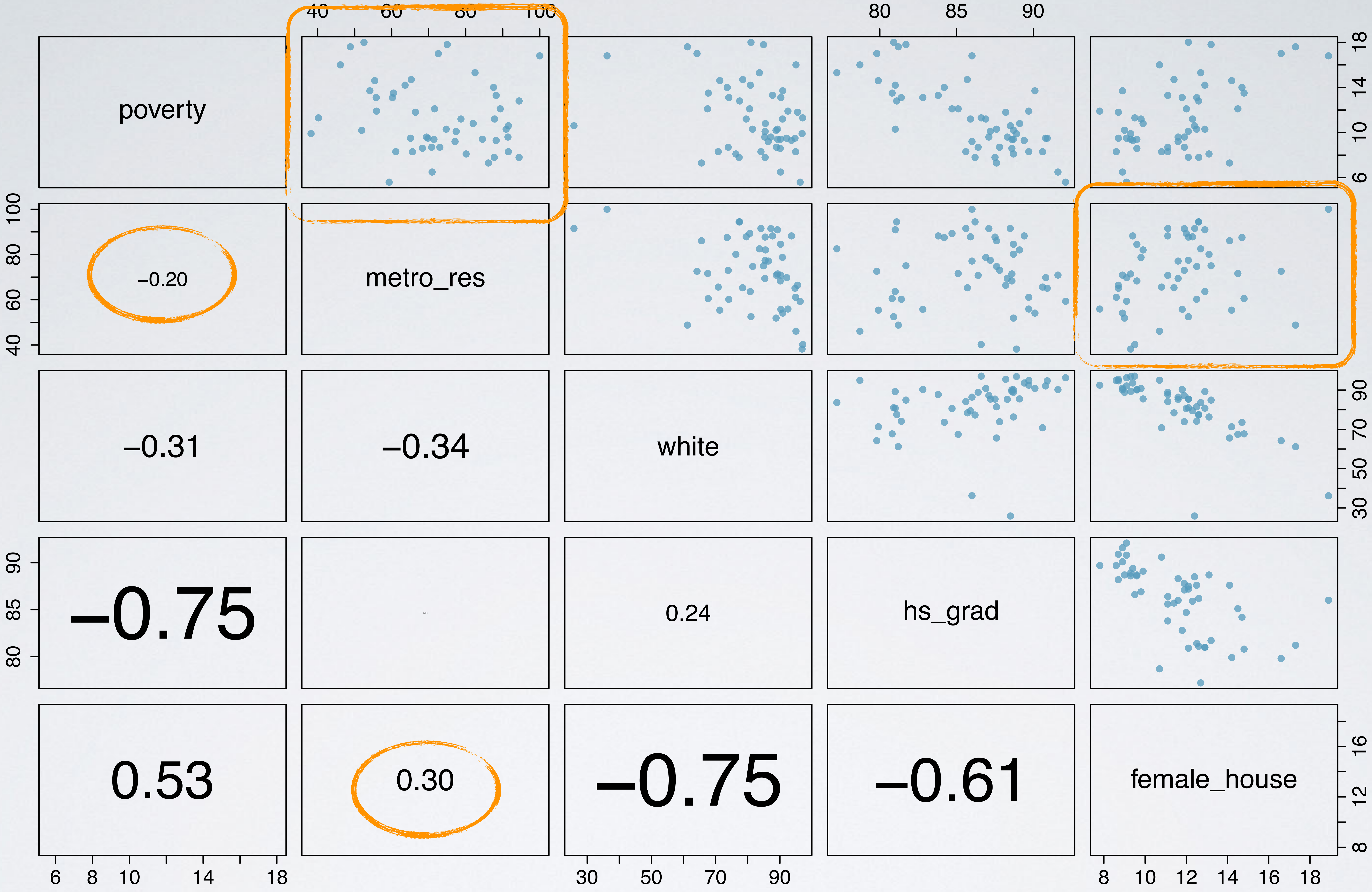
adjusted R^2

- calculation
- uses

Duke
UNIVERSITY

Dr. Mine Çetinkaya-Rundel
Duke University


poverty
metro_res
white
hs_grad
female_house
−0.20
−0.31
−0.34
−0.75
0.24
0.53
0.30
−0.75
−0.61

R

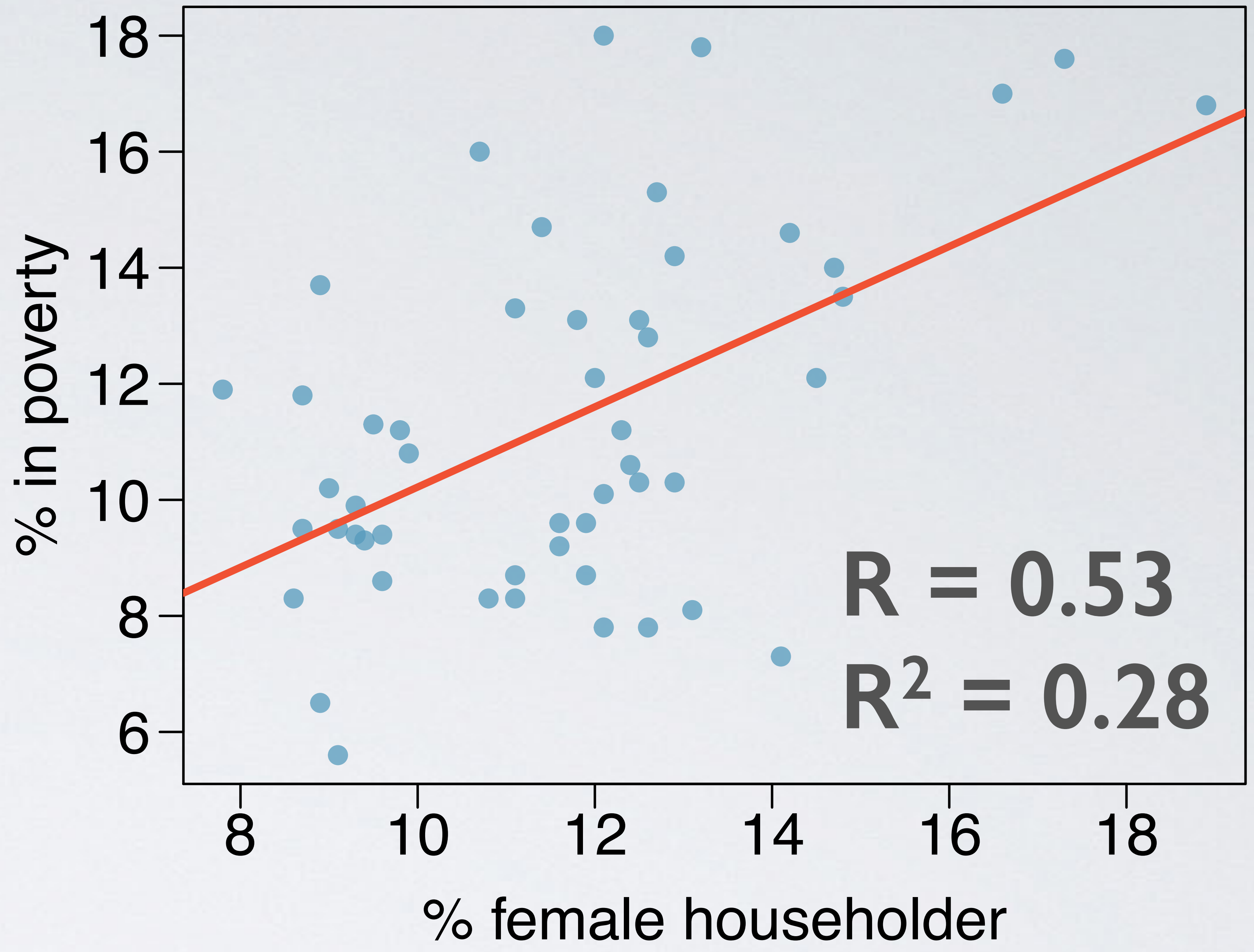
```
# load data
> states = read.csv("http://bit.ly/dasi_states")

# fit model
> pov_slr = lm(poverty ~ female_house, data = states)
> summary(pov_slr)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.3094     1.8970   1.745   0.0873 .
female_house   0.6911     0.1599   4.322 7.53e-05 ***

Residual standard error: 2.664 on 49 degrees of freedom
Multiple R-squared:  0.276,  Adjusted R-squared:  0.2613
F-statistic: 18.68 on 1 and 49 DF,  p-value: 7.534e-05
```

predicting poverty from % female householder



Linear model:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	0.69	0.16	4.32	0.00

another look at R^2

ANOVA:	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			

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

predicting poverty from % female householder + % white

R

```
> pov_mlr = lm(poverty ~ female_house + white, data = states)
> summary(pov_mlr)
```

	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

R

```
> anova(pov_mlr)
```

	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			

$$R^2 = \frac{132.57 + 8.21}{480.25} = 0.29$$

adjusted R^2

adjusted R^2: $$R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-k-1}\right)$$

k : number of predictors

practice

Calculate adjusted R^2 for the multiple linear regression model predicting % living in poverty from % female householders and % white.
Remember n = 51 (50 states + DC).

	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			

$$R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-k-1} \right)$$

$$= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) = 0.26$$

R^2 vs. adjusted R^2

	R^2	adjusted R^2
Model 1 (poverty vs. female_house)	0.28	0.26
Model 2 (poverty vs. female_house + white)	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.

properties of adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-k-1} \right)$$

- k is never negative → adjusted $R^2 < R^2$
- adjusted R^2 applies a penalty for the number of predictors included in the model
- we choose models with higher adjusted R^2 over others