Advanced Statistics
DS2003 (BDS-4A)
Lecture 19

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

26 April, 2022

Previous Lecture

- Correlation coefficient calculation (r or R)
- Calculating slope using the correlation coefficient and the sample standard deviations
- Types of outliers in linear regression
  - *high leverage* points
  - *influential* points

Today

- Use of Python for linear regression analysis:
  - Example with real sample data
  - Example with randomly generated values
- Multiple Linear Regression

Python: Simple Linear Regression

- Google Colab Notebooks:
  - https://colab.research.google.com/drive/17TKVgRndR6NH5Dyn2aQdQ3kC8_nBPlsD?usp=sharing (Real Data Items)
  - https://colab.research.google.com/drive/1OQBmyQrHvLOkEctXLt76ntu_ghzMbylL?usp=sharing (Randomly Generated Data Points)

Introduction to multiple regression

Multiple regression

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

Poverty vs. region (east, west)

$$\widehat{poverty} = 11.17 + 0.38 \times west$$

- Explanatory variable: region, reference level: east
- Intercept: The estimated average poverty percentage in eastern states is 11.17%
  - This is the value we get if we plug in **0** for the explanatory variable
- Slope: The estimated average poverty percentage in western states is 0.38% higher than eastern states.
  - Then, the estimated average poverty percentage in western states is 11.17 + 0.38 = 11.55%.
  - This is the value we get if we plug in **1** for the explanatory variable

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) is the reference level?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

(a) northeast
(b) midwest
(c) west
(d) south
(e) cannot tell

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) is the reference level?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

(a) northeast
(b) midwest
(c) west
(d) south
(e) cannot tell

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) has the lowest poverty percentage?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

(a) northeast
(b) midwest
(c) west
(d) south
(e) cannot tell

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) has the lowest poverty percentage?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

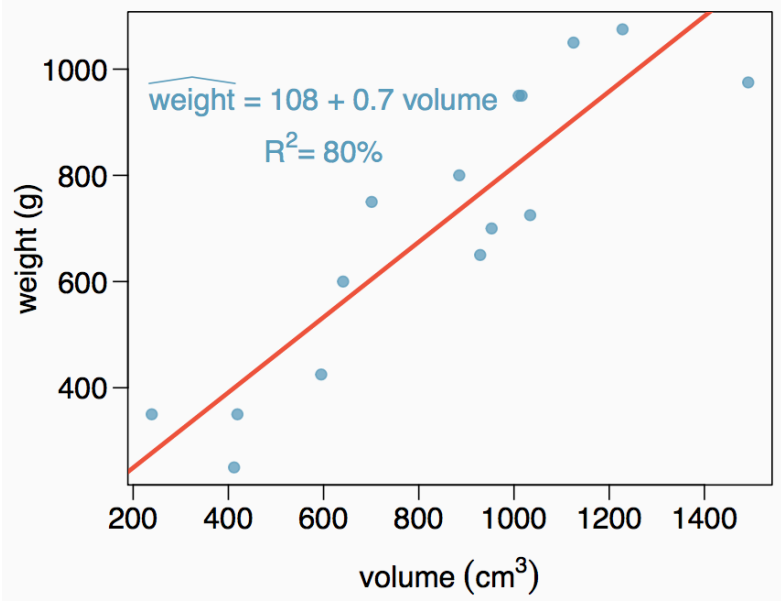
(a) northeast
(b) midwest
(c) west
(d) south
(e) cannot tell

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.)

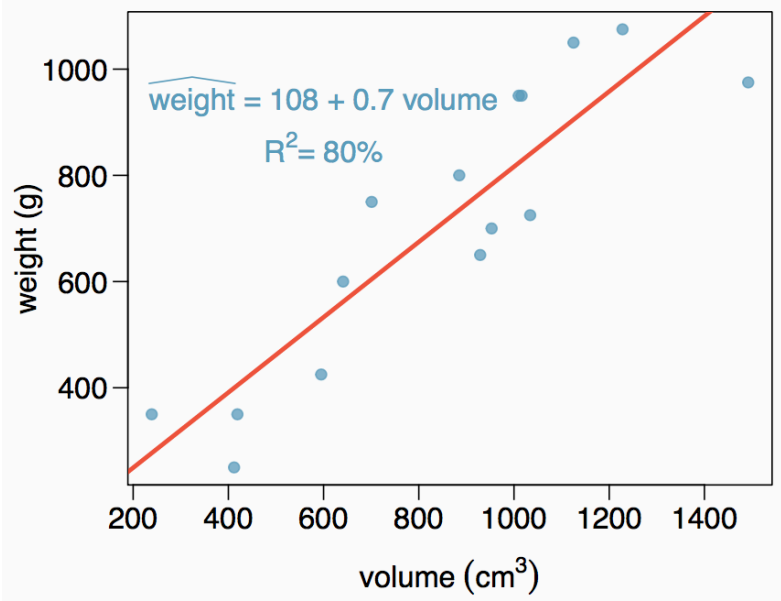
The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



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 7g 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.

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 7g 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

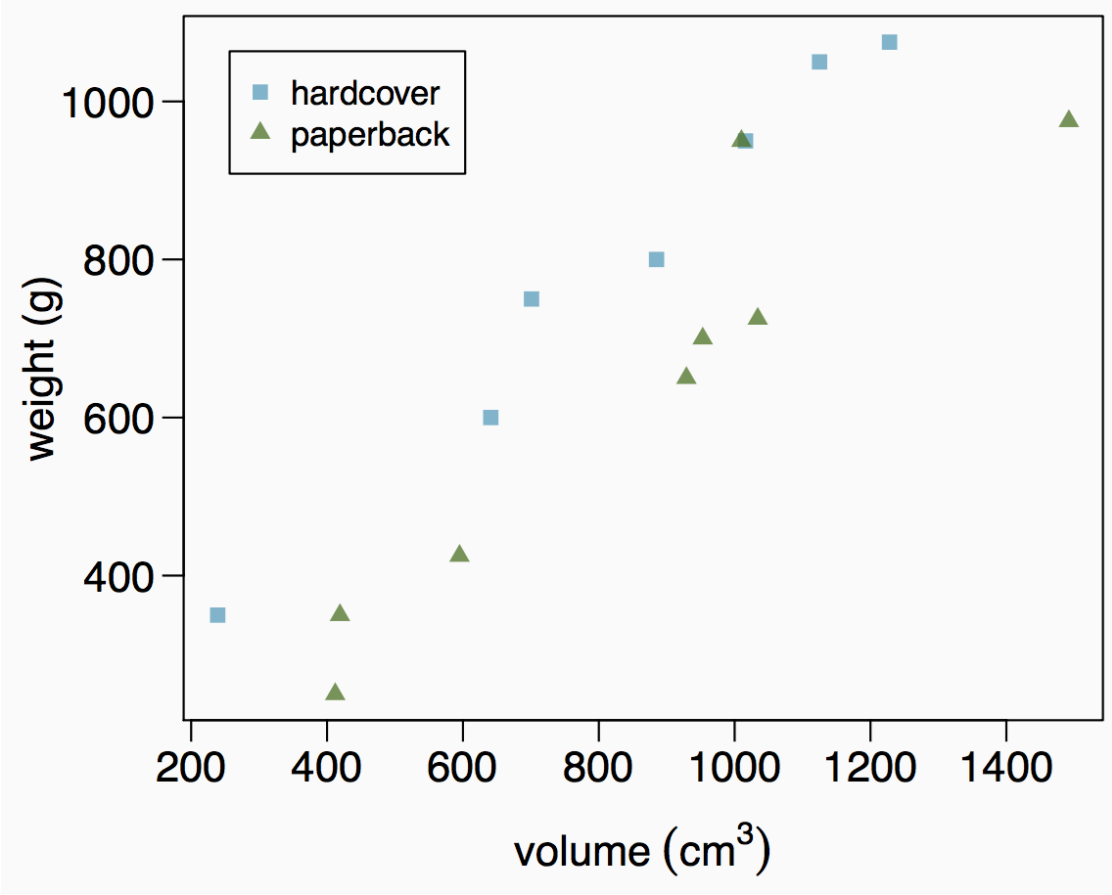
somewhat abbreviated output...

```
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

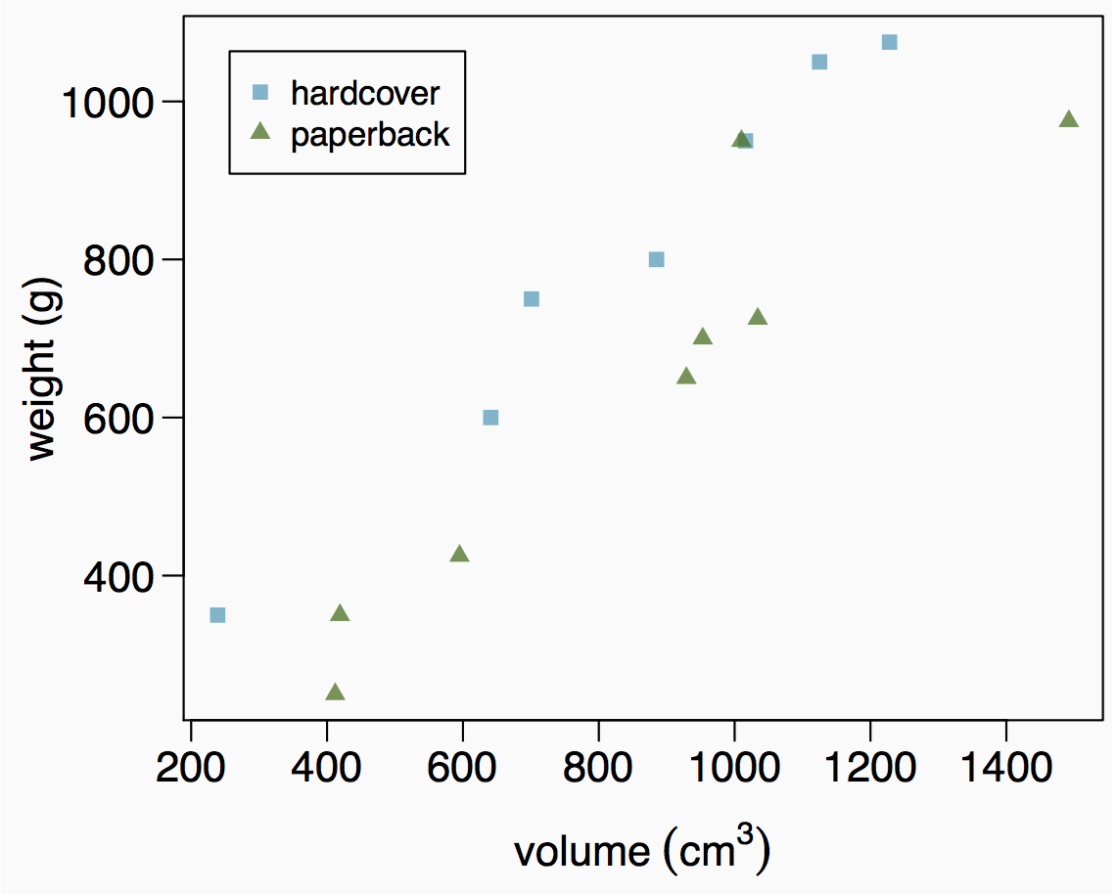
Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?



Weights of hardcover and paperback books

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

Paperbacks generally weigh less than hardcover books after controlling for the book's volume.



Modeling weights of books using volume and cover type

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

Based on the regression output below, which level of cover is the reference level? Note that pb: paperback.

	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. paperback

b. hardcover

Determining the reference level

Based on the regression output below, which level of cover is the reference level? Note that pb: paperback.

	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. paperback

b. *hardcover*

Sources

- openintro.org/os (Chapter 9, Section 9.1)

Google Colab Notebooks:

- https://colab.research.google.com/drive/17TKVgRndR6NH5Dyn2aQdQ3kC8_nBPlsD?usp=sharing (Real Data Items)
- https://colab.research.google.com/drive/1OQBmyQrHvLOkEctXLt76ntu_ghzMbylL?usp=sharing (Randomly Generated Data Points)