Multiple regression

- Simple linear regression: Bivariate - two variables: y and x
- Multiple linear regression: Multiple variables: y and $x_1, x_2, \cdots$

Multiple regression

$$\widehat{\%\ in\ poverty} = \beta_0 + \beta_1 * \%\ HS\ grad + \beta_2 * \%\ Metropolitan\ Residence$$

We can try to predict poverty with HS graduation rate AND Metropolitan Residence ratio

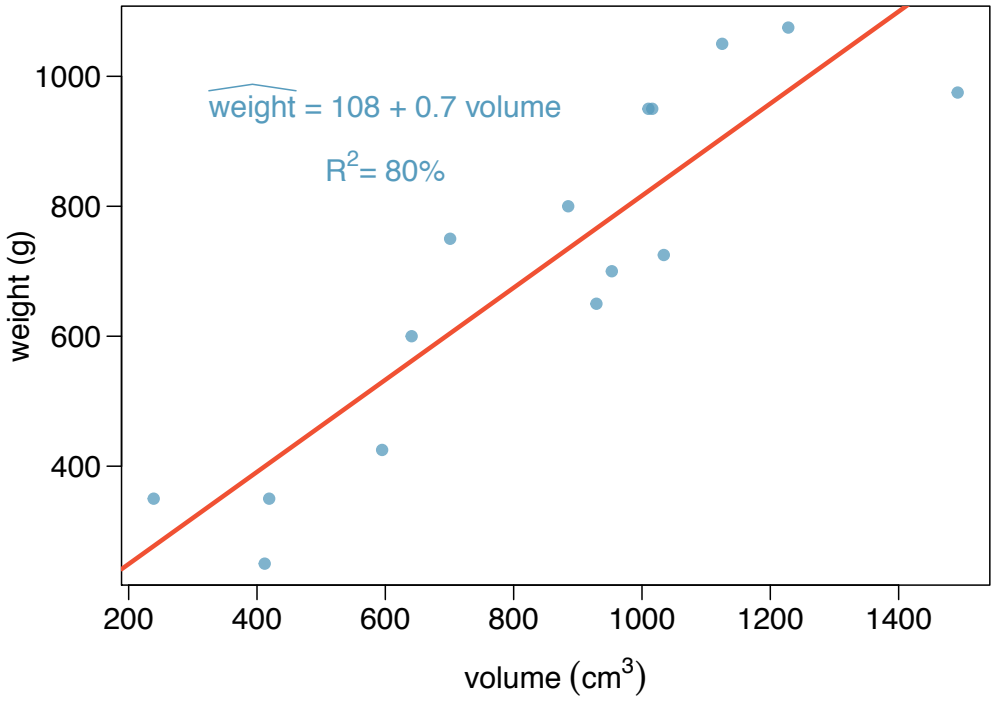
```
my_lm=lm(Poverty~Graduates+Metropolitan.Residence,
         data=poverty)
```

Weights of books

	weight (g)	volume (cm^3)	cover
1	800	885	hc
2	950	1016	hc
3	1050	1125	hc
4	350	239	hc
5	750	701	hc
6	600	641	hc
7	1075	1228	hc
8	250	412	pb
9	700	953	pb
10	650	929	pb
11	975	1492	pb
12	350	419	pb
13	950	1010	pb
14	425	595	pb
15	725	1034	pb

Weights of books (cont.)

The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



(a) Weights of 80% of the books can be predicted accurately using this model.

(b) Books that are 10 cm^3 over average are expected to weigh 7 g over average.

(c) The correlation between weight and volume is $R = 0.80^2 = 0.64$.

(d) The model underestimates the weight of the book with the highest volume.

Modeling weights of books using volume

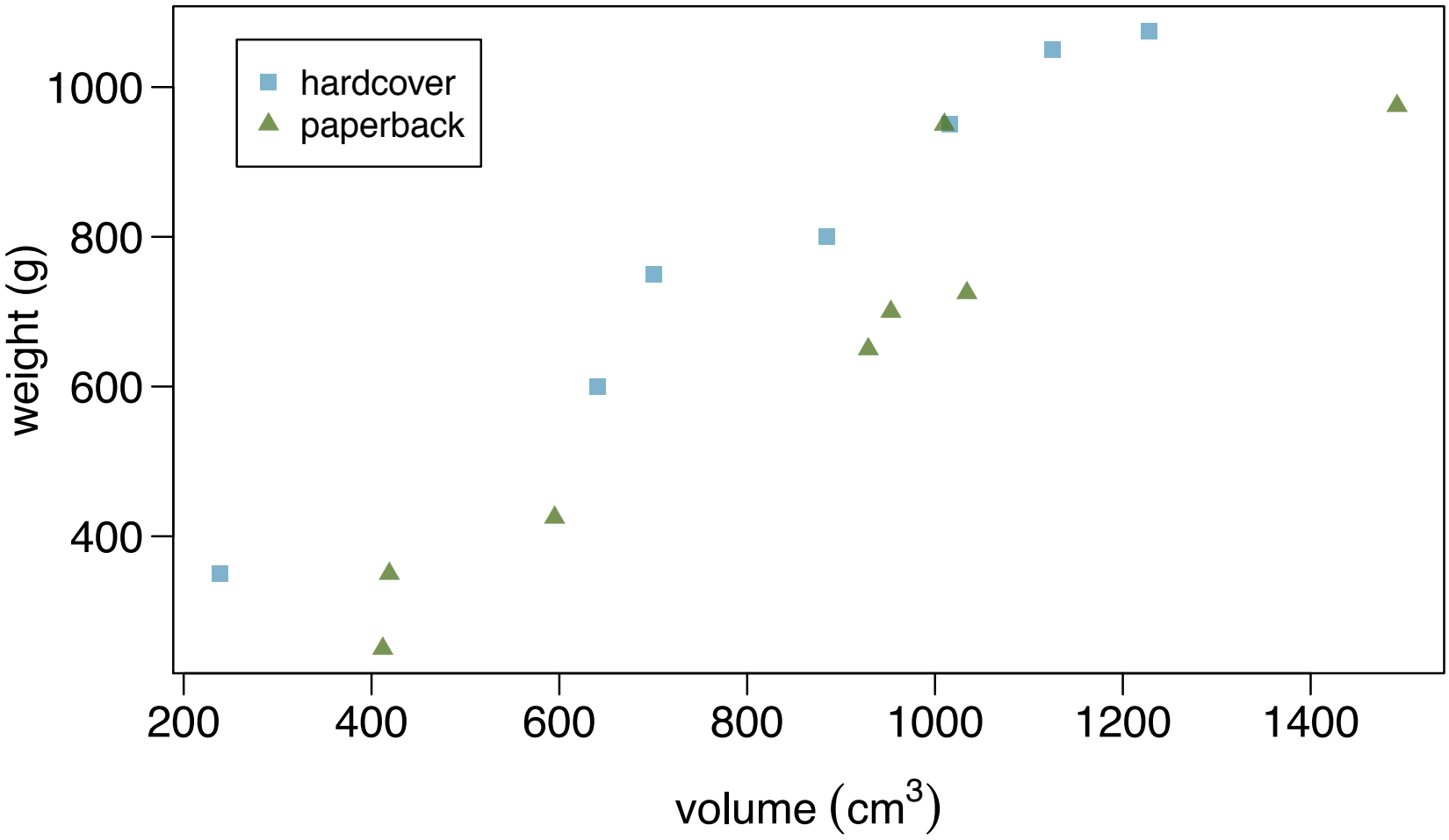
```
data(allbacks)
m1 = lm(weight ~ volume, data = allbacks)
summary(m1)
```

somewhat abbreviated output...

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 107.67931   88.37758   1.218    0.245
volume        0.70864    0.09746   7.271 6.26e-06 ***


Residual standard error: 123.9 on 13 degrees of freedom
Multiple R-squared:  0.8026,  Adjusted R-squared:  0.7875
F-statistic: 52.87 on 1 and 13 DF,  p-value: 6.262e-06
```

Weights of hardcover and paperback books

Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?



Modeling weights of books using volume and cover type

```
m2 = lm(weight ~ volume+cover, data = allbacks)
summary(m2)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  197.96284   59.19274   3.344 0.005841 **
volume         0.71795    0.06153  11.669  6.6e-08 ***
cover:pb    -184.04727   40.49420  -4.545 0.000672 ***


Residual standard error: 78.2 on 12 degrees of freedom
Multiple R-squared: 0.9275,  Adjusted R-squared: 0.9154
F-statistic: 76.73 on 2 and 12 DF,  p-value: 1.455e-07
```

Determining the reference level

Which of the below correctly describes the roles of variables in this regression model?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) response: weight, explanatory: volume, paperback cover

(b) response: weight, explanatory: volume, hardcover cover

(c) response: volume, explanatory: weight, cover type

(d) response: weight, explanatory: volume, cover type

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover:pb$$

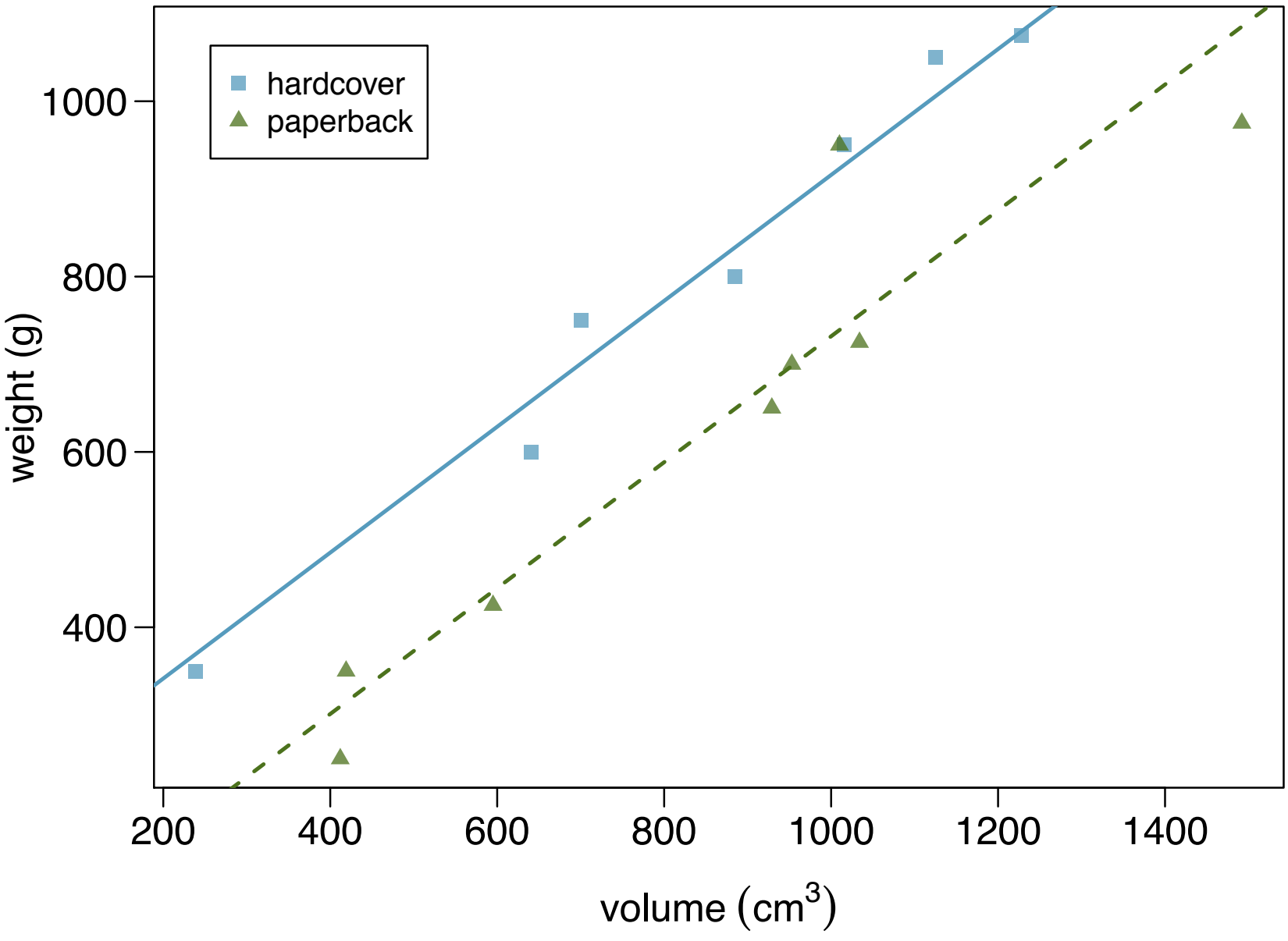
1. For *hardcover* books: plug in *0* for `cover`

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 0 \\ &= 197.96 + 0.72\ volume \end{aligned}$$

2. For *paperback* books: plug in *1* for `cover`

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 1 \\ &= 13.91 + 0.72\ volume \end{aligned}$$

Visualising the linear model



Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.
- *Slope of cover:* All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books.
- *Intercept:* Hardcover books with no volume are expected on average to weigh 198 grams.
  - Obviously, the intercept does not make sense in context. It only serves to adjust the height of the line.

Prediction

Which of the following is the correct calculation for the predicted weight of a paperback book that is 600 cm^3?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

(a) 197.96 + 0.72 * 600 - 184.05 * 1 $=$ *445.91 grams*

(b) 184.05 + 0.72 * 600 - 197.96 * 1

(c) 197.96 + 0.72 * 600 - 184.05 * 0

(d) 197.96 + 0.72 * 1 - 184.05 * 600

Prediction

the predicted weight of a paperback book that is 600 cm^3

```
mul_lm=lm(weight~volume+cover, data=allbacks)

my_m_prediction=function(x, my_model){
  x=c(1,x)
  pr=sum(as.vector(x*my_model$coefficients))
  return(as.numeric(pr))
}

x=c(600,1) #values entered in the order
           #as they appear in the model
my_m_prediction(x, my_m_prediction)
```

Prediction ctnd.

the predicted weight of a paperback book that is 600 cm^3

```
volume=c(800,950,1050)
cover=c('hb','pb','hb')
newdata=data.frame(volume, cover)
predict(mul_lm, newdata)
```

Another example: Modeling kid's test scores

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are from a survey of adult American women and their children - a subsample from the National Longitudinal Survey of Youth.

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
⋮					
5	115	yes	92.75	yes	27
6	98	no	107.90	no	18
⋮					
434	70	yes	91.25	yes	25

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. (2007) Cambridge University Press.

Interpreting the slope

What is the correct interpretation of the slope for mom's IQ?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

All else held constant, kids with mothers whose IQs are one point higher tend to score on average 0.56 points higher.

Interpreting the slope

What is the correct interpretation of the intercept?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

Kids whose moms haven't gone to HS, did not work during the first three years of the kid's life, have an IQ of 0 and are 0 yrs old are expected on average to score 19.59. Obviously, the intercept does not make any sense in context.

Another look at R^2

R^2 can be calculated as:

1. square the correlation coefficient of x and y (that is R) (how we have been calculating it)
2. based on definition:

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

Another look at R^2

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

```
aov(lm_pov_grad)

## Call:
##    aov(formula = lm_pov_grad)
##
## Terms:
##                 Graduates Residuals
## Sum of Squares   267.8807  212.3667
## Deg. of Freedom         1        49
##
## Residual standard error: 2.08183
## Estimated effects may be unbalanced
```

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{267.8807}{267.8807 + 212.3667} = 0.5578 \checkmark$$

Why bother?

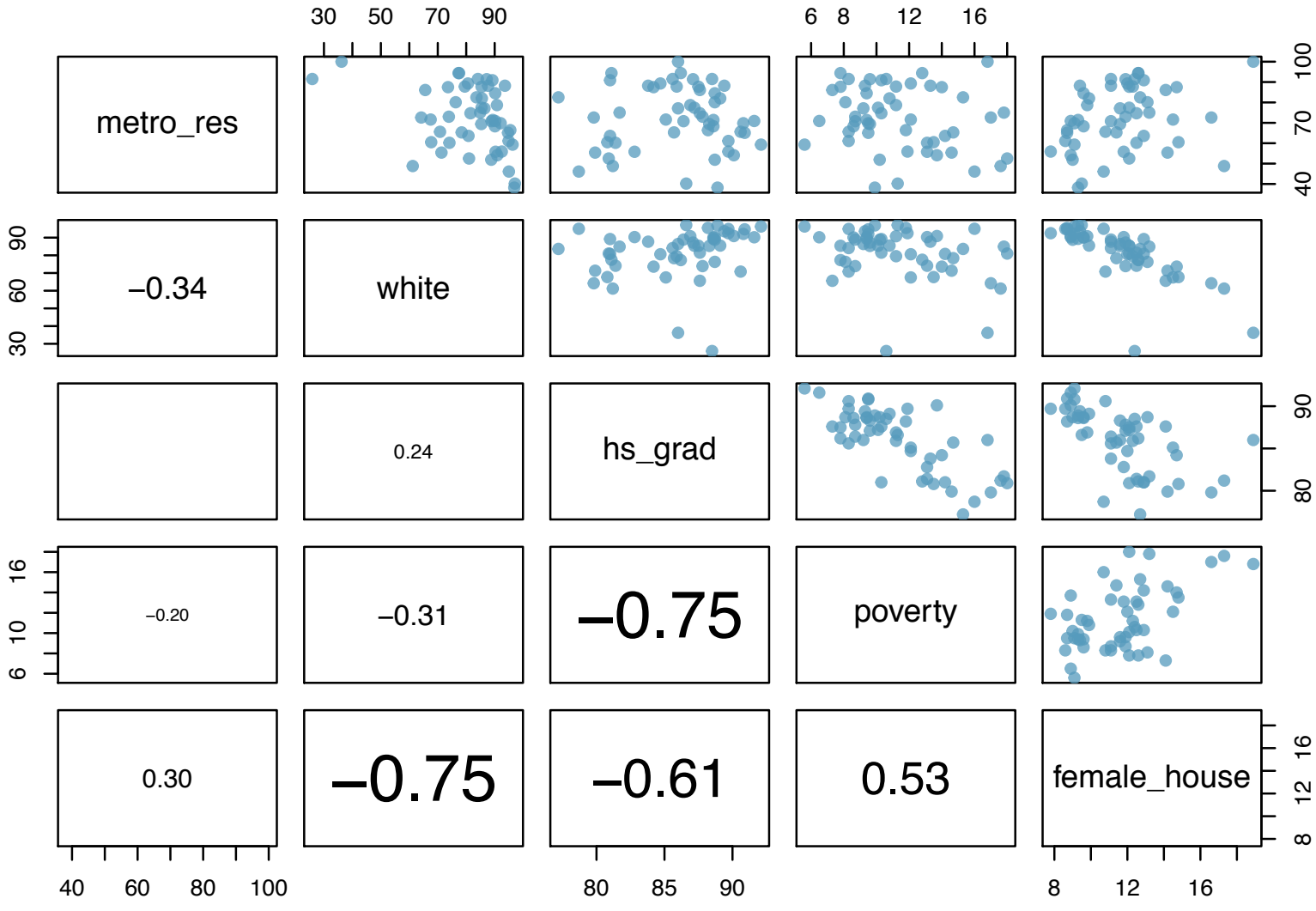
Why bother with another approach for calculating R^2 when we had a perfectly good way to calculate it as the correlation coefficient squared?

- For single-predictor linear regression, having multiple 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

```
names(poverty) = c("state", "metro_res", "white", "hs_grad", "po
pov_slr = lm(poverty ~ female_house+white, data = poverty)
summary(pov_slr)

##
## Call:
## lm(formula = poverty ~ female_house + white, data = poverty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5245 -1.8526 -0.0381  1.3770  6.2689
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.57894    5.78491  -0.446 0.657743
## female_house  0.88689    0.24191   3.666 0.000615 ***
## white         0.04418    0.04101   1.077 0.286755
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
```

Does adding the variable `white` to the model add valuable information that wasn't provided by `female_house`?



Collinearity between explanatory variables (cont.)

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* complicates model estimation.

  Remember: Predictors are also called explanatory or independent variables. Ideally, they would be independent of each other.
- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.
- While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to prevent correlation among predictors.

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.

Model Selection

Backward Elimination

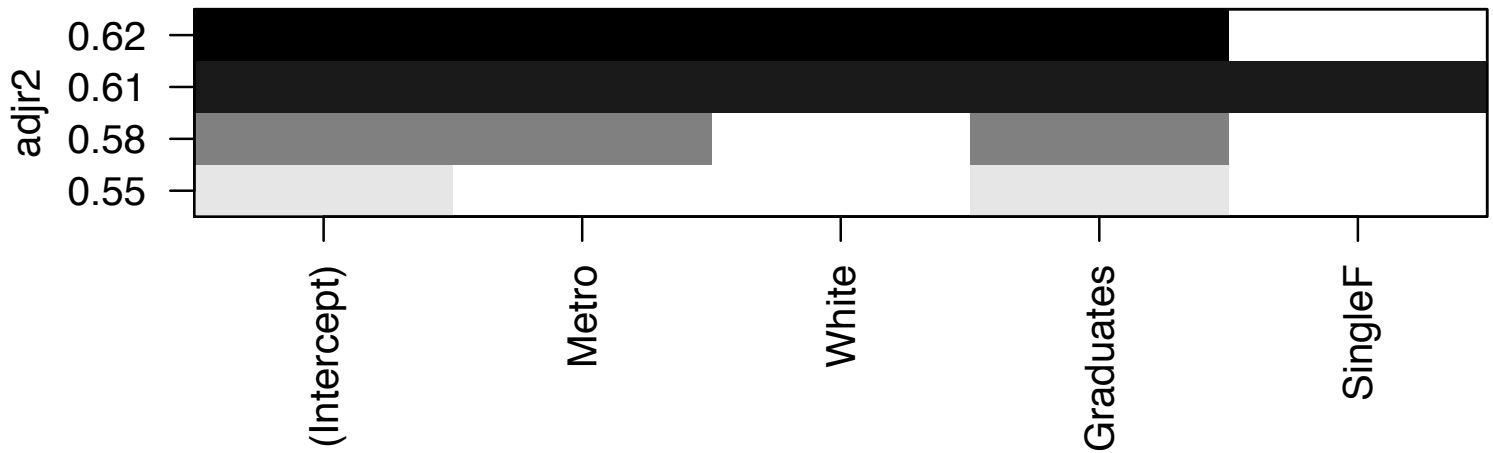
1. R^2_{adj} approach:
   - Start with the full model
   - Drop one variable at a time and record R^2_{adj} of each smaller model
   - Pick the model with the highest increase in R^2_{adj}
   - Repeat until none of the models yield an increase in R^2_{adj}
2. p-value approach:
   - Start with the full model
   - Drop the variable with the highest p-value and refit a smaller model
   - Repeat until all variables left in the model are significant

Model Selection

```
poverty=read.csv('../data_poverty.txt', sep='\t')
names(poverty)<-c('State', 'Metro', 'White', 'Graduates', 'Poverty','SingleF')

my_lm=lm(Poverty~Metro+White+Graduates+SingleF, data=poverty)

sum=summary(my_lm)
#read adj.r.squared
sum$adj.r.squared
slm1=lm(Poverty~White+Graduates+SingleF, data=poverty) # -Metro
slm2=lm(Poverty~Metro+Graduates+SingleF, data=poverty) # -White
slm3=lm(Poverty~Metro+White+SingleF, data=poverty)     # -Graduates
slm4=lm(Poverty~Metro+White+Graduates, data=poverty)   # -SingleF

summary(slm1)
summary(slm2)
summary(slm3)
summary(slm4)

sslm1=lm(Poverty~White+Graduates, data=poverty)  #slm4 - Metro
sslm2=lm(Poverty~Metro+Graduates, data=poverty)  #slm4 - White
sslm3=lm(Poverty~Metro+White, data=poverty)   # slm4 - Graduate
summary(sslm1)
summary(sslm2)
summary(sslm3)
```

Model Selection

```
#install.packages('leaps')
require(leaps)
poverty=read.csv('../data_poverty.txt', sep='\t')
names(poverty)<-c('State', 'Metro', 'White', 'Graduates', 'Poverty','SingleF')

lps=regsubsets(Poverty~Metro+White+Graduates+SingleF, data=poverty, method='backward')
plot(lps, scale="adjr2" )
```




```
#
# sm=summary(lps); sm$adjr2
```

Model Selection

Forward-selection

1. R^2_{adj} approach:
   - Start with regressions of response vs. each explanatory variable
   - Pick the model with the highest R^2_{adj}
   - Add the remaining variables one at a time to the existing model, and once again pick the model with the highest R^2_{adj}
   - Repeat until the addition of any of the remanning variables does not result in a higher R^2_{adj}
2. p – *value* approach:
   - Start with regressions of response vs. each explanatory variable
   - Pick the variable with the lowest significant p-value
   - Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value
   - Repeat until any of the remaining variables does not have a significant p-value

In forward-selection the p-value approach isn't any simpler (you still need to fit a bunch of models), so there's almost no incentive to use it.