Lecture 8: Multiple Regression

Heidi Perry, PhD

Hack University

heidiperryphd@gmail.com

11/3/2016

Overview

1. Categorical explanatory variables
2. Introduction to multiple regression
   - Many variables in a model
   - Adjusted R^2
3. Model selection
4. Checking model conditions using graphs

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) *northeast*

(c) midwest

(d) west

(e) south

(f) 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	$\Pr(>\lvert t\rvert)$
(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) *northeast*

(c) midwest

(d) west

(e) south

(f) cannot tell

Multiple regression

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

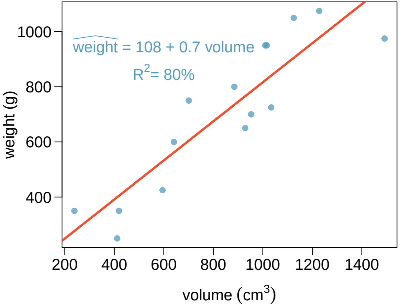
Weights of books

	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

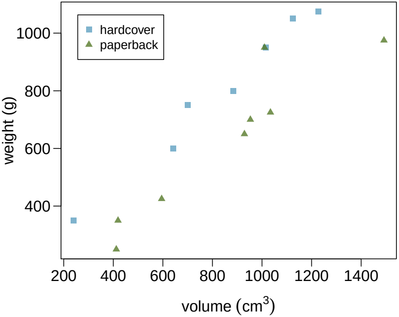
From: Maindonald, J.H. and Braun, W.J. (2nd ed., 2007) "Data Analysis and Graphics Using R"

Weights of books

The scatterplot shows the relationship between weights and volumes of books as well as the regression output.



Distinguishing between hardcover and 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 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
```

Linear model

	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

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

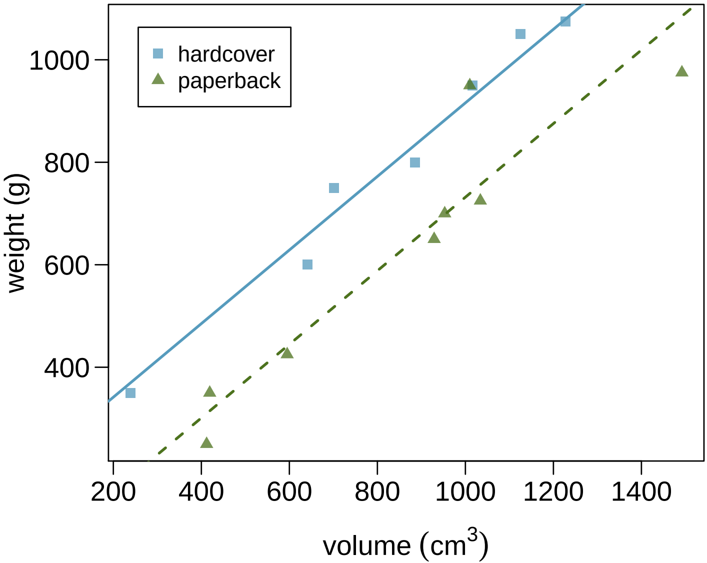
1. For *hardcover* books: plug in **0** for `cover`

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

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

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

Visualising the linear model

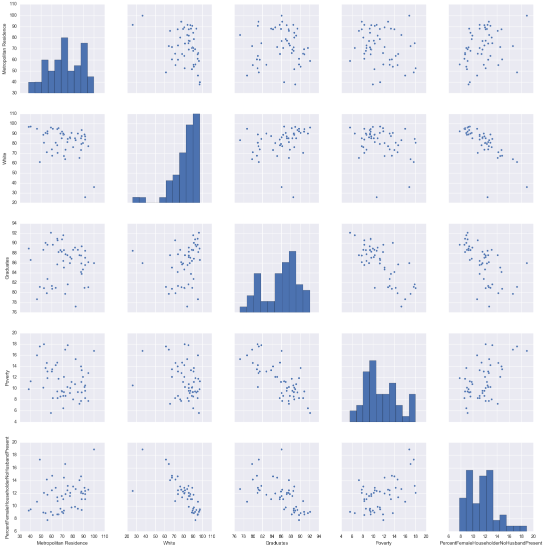


Interpretation of the regression coefficients

	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

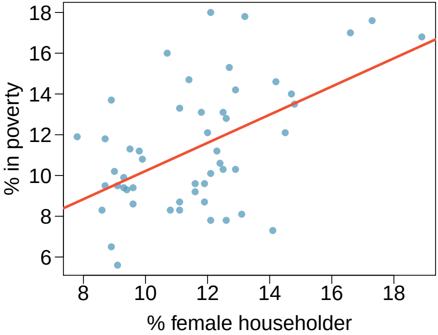
- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.
- *Slope of cover:* All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books.
- *Intercept:* Hardcover books with no volume 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.

Revisit: Modeling poverty



Predicting poverty using % female householder

	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



$R = 0.53$

$R^2 = 0.53^2 = 0.28$

Another look at R^2

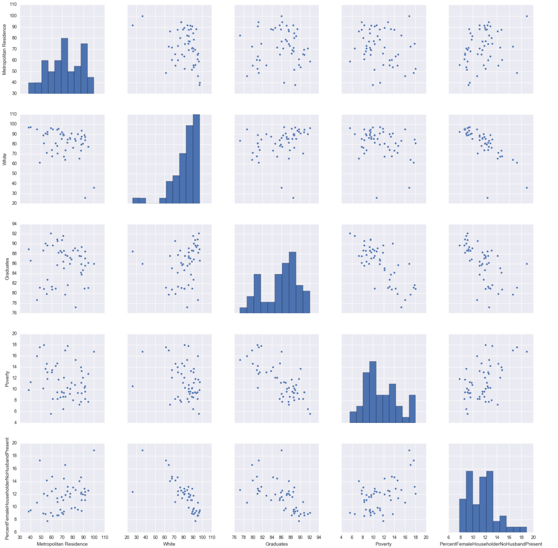
R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y.
2. square the correlation coefficient of y and $\hat{y}$
3. based on the following definition, using ANOVA:

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in } y}$$

- For single-predictor linear regression, having three ways to calculate the same value may seem like overkill.
- However, in multiple linear regression, we can't calculate R^2 as the square of the correlation between x and y because we have multiple xs.
- And next we'll learn another measure of explained variability, *adjusted R^2*, that requires the use of the third approach, ratio of explained and unexplained variability.

Does adding the variable `white` to the model add valuable information that wasn't provided by `female_house`?



Collinearity between explanatory variables

poverty vs. % female head of household

	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

poverty vs. % female head of household and % female hh

	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

Collinearity between explanatory variables (cont.)

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* complicates model estimation. Predictors are also called explanatory or independent variables. Ideally, they would be independent of each other.
- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.
- While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to prevent correlation among predictors.

R^2 vs. adjusted R^2

	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.

Adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right)$$

where n is the number of cases and p is the number of predictors in the model.

Adjusted R^2 applies a penalty for the number of predictors included in the model. When determining which predictors to include, choose the model with the highest R^2_{adj}.

Model selection strategies

Backward-elimination

1. R^2_{adj} approach:
    - Start with the full model
    - Drop one variable at a time and record R^2_{adj} of each smaller model
    - Pick the model with the highest increase in R^2_{adj}
    - Repeat until none of the models yield an increase in R^2_{adj}
2. p-value approach:
    - Start with the full model
    - Drop the variable with the highest p-value and refit a smaller model
    - Repeat until all variables left in the model are significant

Forward-selection

1. R^2_{adj} approach:
   - Start with regressions of response vs. each explanatory variable
   - Pick the model with the highest R^2_{adj}
   - Add the remaining variables one at a time to the existing model, and once again pick the model with the highest R^2_{adj}
   - Repeat until the addition of any of the remanning variables does not result in a higher R^2_{adj}
2. $p - value$ approach:
   - Start with regressions of response vs. each explanatory variable
   - Pick the variable with the lowest significant p-value
   - Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value
   - Repeat until any of the remaining variables does not have a significant p-value

In forward-selection the p-value approach isn't any simpler (you still need to fit a bunch of models), so there's almost no incentive to use it.

Modeling conditions

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

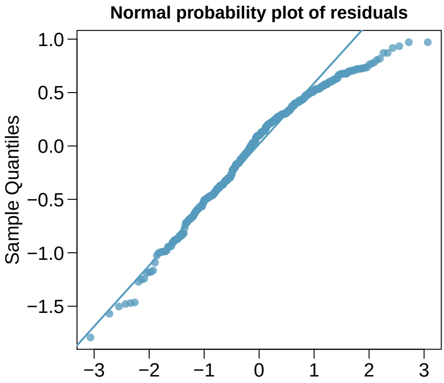
The model depends on the following conditions

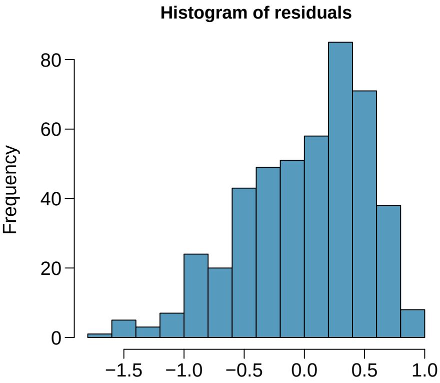
1. residuals are nearly normal (primary concern relates to residuals that are outliers)
2. residuals have constant variability
3. residuals are independent
4. each variable is linearly related to the outcome

We often use graphical methods to check the validity of these conditions, which we will go through in detail in the following slides.

(1) nearly normal residuals

normal probability plot and/or histogram of residuals:

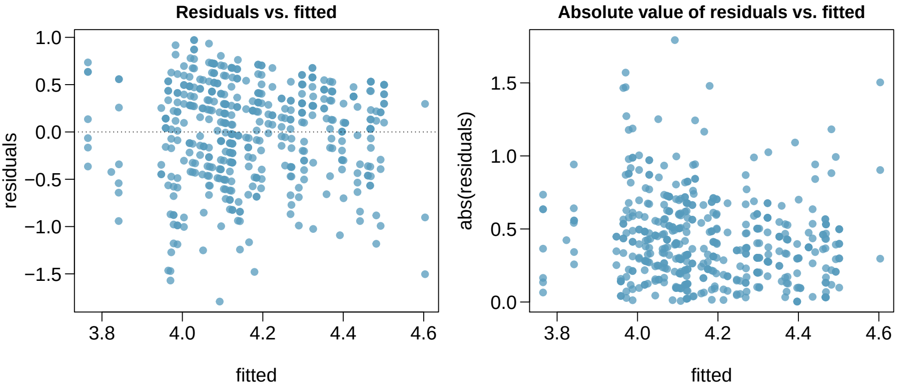




Does this condition appear to be satisfied?

(2) constant variability in residuals

scatterplot of residuals and/or absolute value of residuals vs. fitted (predicted):



Does this condition appear to be satisfied?

Checking constant variance - recap

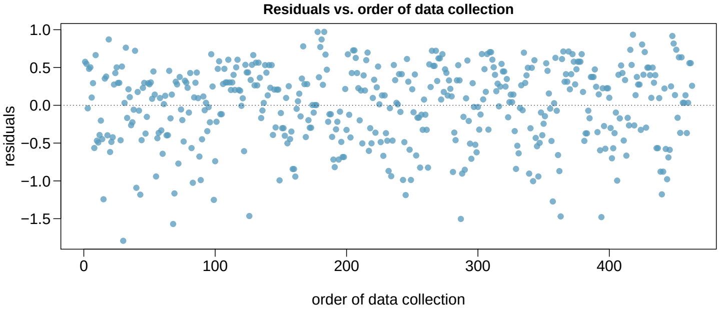
- When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.
- With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

In multiple linear regression there are many explanatory variables, so a plot of residuals vs. one of them wouldn't give us the complete picture.

(3) independent residuals

scatterplot of residuals vs. order of data collection:



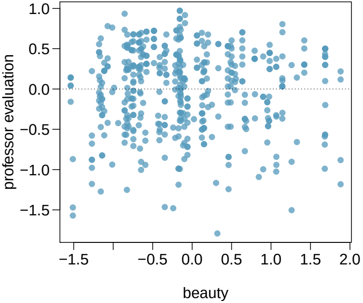
Does this condition appear to be satisfied?

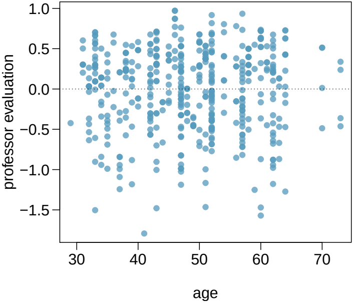
More on the condition of independent residuals

- Checking for independent residuals allows us to indirectly check for independent observations.
- If observations and residuals are independent, we would not expect to see an increasing or decreasing trend in the scatterplot of residuals vs. order of data collection.
- This condition is often violated when we have time series data. Such data require more advanced time series regression techniques for proper analysis.

(4) linear relationships

scatterplot of residuals vs. each (numerical) explanatory variable:





Does this condition appear to be satisfied?

We use residuals instead of the predictors on the y-axis so that we can still check for linearity without worrying about other possible violations like collinearity between the predictors.

References

📄 David Diez, Christopher Barr, & Mine Çetinkaya-Rundel (2015)
OpenIntro Statistics, OpenIntro

Recommended Reading

OpenIntro Statistics, Chapter 8
Data Science from Scratch, Chapter 15

Articles for discussion:
If Correlation Doesn't Imply Causation, Then What Does?
Workflows in Python: Curating Features and Thinking Scientifically about Algorithms

Exercise

Lesson7_MultipleRegression.ipynb