Statistical Inference

Introduction to Linear Regression

Behnam Bahrak
Spring 2020

Multiple Linear Regression

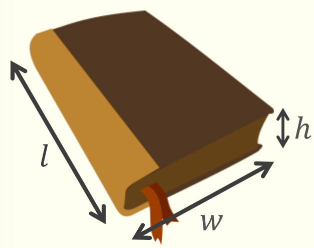
- Predicting birth weight of babies from a **variety** of variables:

	bwt	gestation	parity	age	height	weight	smoke
1	120	284	0	27	62	100	0
2	113	282	0	33	64	135	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1236	117	297	0	38	65	129	0

$$y \sim x_1 + x_2 + x_3 + x_4 + x_5 + x_6$$

Multiple Predictors

weights of books

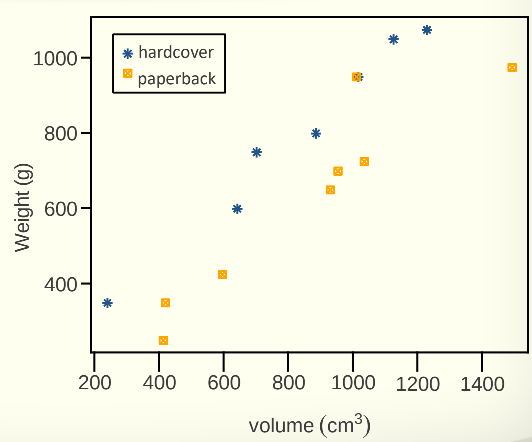


	weight (g)	volume (cm^3)	cover
1	800	885	hb
2	950	1016	hb
3	1050	1125	hb
4	350	239	hb
5	750	701	hb
6	600	641	hb
7	1075	1228	hb
8	250	412	pb
9	700	953	pb
10	650	929	pb
11	975	1492	pb
12	350	419	pb
13	950	1010	pb
14	425	595	pb
15	725	1034	pb

Hardcover vs. Paperback

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

Paperbacks generally weigh less than hardcover books.



Multiple Linear Regression in R

R

```
# 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
```


Example

	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

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

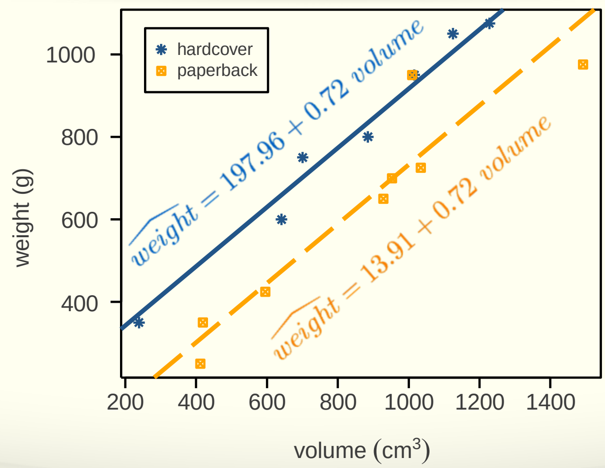
- For hardcover books: plug in **0** for cover:

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

- For paperback books: plug in **1** for cover:

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

Example



Interpreting the regression parameters: slope

	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

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

- Slope of **volume**: All else held constant, for each 1 cm^3 increase in volume the model predicts the books to be heavier on average by 0.72 grams.
- Slope of **cover**: All else held constant, the model predicts that paperback books weigh 184.05 grams lower than hardcover books, on average.

Interpreting the regression parameters: intercept

	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

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

Intercept: Hardcover books with no volume are expected on average to weigh 198 grams.

- Meaningless in context, serves to adjust the height of the line.

Prediction

- Predict the weight of a paperback book that is 600 cm^3 in volume.

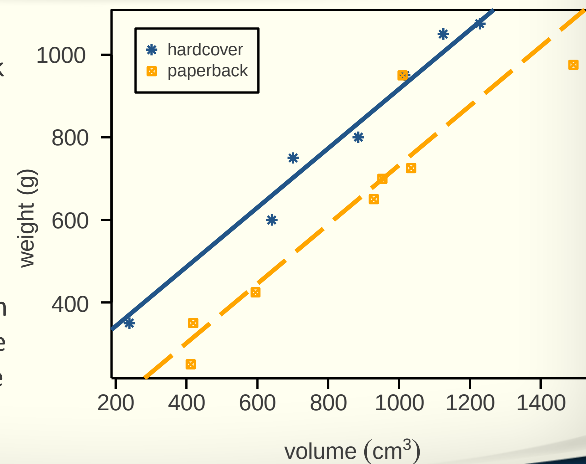
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

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

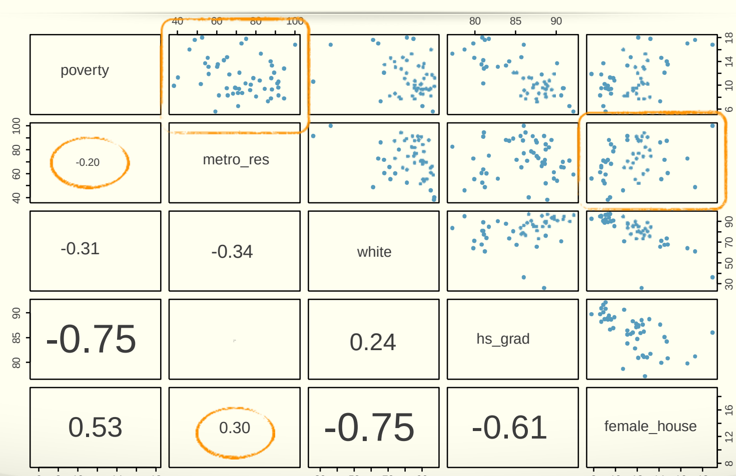
$$197.96 + 0.72 \times 600 - 184.05 \times 1 = 445.91\ grams$$

Interaction Variables

- Model assumes 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).



Adjusted R^2



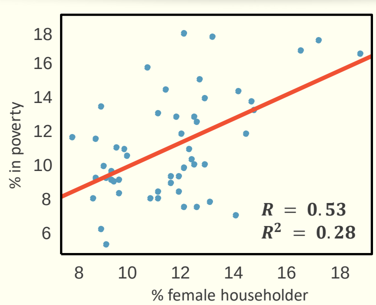
Adjusted R^2

```
R
# 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

Example

➢ 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