Multiple linear regression

Ripped from:

- Gemini-generated small dataset
- Statistics 102 Slides by Colin Rundel: `https://www2.stat.duke.edu/courses/Spring13/sta102.001/Lec/Lec18.pdf` — in the first half (in turn: Data Analysis and Graphics Using R, 2nd ed.)

Many variables in a linear model

Many variables in a linear model: theory

Many variables in a linear model: application

R^2 and Adjusted R^2

Collinearity and parsimony

Inference for MLR

Many variables in a linear model: theory

Example: Housing Dataset

Consider a small dataset predicting house price (Y) using area (X_1) and age (X_2):

Obs (i)	**Price (y_i)**	**Area (x_{i1})**	**Age (x_{i2})**
1	300	1.5	10
2	400	2.0	5
3	350	1.8	8
4	500	2.4	2

Regression Model:

$$y = b_0 + b_1 x_1 + b_2 x_2$$

Example: Housing Dataset — System of Linear Equations

Writing out the equation for each observation explicitly:

$$
\begin{aligned}
300 &= b_0(1) + b_1(1.5) + b_2(10) + \epsilon_1 \\
400 &= b_0(1) + b_1(2.0) + b_2(5) + \epsilon_2 \\
350 &= b_0(1) + b_1(1.8) + b_2(8) + \epsilon_3 \\
500 &= b_0(1) + b_1(2.4) + b_2(2) + \epsilon_4
\end{aligned}
$$

Notice that the intercept b_0 is implicitly multiplied by a column of 1s.

Example: Housing Dataset — Matrix Setup

Rewrite as a matrix equation:

$$\begin{bmatrix} 300 \\ 400 \\ 350 \\ 500 \end{bmatrix} = \begin{bmatrix} 1 & 1.5 & 10 \\ 1 & 2.0 & 5 \\ 1 & 1.8 & 8 \\ 1 & 2.4 & 2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix}$$

Example: Housing Dataset — Matrix Setup

Rewrite as a matrix equation:

$$\begin{bmatrix} 300 \\ 400 \\ 350 \\ 500 \end{bmatrix} = \begin{bmatrix} 1 & 1.5 & 10 \\ 1 & 2.0 & 5 \\ 1 & 1.8 & 8 \\ 1 & 2.4 & 2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix}$$

This has the form $\mathbf{y} = \mathbf{X}\boldsymbol{b} + \boldsymbol{\epsilon}$, where:

- $\mathbf{y}$ is the 4×1 response vector
- $\mathbf{X}$ is the 4×3 **design matrix**
- $\boldsymbol{b}$ is the 3×1 parameter vector
- $\boldsymbol{\epsilon}$ is the 4×1 error vector

Example: Housing Dataset — Matrix Setup

Rewrite as a matrix equation:

$$\begin{bmatrix} 300 \\ 400 \\ 350 \\ 500 \end{bmatrix} = \begin{bmatrix} 1 & 1.5 & 10 \\ 1 & 2.0 & 5 \\ 1 & 1.8 & 8 \\ 1 & 2.4 & 2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix}$$

Linear regression minimizes $\|\boldsymbol{\epsilon}\|$ to give:

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\boldsymbol{b} \qquad \Longrightarrow \qquad \boxed{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \boldsymbol{b}}$$

Many variables in a linear model: application

Many variables in a linear model with categorical variables

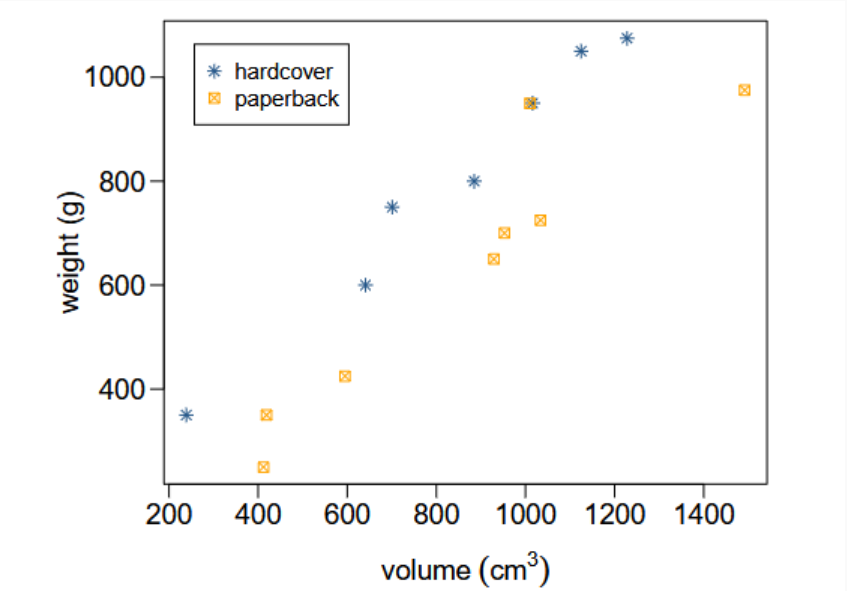
	weight (g)	volume (cm^3)	cover
1	800	885	hc
2	950	1016	hc
3	1050	1125	hc
4	350	239	hc
5	750	701	hc
6	600	641	hc
7	1075	1228	hc
8	250	412	pb
9	700	953	pb
10	650	929	pb
11	975	1492	pb
12	350	419	pb
13	950	1010	pb
14	425	595	pb
15	725	1034	pb

Many variables in a linear model with dummy variables

	weight (g)	volume (cm^3)	paperback
1	800	885	0
2	950	1016	0
3	1050	1125	0
4	350	239	0
5	750	701	0
6	600	641	0
7	1075	1228	0
8	250	412	1
9	700	953	1
10	650	929	1
11	975	1492	1
12	350	419	1
13	950	1010	1
14	425	595	1
15	725	1034	1

Many variables in a linear model

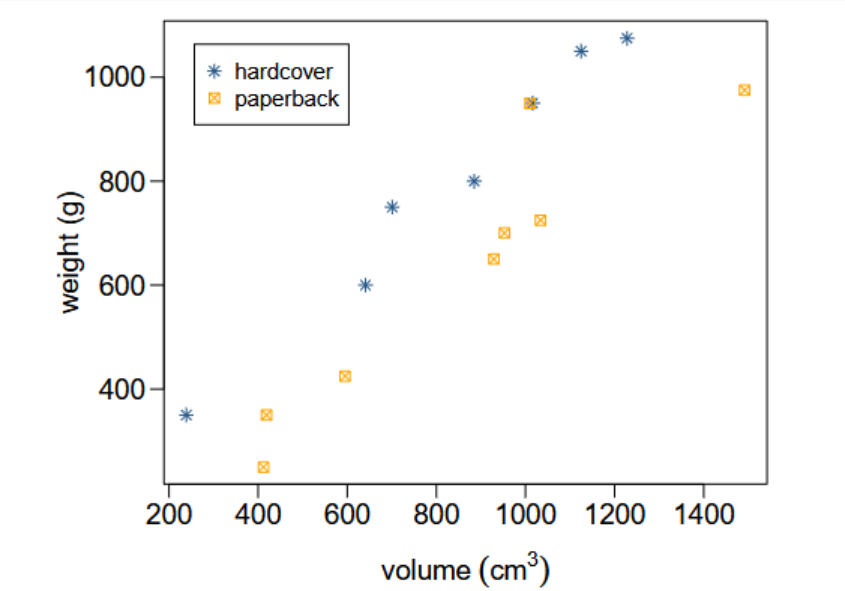
Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?



Paperbacks generally weigh less than hardcover books.

Many variables in a linear model

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



Paperbacks generally weigh less than hardcover books.

Paperbacks generally weigh less than hardcover books.

Many variables in a linear model

```
book_mlr = lm(weight ~ volume + paperback, 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 ***
## paperback   -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
```

Many variables in a linear model

SAS code:

```
filename myurl url "https://manctsui.github.io/datasets/book_weights.cs
proc import
 datafile=myurl out=book_weights dbms=csv replace;
 getnames = yes;
run;

data book_weights;
 set book_weights;
 if cover = "hc" then paperback = 0;
 if cover = "pb" then paperback = 1;
run;

proc reg data=book_weights;
   model weight = volume paperback;
   title "Multiple Linear Regression: Book Weight vs. Volume and Cover"
run;
```

Many variables in a linear model

Total rows: 15 Total columns: 5

	book_id	weight_g	volume_cm3	cover	paperback
1	1	800	885	hc	0
2	2	950	1016	hc	0
3	3	1050	1125	hc	0
4	4	350	239	hc	0
5	5	750	701	hc	0
6	6	600	641	hc	0
7	7	1075	1228	hc	0
8	8	250	412	pb	1
9	9	700	953	pb	1
10	10	650	929	pb	1
11	11	975	1492	pb	1
12	12	350	419	pb	1
13	13	950	1010	pb	1
14	14	425	595	pb	1
15	15	725	1034	pb	1

Many variables in a linear model

Multiple Linear Regression: Book Weight vs. Volume and Cover

The REG Procedure
Model: MODEL1
Dependent Variable: weight_g

Number of Observations Read	15
Number of Observations Used	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	938453	469226	76.73	<.0001
Error	12	73381	6115.05053		
Corrected Total	14	1011833			

Root MSE	78.19879	R-Square	0.9275
Dependent Mean	706.66667	Adj R-Sq	0.9154
Coeff Var	11.06587		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	197.96284	59.19274	3.34	0.0058
volume_cm3	1	0.71795	0.06153	11.67	<.0001
paperback	1	-184.04727	40.49420	-4.55	0.0007

Many variables in a linear model

	DF	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	1	197.96	59.19	3.34	0.01
volume	1	0.72	0.06	11.67	0.00
paperback	1	-184.05	40.49	-4.55	0.00

$$\widehat{\text{weight}} = 197.96 + 0.72 \times \text{volume} - 184.05 \times \text{paperback}$$

Many variables in a linear model

	DF	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	1	197.96	59.19	3.34	0.01
volume	1	0.72	0.06	11.67	0.00
paperback	1	-184.05	40.49	-4.55	0.00

$$\widehat{\text{weight}} = 197.96 + 0.72 \times \text{volume} - 184.05 \times \text{paperback}$$

For hardcover books: plug in paperback $= 0$:

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

Many variables in a linear model

	DF	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	1	197.96	59.19	3.34	0.01
volume	1	0.72	0.06	11.67	0.00
paperback	1	-184.05	40.49	-4.55	0.00

$$\widehat{\text{weight}} = 197.96 + 0.72 \times \text{volume} - 184.05 \times \text{paperback}$$

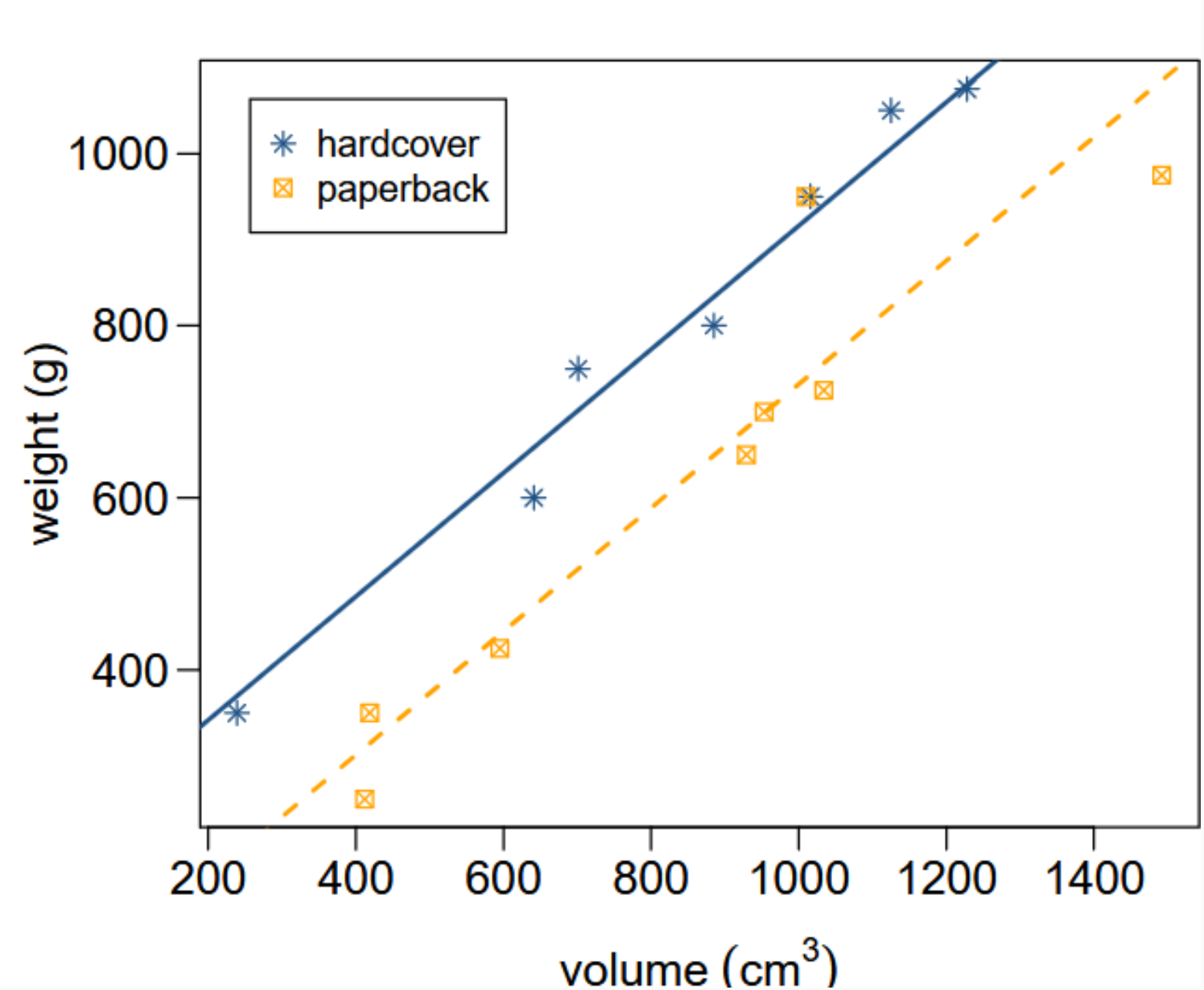
For hardcover books: plug in paperback $= 0$:

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

For paperback books: plug in paperback $= 1$:

$$\widehat{\text{weight}} = 197.96 + 0.72 \times \text{volume} - 184.05 \times 1 = 13.91 + 0.72 \times \text{volume}$$

Visualizing the linear model



Many variables in a 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
paperback	-184.05	40.49	-4.55	0.00

Many variables in a 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
paperback	-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.

Many variables in a 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
paperback	-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.

Many variables in a 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
paperback	-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.

Many variables in a linear model

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
paperback	-184.05	40.49	-4.55	0.00

Many variables in a linear model

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
paperback	-184.05	40.49	-4.55	0.00

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

Many variables in a linear model

$$\widehat{\text{weight}} = 197.96 + 0.72 \times \text{volume} - 184.05 \times \text{paperback}$$

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.

Many variables in a linear model

```
/* 1. Create dataset with interaction variable */
data book_interaction;
   set book_weights;
   interaction = volume_cm3 * paperback; /* Multiply predictor variable
run;

/* 2. Run PROC REG with the explicit interaction variable */
proc reg data=book_interaction;
   model weight_g = volume_cm3 paperback interaction;
run;
```

Many variables in a linear model

The REG Procedure
Model: MODEL1
Dependent Variable: weight_g

Number of Observations Read	15
Number of Observations Used	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	940715	313572	48.50	<.0001
Error	11	71118	6465.27512		
Corrected Total	14	1011833			

Root MSE	80.40693	R-Square	0.9297
Dependent Mean	706.66667	Adj R-Sq	0.9105
Coeff Var	11.37834		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	161.58654	86.51918	1.87	0.0887
volume_cm3	1	0.76159	0.09718	7.84	<.0001
paperback	1	-120.21407	115.65899	-1.04	0.3209
interaction	1	-0.07573	0.12802	-0.59	0.5661

Many variables in a linear model

	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
paperback	-120.2141	115.6590	-1.04	0.3209
interaction	-0.0757	0.1280	-0.59	0.5661

$$\widehat{\text{weight}} = 161.58 + 0.76\,\text{volume} - 120.21\,\text{paperback} - 0.076\,\text{volume} \times \text{paperback}$$

Regression equations for hardbacks:

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

Regression equations for paperbacks:

$$\begin{aligned}\widehat{\text{weight}} &= 161.58 + 0.76\,\text{volume} - 120.21 \times 1 - 0.076\,\text{volume} \times 1 \\ &= 41.37 + 0.684\,\text{volume}\end{aligned}$$

Many variables in a linear model

	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
paperback	-120.2141	115.6590	-1.04	0.3209
interaction	-0.0757	0.1280	-0.59	0.5661

Caveat: Interaction terms remove the additive assumption in the regression model.

R^2 and Adjusted R^2

R^2 and Adjusted R^2

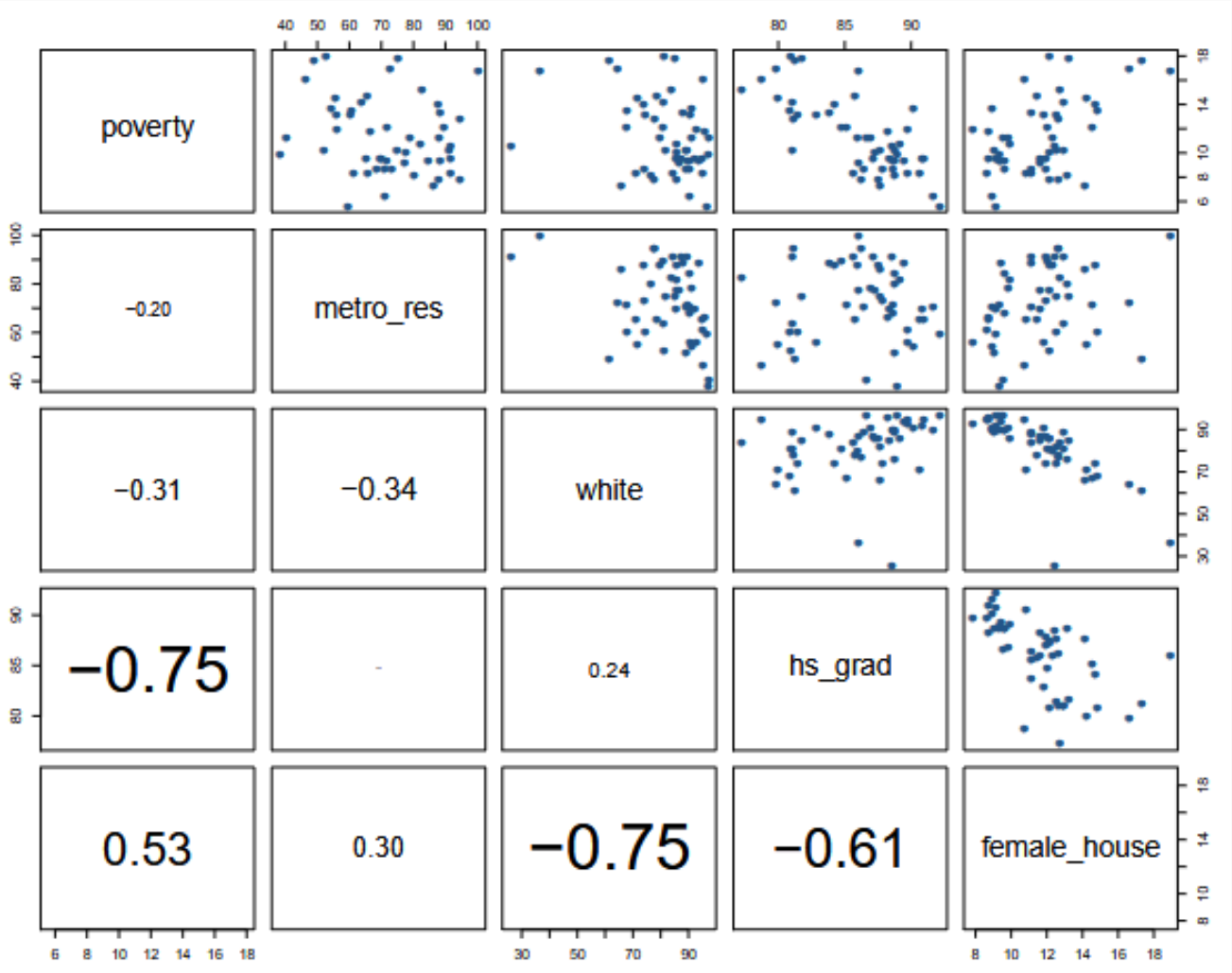
```
poverty <- read.table("https://raw.githubusercontent.com/jbryer/DATA606Spring2021/master/course_data/poverty.txt", h = T, sep = "\t")
head(poverty)
```

A data.frame: 6 × 6

	State	Metropolitan.Residence	White	Graduates	Poverty	PercentFemaleHouseholderNoHusbandPresent
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Alabama	55.4	71.3	79.9	14.6	14.2
2	Alaska	65.6	70.8	90.6	8.3	10.8
3	Arizona	88.2	87.7	83.8	13.3	11.1
4	Arkansas	52.5	81.0	80.9	18.0	12.1
5	California	94.4	77.5	81.1	12.8	12.6
6	Colorado	84.5	90.2	88.7	9.4	9.6

Data from e.g. Gelman, Data Analysis using Regression and Multilevel/Hierarchial Models.

R^2 and Adjusted R^2



R^2 and 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

R^2 and 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.

R^2 and 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.

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

Collinearity and parsimony

Collinearity and parsimony

- *Multicollinearity* occurs when a subset of x_i's are highly correlated: a linear model built from some x_i's already predict another x_j with $R^2 \approx 1$.

Collinearity and parsimony

- *Multicollinearity* occurs when a subset of x_i's are highly correlated: a linear model built from some x_i's already predict another x_j with $R^2 \approx 1$.

  Remember: Predictors are also called explanatory or independent variables, so ideally they should be independent of each other.

Collinearity and parsimony

- *Multicollinearity* occurs when a subset of x_i's are highly correlated: a linear model built from some x_i's already predict another x_j with $R^2 \approx 1$.

  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.

Collinearity and parsimony

- *Multicollinearity* occurs when a subset of x_i's are highly correlated: a linear model built from some x_i's already predict another x_j with $R^2 \approx 1$.

  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.

Inference for MLR

Inference for MLR

- Linearity: y is linearly related to $x_1, \ldots, x_k$
- Independence of residuals.
- Normality of residuals.
- Equal variance of residuals acrossed any fixed values of $x_1, \ldots, x_k$.
- No multicollinearity: some x_i's cannot predict another x_j via a linear model with R^2 close to 1. This ensures $X^T X$ can be inverted in a stable numeric sense.

Inference for MLR

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
65	yes	121.12	yes	27
115	yes	92.75	yes	27
98	no	107.90	no	18
⋮	⋮	⋮	⋮	⋮
43	yes	91.25	yes	25

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

Inference for MLR

```
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 MLR

Is the model as a whole significant?

Inference for MLR

Is the model as a whole significant?

$$H_0 : b_1 = b_2 = \cdots = b_k = 0$$

$$H_a : \text{At least one } b_i \neq 0$$

Inference for MLR

Is the model as a whole significant?

$$H_0 : b_1 = b_2 = \cdots = b_k = 0$$

$$H_a : \text{At least one } b_i \neq 0$$

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

Inference for MLR

Is the model as a whole significant?

$$H_0 : b_1 = b_2 = \cdots = b_k = 0$$

$$H_a : \text{At least one } b_i \neq 0$$

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.

Inference for MLR

Is the model as a whole significant?

$$H_0 : b_1 = b_2 = \cdots = b_k = 0$$

$$H_a : \text{At least one } b_i \neq 0$$

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 at least one of the bs is non-zero.

Inference for MLR

Is the model as a whole significant?

$$H_0 : b_1 = b_2 = \cdots = b_k = 0$$

$$H_a : \text{At least one } b_i \neq 0$$

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 at least one of the bs is non-zero.
- The F test not yielding a significant result doesn't mean individual 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 MLR

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?

Inference for MLR

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?

$H_0 : b_1 = 0$ when all other variables are included in the model

$H_a : b_1 \neq 0$ when all other variables are included in the model

Inference for MLR

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?

$H_0 : b_1 = 0$ when all other variables are included in the model

$H_a : b_1 \neq 0$ 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

Inference for MLR

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?

$H_0 : b_1 = 0$ when all other variables are included in the model

$H_a : b_1 \neq 0$ 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

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.

Inference for MLR

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

Inference for MLR

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

All else being equal, kids whose moms worked during the first three years of the kid's life are estimated to score 2.54 points higher than those whose moms did not work.

Inference for MLR

Inference for a slope for a MLR model:

Inference for MLR

Inference for a slope for a MLR model:

- **Hypothesis test:**

$$t = \frac{b_i - \text{null value}}{SE_{b_i}}, \quad df = n - k - 1$$

Inference for MLR

Inference for a slope for a MLR model:

- **Hypothesis test:**

$$t = \frac{b_i - \text{null value}}{SE_{b_i}}, \quad df = n - k - 1$$

- **Confidence interval:**

$$b_i \pm t^*_{df} \times SE_{b_i}$$

Inference for MLR

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

Inference for MLR

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

$$df = n - k - 1 = 434 - 4 - 1 = 429$$

Inference for MLR

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

$$df = n - k - 1 = 434 - 4 - 1 = 429$$

$$b_k \pm t^*_{c=95\%, df=429} SE_{b_k}$$

Inference for MLR

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

$$df = n - k - 1 = 434 - 4 - 1 = 429$$

$$b_k \pm t^*_{c=95\%, df=429} SE_{b_k}$$

$$2.54 \pm 1.97 \times 1.08 = 2.54 \pm 2.13 \implies (0.41, 4.67)$$

Inference for MLR

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

$$df = n - k - 1 = 434 - 4 - 1 = 429$$

$$b_k \pm t^*_{c=95\%, df=429} SE_{b_k}$$

$$2.54 \pm 1.97 \times 1.08 = 2.54 \pm 2.13 \implies (0.41, 4.67)$$

Interpretation? We are 95% confident that, all else being equal, kids whose moms worked during the first three years of the kid's life are estimated to score 0.41 to 4.67 points higher than those whose moms did not work.

Inference for MLR

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

$$df = n - k - 1 = 434 - 4 - 1 = 429$$

$$b_k \pm t^*_{c=95\%, df=429} SE_{b_k}$$

$$2.54 \pm 1.97 \times 1.08 = 2.54 \pm 2.13 \implies (0.41, 4.67)$$

SAS code:

```
proc reg data=mydata;
   model y = x1 x2 x3 / clb alpha=0.05;
run;
```

Inference for MLR

Multiple Linear Regression: Book Weight vs. Volume and Cover

The REG Procedure
Model: MODEL1
Dependent Variable: weight_g

Number of Observations Read	15
Number of Observations Used	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	938453	469226	76.73	<.0001
Error	12	73381	6115.05053		
Corrected Total	14	1011833			

Root MSE	78.19879	R-Square	0.9275
Dependent Mean	706.66667	Adj R-Sq	0.9154
Coeff Var	11.06587		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	197.96284	59.19274	3.34	0.0058	68.99295	326.93274
volume_cm3	1	0.71795	0.06153	11.67	<.0001	0.58390	0.85201
paperback	1	-184.04727	40.49420	-4.55	0.0007	-272.27655	-95.81799

Inference for MLR

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

Inference for MLR

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

`mom_hs` and `mom_iq` are significant.

`mom_work` and `mom_age` are not.