***COURSERA: STATS W/ R SPECIALIZATION***

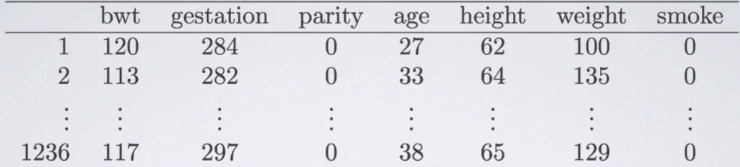
***COURSE 3 - Linear Regression and Modeling***

**WEEK 3 – Multiple Linear Regression**

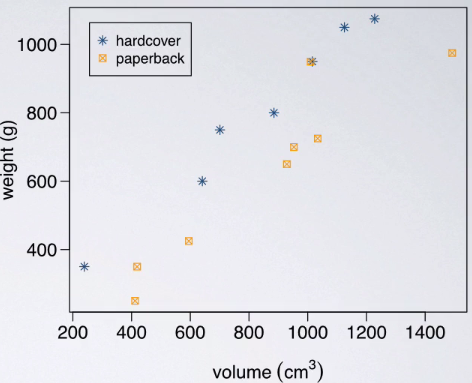
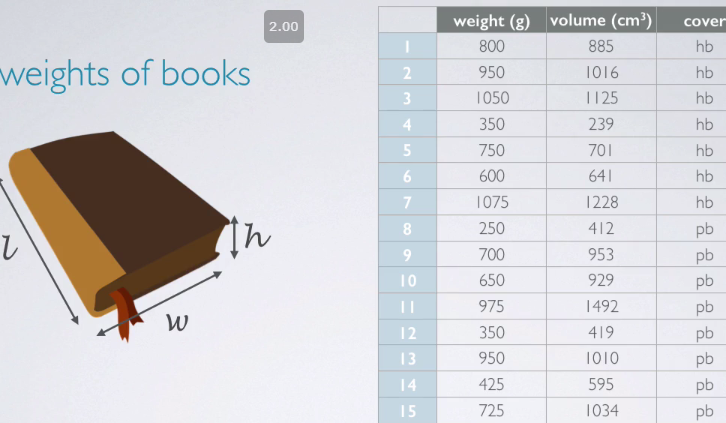
***5.3.1 Regression with multiple predictors***

**Multiple Predictors**

* Set of 1236 observations w/ data from birth weight of babies + a variety of variables on the baby, the birth, or the mother.



* Want to predict birth weight of babies from a **set** of these variables = multiple linear regression
* Explanatory variables in set can be numerical or categorical.
* Build a model predicting weight of a book using its volume + cover type



* It appears that paperbacks orange squares (PB) generally weigh less than HC books.
* Relationship = similar for 2 types of books (this is expected) = As volume increases, so does weight
* Also noting that PB generally weigh less than HC books.
* Next, fit the model w/ data from DAAG library 🡪 data set = allbacks,

> model1 <- lm(weight ~ volume + cover, data = allbacks)

> summary(model1)

Call:

lm(formula = weight ~ volume + cover, data = allbacks)

Residuals:

Min 1Q Median 3Q Max

-110.10 -32.32 -16.10 28.93 210.95

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 \*\*\*

coverpb -184.04727 40.49420 -4.545 0.000672 \*\*\*

Residual standard error: 78.2 on 12 degrees of freedom

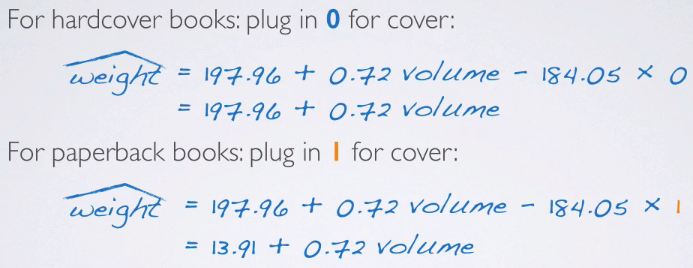
Multiple R-squared: 0.9275, Adjusted R-squared: 0.9154

F-statistic: 76.73 on 2 and 12 DF, p-value: 1.455e-07

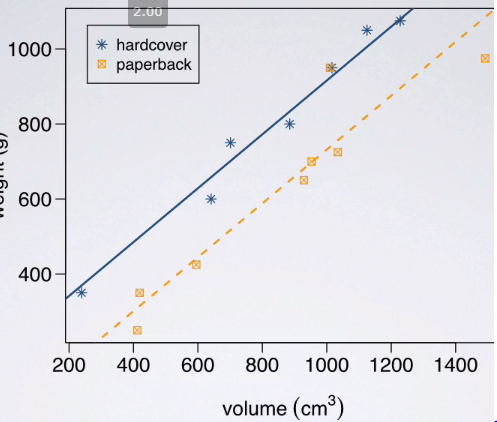
* See estimate for the intercept = 197.96 + estimates for slope of volume + of cover variables.
* Also note only 1 level of **cover** is in the output = the **NON-reference level** 🡪 hardcover books must be the **reference level**.
* Multiple R2 = 92.75% 🡺 92.75% of variability of book weights is explained by volume + cover type
* Pretty high R2 but we’d expect that b/c *what else could the weight of a book depend on*?
* Perhaps paper type + something like that makes up the remaining 7% of unexplained variability
* Using these estimates, we can easily write the linear model.



* Remember PB = non-reference level, meaning HC books = reference level
* **For reference levels, always plug in 0 in our linear model**.
* To simplify this linear model to see what it’d look like only for HC books, plug in a 0 for cover
* For a PB, plug in 1 for cover



* So same slope, different intercepts.



* Multiple linear regression has allowed us to fit these separate parallel lines for the 2 types of books, as opposed to imposing only 1 single line describing all books
* That line would be somewhere in-between these 2 in order to be able minimize the residuals, but it really wouldn't do a good job explaining either the HC or PB books.
* Slope estimate for volume = 0.72 🡺 **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 for cover variable = -184.05 + remember since HC = reference level + PB = non-reference level.
* In this case, we can think of this value as when going from HC to PB, there is an expected decrease in weight of 184.05 g
* In context of the data (keeping in mind this is an *observational* study), interpret this as *all else held constant*, the model predicts PB books weigh 184.05 grams lower than HC books, on average.
* The intercept = 197.96 = predicted value of the response variable when x = 0.
* In this case, we have 2 x's, so when volume = 0 + when cover = 0 (i.e. reference level for cover is being considered), weight = 198 g.
* In other words, HC books w/ no volume are expected on average to weigh 198 grams.
* Obviously, this is completely meaningless in context b/c what is a book cannot NOT have a volume
* Intercept is still useful b/c it serves to adjust the height of the regression line.
* Predict weight of a PB book 600 cm^3 in volume.

> non\_reference\_model(600)

[1] 444.6856

* This model predicts a PB book 600 cm^3 to weigh 444.68 grams.
* Note this model assumes that HC + PB books have the *same slope* for the relationship between volume + weight.
* This is probably not an unreasonable assumption for books, but this may not be the case.
* If trying to predict calories burned from # of minutes of exercise + a categorical variable sex, the relationship between these may not be the same for males + females
* In that case, it wouldn't really make sense to model this using 2 parallel lines.
* If this assumption of parallel lines/idea of the same slope for the 2 levels of the categorical variable is NOT reasonable, introduce an **interaction variable** in the model.

**Adjusted R2**