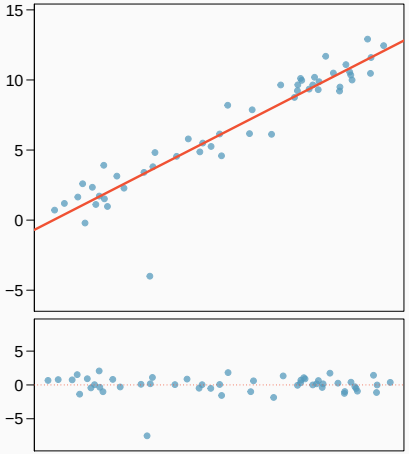
Announcements

- Whatever happened to homework 4?
  - I forgot about it and there wouldn't be a ton of new content on it so we'll just roll those topics into next week's
- Lab 4 write-up still due on Monday though!
- Have read 6.3 by Monday

Warm up!

Which of the below best describes the outlier?

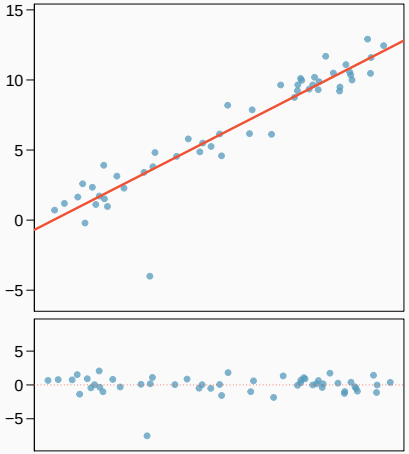
A) influential
B) high leverage
C) normal outlier
D) there are no outliers



Warm up!

Which of the below best describes the outlier?

A) influential
B) high leverage
C) normal outlier
D) there are no outliers



Intro to Multiple Regression

- Simple linear regression: two variables
  - y and x
- Multiple linear regression: multiple variables
  - y and $x_1, x_2, \ldots$
- Models will have the form

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

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

From: Maindonald, J.H. and Braun, W.J. (2nd ed., 2007) “Data Analysis and Graphics Using R”

Understanding Check

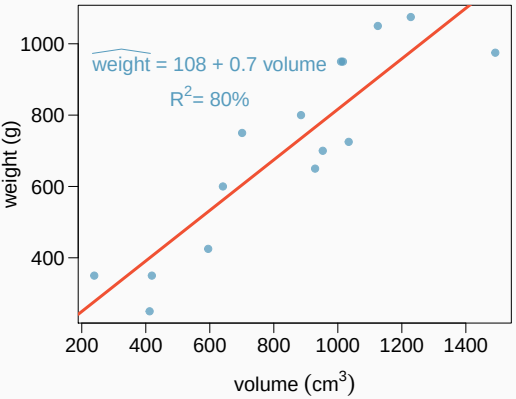
The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?

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.

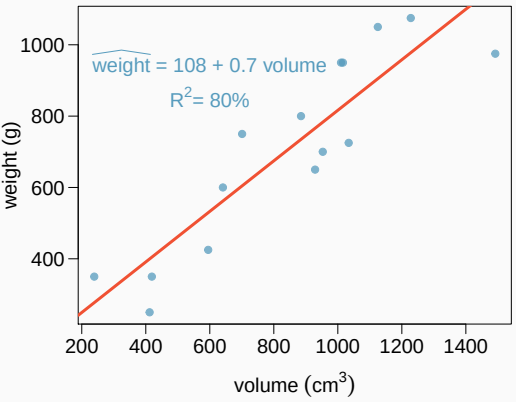
C) The correlation between weight and volume is $R = 0.80^2 = 0.64$.

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



Understanding Check

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.

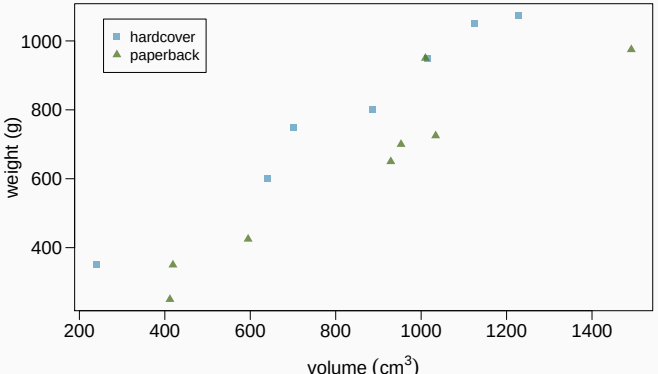
Modeling book weights using volume

```
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
```

Hardcover vs Paperback

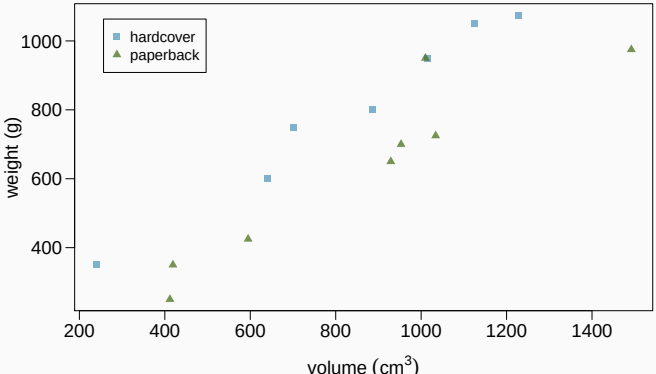
Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?



Hardcover vs Paperback

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

- Paperbacks generally weigh less that hardcover books after accounting for a 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
```

Understanding Reference Levels

- Previously we saw how we can use indicator variables to describe a 2 state categorical variable numerically.
- The variable we set to 0 is the ***reference***.
  - A reference will never have an estimated slope
- Here, the fact that we see `cover:pb` tells us that the reference is a hardcover
  - Hardcover = 0
  - Paperback = 1

Linear model

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

Linear model

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

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

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05 \times 0$$

Linear model

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

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

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

Linear model

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

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`

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05 \times 1$$

Linear model

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

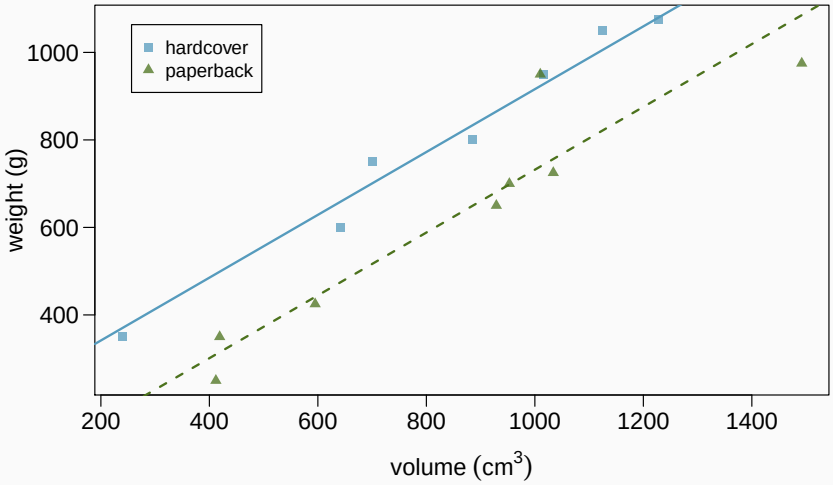
1. For ***hardcover*** books: plug in 0 for `cover`

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

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

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

Visualizing the Model



Interpretation of 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

Interpretation of 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.

Interpretation of 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.

Interpretation of 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.

Interpretation of 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.

A balancing act

- It can be tempting to think that adding more variables into your analysis will improve your predictive powers
  - Especially since your reported R^2 value will increase for each additional predictor you add
- Predictor variables that are correlated with one another though complicates the model estimation
- Try not to add predictors which are associated to one other to the model.
  - Will make your model seem better without actually being better
- Adding too many predictors can result in ***overfitting*** where you are fitting the noise in the data instead of the correlations in the data

Comparing R^2's

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

Comparing R^2's

	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.

Comparing R^2's

	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.
- Can use to compare to see if your adding another predictor really made a difference in your model's predictive powers

Adjusted R^2

- Previously, we had:

$$R^2 = \frac{\sigma_y^2 - \sigma_{res}^2}{\sigma_y^2}$$

- Adjusted R^2 looks like:

$$R^2_{adj} = \frac{\sigma_y^2 - \sigma_{res}^2}{\sigma_y^2} \times \frac{n-1}{n-k-1}$$

  where n is the number of observations and k the number of predictor variables used

- Offsets the fact that using more predictor variables inherently increases the traditional R^2 value.

Understanding Check

We looked at the book weight analysis using just volume and then volume and cover. Given the reported R^2 results, is it safe to say that our second model is doing a better job predicting the weights?

A) Yes

B) No

```
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
```

```
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
```