Models with multiple predictors

Data Science in a Box

datasciencebox.org

The linear model with multiple predictors

Data: Book weight and volume

The `allbacks` data frame gives measurements on the volume and weight of 15 books, some of which are paperback and some of which are hardback

- Volume - cubic centimetres
- Area - square centimetres
- Weight - grams

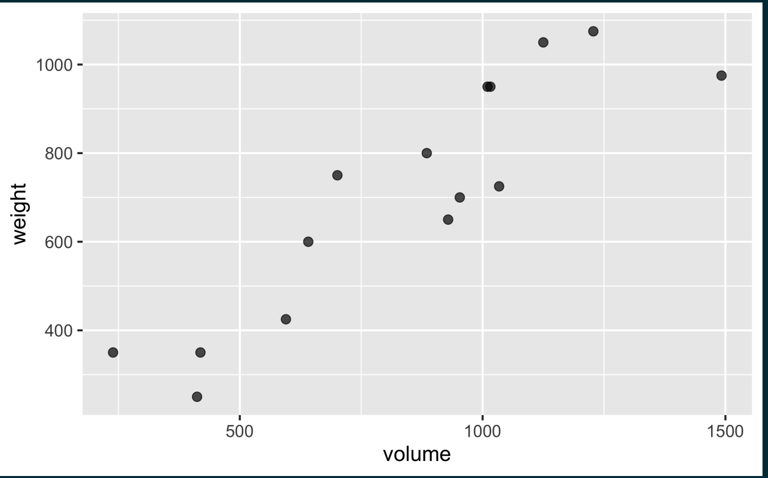
```
## # A tibble: 15 × 4
##    volume  area weight cover
##     <dbl> <dbl>  <dbl> <fct>
##  1    885   382    800 hb
##  2   1016   468    950 hb
##  3   1125   387   1050 hb
##  4    239   371    350 hb
##  5    701   371    750 hb
##  6    641   367    600 hb
##  7   1228   396   1075 hb
##  8    412     0    250 pb
##  9    953     0    700 pb
## 10    929     0    650 pb
## 11   1492     0    975 pb
## 12    419     0    350 pb
## 13   1010     0    950 pb
## 14    595     0    425 pb
## 15   1034     0    725 pb
```

These books are from the bookshelf of J. H. Maindonald at Australian National University.

Book weight vs. volume

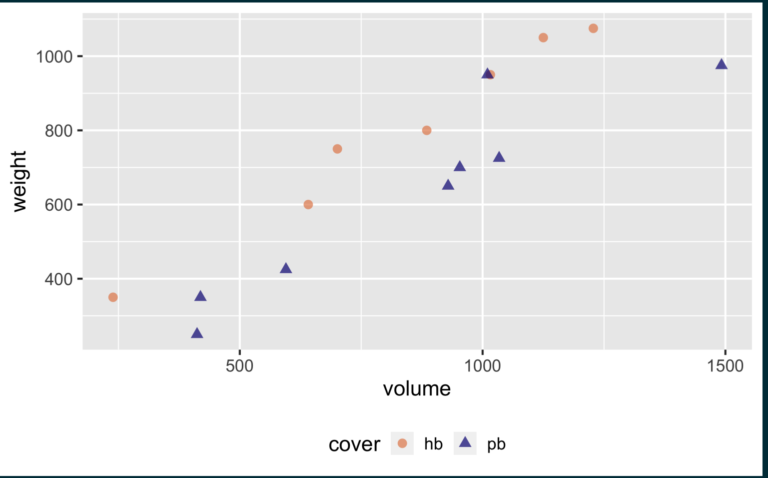
```
linear_reg() %>%
  set_engine("lm") %>%
  fit(weight ~ volume, data = allbacks) %>
  tidy()
```

```
## # A tibble: 2 × 5
##   term        estimate std.error statistic    p.value
##   <chr>          <dbl>     <dbl>     <dbl>      <dbl>
## 1 (Intercept)  108.      88.4         1.22 0.245
## 2 volume         0.709    0.0975      7.27 0.00000626
```



Book weight vs. volume and cover

```
linear_reg() %>%
  set_engine("lm") %>%
  fit(weight ~ volume + cover, data = allb
  tidy()
```




```
## # A tibble: 3 × 5
##   term        estimate std.error statistic      p.value
##   <chr>          <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)  198.       59.2        3.34 0.00584
## 2 volume         0.718     0.0615    11.7  0.0000000660
## 3 coverpb     -184.       40.5       -4.55 0.000672
```

Interpretation of estimates

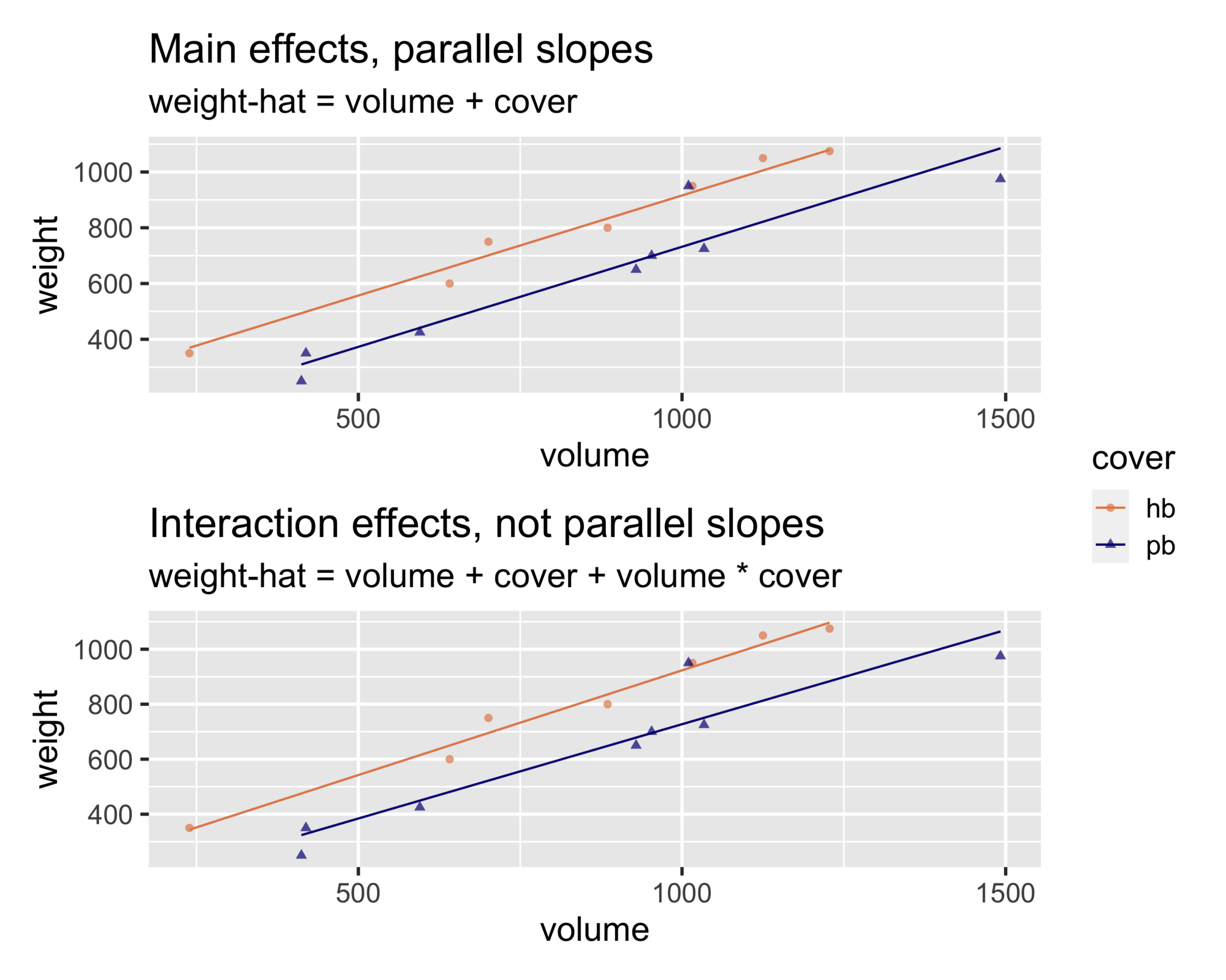
```
## # A tibble: 3 × 5
##   term        estimate std.error statistic      p.value
##   <chr>          <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)  198.      59.2         3.34 0.00584
## 2 volume         0.718    0.0615     11.7  0.0000000660
## 3 coverpb     -184.      40.5        -4.55 0.000672
```

- **Slope - volume:** *All else held constant*, for each additional cubic centimetre books are larger in volume, we would expect the weight to be higher, on average, by 0.718 grams.
- **Slope - cover:** *All else held constant*, paperback books are weigh, on average, by 184 grams less than hardcover books.
- **Intercept:** Hardcover books with 0 volume are expected to weigh 198 grams, on average. (Doesn't make sense in context.)

Main vs. interaction effects

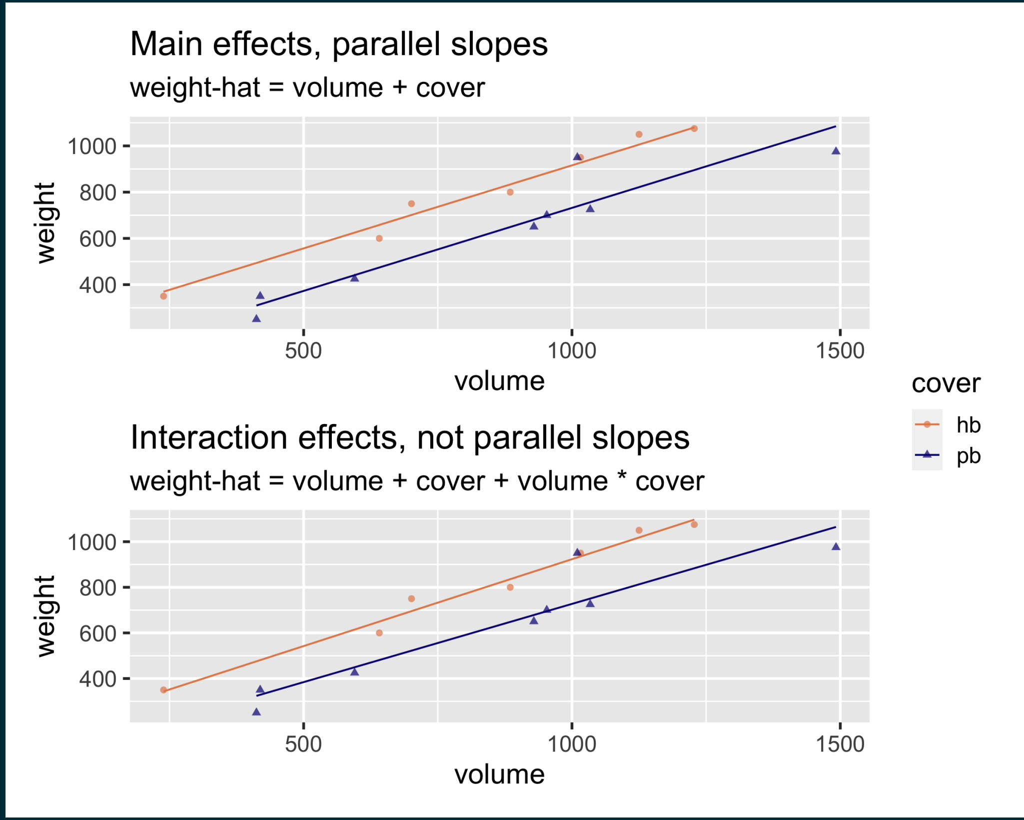
Suppose we want to predict weight of books from their volume and cover type (hardback vs. paperback). Do you think a model with main effects or interaction effects is more appropriate? Explain your reasoning.

Hint: Main effects would mean rate at which weight changes as volume increases would be the same for hardback and paperback books and interaction effects would mean the rate at which weight changes as volume increases would be different for hardback and paperback books.


Main effects, parallel slopes
weight-hat = volume + cover
weight
1000
800
600
400
500
1000
1500
volume
Interaction effects, not parallel slopes
weight-hat = volume + cover + volume * cover
weight
1000
800
600
400
500
1000
1500
volume
cover
hb
pb

In pursuit of Occam's razor

- Occam's Razor states that among competing hypotheses that predict equally well, the one with the fewest assumptions should be selected.
- Model selection follows this principle.
- We only want to add another variable to the model if the addition of that variable brings something valuable in terms of predictive power to the model.
- In other words, we prefer the simplest best model, i.e. **parsimonious** model.



Visually, which of the two models is preferable under Occam's razor?

R-squared

- R^2 is the percentage of variability in the response variable explained by the regression model.

```r
glance(book_main_fit)$r.squared
```

```
## [1] 0.9274776
```

```r
glance(book_int_fit)$r.squared
```

```
## [1] 0.9297137
```

- Clearly the model with interactions has a higher R^2.
- However using R^2 for model selection in models with multiple explanatory variables is not a good idea as R^2 increases when **any** variable is added to the model.

Adjusted R-squared

... a (more) objective measure for model selection

- Adjusted R^2 doesn't increase if the new variable does not provide any new informaton or is completely unrelated, as it applies a penalty for number of variables included in the model.
- This makes adjusted R^2 a preferable metric for model selection in multiple regression models.

Comparing models

```r
glance(book_main_fit)$r.squared
```

```
## [1] 0.9274776
```

```r
glance(book_int_fit)$r.squared
```

```
## [1] 0.9297137
```

```r
glance(book_main_fit)$adj.r.squared
```

```
## [1] 0.9153905
```

```r
glance(book_int_fit)$adj.r.squared
```

```
## [1] 0.9105447
```

- Is R-sq higher for int model?

```r
glance(book_int_fit)$r.squared > glance(book_main_fit)$r.squared
```

```
## [1] TRUE
```

- Is R-sq adj. higher for int model?

```r
glance(book_int_fit)$adj.r.squared > glance(book_main_fit)$adj.r.squared
```

```
## [1] FALSE
```