Introduction to Data Visualization

Author: Nicholas G Reich

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

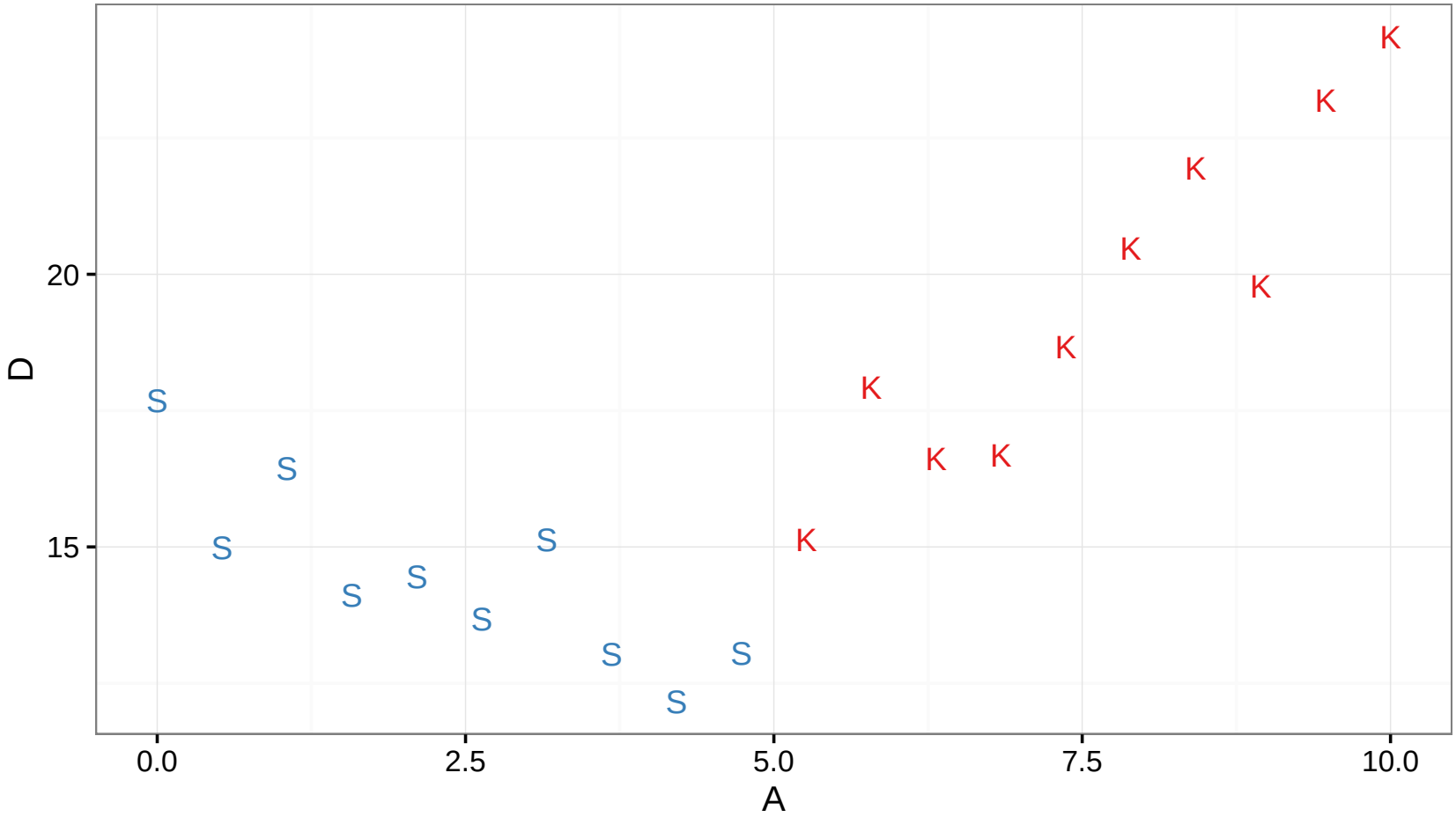
Outline

- Review of model syntax
- Summary of R^2 and R^2_{adj}
- Outlier classification and influence
- Model selection

model syntax

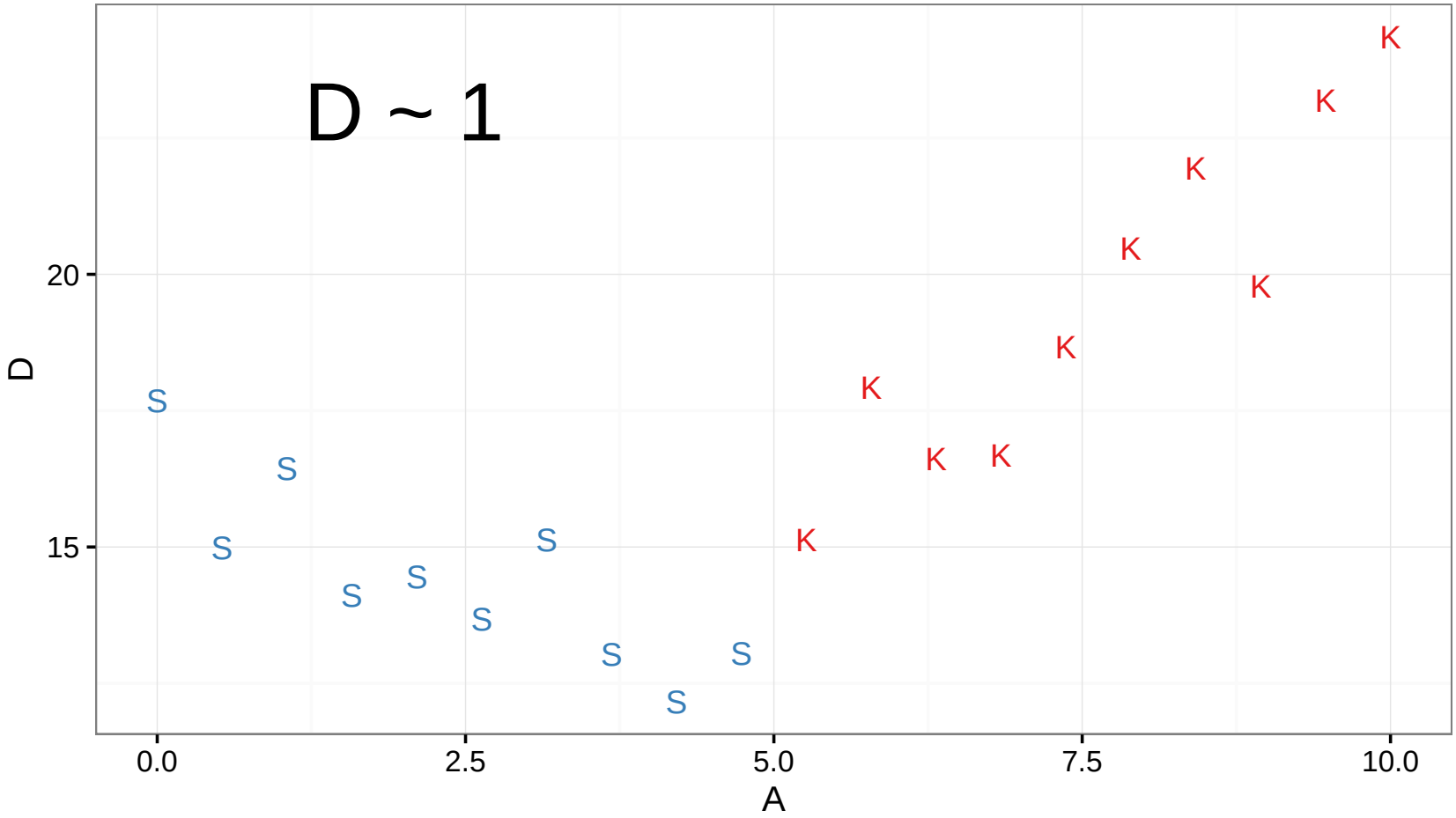
Example data

- D = a quantitative variable
- A = a quantitative variable
- G = a categorical variable with two levels, S and K



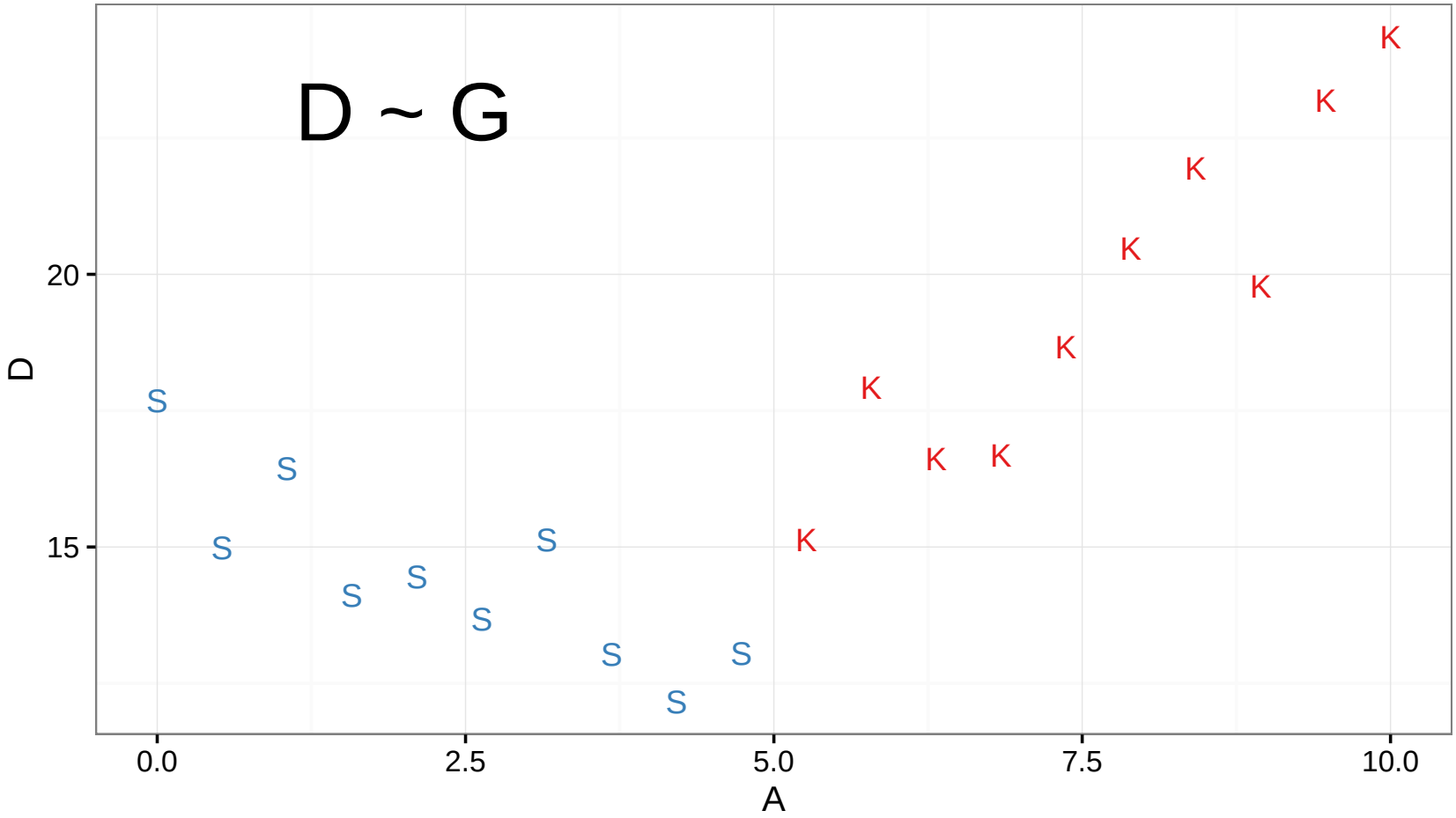
Draw the model...

- D = a quantitative variable
- A = a quantitative variable
- G = a categorical variable with two levels, S and K



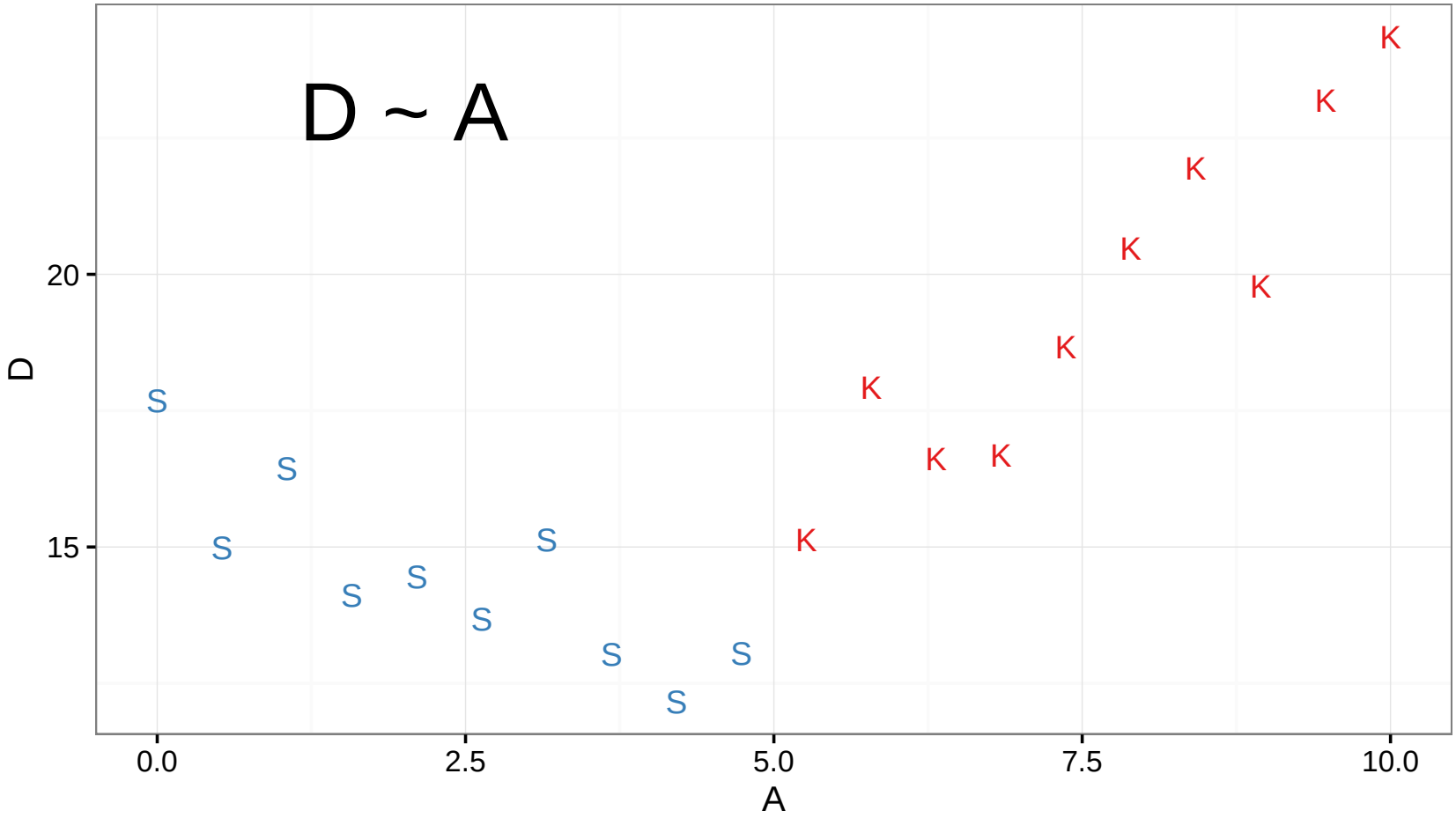
Draw the model...

- D = a quantitative variable
- A = a quantitative variable
- G = a categorical variable with two levels, S and K



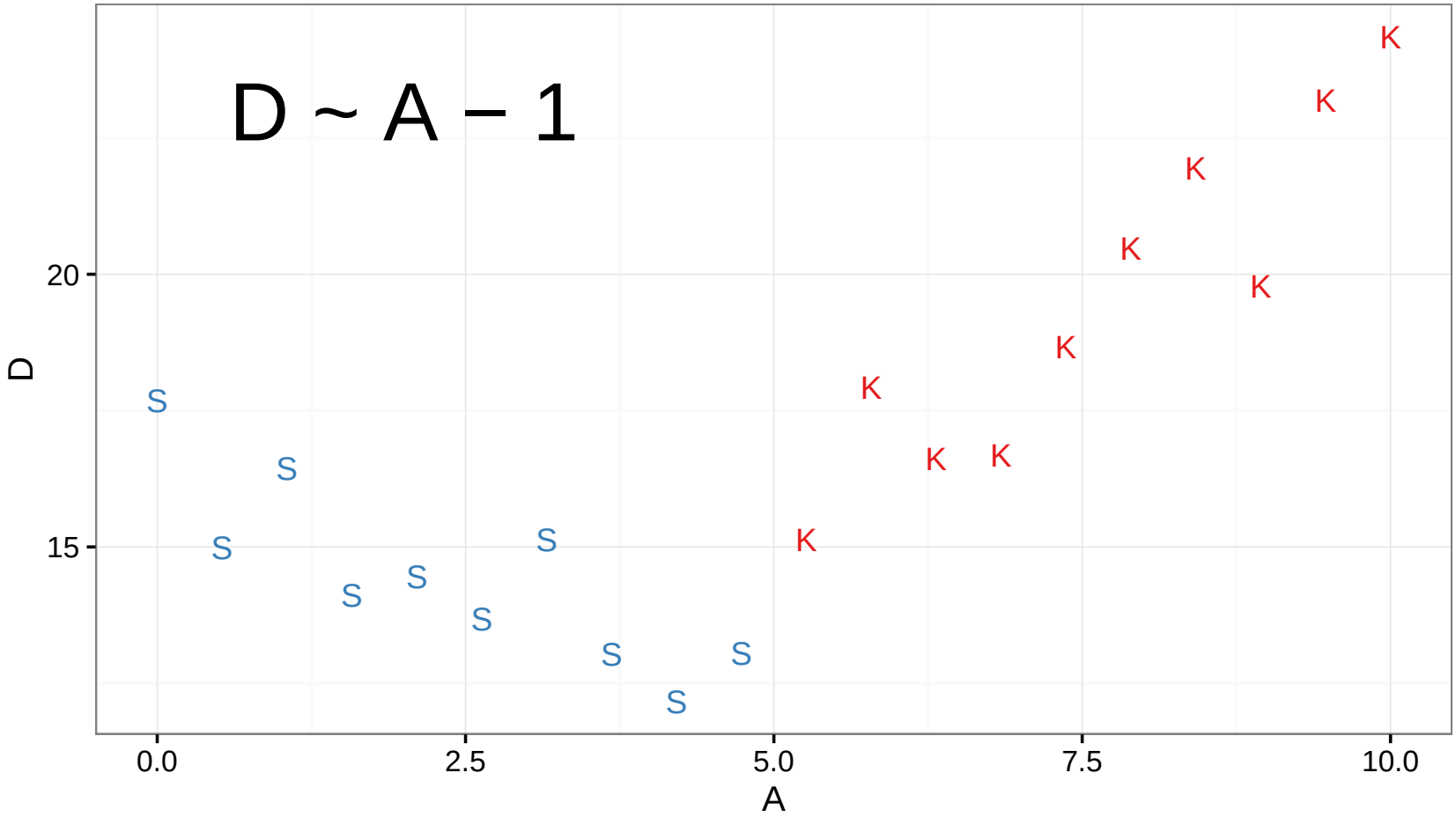
Draw the model...

- D = a quantitative variable
- A = a quantitative variable
- G = a categorical variable with two levels, S and K



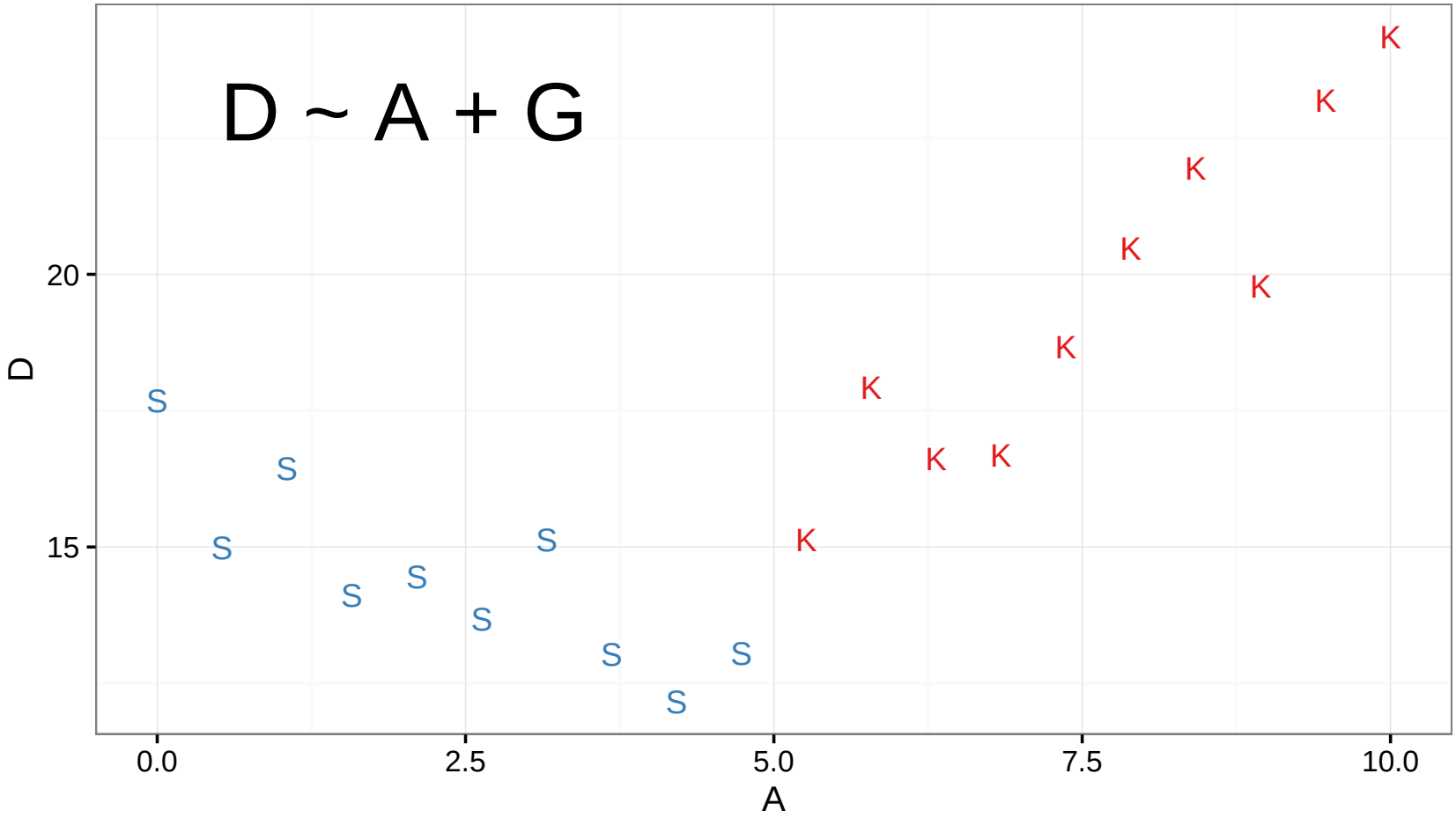
Draw the model...

- D = a quantitative variable
- A = a quantitative variable
- G = a categorical variable with two levels, S and K



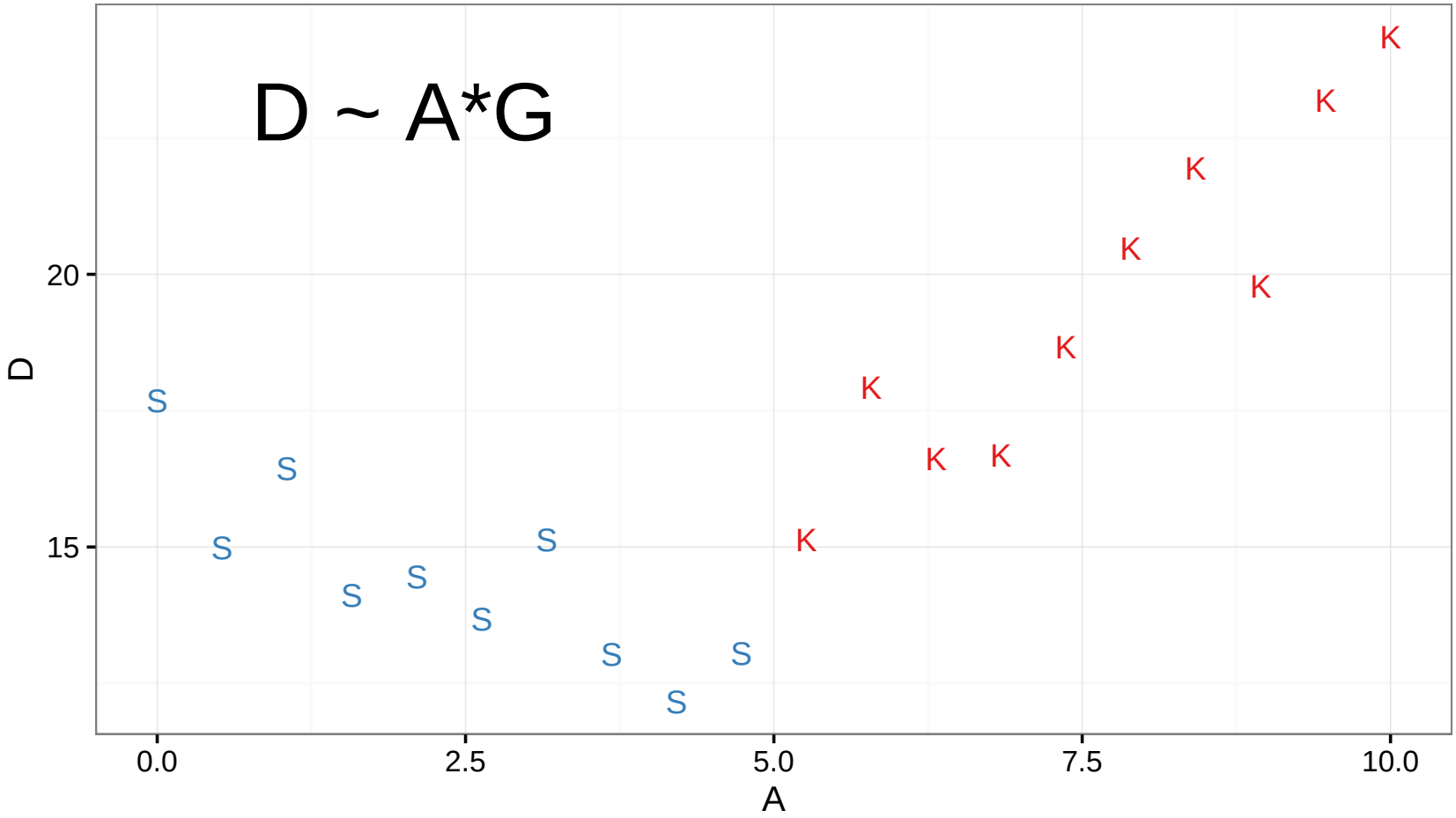
Draw the model...

- D = a quantitative variable
- A = a quantitative variable
- G = a categorical variable with two levels, S and K



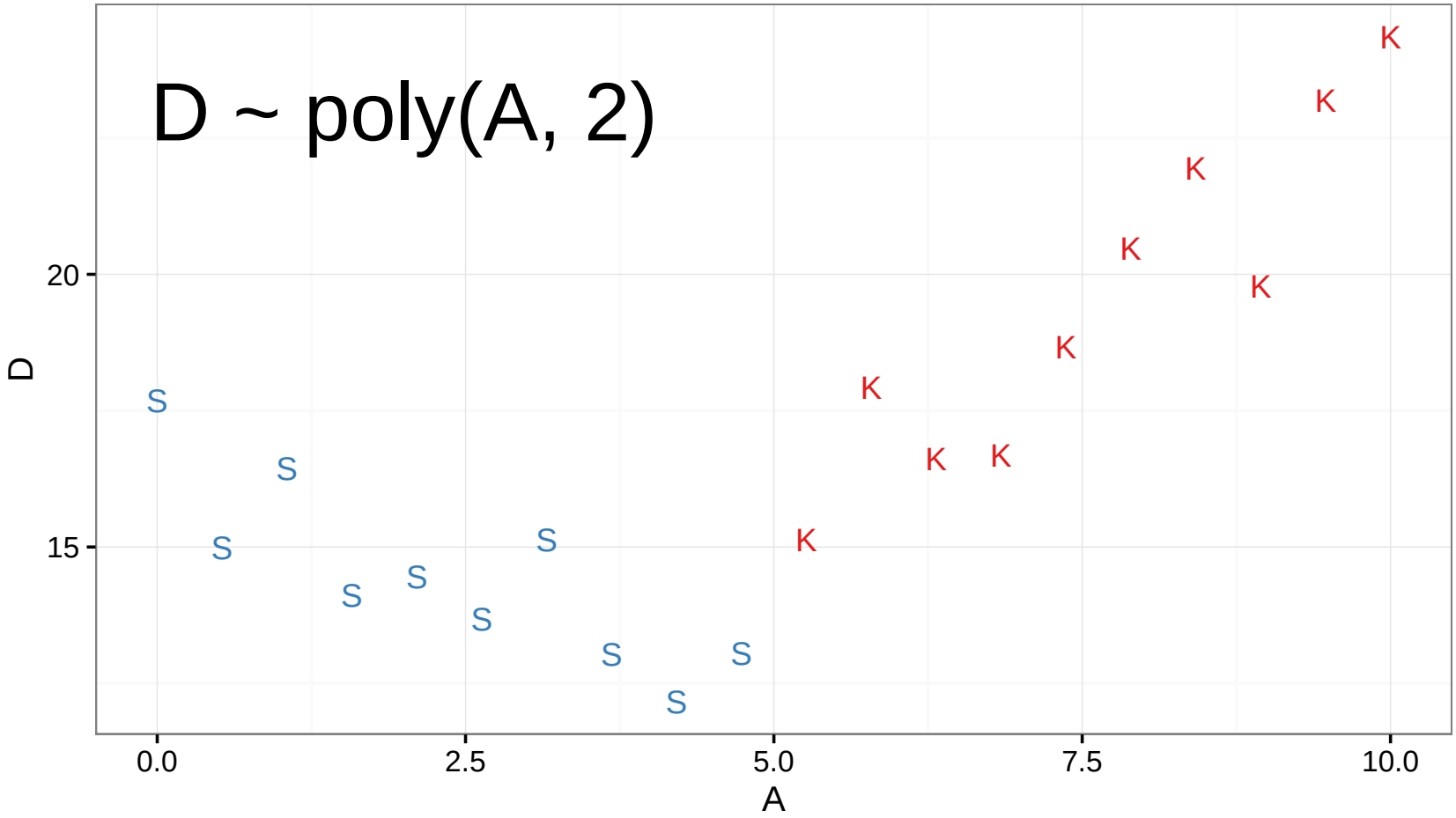
Draw the model...

- D = a quantitative variable
- A = a quantitative variable
- G = a categorical variable with two levels, S and K



Draw the model...

- D = a quantitative variable
- A = a quantitative variable
- G = a categorical variable with two levels, S and K



summary of R^2 and R^2_{adj}

Another look at R^2

R^2 can be calculated in three ways:

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)
2. square the correlation coefficient of y and $\hat{y}$

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)
2. square the correlation coefficient of y and $\hat{y}$
3. based on definition:

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in } y}$$

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)
2. square the correlation coefficient of y and $\hat{y}$
3. based on definition:

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in } y}$$

Using an *ANOVA* table we can calculate the explained variability and total variability in y.

Sum of squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

$$
\begin{aligned}
\text{Sum of squares of } y\text{: } SS_{Total} &= \sum (y - \bar{y})^2 = 480.25 \\
\text{Sum of squares of residuals: } SS_{Error} &= \sum e_i^2 = 347.68 \\
\text{Sum of squares of } x\text{: } SS_{Model} &= SS_{Total} - SS_{Error} \\
&= 480.25 - 347.68 = 132.57
\end{aligned}
$$

$$
R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57}{480.25} = 0.28
$$

Why bother?

Why bother with another approach for calculating R^2 when we had a perfectly good way to calculate it as the correlation coefficient squared?

Why bother?

Why bother with another approach for calculating R^2 when we had a perfectly good way to calculate it as the correlation coefficient squared?

- For single-predictor linear regression, having three ways to calculate the same value may seem like overkill.
- However, in multiple linear regression, we can't calculate R^2 as the square of the correlation between x and y because we have multiple xs.
- And next we'll learn another measure of explained variability, *adjusted* R^2, that requires the use of the third approach, ratio of explained and unexplained variability.

Predicting poverty using % female hh + % white

Linear model:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

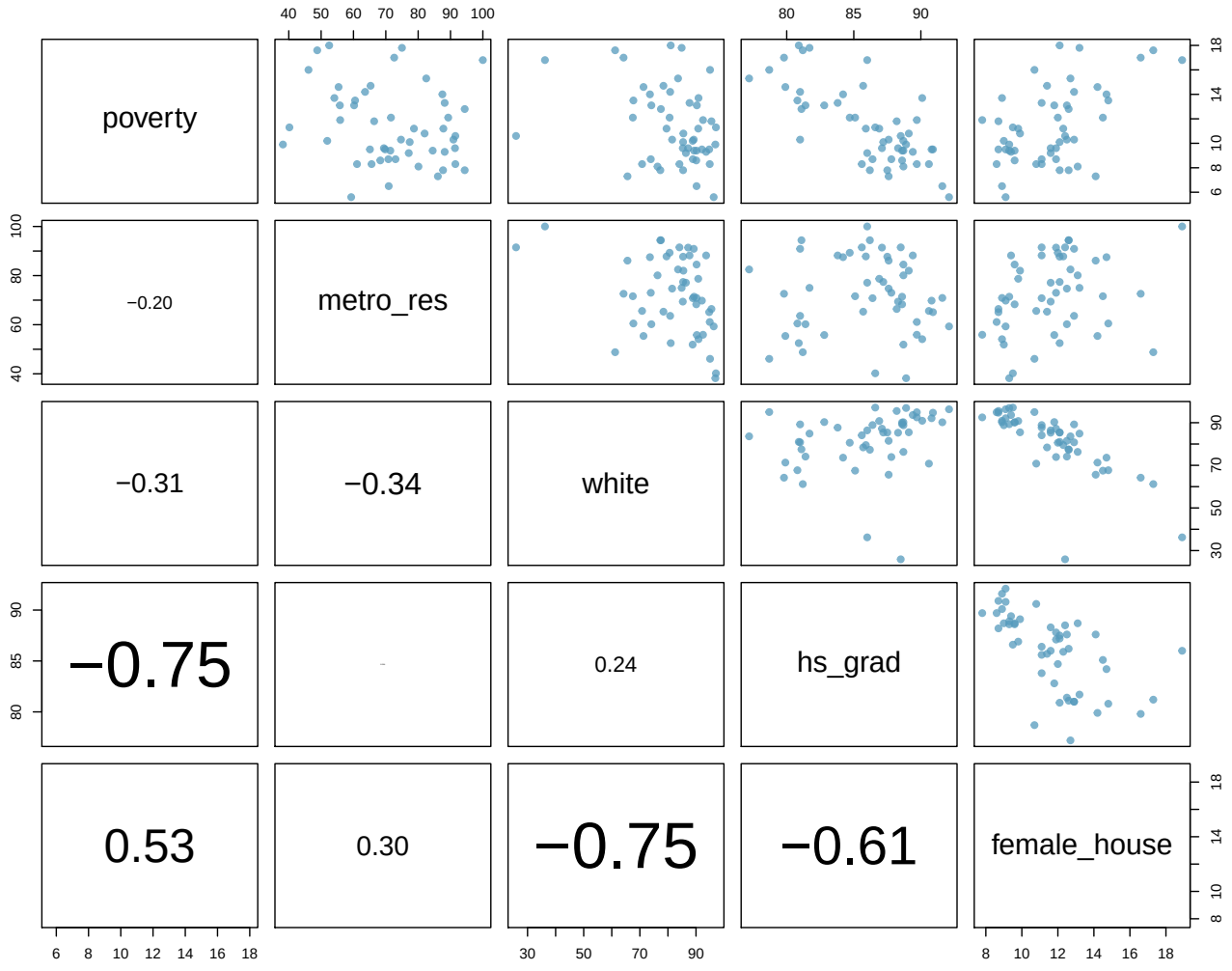
Predicting poverty using % female hh + % white

Linear model:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57 + 8.21}{480.25} = 0.29$$

Does adding the variable Varwhite to the model add valuable information that wasn't provided by Varfemale_house?



R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

- When any variable is added to the model R^2 increases.

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

- When any variable is added to the model R^2 increases.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adjusted R^2 does not increase.

Adjusted R^2

Adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right)$$

where n is the number of cases and p is the number of predictors (explanatory variables) in the model.

- Because p is never negative, R^2_{adj} will always be smaller than R^2.
- R^2_{adj} applies a penalty for the number of predictors included in the model.
- Therefore, we choose models with higher R^2_{adj} over others.

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right)$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right)
\end{aligned}
$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48} \right)
\end{aligned}
$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48} \right) \\
&= 1 - 0.74
\end{aligned}
$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48} \right) \\
&= 1 - 0.74 \\
&= 0.26
\end{aligned}
$$

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

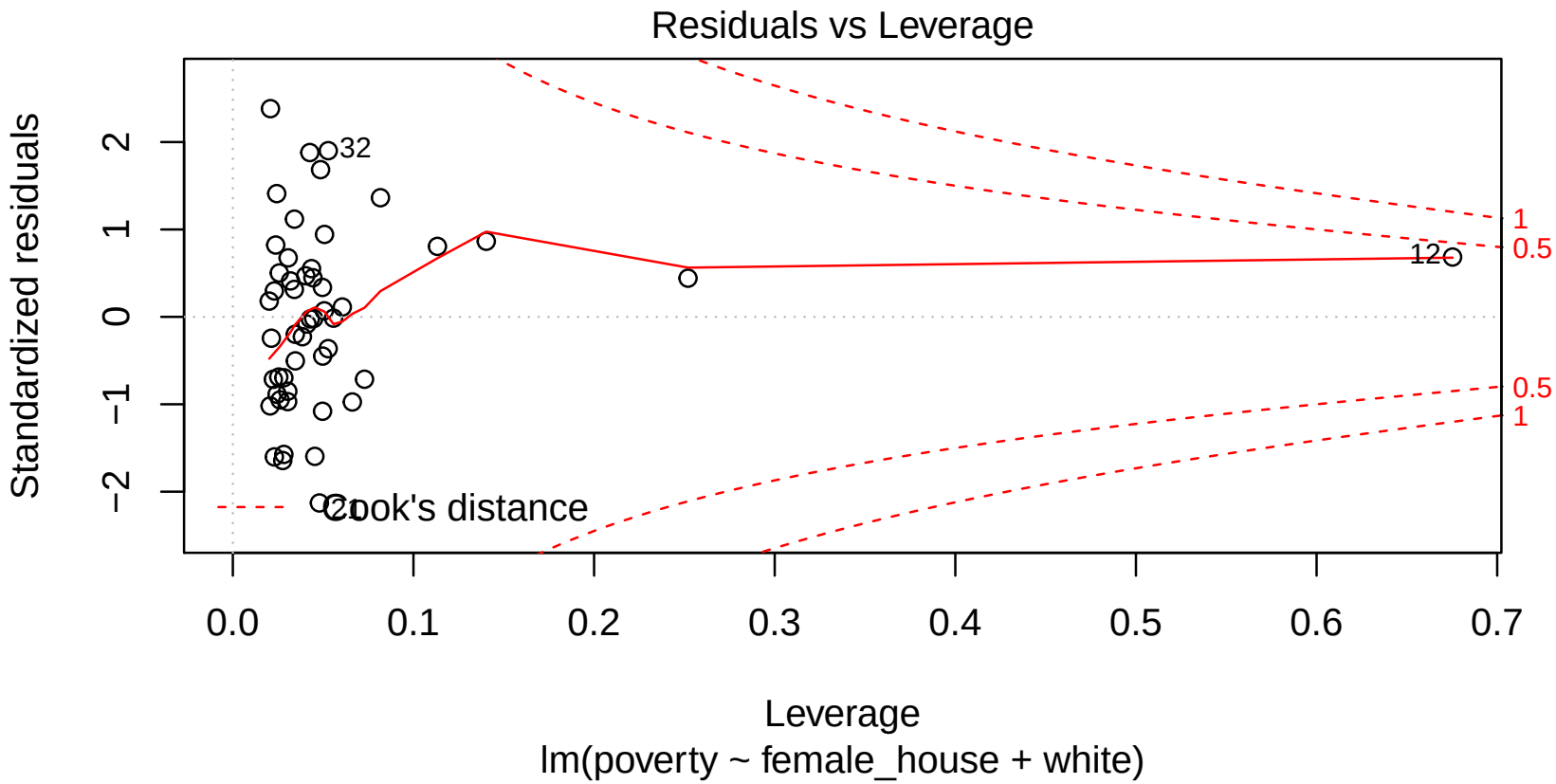
Intuitively, "influence" is a combination of outlying-ness and leverage. More specifically, we can measure the "deletion influence" of each observation: quantify how much $\hat{\boldsymbol{\beta}}$ changes if an observation is left out.

- 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.)

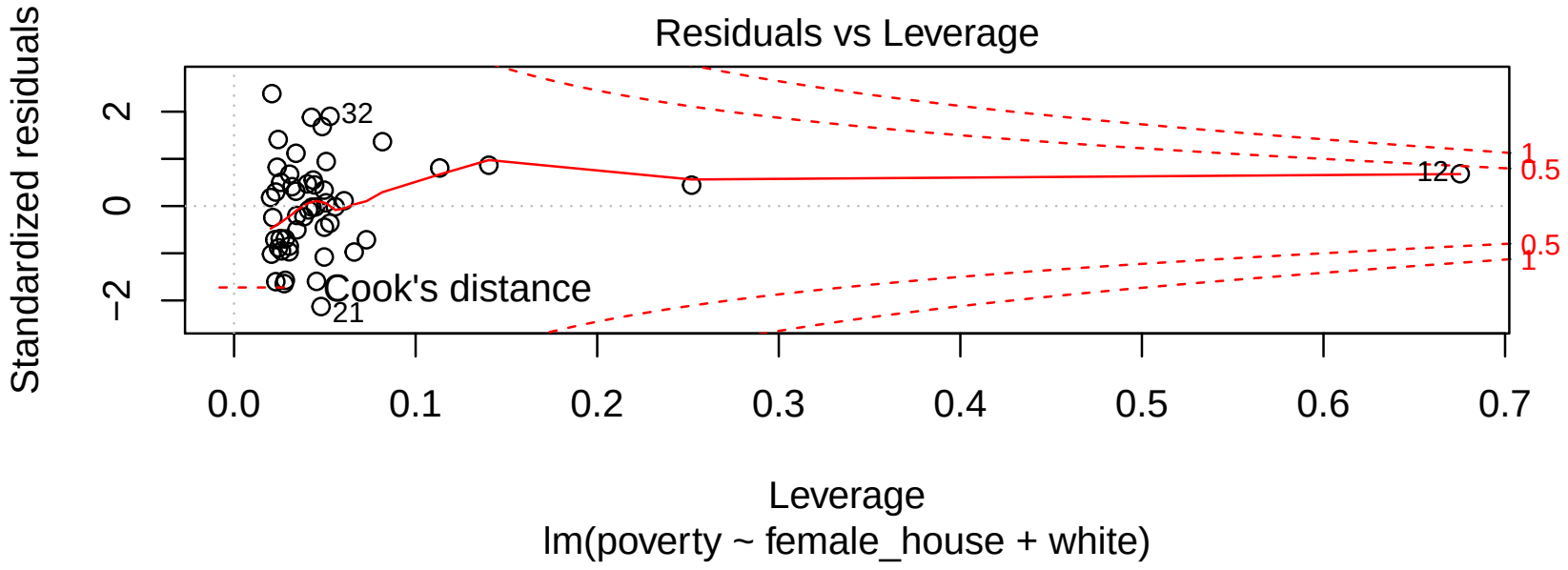
Example diagnostic plots with poverty data

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=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 checking 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, and 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.