Unit 7: Multiple Linear Regression
Lecture 1: Introduction to MLR

Statistics 104

Mine Çetinkaya-Rundel

November 19, 2013

Announcements

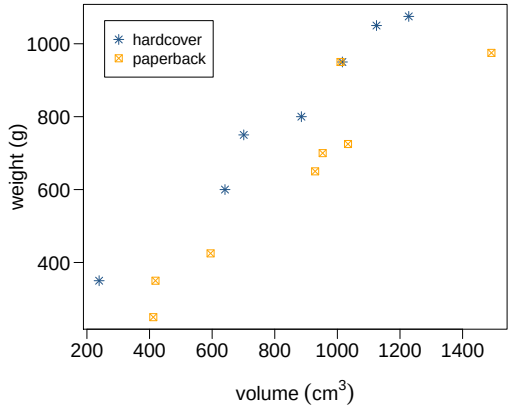
- OH this week:
  - Tue: 4:30-6:30pm (today)
  - Wed: 1:30 - *4pm*
  - *Fri: 10-11:30am & 12:30-1:30pm*
  - Can linger after class on Thursday to answer questions
- Project 2:
  - Sample posters at my office if you'd like to see
  - Schedule team meetings (in person + GoogleHangouts/Skype/Facetime/?)
  - Poster session will be in Link classrooms, I'll email with details as we get closer
- Project 1:
  - Make sure you pick up your paper and go through the comments
  - Questions on content: TAs or myself
  - Regreade request: myself only – write up your reasoning for why you believe you had the right answer/approach but points were taken off anyway, especially useful if you can highlight relevant sections. I will regrade the entire project.

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 hard cover 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

```
# load data
library(DAAG)
data(allbacks)

# fit model
book_mlr = lm(weight ~ volume + cover, data = allbacks)
summary(book_mlr)
```

```
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
```


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

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

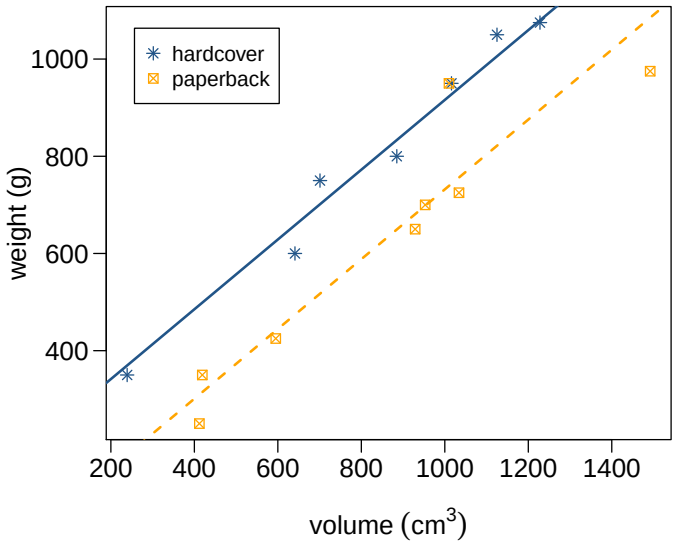
1. For *hardcover* books: plug in *0* for `cover`

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

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, for each 1 cm^3 increase in volume we would expect weight to increase on average by 0.72 grams.
- *Slope of cover:* All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books, on average.
- *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

Clicker question

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) $197.96 + 0.72 \times 600 - 184.05 \times 1$

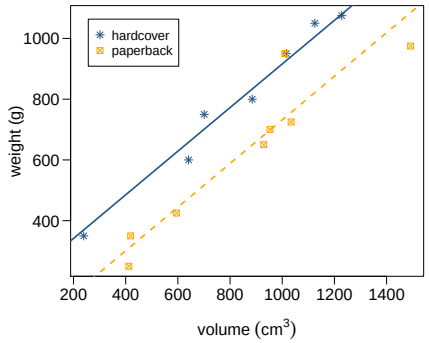
(b) $184.05 + 0.72 \times 600 - 197.96 \times 1$

(c) $197.96 + 0.72 \times 600 - 184.05 \times 0$

(d) $197.96 + 0.72 \times 1 - 184.05 \times 600$

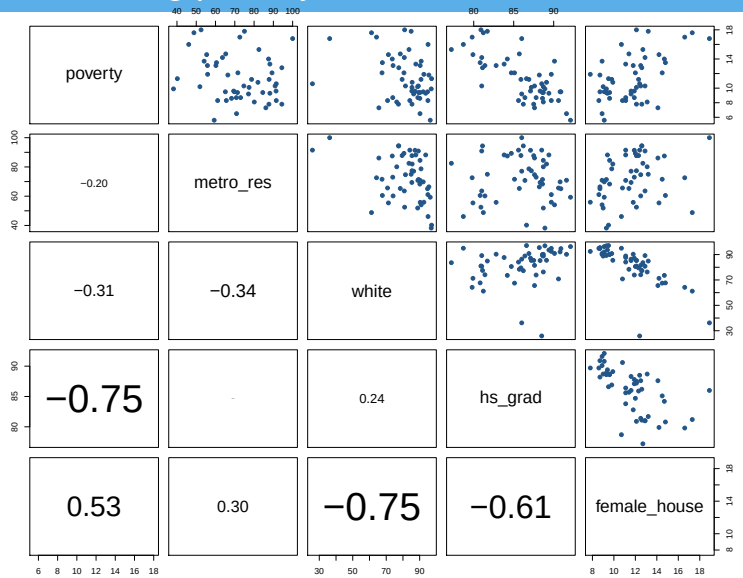
A note on "interaction" variables

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



This model assumes that hardcover and paperback books have the same slope for the relationship between their volume and weight. If this isn't reasonable, then we would include an "interaction" variable in the model (beyond the scope of this course).

Revisit: Modeling poverty

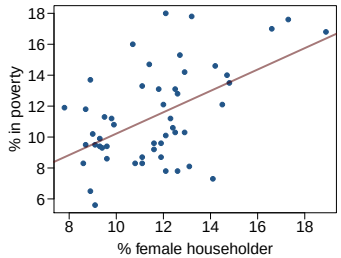


Predicting poverty using % female householder

```
# load data
poverty = read.csv("http://stat.duke.edu/~mc301/data/poverty.csv")

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

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



$R = 0.53$

$R^2 = 0.53^2 = 0.28$

Another look at R^2 - from last time

```
anova(pov_slr)
```

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			

$$\begin{aligned} \text{SS of } y\text{: } SS_{Tot} &= \sum (y - \bar{y})^2 = 480.25 \rightarrow \textit{total variability} \\ \text{SS of residuals: } SS_{Err} &= \sum e_i^2 = 347.68 \rightarrow \textit{unexplained variability} \\ \text{SS of regression: } SS_{Reg} &= SS_{Total} - SS_{Error} \rightarrow \textit{explained variability} \\ &= 480.25 - 347.68 = 132.57 \end{aligned}$$

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

Predicting poverty using % female hh + % white

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

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(pov_mlr)
```

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			

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

Adjusted R^2

Adjusted R^2

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

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

Application exercise: Adjusted R^2

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.

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			

(a) 0.26

(b) 0.29

(c) 0.32

(d) 0.71

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	

- 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 - properties

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

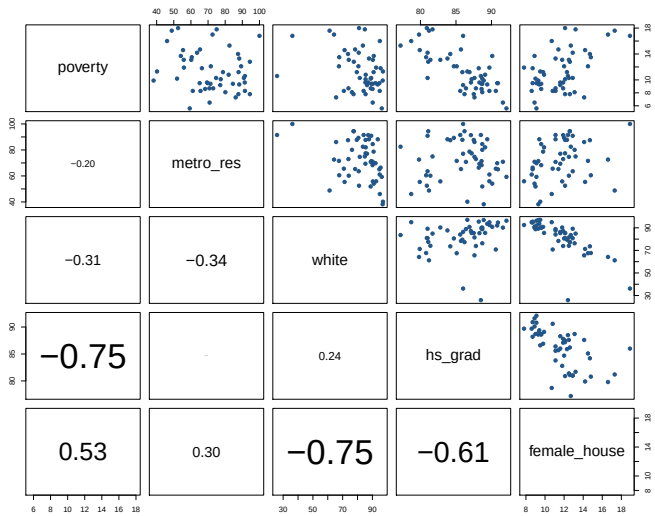
- Because k is never negative, R^2_{adj} will always be smaller than R^2.
- R^2_{adj} applies a penalty for the number of predictors included in the model.
- Therefore, we choose models with higher R^2_{adj} over others.

Clicker question

True or false: Adjusted R^2 tells us the percentage of variability in the response variable explained by the model.

(a) True
(b) False

We saw that adding the variable `white` to the model did not increase adjusted R^2, i.e. did not add any valuable information to the model. Why?



Collinearity between explanatory variables (cont.)

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

  Remember: Predictors are also called explanatory or independent variables, so they should 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.
- In addition, addition of collinear variables can result in biased estimates of the slope parameters.
- While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to control for correlated predictors.