Lecture 22 - Multiple linear regression

Sta 111

Colin Rundel

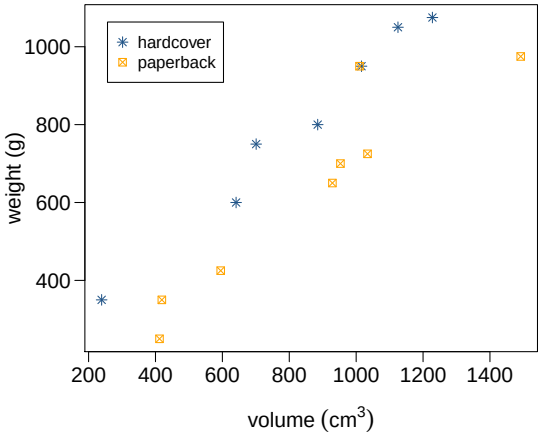
June 16, 2014

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 hard cover and paperback books

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



Modeling weights of books using volume and cover type

```
book_mlr = lm(weight ~ volume + cover, data = allbacks)
summary(book_mlr)

## 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

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

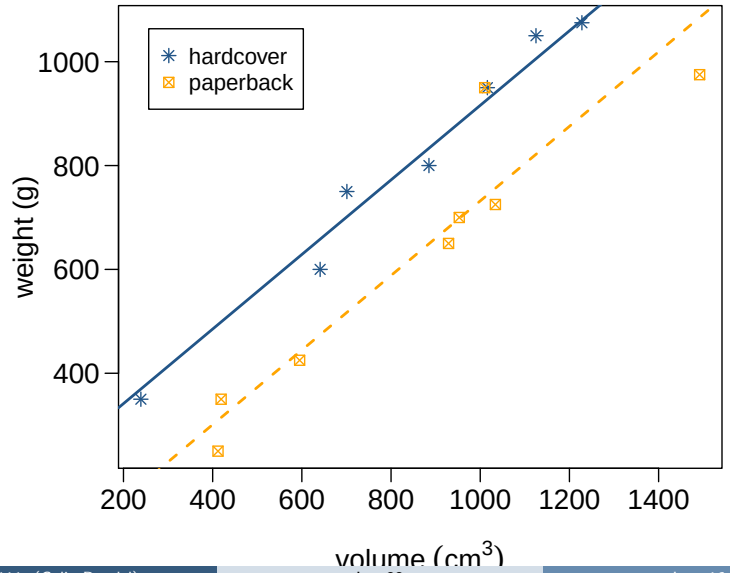
1. For *hardcover* books: plug in *0* for `cover`

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

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

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

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, for each 1 cm^3 increase in volume we would expect weight to increase on average by 0.72 grams.
- *Slope of cover:* All else held constant, the model predicts that paperback books weigh 184 grams less than hardcover books, on average.
- *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.

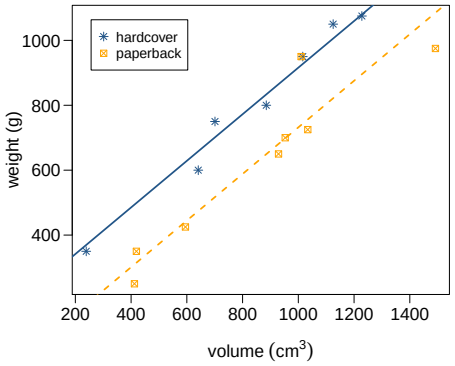
Prediction

What is the correct calculation for the predicted weight of a paperback book that has a volume of 600 cm^3?

	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

A note on interactions

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



This model assumes that hardcover and paperback books have the same slope for the relationship between their volume and weight. If this isn't reasonable, then we would include an "interaction" variable in the model.

Example of an interaction

```
summary( lm(weight ~ volume + cover + volume:cover, data = allbacks) )

## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     161.58654   86.51918   1.868   0.0887 .
## volume            0.76159    0.09718   7.837 7.94e-06 ***
## coverpb        -120.21407  115.65899  -1.039   0.3209
## volume:coverpb   -0.07573    0.12802  -0.592   0.5661
##
## Residual standard error: 80.41 on 11 degrees of freedom
## Multiple R-squared: 0.9297, Adjusted R-squared: 0.9105
## F-statistic:  48.5 on 3 and 11 DF,  p-value: 1.245e-06
```

$$\widehat{\texttt{weight}} = 161.58 + 0.76\ \texttt{volume} - 120.21\ \texttt{cover:pb} - 0.076\ \texttt{volume} \times \texttt{cover:pb}$$

Example of an interaction - interpretation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	161.5865	86.5192	1.87	0.0887
volume	0.7616	0.0972	7.84	0.0000
coverpb	-120.2141	115.6590	-1.04	0.3209
volume:coverpb	-0.0757	0.1280	-0.59	0.5661

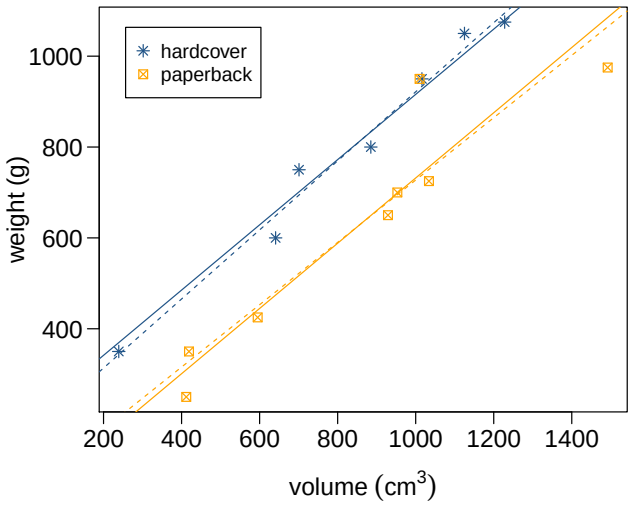
Regression equations for hardbacks:

$$\begin{aligned}\widehat{\texttt{weight}} &= 161.58 + 0.76\ \texttt{volume} - 120.21 \times \textit{0} - 0.076\ \texttt{volume} \times \textit{0} \\ &= 161.58 + 0.76\ \texttt{volume}\end{aligned}$$

Regression equations for paperbacks:

$$\begin{aligned}\widehat{\texttt{weight}} &= 161.58 + 0.76\ \texttt{volume} - 120.21 \times \textit{1} - 0.076\ \texttt{volume} \times \textit{1} \\ &= 41.37 + 0.686\ \texttt{volume}\end{aligned}$$

Example of an interaction - Results



Poverty vs. region (east, west)

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

- *Explanatory variable*: region
- *Reference level*: east
- *Intercept:* estimated average % poverty in eastern states is 11.17%
  - This is the value we get if we plug in *0* for the explanatory variable
- *Slope:* estimated average % poverty in western states is 0.38% higher than eastern states.
  - Estimated average % poverty 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)

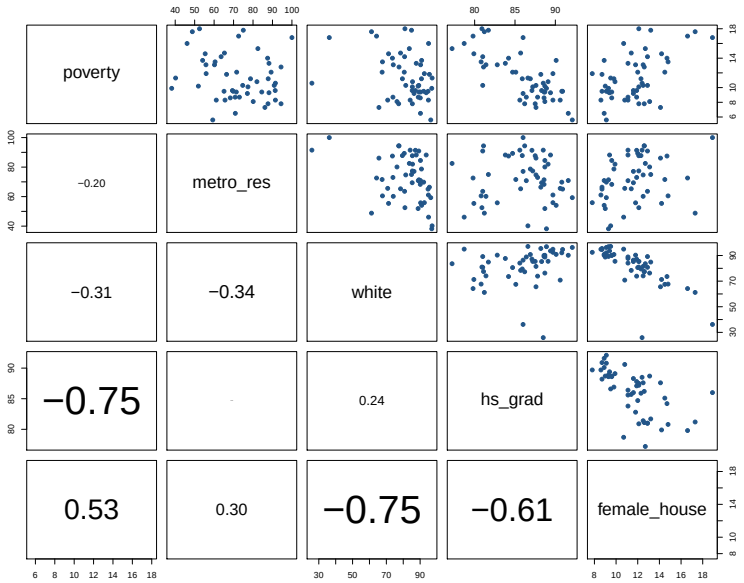
Which region (Northeast, Midwest, West, South) is the reference level?

	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

Interpretation:

- Predict 9.50% poverty in Northeast
- Predict 9.53% poverty in Midwest
- Predict 11.29% poverty in West
- Predict 13.66% poverty in South

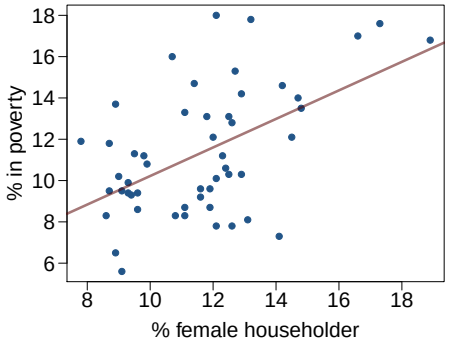
Revisit: Modeling poverty



Predicting poverty using % female householder

```
summary(lm(poverty ~ female_house, data = poverty))
```

	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 - from last week

```
anova(lm(poverty ~ female_house, data = poverty))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

$$
\begin{aligned}
SS_{Tot} &= \sum (y_i - \bar{y})^2 = 480.25 \rightarrow \textit{total variability} \\
SS_{Err} &= \sum (y_i - \hat{y}_i)^2 = 347.68 \rightarrow \textit{unexplained variability} \\
SS_{Reg} &= \sum (\hat{y}_i - \bar{y})^2 \rightarrow \textit{explained variability} \\
&= SS_{Total} - SS_{Error} \\
&= 480.25 - 347.68 = 132.57
\end{aligned}
$$

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57}{480.25} = 0.28 \checkmark$$

Predicting poverty using % female hh + % white

```
pov_mlr = lm(poverty ~ female_house + white, data = poverty)
summary(pov_mlr)
```

	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

```
anova(pov_mlr)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57 + 8.21}{480.25} = 0.29$$

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (poverty vs. female_house)	0.2760	0.2613
Model 2 (poverty vs. female_house + white)	0.2931	0.2637

- We would like to have some criteria to evaluate if adding an additional variable makes a difference in the explanatory power of the model.
- When any variable is added to the model R^2 increases (or stays the same).
- Adjusted R^2 is based on R^2 but it penalizes the addition of variables.

Adjusted R^2

Adjusted R^2

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

where n is the number of cases and k is the number of predictors (explanatory variables) in the model.

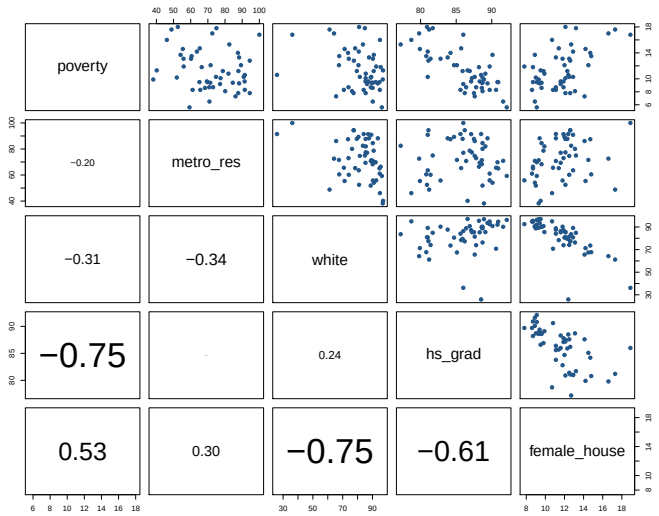
- Because k is never negative, R^2_{adj} will always be smaller than R^2.
- R^2_{adj} applies a penalty for the number of predictors included in the model.
- Therefore, we prefer models with higher R^2_{adj}

Calculate adjusted R^2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-k-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48} \right) \\
&= 1 - 0.74 \\
&= 0.26
\end{aligned}
$$

We saw that adding the variable `white` to the model only marginally increased adjusted R^2, i.e. did not add much useful information to the model. Why?



Collinearity between explanatory variables (cont.)

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* (also called *multicollinearity*) complicates model estimation.

  Remember: Predictors are also called explanatory or independent variables, so ideally they should 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 model that explains as much as possible - the most *parsimonious* model.
- In addition, inclusion of collinear variables can result in biased estimates of the slope parameters.
- While it's impossible to avoid all collinearity, often experiments are designed to control for correlated predictors.

Modeling children's test scores

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are from a survey of adult American women and their children - a subsample from the National Longitudinal Survey of Youth.

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
⋮					
5	115	yes	92.75	yes	27
6	98	no	107.90	no	18
⋮					
434	70	yes	91.25	yes	25

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* (2007) Cambridge University Press.

Model output

```
cog_full = lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age,
       data = cognitive)
summary(cog_full)

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.59241    9.21906   2.125   0.0341
## mom_hsyes    5.09482    2.31450   2.201   0.0282
## mom_iq       0.56147    0.06064   9.259   <2e-16
## mom_workyes  2.53718    2.35067   1.079   0.2810
## mom_age      0.21802    0.33074   0.659   0.5101
##
## Residual standard error: 18.14 on 429 degrees of freedom
## Multiple R-squared: 0.2171, Adjusted R-squared: 0.2098
## F-statistic: 29.74 on 4 and 429 DF,  p-value: < 2.2e-16
```


Inference for the model as a whole

Is the model as a whole significant?

$$H_0 : \beta_0 = \beta_1 = \cdots = \beta_k = 0$$
$$H_A : \text{At least one of the } \beta_i \neq 0$$

```
F-statistic: 29.74 on 4 and 429 DF,  p-value: < 2.2e-16
```

Since p-value < 0.05, the model as a whole is significant.

- The F test yielding a significant result doesn't mean the model fits the data well, it just means the model has some explanatory power.
- The F test not yielding a significant result doesn't mean individuals variables included in the model are not good predictors of y, it just means that the combination of these variables doesn't yield a good model.

Inference for the slope(s)

Is whether or not mom went to high school a significant predictor of kid's cognitive test score, given all other variables in the model?

$$H_0 : \beta_1 = 0, \text{ when all other variables are included in the model}$$
$$H_A : \beta_1 \neq 0, \text{ when all other variables are included in the model}$$

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.59241    9.21906   2.125   0.0341
mom_hsyes    5.09482    2.31450   2.201   0.0282
mom_iq       0.56147    0.06064   9.259   <2e-16
mom_workyes  2.53718    2.35067   1.079   0.2810
mom_age      0.21802    0.33074   0.659   0.5101

Residual standard error: 18.14 on 429 degrees of freedom
```

$$T = 2.201, \; df = n - k = 434 - 5 = 429, \; \text{p-value} = 0.0282$$

Since p-value < 0.05, whether or not mom went to high school is a significant predictor of kid's test score, given all other variables in the model.

Interpreting the slope

What is the correct interpretation of the slope for `mom_work`?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

CI Recap from last time

Inference for the slope for a SLR model (only one explanatory variable):

- Hypothesis test:

$$T = \frac{b_1 - \textit{null value}}{SE_{b_1}} \qquad df = n - 2$$

- Confidence interval:

$$b_1 \pm t^\star_{df} \times SE_{b_1}$$

The only difference for MLR is that we use b_i instead of b_1, and use $df = n - k$

CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$\begin{aligned} b_i &\pm t^\star SE_{b_k} \\ df &= n - k = 434 - 5 = 429 \rightarrow 400 \\ 2.54 &\pm 1.97 \times 2.35 \\ 2.54 &\pm 4.63 \\ (-2.0895 &, 7.1695) \end{aligned}$$

Interpretation?

Inference for the slope(s) (cont.)

Given all variables in the model, which variables are significant predictors of kid's cognitive test score?

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.59241    9.21906   2.125   0.0341
mom_hsyes    5.09482    2.31450   2.201   0.0282
mom_iq       0.56147    0.06064   9.259   <2e-16
mom_workyes  2.53718    2.35067   1.079   0.2810
mom_age      0.21802    0.33074   0.659   0.5101
```


Modeling kid's test scores (revisited)

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are from a survey of adult American women and their children - a subsample from the National Longitudinal Survey of Youth.

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
⋮	⋮	⋮	⋮	⋮	⋮
5	115	yes	92.75	yes	27
6	98	no	107.90	no	18
⋮	⋮	⋮	⋮	⋮	⋮
434	70	yes	91.25	yes	25

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* (2007) Cambridge University Press.

Model output

```
cog_full = lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age,
              data = cognitive)
summary(cog_full)

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.59241    9.21906   2.125   0.0341
## mom_hsyes    5.09482    2.31450   2.201   0.0282
## mom_iq       0.56147    0.06064   9.259   <2e-16
## mom_workyes  2.53718    2.35067   1.079   0.2810
## mom_age      0.21802    0.33074   0.659   0.5101
##
## Residual standard error: 18.14 on 429 degrees of freedom
## Multiple R-squared: 0.2171, Adjusted R-squared: 0.2098
## F-statistic: 29.74 on 4 and 429 DF,  p-value: < 2.2e-16
```


Backward-elimination

- Adjusted R^2 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 largest increase in R^2_{adj}
  - Repeat until none of the reduced models yield an increase in R^2_{adj}
- When removing a categorical variable all levels should be included or removed

Backward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	*0.2098*
Step 1	kid_score ~ mom_iq + mom_work + mom_age	0.2027
	kid_score ~ mom_hs + mom_work + mom_age	0.0541
	kid_score ~ mom_hs + mom_iq + mom_age	0.2095
	kid_score ~ mom_hs + mom_iq + mom_work	*0.2109*
Step 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_hs + mom_work	0.0546
	kid_score ~ mom_hs + mom_iq	*0.2105*
Step 3*	kid_score ~ mom_hs	0.2024
	kid_score ~ mom_iq	0.0546

Backward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	0.2098
Step 1	kid_score ~ mom_iq + mom_work + mom_age	0.2027
	kid_score ~ mom_hs + mom_work + mom_age	0.0541
	kid_score ~ mom_hs + mom_iq + mom_age	0.2095
	kid_score ~ mom_hs + mom_iq + mom_work	0.2109
Step 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_hs + mom_work	0.0546
	kid_score ~ mom_hs + mom_iq	0.2105
Step 3*	kid_score ~ mom_hs	0.2024
	kid_score ~ mom_iq	0.0546

Forward-selection

1. Adjusted R^2 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}

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Step 1	kid_score ~ mom_hs	0.0539
	kid_score ~ mom_work	0.0097
	kid_score ~ mom_age	0.0062
	kid_score ~ mom_iq	*0.1991*
Step 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_iq + mom_age	0.1999
	kid_score ~ mom_iq + mom_hs	*0.2105*
Step 3	kid_score ~ mom_iq + mom_hs + mom_age	0.2095
	kid_score ~ mom_iq + mom_hs + mom_work	*0.2109*
Step 4*	kid_score ~ mom_iq + mom_hs + mom_age + mom_work	0.2098

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Step 1	kid_score ~ mom_hs	0.0539
	kid_score ~ mom_work	0.0097
	kid_score ~ mom_age	0.0062
	kid_score ~ mom_iq	0.1991
Step 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_iq + mom_age	0.1999
	kid_score ~ mom_iq + mom_hs	0.2105
Step 3	kid_score ~ mom_iq + mom_hs + mom_age	0.2095
	kid_score ~ mom_iq + mom_hs + mom_work	0.2109
Step 4*	kid_score ~ mom_iq + mom_hs + mom_age + mom_work	0.2098

Expert opinion as criterion for model selection

In addition to the quantitative approaches we discussed, variables can be included in (or eliminated from) the model based on expert opinion.

Final model choice

```
cog_final = lm(kid_score ~ mom_hs + mom_iq, data = kid)
summary(cog_final)

## Call:
## lm(formula = kid_score ~ mom_hs + mom_iq, data = kid)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.73154    5.87521   4.380 1.49e-05 ***
## mom_hsyes    5.95012    2.21181   2.690  0.00742 **
## mom_iq       0.56391    0.06057   9.309  < 2e-16 ***
##
## Residual standard error: 18.14 on 431 degrees of freedom
## Multiple R-squared: 0.2141, Adjusted R-squared: 0.2105
## F-statistic: 58.72 on 2 and 431 DF,  p-value: < 2.2e-16
```


Conditions for MLR

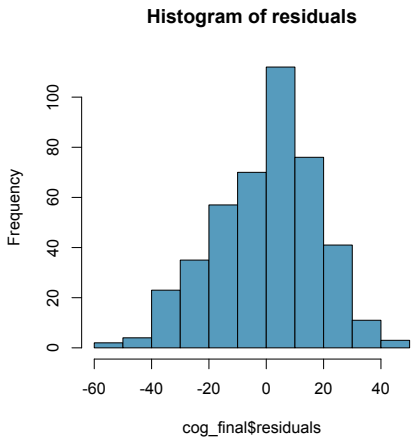
In order to perform inference for multiple regression we require the following conditions:

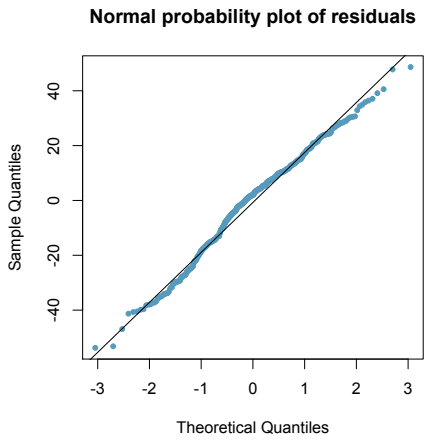
(1) Nearly normal residuals

(2) Constant variability of residuals

(3) Independent residuals

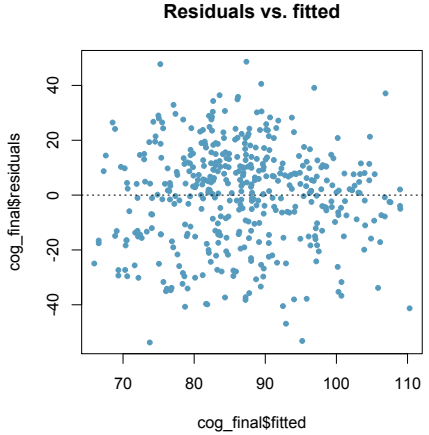
Nearly normal residuals

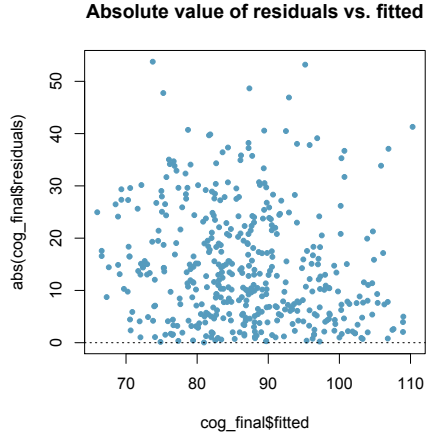




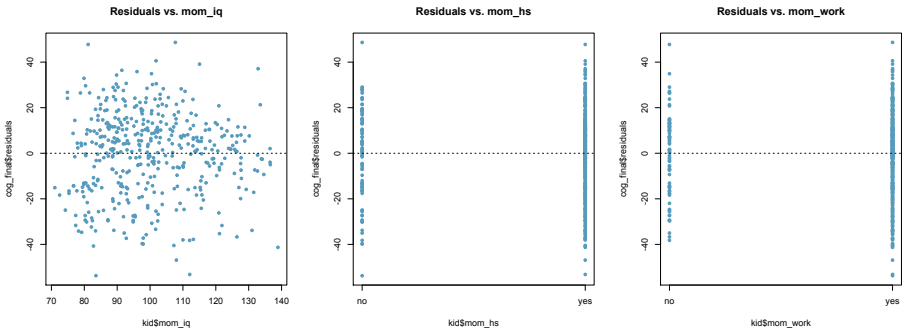
Constant variability of residuals

Why do we use the fitted (predicted) values in MLR?



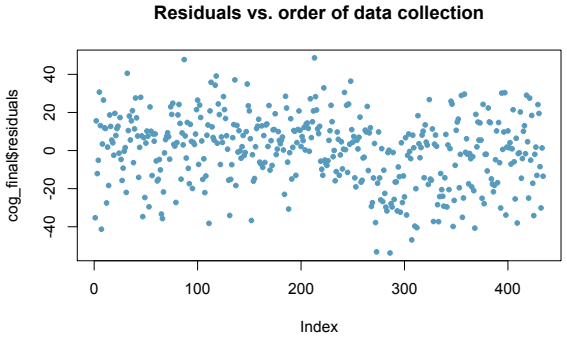


Constant variability of residuals (cont.)



Independent residuals

- If we suspect that order of data collection may influence the outcome (mostly in time series data):



- If not, think about how data are sampled.