Lecture 19 - Introduction to Multiple Regression

Sta102

June 13, 2016

Colin Rundel & Mine Çetinkaya-Rundel

Linear regression with categorical predictors

Poverty vs. region (east, west)

```r
str(poverty)
```

```
## 'data.frame':ȅȅI51 obs. of  7 variables:
##  $ State    : Factor w/ 51 levels "Alabama","Alaska",..: 1 2 3 4 5
##  $ Metro    : num  55.4 65.6 88.2 52.5 94.4 84.5 87.7 80.1 100 89.3
##  $ Graduates: num  79.9 90.6 83.8 80.9 81.1 88.7 87.5 88.7 86 84.7
##  $ Poverty  : num  14.6 8.3 13.3 18 12.8 9.4 7.8 8.1 16.8 12.1 ...
##  $ FemaleHH : num  14.2 10.8 11.1 12.1 12.6 9.6 12.1 13.1 18.9 12 .
##  $ region2  : Factor w/ 2 levels "east","west": 1 2 2 2 2 2 1 1 1 1
##  $ region4  : Factor w/ 4 levels "northeast","midwest",..: 4 3 3 4
```

Poverty vs. region (east, west)

```
poverty %>%
  group_by(region2) %>%
  summarize(mean=mean(Poverty),
            med=median(Poverty),
            sd=sd(Poverty),
            iqr=IQR(Poverty))
```

```
## Source: local data frame [2 x 5]
##
##   region2     mean   med       sd   iqr
##    (fctr)    (dbl) (dbl)    (dbl) (dbl)
## 1    east 11.17037  10.3 3.085427   4.6
## 2    west 11.55000  10.7 3.168459   4.0
```

Poverty vs. region (east, west)

```
##
## Call:
## lm(formula = Poverty ~ region2, data = poverty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5704 -2.2000 -0.8704  2.0398  6.4500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>♦t♦)
## (Intercept)  11.1704     0.6013  18.576   <2e-16 ***
## region2west   0.3796     0.8766   0.433    0.667
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.125 on 49 degrees of freedom
## Multiple R-squared:  0.003813,ℯℯIAdjusted R-squared:  -0.01652
## F-statistic: 0.1875 on 1 and 49 DF,  p-value: 0.6669
```

Poverty vs. region (east, west)

$$\% \widehat{poverty} = 11.17 + 0.38 \times \mathbb{1}_{west}$$

- *Explanatory variable*: region

Poverty vs. region (east, west)

$$\% \widehat{poverty} = 11.17 + 0.38 \times \mathbb{1}_{west}$$

- *Explanatory variable*: region
- *Reference level*: east

Poverty vs. region (east, west)

$$\% \widehat{poverty} = 11.17 + 0.38 \times \mathbb{1}_{west}$$

- *Explanatory variable*: region
- *Reference level*: east
- *Intercept:* estimated average % poverty in eastern states is 11.17%

Poverty vs. region (east, west)

$$\widehat{\% \text{ } poverty} = 11.17 + 0.38 \times \mathbb{1}_{west}$$

- *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

Poverty vs. region (east, west)

$$\widehat{\%\ poverty} = 11.17 + 0.38 \times \mathbb{1}_{west}$$

- *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.

Poverty vs. region (east, west)

$$\widehat{\%\ poverty} = 11.17 + 0.38 \times \mathbb{1}_{west}$$

- *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%.

Poverty vs. Region (Northeast, Midwest, West, South)

```
##
## Call:
## lm(formula = Poverty ~ region4, data = poverty)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.359 -1.559 -0.025  1.574  6.508
##
## Coefficients:
##                Estimate Std. Error t value Pr(>♦t♦)
## (Intercept)      9.5000     0.8682  10.943 1.62e-14 ***
## region4midwest   0.0250     1.1485   0.022 0.982725
## region4west      1.7923     1.1294   1.587 0.119220
## region4south     4.1588     1.0736   3.874 0.000331 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.604 on 47 degrees of freedom
## Multiple R-squared:  0.3361,ееIAdjusted R-squared:  0.2938
## F-statistic: 7.933 on 3 and 47 DF,  p-value: 0.0002205
```

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

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

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

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

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

Poverty vs. Region (Northeast, Midwest, West, South)

```
poverty %>%
  group_by(region4) %>%
  summarize(mean=mean(Poverty),
            med=median(Poverty),
            sd=sd(Poverty),
            iqr=IQR(Poverty))
```

```
## Source: local data frame [4 x 5]
##
##     region4     mean   med       sd   iqr
##      (fctr)    (dbl) (dbl)    (dbl) (dbl)
## 1 northeast  9.50000  9.60 2.381701  2.50
## 2   midwest  9.52500  9.55 1.415579  1.55
## 3      west 11.29231 10.80 2.647471  3.40
## 4     south 13.65882 14.20 3.233431  3.90
```

Poverty vs. Region (Northeast, Midwest, West, South)

```r
summary(aov(poverty$Poverty~poverty$region4))
```

```
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## poverty$region4  3  161.4   53.81   7.933 0.00022 ***
## Residuals       47  318.8    6.78
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(lm(Poverty ~ region4, data=poverty))
```

```
...
## Residual standard error: 2.604 on 47 degrees of freedom
## Multiple R-squared:  0.3361,  Adjusted R-squared:  0.2938
## F-statistic: 7.933 on 3 and 47 DF,  p-value: 0.0002205
```

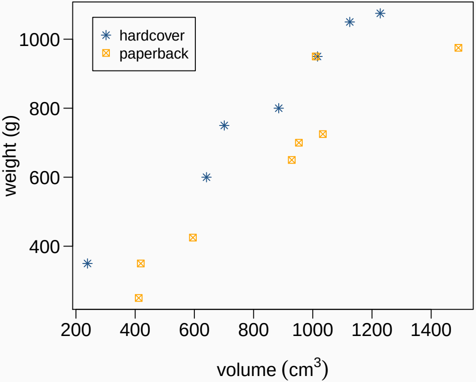
Linear models with multiple predictors

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

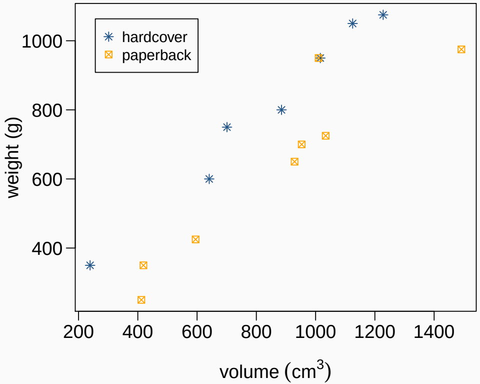
Weights of hard cover and paperback books

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



Weights of hard cover and paperback books

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



Paperbacks generally weigh less than hardcover 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

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

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

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`

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

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`

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

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`

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

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

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

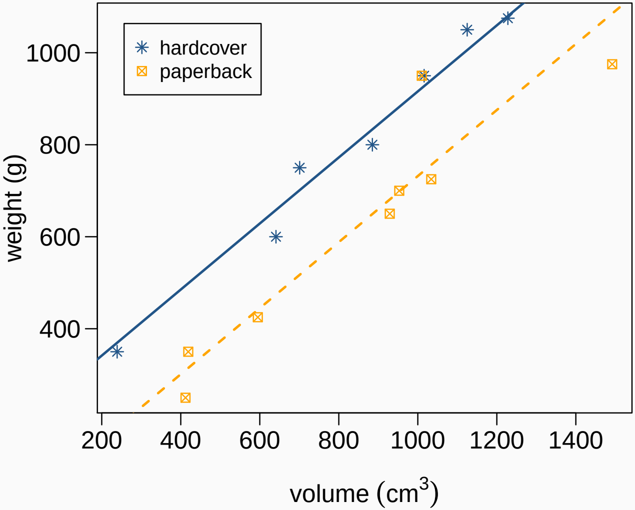
1. For *hardcover* books: plug in *0* for `cover`

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

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

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.

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.

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.

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

Prediction

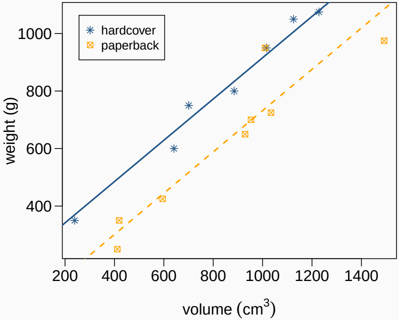
What is the correct calculation for the predicted weight of a paperback book that has a volume of 600 cm^3?

	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

$$197.96 + 0.72 \times 600 - 184.05 \times 1 = \mathit{445.91\ grams}$$

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

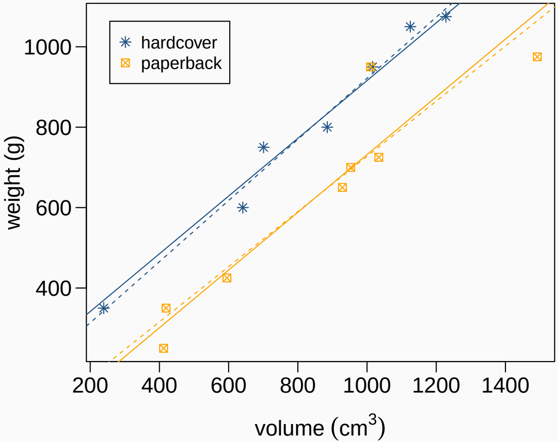
Regression equations for hardbacks:

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

Regression equations for paperbacks:

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

Example of an interaction - Results



R^2 and Adjusted R^2

Another look at R

For a linear regression we have defined the correlation coefficient to be

$$R = \text{Cor}(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

This definition works fine for the simple linear regression case where X and Y are numeric variables, but does not work for regression with a categorical predictor or for multiple regression.

A more useful, and equivalent, definition is $R = \text{Cor}(Y, \hat{Y})$, which will work for all regression examples we will see in this class.

Another look at R, cont.

Claim: $\text{Cor}(X, Y) = \text{Cor}(Y, \hat{Y})$

Another look at R, cont.

Claim: $\text{Cor}(X, Y) = \text{Cor}(Y, \hat{Y})$

Remember: $\text{Cor}(X, Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}, \quad \hat{Y} = b_0 + b_1\, X,$

$\text{Var}(aX + b) = a^2\, \text{Var}(X),$

$\text{Cov}(aX + b, Y) = a\, \text{Cov}(X, Y)$

Another look at R, cont.

Claim: $\text{Cor}(X, Y) = \text{Cor}(Y, \hat{Y})$

Remember: $\text{Cor}(X, Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}, \quad \hat{Y} = b_0 + b_1\, X,$
$\text{Var}(aX + b) = a^2\ \text{Var}(X),$
$\text{Cov}(aX + b, Y) = a\ \text{Cov}(X, Y)$

$$
\begin{aligned}
\text{Cor}(Y, \hat{Y}) &= \frac{\text{Cov}(Y, \hat{Y})}{\sqrt{\text{Var}(Y)\text{Var}(\hat{Y})}} \\
&= \frac{\text{Cov}(Y, b_0 + b_1\, X)}{\sqrt{\sigma_Y^2 \text{Var}(b_0 + b_1\, X)}} \\
&= \frac{b_1\ \text{Cov}(Y, X)}{\sigma_Y \sqrt{b_1^2 \text{Var}(X)}} \\
&= \frac{b_1\ \text{Cov}(Y, X)}{b_1\ \sigma_Y \sigma_X} \\
&= \text{Cor}(X, Y)
\end{aligned}
$$

Another look at R^2

Can we still claim that R^2 for a MLR is still a measure of variability "explained" by the model?

Another look at R^2

Can we still claim that R^2 for a MLR is still a measure of variability "explained" by the model?

This definition comes from an ANOVA-like approach where we partition total uncertainty into model uncertainty and residual (error) uncertainty.

$$SST = SSG + SSE$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i-\bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

For a MLR we can do the same thing we did with SLR just using the more complex $\hat{y}_i$

$$SST = SSR + SSE$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

Another look at R^2

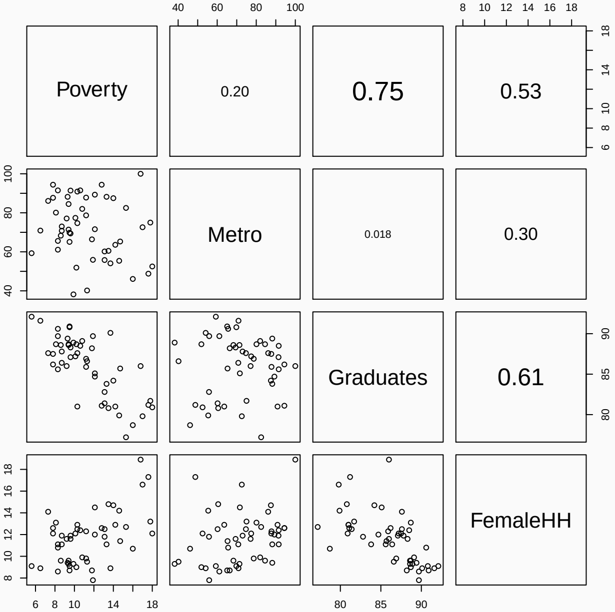
After a fair bit of algebra we can show that,

$$R^2 = \text{Cor}(Y, \hat{Y})^2 = \frac{Cov(Y, \hat{Y})^2}{Var(Y)Var(\hat{Y})}$$

$$= \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{SSR}{SST}$$

$$= \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

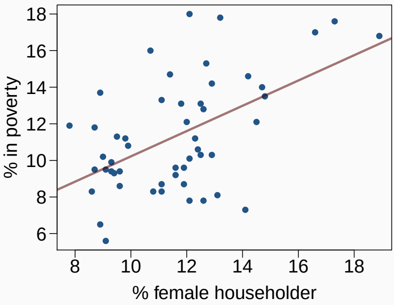
Revisit: Modeling poverty



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			

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			

$$SS_{Tot} = \sum(y - \bar{y})^2 = 480.25 \rightarrow \textit{total variability}$$

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			

$$SS_{Tot} = \sum (y - \bar{y})^2 = 480.25 \rightarrow \textit{total variability}$$

$$SS_{Err} = \sum e_i^2 = 347.68 \rightarrow \textit{unexplained variability}$$

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			

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

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			

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

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

Predicting poverty using % female hh + % metro

```
pov_mlr = lm(Poverty ~ FemaleHH + Graduates, data = poverty)
summary(pov_mlr)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	58.3203	9.8470	5.923	3.29e-07
FemaleHH	0.1439	0.1583	0.909	0.368
Graduates	-0.5656	0.1001	-5.651	8.51e-07

```
anova(pov_mlr)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FemaleHH	1	132.57	132.568	30.479	1.341e-06
Graduates	1	138.91	138.906	31.936	8.511e-07
Residuals	48	208.77	4.349		

Predicting poverty using % female hh + % metro

```
pov_mlr = lm(Poverty ~ FemaleHH + Graduates, data = poverty)
summary(pov_mlr)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	58.3203	9.8470	5.923	3.29e-07
FemaleHH	0.1439	0.1583	0.909	0.368
Graduates	-0.5656	0.1001	-5.651	8.51e-07

```
anova(pov_mlr)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FemaleHH	1	132.57	132.568	30.479	1.341e-06
Graduates	1	138.91	138.906	31.936	8.511e-07
Residuals	48	208.77	4.349		

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57 + 138.91}{480.25} = 0.565$$

R^2 vs. adjusted R^2

	R^2
Model 1 (poverty vs. FemaleHH)	0.276
Model 2 (poverty vs. Graduates)	0.5578
Model 3 (poverty vs. FemaleHH + Graduates)	0.565

- We would like to have some criteria to evaluate if adding an additional variable makes a difference in the explanatory power of the model.

R^2 vs. adjusted R^2

	R^2
Model 1 (poverty vs. FemaleHH)	0.276
Model 2 (poverty vs. Graduates)	0.5578
Model 3 (poverty vs. FemaleHH + Graduates)	0.565

- 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.

R^2 vs. adjusted R^2

	R^2
Model 1 (poverty vs. FemaleHH)	0.276
Model 2 (poverty vs. Graduates)	0.5578
Model 3 (poverty vs. FemaleHH + Graduates)	0.565

- 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 <u>any</u> variable is added to the model R^2 increases.
- Adjusted R^2 is based on R^2 but it penalizes the addition of variables.

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (poverty vs. FemaleHH)	0.276	0.261
Model 2 (poverty vs. Graduates)	0.5578	0.549
Model 3 (poverty vs. FemaleHH + Graduates)	0.565	0.547

- 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.
- Adjusted R^2 is based on R^2 but it penalizes the addition of variables.

Adjusted R^2

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

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

- Because k is never negative, R^2_{adj} will always be less than or equal to 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)
FemaleHH	1	132.57	132.57	22.84	0.0000
Graduates	1	138.91	138.906	31.936	8.511e-07
Residuals	48	208.77	4.349		
Total	50	480.25			

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

Calculate adjusted R^2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FemaleHH	1	132.57	132.57	22.84	0.0000
Graduates	1	138.91	138.906	31.936	8.511e-07
Residuals	48	208.77	4.349		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-k-1} \right) \\
&= 1 - \left(\frac{208.77}{480.25} \times \frac{51-1}{51-2-1} \right)
\end{aligned}
$$

Calculate adjusted R^2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FemaleHH	1	132.57	132.57	22.84	0.0000
Graduates	1	138.91	138.906	31.936	8.511e-07
Residuals	48	208.77	4.349		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-k-1} \right) \\
&= 1 - \left(\frac{208.77}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{208.77}{480.25} \times \frac{50}{48} \right)
\end{aligned}
$$

Calculate adjusted R^2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FemaleHH	1	132.57	132.57	22.84	0.0000
Graduates	1	138.91	138.906	31.936	8.511e-07
Residuals	48	208.77	4.349		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-k-1} \right) \\
&= 1 - \left(\frac{208.77}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{208.77}{480.25} \times \frac{50}{48} \right) \\
&= 1 - 0.453
\end{aligned}
$$

Calculate adjusted R^2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FemaleHH	1	132.57	132.57	22.84	0.0000
Graduates	1	138.91	138.906	31.936	8.511e-07
Residuals	48	208.77	4.349		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-k-1} \right) \\
&= 1 - \left(\frac{208.77}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{208.77}{480.25} \times \frac{50}{48} \right) \\
&= 1 - 0.453 \\
&= 0.547
\end{aligned}
$$

Predicting poverty using % female hh + % metro

```
pov_mlr = lm(Poverty ~ FemaleHH + Metro, data = poverty)
summary(pov_mlr)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.3127	2.0710	3.53	0.0009
FemaleHH	0.8480	0.1516	5.59	0.0000
Metro	-0.0807	0.0234	-3.45	0.0012

```
anova(pov_mlr)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FemaleHH	1	132.57	132.57	22.84	0.0000
Metro	1	69.12	69.12	11.91	0.0012
Residuals	48	278.56	5.80		

Predicting poverty using % female hh + % metro

```
pov_mlr = lm(Poverty ~ FemaleHH + Metro, data = poverty)
summary(pov_mlr)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.3127	2.0710	3.53	0.0009
FemaleHH	0.8480	0.1516	5.59	0.0000
Metro	-0.0807	0.0234	-3.45	0.0012

```
anova(pov_mlr)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FemaleHH	1	132.57	132.57	22.84	0.0000
Metro	1	69.12	69.12	11.91	0.0012
Residuals	48	278.56	5.80		

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57 + 69.12}{480.25} = 0.42$$

R^2 vs. adjusted R^2

	R^2	
Model 1 (poverty vs. FemaleHH)	0.276	0.261
Model 2 (poverty vs. Metro)	0.042	0.022
Model 3 (poverty vs. FemaleHH + Metro)	0.420	0.396

R^2 vs. adjusted R^2

	R^2	
Model 1 (poverty vs. FemaleHH)	0.276	0.261
Model 2 (poverty vs. Metro)	0.042	0.022
Model 3 (poverty vs. FemaleHH + Metro)	0.420	0.396

R^2 vs. adjusted R^2

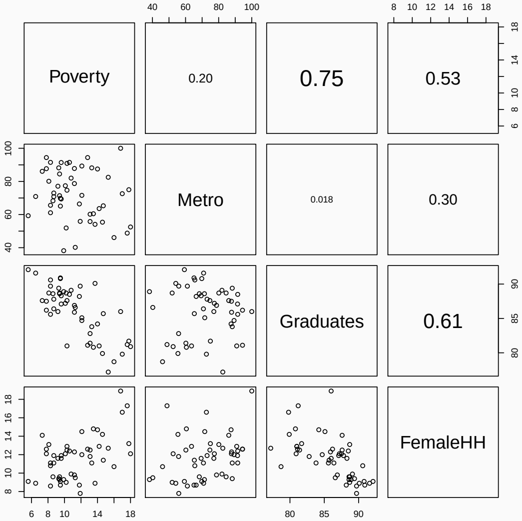
	R^2	
Model 1 (poverty vs. FemaleHH)	0.276	0.261
Model 2 (poverty vs. Metro)	0.042	0.022
Model 3 (poverty vs. FemaleHH + Metro)	0.420	0.396

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (poverty vs. FemaleHH)	0.276	0.261
Model 2 (poverty vs. Metro)	0.042	0.022
Model 3 (poverty vs. FemaleHH + Metro)	0.420	0.396

Collinearity and parsimony

We saw that adding the variable `FemaleHH` to the model with `Graduates` 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.

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.
- All else being equal we want the simplest model that explains as much as possible - what we call the most *parsimonious* model.

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.
- All else being equal we want the simplest model that explains as much as possible - what we call the most *parsimonious* model.
- Adding collinear variables rarely adds much to the model in terms of explanatory power, and in some cases inclusion of collinear variables can result in biased estimates of the slope parameters.

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.
- All else being equal we want the simplest model that explains as much as possible - what we call the most *parsimonious* model.
- Adding collinear variables rarely adds much to the model in terms of explanatory power, and in some cases 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.

Model diagnostics

Modeling children's test scores

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are 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

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

```
##
## Call:
## lm(formula = kid_score ~ mom_hs + mom_iq + mom_work + mom_age,
##     data = cognitive)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.134 -12.624   2.293  11.250  50.206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>♦t♦)
## (Intercept) 20.82261    9.18765   2.266   0.0239 *
## mom_hs       5.56118    2.31345   2.404   0.0166 *
## mom_iq       0.56208    0.06077   9.249   <2e-16 ***
## mom_work     0.13373    0.76763   0.174   0.8618
## mom_age      0.21986    0.33231   0.662   0.5086
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.17 on 429 degrees of freedom
## Multiple R-squared:  0.215,ĕĕIAdjusted R-squared:  0.2077
## F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```

Conditions for MLR Inference

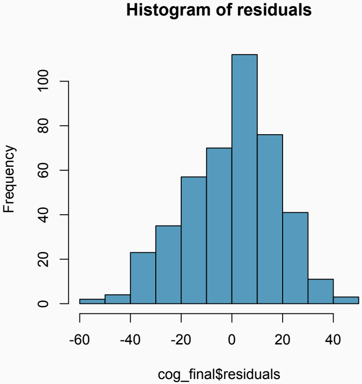
In order to conduct inference for multiple regression we require the following conditions:

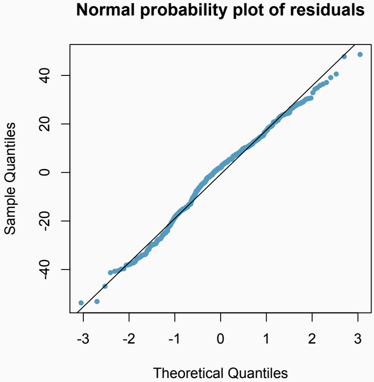
(1) Unstructured / nearly normal residuals

(2) Constant variability of residuals

(3) Independent residuals

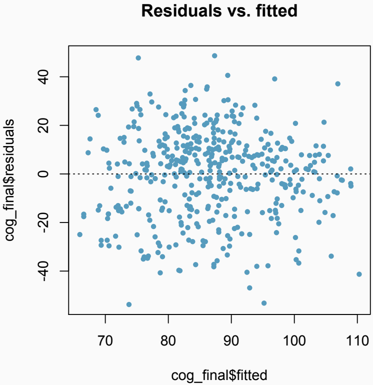
Nearly normal residuals

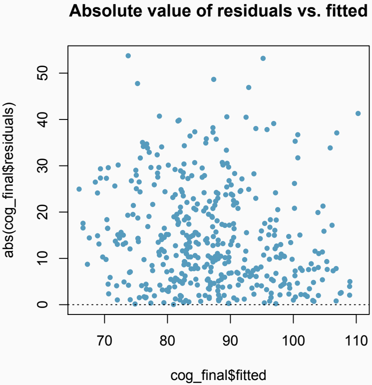




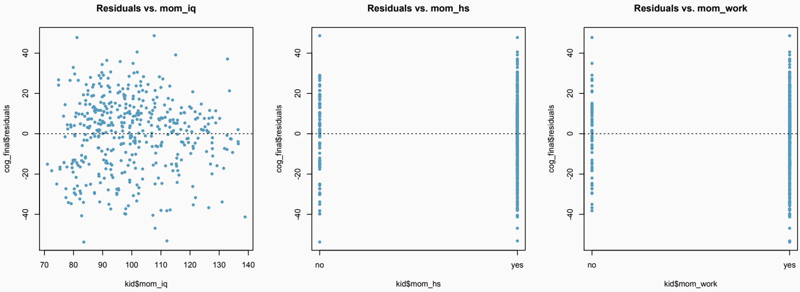
Unstructured / Constant variability of residuals

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



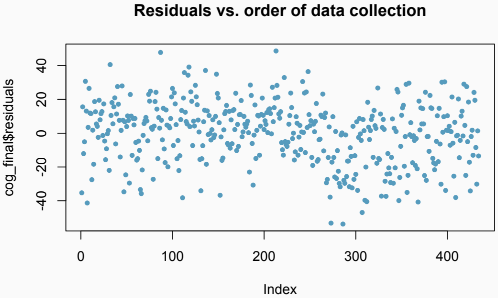


Constant variability of residuals (cont.)



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.