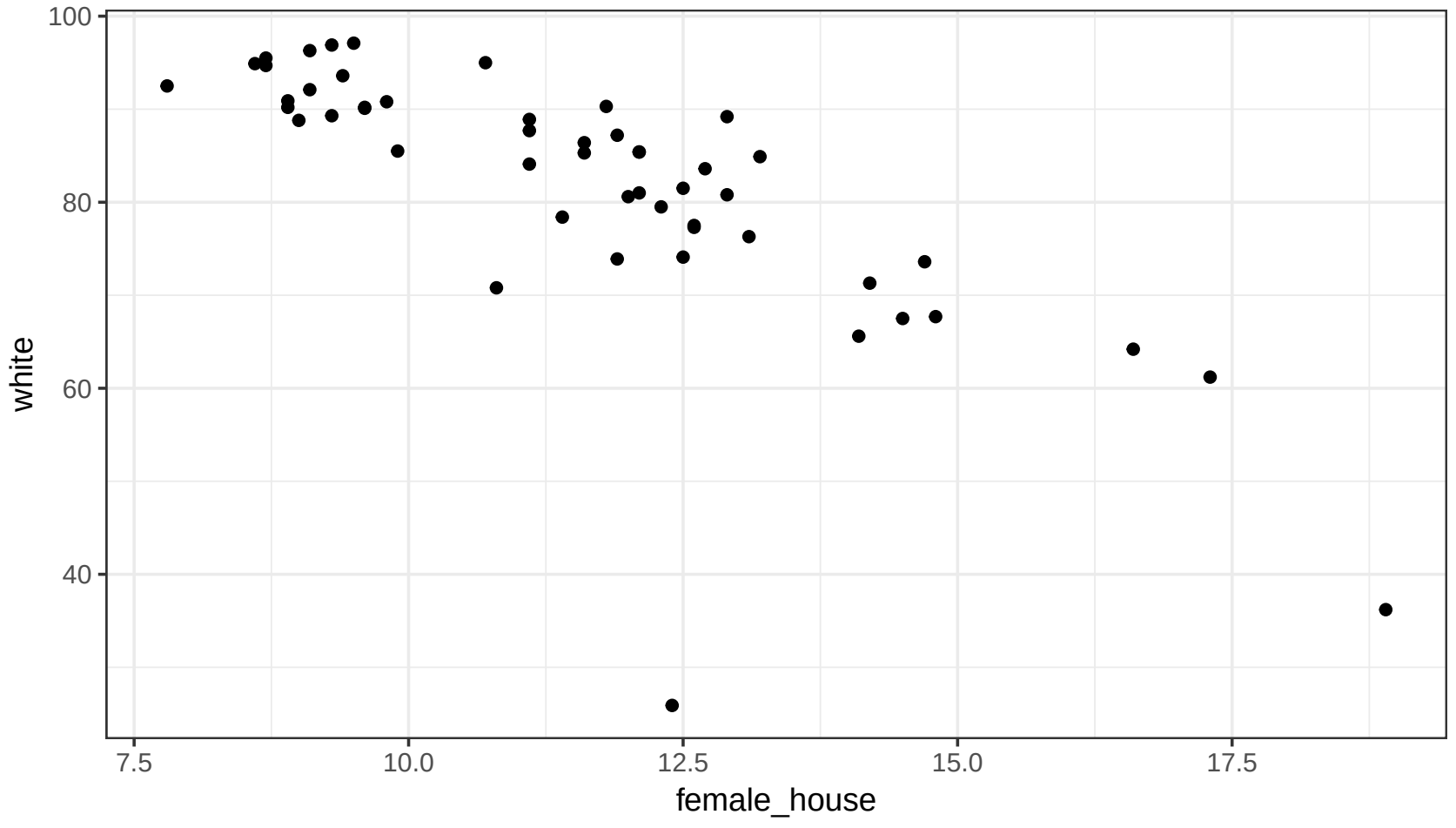
Poverty data: poverty vs % white

```
ggplot(poverty, aes(x=white, y=poverty, label=state)) + geom_text()
```



Poverty data: % white vs % of female HH head

```
ggplot(poverty, aes(x=female_house, y=white)) + geom_point()
```



Example multiple linear regression model

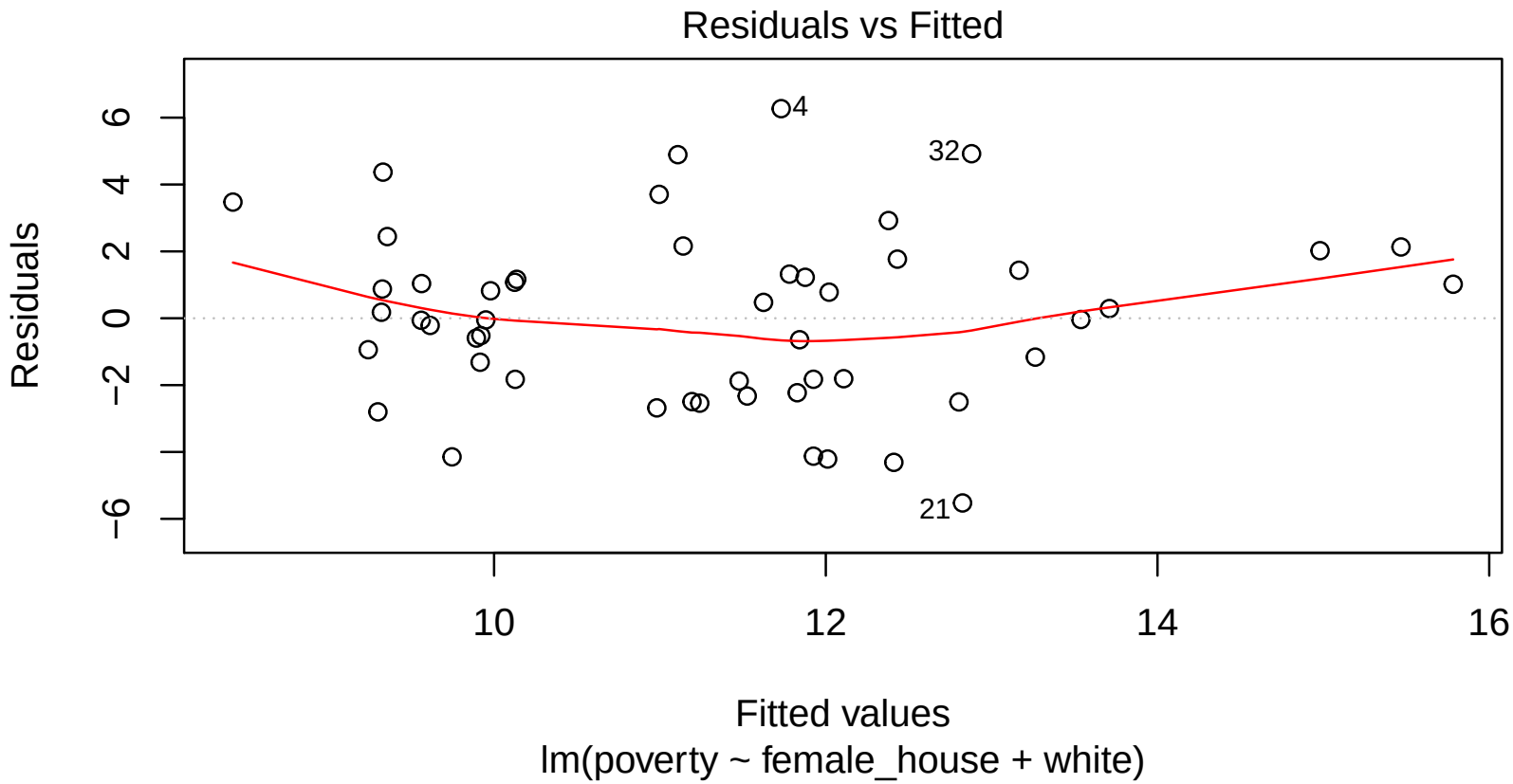
Let's say we are interested in looking at the associations of race and household makeup on poverty at the state level. We could start with a regression model as follows:

$$\widehat{poverty} = \beta_0 + \beta_1 \cdot pct\_white + \beta_2 \cdot pct\_female\_hh$$

Example diagnostic plots with poverty data

You can use the `plot.lm()` function to look at leverage, outlying-ness, and influence all together.

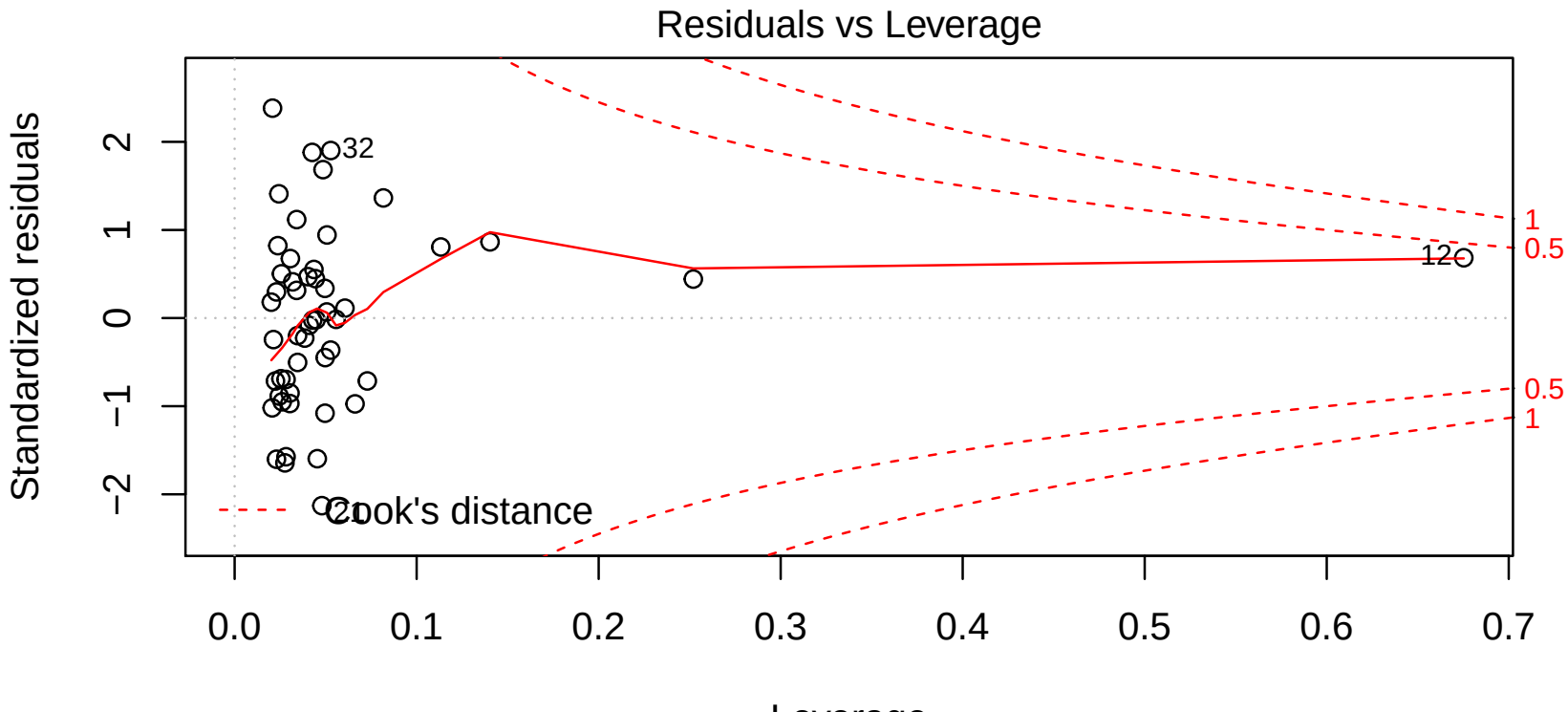
```
mlr = lm(poverty ~ female_house + white, data = poverty)
plot(mlr, which=1)
```



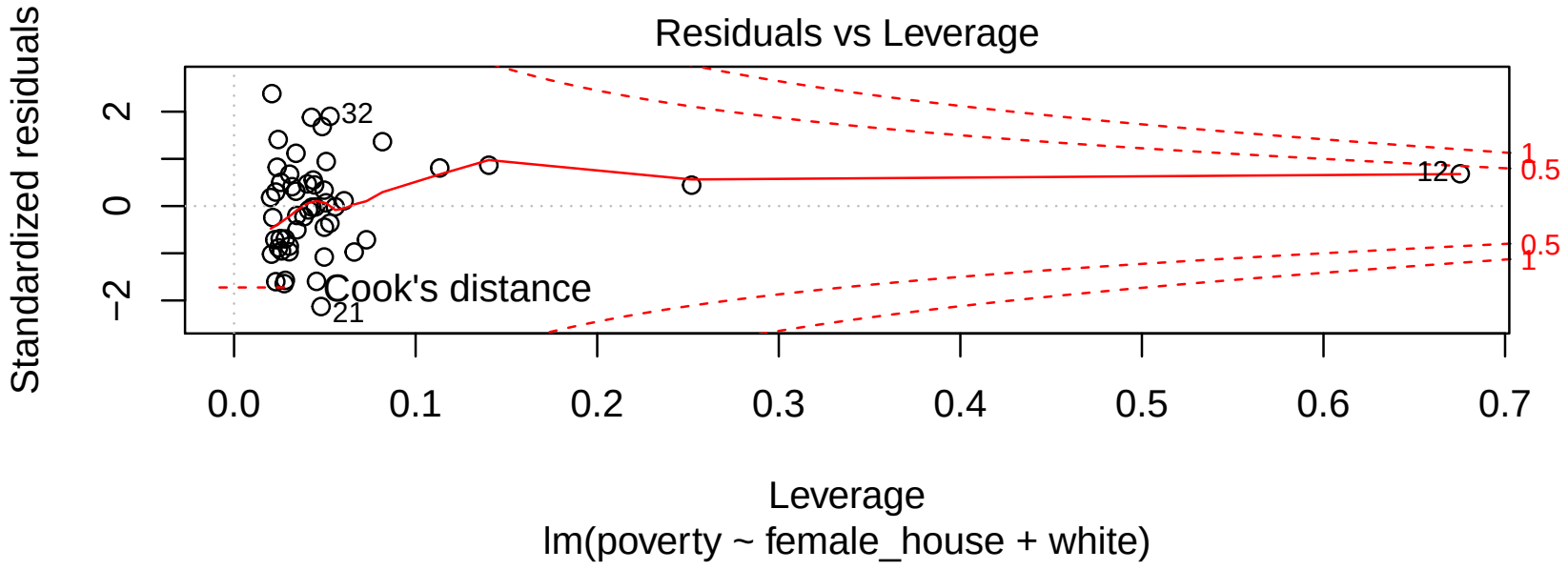
Example diagnostic plots with poverty data

You can use the `plot.lm()` function to look at leverage, outlying-ness, and influence all together. (Note: when you call `plot()` on an `lm` object, i.e. the output from a call to `lm` then you are really using the function `plot.lm()`.)

```
mlr = lm(poverty ~ female_house + white, data = poverty)
plot(mlr, which=5)
```



Investigate identified points!




```
poverty[12,]
```

```
##     state metro_res white hs_grad poverty female_house
## 12 Hawaii      91.5  25.9    88.5    10.6          12.4
```

```
colMeans(poverty[,2:6])
```

```
##    metro_res        white      hs_grad      poverty female_house
##     72.24902     81.71961     86.01176     11.34902     11.63333
```

Model diagnostics summary

You are looking for...

- Points that show worrisome level of influence $\Longrightarrow$ sensitivity analysis!
- Systematic departures from model assumptions $\Longrightarrow$ transformations, different model structure
- Unrealistic outliers $\Longrightarrow$ check your data!

No points show worrisome influence in this poverty data analysis, although observation 12 was high leverage.

model selection

Model selection

Why are you building a model in the first place?

Model selection: considerations

Things to keep in mind...

- **Why am I building a model?** Some common answers
  - Estimate an association
  - Test a particular hypothesis
  - Predict new values
- What predictors will I allow?
- What predictors are needed?

Different answers to these questions will yield different final models.

Model selection: realities

All models are wrong. Some are more useful than others.
- George Box

- In practice, issues with sample size, collinearity (highly correlated predictor variables), and limited available predictors are real problems.
- There is not a single best algorithm for model selection! It pretty much always requires thoughful reasoning and knowledge about the data at hand.
- When in doubt (unless you are specifically "data mining"), err on the side creating a process that does not require choices being made (by you or the computer) about which covariates to include.

Basic ideas for model selection

For association studies, when your sample size is large

- Include key covariates of interest.
- Include covariates needed because they might be confounders.
- Include covariates that your colleagues/reviewers/collaborators will demand be included for face validity.
- Do NOT go on a fishing expedition for significant results!
- Do NOT use "stepwise selection" methods!
- Subject the selected model to model checking/diagnostics, possibly adjust model structure (i.e. include non-linear relationships with covariates) as needed.

Basic ideas for model selection

For association studies, when your sample size is small

- Same as above, but may need to be more frugal with how many predictors you include.
- Rule of thumb for multiple linear regression is to have at least 15 observations for each covariate you include in your model.