Multiple Regression I

Zhenisbek Assylbekov

Department of Mathematics

Regression Analysis

Multiple regression models

We now add more predictors, linearly, to the model.

Multiple regression models

We now add more predictors, linearly, to the model. For example let's add one more to the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Multiple regression models

We now add more predictors, linearly, to the model. For example let's add one more to the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

For *any* Y in this population with predictors (x_1, x_2) we have

$$\mathrm{E}[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Multiple regression models

We now add more predictors, linearly, to the model. For example let's add one more to the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

For *any* Y in this population with predictors (x_1, x_2) we have

$$\mathrm{E}[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

This is an equation of a plane in $\mathbb{R}^3$.

Multiple regression models

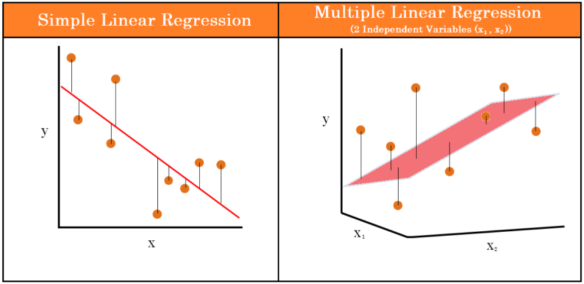
We now add more predictors, linearly, to the model. For example let's add one more to the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

For *any* Y in this population with predictors (x_1, x_2) we have

$$\mathrm{E}[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

This is an equation of a plane in $\mathbb{R}^3$.



Multiple regression models

Generally, for $k = p - 1$ predictors $x_1, \ldots, x_k$ our model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

where

Multiple regression models

Generally, for $k = p - 1$ predictors $x_1, \ldots, x_k$ our model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

where

- β_0 is

Multiple regression models

Generally, for $k = p - 1$ predictors $x_1, \ldots, x_k$ our model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

where

- β_0 is mean response when all predictors equal zero (if this makes sense).

Multiple regression models

Generally, for $k = p - 1$ predictors $x_1, \ldots, x_k$ our model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

where

- β_0 is mean response when all predictors equal zero (if this makes sense).
- β_j is

Multiple regression models

Generally, for $k = p - 1$ predictors $x_1, \ldots, x_k$ our model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

where

- β_0 is mean response when all predictors equal zero (if this makes sense).
- β_j is the change in mean response when x_j is increased by one unit *but the remaining predictors are held constant*.

Multiple regression models

Generally, for $k = p - 1$ predictors $x_1, \ldots, x_k$ our model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

where

- β_0 is mean response when all predictors equal zero (if this makes sense).
- β_j is the change in mean response when x_j is increased by one unit *but the remaining predictors are held constant*.
- We will assume normal errors:

$$\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

Example: Dwayne Portrait Studio data

Dwaine Studios, Inc., operates portrait studios in 21 cities of medium size.

Example: Dwayne Portrait Studio data

Dwaine Studios, Inc., operates portrait studios in 21 cities of medium size. These studios specialize in portraits of children.

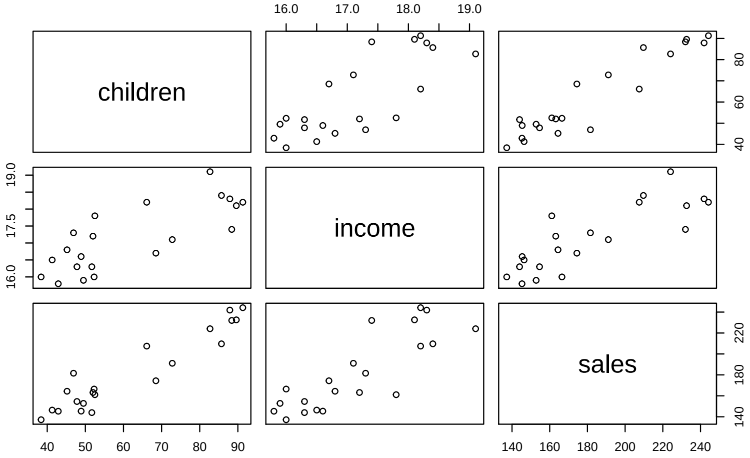
Example: Dwayne Portrait Studio data

Dwaine Studios, Inc., operates portrait studios in 21 cities of medium size. These studios specialize in portraits of children. The company is considering an expansion into other cities of medium size and wishes to investigate whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (x_1) and the per capita disposable personal income in the community (x_2).

Assume the linear model is appropriate. One way to check marginal relationships is through a scatterplot matrix.

Scatterplot matrix

https://github.com/zh3nis/MATH440/blob/main/chp06/scatter_matrix.R



The general linear model encompasses. . .

Qualitative predictors

We can use dummy variables to represent qualitative predictors.

The general linear model encompasses. . .

Qualitative predictors

We can use dummy variables to represent qualitative predictors.

Example. Binary predictor

The general linear model encompasses...

Qualitative predictors

We can use dummy variables to represent qualitative predictors.

Example. Binary predictor

- Y = length of hospital stay

The general linear model encompasses...

Qualitative predictors

We can use dummy variables to represent qualitative predictors.

Example. Binary predictor

- Y = length of hospital stay
- x_1 = gender of patient ($x_1 = 0$ male, $x_1 = 1$ female)

The general linear model encompasses. . .

Qualitative predictors

We can use dummy variables to represent qualitative predictors.

Example. Binary predictor

- Y = length of hospital stay
- x_1 = gender of patient ($x_1 = 0$ male, $x_1 = 1$ female)
- x_2 = severity of a disease on 100 point scale

The general linear model encompasses...

Qualitative predictors

We can use dummy variables to represent qualitative predictors.

Example. Binary predictor

- Y = length of hospital stay
- x_1 = gender of patient ($x_1 = 0$ male, $x_1 = 1$ female)
- x_2 = severity of a disease on 100 point scale

$$\mathrm{E}[Y] = \begin{cases} \beta_0 + \beta_1 \cdot 0 + \beta_2 x_2 & \text{males} \\ \beta_0 + \beta_1 \cdot 1 + \beta_2 x_2 & \text{females.} \end{cases}$$

The general linear model encompasses...

Qualitative predictors

We can use dummy variables to represent qualitative predictors.

Example. Binary predictor

- Y = length of hospital stay
- x_1 = gender of patient ($x_1 = 0$ male, $x_1 = 1$ female)
- x_2 = severity of a disease on 100 point scale

$$\mathrm{E}[Y] = \begin{cases} \beta_0 + \beta_1 \cdot 0 + \beta_2 x_2 & \text{males} \\ \beta_0 + \beta_1 \cdot 1 + \beta_2 x_2 & \text{females.} \end{cases}$$

Response functions are two parallel lines, shifted by β_1 units.

The general linear model encompasses...

Polynomial regression

The general linear model encompasses...

Polynomial regression

Often appropriate for curvilinear relationship between response and predictor

The general linear model encompasses...

Polynomial regression

Often appropriate for curvilinear relationship between response and predictor

Example.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon.$$

The general linear model encompasses. . .

Polynomial regression

Often appropriate for curvilinear relationship between response and predictor

Example.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon.$$

Letting $x_2 = x_1^2$ this is in the form of the general linear model.

The general linear model encompasses. . .

Polynomial regression

Often appropriate for curvilinear relationship between response and predictor

Example.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon.$$

Letting $x_2 = x_1^2$ this is in the form of the general linear model.

Transformed response

The general linear model encompasses. . .

Polynomial regression

Often appropriate for curvilinear relationship between response and predictor

Example.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon.$$

Letting $x_2 = x_1^2$ this is in the form of the general linear model.

Transformed response

Example.

$$\log Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

The general linear model encompasses...

Polynomial regression

Often appropriate for curvilinear relationship between response and predictor

Example.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon.$$

Letting $x_2 = x_1^2$ this is in the form of the general linear model.

Transformed response

Example.

$$\log Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

Let $Y^* = \log Y$ and get general linear model.

The general linear model encompasses...

Interaction effects

The general linear model encompasses...

Interaction effects

Example.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

The general linear model encompasses...

Interaction effects

Example.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

Let $x_3 = x_1 x_2$ and get general linear model.

The general linear model encompasses...

Interaction effects

Example.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

Let $x_3 = x_1 x_2$ and get general linear model.

All of these models are *linear in the coefficients*, the β_j terms.

The general linear model encompasses...

Interaction effects

Example.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

Let $x_3 = x_1 x_2$ and get general linear model.

All of these models are *linear in the coefficients*, the β_j terms. An example of a model that is *not* in general linear model form is exponential growth:

$$Y = \beta_0 \exp(\beta_1 x) + \epsilon.$$

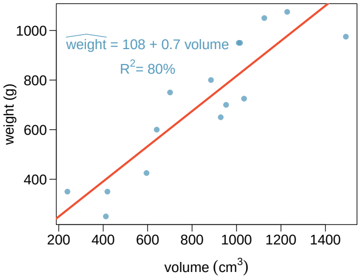
Another example with a binary predictor – weights of books

	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")

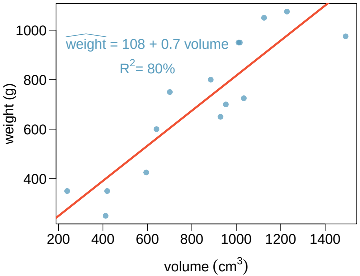
Weights of books (cont.)

The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



Weights of books (cont.)

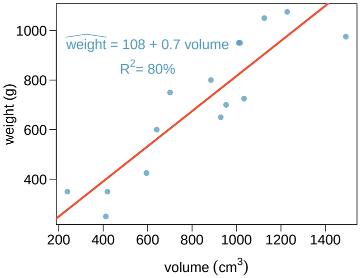
The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



(a) Weights of 80% of the books can be predicted accurately using this model.

Weights of books (cont.)

The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?

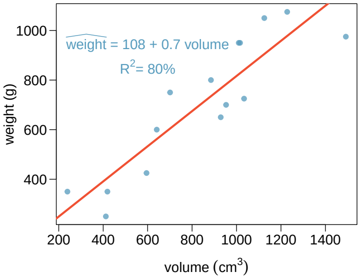


(a) Weights of 80% of the books can be predicted accurately using this model.

(b) Books that are 10 cm^3 over average are expected to weigh 7 g over average.

Weights of books (cont.)

The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



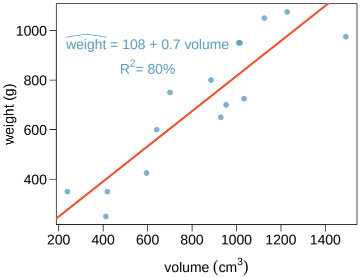
(a) Weights of 80% of the books can be predicted accurately using this model.

(b) Books that are 10 cm^3 over average are expected to weigh 7 g over average.

(c) The correlation between weight and volume is $R = 0.80^2 = 0.64$.

Weights of books (cont.)

The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



(a) Weights of 80% of the books can be predicted accurately using this model.

(b) Books that are 10 cm^3 over average are expected to weigh 7 g over average.

(c) The correlation between weight and volume is $R = 0.80^2 = 0.64$.

(d) The model underestimates the weight of the book with the highest volume.

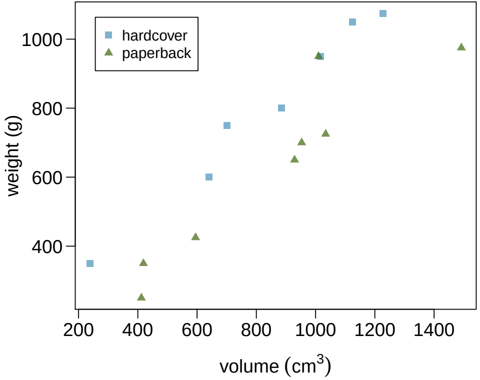
Modeling weights of books using only volume

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 107.67931   88.37758   1.218    0.245
volume        0.70864    0.09746   7.271 6.26e-06

Residual standard error: 123.9 on 13 degrees of freedom
Multiple R-squared: 0.8026,Adjusted R-squared: 0.7875
F-statistic: 52.87 on 1 and 13 DF,  p-value: 6.262e-06
```

Weights of hardcover and paperback books

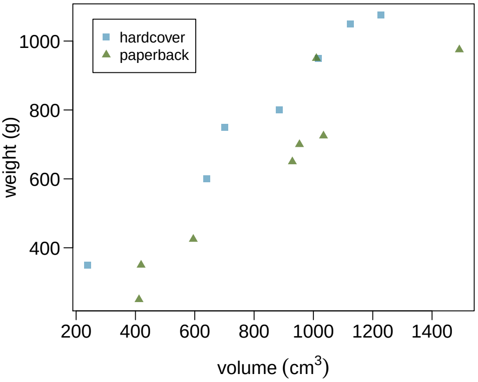
Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?



Weights of hardcover 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 after controlling for the books 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
```

Determining the reference level

Based on the regression output below, which level of `cover` is `the reference level? Note that pb: paperback.`

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) `paperback`

(b) `hardcover`

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

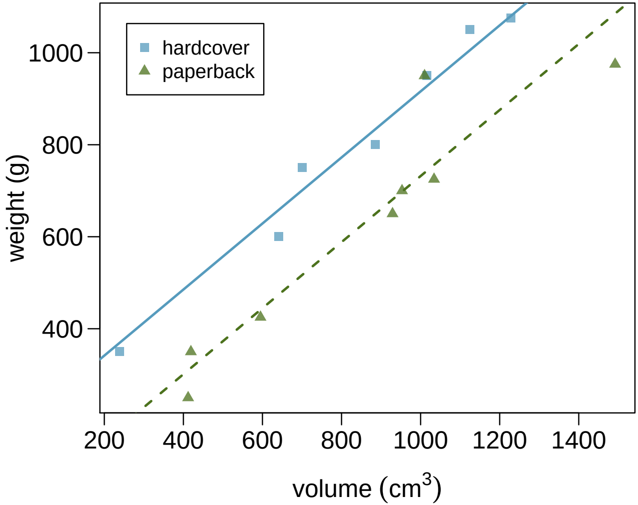
1. For *hardcover* books: plug in 0 for `cover`

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

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

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

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

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

- β_1:

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

- β_1: <u>All else held constant</u>, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more, 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

- β_1: All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more, on average.
- β_2:

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

- β_1: All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more, on average.
- β_2: All else held constant, the model predicts that paperback books weigh 184 grams lower 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

- β_1: All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more, on average.
- β_2: All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books, on average.
- β_0:

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

- β_1: All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more, on average.
- β_2: All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books, on average.
- β_0: 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

- β_1: <u>All else held constant</u>, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more, on average.
- β_2: <u>All else held constant</u>, the model predicts that paperback books weigh 184 grams lower than hardcover books, on average.
- β_0: 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

- β_1: All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more, on average.
- β_2: All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books, on average.
- β_0: Hardcover books with no volume are expected on average to weigh 198 grams.
  - Obviously, the intercept does not make sense in the context.

Prediction

Which of the following is the correct calculation for the predicted weight of a paperback book that is 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) 197.96 + 0.72 * 600 - 184.05 * 1

(b) 184.05 + 0.72 * 600 - 197.96 * 1

(c) 197.96 + 0.72 * 600 - 184.05 * 0

(d) 197.96 + 0.72 * 1 - 184.05 * 600

Prediction

Which of the following is the correct calculation for the predicted weight of a paperback book that is 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) 197.96 + 0.72 * 600 - 184.05 * 1 = 445.91 grams

(b) 184.05 + 0.72 * 600 - 197.96 * 1

(c) 197.96 + 0.72 * 600 - 184.05 * 0

(d) 197.96 + 0.72 * 1 - 184.05 * 600

6.2 General linear model in matrix terms

Response vector:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

6.2 General linear model in matrix terms

Response vector:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Design matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

The first column is a place-holder for the intercept term.

6.2 General linear model in matrix terms

Response vector:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Design matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

The first column is a place-holder for the intercept term.
What does each column represent?

6.2 General linear model in matrix terms

Response vector:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Design matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

The first column is a place-holder for the intercept term. What does each column represent? What does each row represent?

General linear model in matrix terms

(Unknown) regression coefficients:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

(Unobserved) error vector:

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

General linear model in matrix terms

The general linear model is written in matrix terms as

$$\mathbf{Y} = \underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}_{n\times 1} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}}_{n\times p} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{p\times 1} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{n\times 1},$$

where $p = k + 1$, or in short as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon.$$

General linear model in matrix terms

Minimal assumptions about the random error vector $\boldsymbol{\epsilon}$ are

$$\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0} \quad \text{and} \quad \mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma^2 \mathbf{I}_n,$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

General linear model in matrix terms

Minimal assumptions about the random error vector $\boldsymbol{\epsilon}$ are

$$\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0} \quad \text{and} \quad \mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma^2 \mathbf{I}_n,$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

In general, we will require more and assume

$$\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

General linear model in matrix terms

Minimal assumptions about the random error vector $\boldsymbol{\epsilon}$ are

$$\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0} \quad \text{and} \quad \mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma^2 \mathbf{I}_n,$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

In general, we will require more and assume

$$\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

This allows us to construct t and F tests, obtain confidence intervals, etc.

Fitting the model

Estimating $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$

Fitting the model

Estimating $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$

Recall the least-squares method:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})]^2 \longrightarrow \min_{\boldsymbol{\beta}}$$

Fitting the model

Estimating $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$

Recall the least-squares method:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})]^2 \longrightarrow \min_{\boldsymbol{\beta}}$$

Using matrix calculus we can show that the LSEs are

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

Fitting the model

Estimating $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$

Recall the least-squares method:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})]^2 \longrightarrow \min_{\boldsymbol{\beta}}$$

Using matrix calculus we can show that the LSEs are

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

These are also MLEs.

Fitted values, residuals, and hat matrix

The *fitted values* are

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \mathbf{Xb}$$

Fitted values, residuals, and hat matrix

The *fitted values* are

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \mathbf{X}\mathbf{b} = \underbrace{[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]}_{\text{projection matrix}}\mathbf{Y} = \mathbf{H}\mathbf{Y}.$$

Fitted values, residuals, and hat matrix

The *fitted values* are

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \mathbf{X}\mathbf{b} = \underbrace{[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]}_{\text{projection matrix}}\mathbf{Y} = \mathbf{H}\mathbf{Y}.$$

The *residuals* are

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Fitted values, residuals, and hat matrix

The *fitted values* are

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \mathbf{X}\mathbf{b} = \underbrace{[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]}_{\text{projection matrix}}\mathbf{Y} = \mathbf{H}\mathbf{Y}.$$

The *residuals* are

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{Y} - \hat{\mathbf{Y}}$$

Fitted values, residuals, and hat matrix

The *fitted values* are

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \mathbf{X}\mathbf{b} = \underbrace{[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]}_{\text{projection matrix}}\mathbf{Y} = \mathbf{H}\mathbf{Y}.$$

The *residuals* are

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b}$$

Fitted values, residuals, and hat matrix

The *fitted values* are

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \mathbf{X}\mathbf{b} = \underbrace{[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]}_{\text{projection matrix}}\mathbf{Y} = \mathbf{H}\mathbf{Y}.$$

The *residuals* are

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b} = [\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]\mathbf{Y}.$$

Fitted values, residuals, and hat matrix

The *fitted values* are

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \mathbf{X}\mathbf{b} = \underbrace{[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]}_{\text{projection matrix}}\mathbf{Y} = \mathbf{H}\mathbf{Y}.$$

The *residuals* are

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b} = [\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]\mathbf{Y}.$$

$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is called the **hat matrix** or **projection matrix**. We'll use it shortly when we talk about diagnostics. Notice that $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$.

Example: Dwayne Studios

In the **Dwayne Portrait Studio data,** we have

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix},$$

Example: Dwayne Studios

In the **Dwayne Portrait Studio data,** we have

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix},$$

so the fitted regression line is

$$\hat{Y} = -68.857 + 1.455 \cdot x_1 + 9.366 \cdot x_2,$$

Example: Dwayne Studios

In the **Dwayne Portrait Studio data,** we have

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix},$$

so the fitted regression line is

$$\hat{Y} = -68.857 + 1.455 \cdot x_1 + 9.366 \cdot x_2,$$

where x_1 is # children in thousands, x_2 is income in thousand \$'s.

Example: Dwayne Studios

In the **Dwayne Portrait Studio data,** we have

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix},$$

so the fitted regression line is

$$\hat{Y} = -68.857 + 1.455 \cdot x_1 + 9.366 \cdot x_2,$$

where x_1 is # children in thousands, x_2 is income in thousand $'s.

- b_1:

Example: Dwayne Studios

In the **Dwayne Portrait Studio data,** we have

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix},$$

so the fitted regression line is

$$\hat{Y} = -68.857 + 1.455 \cdot x_1 + 9.366 \cdot x_2,$$

where x_1 is # children in thousands, x_2 is income in thousand \$'s.

- b_1: For 1000 increase in number of children, mean sales increase by \$1,455 holding per capita income constant.

Example: Dwayne Studios

In the **Dwayne Portrait Studio data,** we have

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix},$$

so the fitted regression line is

$$\hat{Y} = -68.857 + 1.455 \cdot x_1 + 9.366 \cdot x_2,$$

where x_1 is # children in thousands, x_2 is income in thousand $'s.

- b_1: For 1000 increase in number of children, mean sales increase by $1,455 holding per capita income constant.
- b_2:

Example: Dwayne Studios

In the **Dwayne Portrait Studio data,** we have

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix},$$

so the fitted regression line is

$$\hat{Y} = -68.857 + 1.455 \cdot x_1 + 9.366 \cdot x_2,$$

where x_1 is # children in thousands, x_2 is income in thousand \$'s.

- b_1: For 1000 increase in number of children, mean sales increase by \$1,455 holding per capita income constant.
- b_2: For each \$1000 increase in per capita income, mean sales increase by \$9,366, holding the number of children constant.

Partitioning SST

Recall the decomposition

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}}$$

Partitioning SST

Recall the decomposition

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}}$$

Notice that

$$\text{SSE} = (\mathbf{Y} - \hat{\mathbf{Y}})^\top(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y}^\top(\mathbf{I} - \mathbf{H})^\top(\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{Y}^\top(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Partitioning SST

Recall the decomposition

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}}$$

Notice that

$$\text{SSE} = (\mathbf{Y} - \hat{\mathbf{Y}})^\top(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y}^\top(\mathbf{I} - \mathbf{H})^\top(\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{Y}^\top(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\text{SSR} = (\hat{\mathbf{Y}} - \bar{\mathbf{Y}})^\top(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}) = \mathbf{Y}^\top\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)^\top\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}$$

$$= \mathbf{Y}^\top\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}$$

Partitioning SST

Recall the decomposition

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}}$$

Notice that

$$\text{SSE} = (\mathbf{Y} - \hat{\mathbf{Y}})^\top(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y}^\top(\mathbf{I} - \mathbf{H})^\top(\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{Y}^\top(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\begin{aligned}\text{SSR} = (\hat{\mathbf{Y}} - \bar{\mathbf{Y}})^\top(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}) &= \mathbf{Y}^\top\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)^\top\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y} \\ &= \mathbf{Y}^\top\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}\end{aligned}$$

where $\frac{1}{n}\mathbf{J} = \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top$

Partitioning SST

Recall the decomposition

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}}$$

Notice that

$$\text{SSE} = (\mathbf{Y} - \hat{\mathbf{Y}})^\top(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y}^\top(\mathbf{I} - \mathbf{H})^\top(\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{Y}^\top(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\begin{aligned}\text{SSR} &= (\hat{\mathbf{Y}} - \bar{\mathbf{Y}})^\top(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}) = \mathbf{Y}^\top\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)^\top\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y} \\ &= \mathbf{Y}^\top\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}\end{aligned}$$

where $\frac{1}{n}\mathbf{J} = \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top$, and we used symmetry and idempotence of $\mathbf{I} - \mathbf{H}$ and $\mathbf{H} - \frac{1}{n}\mathbf{J}$ (Ch 5).

Distribution of SSE

Theorem. $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p}$

Proof.
$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$

Distribution of SSE

Theorem. $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p}$

Proof.

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \quad \Rightarrow \quad \sigma^{-1}\mathbf{Y} \sim \mathcal{N}_n(\sigma^{-1}\mathbf{X}\boldsymbol{\beta}, \mathbf{I}).$$

Distribution of SSE

Theorem. $\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2_{n-p}$

Proof.

$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \quad \Rightarrow \quad \sigma^{-1}\mathbf{Y} \sim \mathcal{N}_n(\sigma^{-1}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$. Hence,

$$\frac{\mathrm{SSE}}{\sigma^2} = (\sigma^{-1}\mathbf{Y})^\top(\mathbf{I} - \mathbf{H})(\sigma^{-1}\mathbf{Y})$$

Distribution of SSE

Theorem. $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p}$

Proof.
$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \quad \Rightarrow \quad \sigma^{-1}\mathbf{Y} \sim \mathcal{N}_n(\sigma^{-1}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$. Hence,

$$\frac{\text{SSE}}{\sigma^2} = (\sigma^{-1}\mathbf{Y})^\top(\mathbf{I} - \mathbf{H})(\sigma^{-1}\mathbf{Y}) \sim \chi^2_{\mathrm{rank}[\mathbf{I}-\mathbf{H}]}(\lambda),$$

Distribution of SSE

Theorem. $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p}$

Proof.

$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \quad \Rightarrow \quad \sigma^{-1}\mathbf{Y} \sim \mathcal{N}_n(\sigma^{-1}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$. Hence,

$$\frac{\text{SSE}}{\sigma^2} = (\sigma^{-1}\mathbf{Y})^\top(\mathbf{I} - \mathbf{H})(\sigma^{-1}\mathbf{Y}) \sim \chi^2_{\mathrm{rank}[\mathbf{I}-\mathbf{H}]}(\lambda),$$

where $\quad \mathrm{rank}[\mathbf{I} - \mathbf{H}] = \mathrm{trace}[\mathbf{I} - \mathbf{H}]$

Distribution of SSE

Theorem. $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p}$

Proof.

$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \quad \Rightarrow \quad \sigma^{-1}\mathbf{Y} \sim \mathcal{N}_n(\sigma^{-1}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$. Hence,

$$\frac{\text{SSE}}{\sigma^2} = (\sigma^{-1}\mathbf{Y})^\top(\mathbf{I} - \mathbf{H})(\sigma^{-1}\mathbf{Y}) \sim \chi^2_{\mathrm{rank}[\mathbf{I}-\mathbf{H}]}(\lambda),$$

where $\quad \mathrm{rank}[\mathbf{I} - \mathbf{H}] = \mathrm{trace}[\mathbf{I} - \mathbf{H}] = \mathrm{trace}[\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]$

Distribution of SSE

Theorem. $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p}$

Proof.

$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \quad \Rightarrow \quad \sigma^{-1}\mathbf{Y} \sim \mathcal{N}_n(\sigma^{-1}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$. Hence,

$$\frac{\text{SSE}}{\sigma^2} = (\sigma^{-1}\mathbf{Y})^\top(\mathbf{I} - \mathbf{H})(\sigma^{-1}\mathbf{Y}) \sim \chi^2_{\mathrm{rank}[\mathbf{I}-\mathbf{H}]}(\lambda),$$

$$\text{where} \quad \mathrm{rank}[\mathbf{I} - \mathbf{H}] = \mathrm{trace}[\mathbf{I} - \mathbf{H}] = \mathrm{trace}[\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]$$
$$= \mathrm{trace}\mathbf{I} - \mathrm{trace}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]$$

Distribution of SSE

Theorem. $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p}$

Proof.

$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \quad \Rightarrow \quad \sigma^{-1}\mathbf{Y} \sim \mathcal{N}_n(\sigma^{-1}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$. Hence,

$$\frac{\text{SSE}}{\sigma^2} = (\sigma^{-1}\mathbf{Y})^\top(\mathbf{I}-\mathbf{H})(\sigma^{-1}\mathbf{Y}) \sim \chi^2_{\mathrm{rank}[\mathbf{I}-\mathbf{H}]}(\lambda),$$

$$\begin{aligned}\text{where} \quad \mathrm{rank}[\mathbf{I}-\mathbf{H}] &= \mathrm{trace}[\mathbf{I}-\mathbf{H}] = \mathrm{trace}[\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top] \\ &= \mathrm{trace}\mathbf{I} - \mathrm{trace}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top] \\ &= n - \mathrm{trace}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}]\end{aligned}$$

Distribution of SSE

Theorem. $\frac{\mathsf{SSE}}{\sigma^2} \sim \chi^2_{n-p}$

Proof.
$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \quad \Rightarrow \quad \sigma^{-1}\mathbf{Y} \sim \mathcal{N}_n(\sigma^{-1}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$. Hence,

$$\begin{aligned}
\frac{\mathsf{SSE}}{\sigma^2} &= (\sigma^{-1}\mathbf{Y})^\top(\mathbf{I}-\mathbf{H})(\sigma^{-1}\mathbf{Y}) \sim \chi^2_{\mathrm{rank}[\mathbf{I}-\mathbf{H}]}(\lambda), \\
\text{where} \quad \mathrm{rank}[\mathbf{I}-\mathbf{H}] &= \mathrm{trace}[\mathbf{I}-\mathbf{H}] = \mathrm{trace}[\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top] \\
&= \mathrm{trace}\mathbf{I} - \mathrm{trace}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top] \\
&= n - \mathrm{trace}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}] = n - p,
\end{aligned}$$

Distribution of SSE

Theorem. $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p}$

Proof.
$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}) \quad \Rightarrow \quad \sigma^{-1}\mathbf{Y} \sim \mathcal{N}_n(\sigma^{-1}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$. Hence,

$$
\begin{aligned}
\frac{\text{SSE}}{\sigma^2} &= (\sigma^{-1}\mathbf{Y})^\top(\mathbf{I}-\mathbf{H})(\sigma^{-1}\mathbf{Y}) \sim \chi^2_{\mathrm{rank}[\mathbf{I}-\mathbf{H}]}(\lambda), \\
\text{where} \quad \mathrm{rank}[\mathbf{I}-\mathbf{H}] &= \mathrm{trace}[\mathbf{I}-\mathbf{H}] = \mathrm{trace}[\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top] \\
&= \mathrm{trace}\,\mathbf{I} - \mathrm{trace}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top] \\
&= n - \mathrm{trace}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}] = n-p, \\
\text{and} \quad \lambda &= (\sigma^{-1}\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{I}-\mathbf{H})(\sigma^{-1}\mathbf{X}\boldsymbol{\beta})
\end{aligned}
$$

Distribution of SSE

Theorem. $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p}$

Proof.

$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \quad \Rightarrow \quad \sigma^{-1}\mathbf{Y} \sim \mathcal{N}_n(\sigma^{-1}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$. Hence,

$$
\begin{aligned}
\frac{\text{SSE}}{\sigma^2} &= (\sigma^{-1}\mathbf{Y})^\top(\mathbf{I}-\mathbf{H})(\sigma^{-1}\mathbf{Y}) \sim \chi^2_{\mathrm{rank}[\mathbf{I}-\mathbf{H}]}(\lambda), \\
\text{where} \quad \mathrm{rank}[\mathbf{I}-\mathbf{H}] &= \mathrm{trace}[\mathbf{I}-\mathbf{H}] = \mathrm{trace}[\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top] \\
&= \mathrm{trace}\mathbf{I} - \mathrm{trace}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top] \\
&= n - \mathrm{trace}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}] = n-p, \\
\text{and} \quad \lambda &= (\sigma^{-1}\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{I}-\mathbf{H})(\sigma^{-1}\mathbf{X}\boldsymbol{\beta}) \\
&= \frac{1}{\sigma^2}\boldsymbol{\beta}^\top\mathbf{X}^\top[\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]\mathbf{X}\boldsymbol{\beta}
\end{aligned}
$$

Distribution of SSE

Theorem. $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p}$

Proof.

$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \quad \Rightarrow \quad \sigma^{-1}\mathbf{Y} \sim \mathcal{N}_n(\sigma^{-1}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$. Hence,

$$
\begin{aligned}
\frac{\text{SSE}}{\sigma^2} &= (\sigma^{-1}\mathbf{Y})^\top(\mathbf{I}-\mathbf{H})(\sigma^{-1}\mathbf{Y}) \sim \chi^2_{\mathrm{rank}[\mathbf{I}-\mathbf{H}]}(\lambda), \\
\text{where} \quad \mathrm{rank}[\mathbf{I}-\mathbf{H}] &= \mathrm{trace}[\mathbf{I}-\mathbf{H}] = \mathrm{trace}[\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top] \\
&= \mathrm{trace}\mathbf{I} - \mathrm{trace}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top] \\
&= n - \mathrm{trace}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}] = n-p, \\
\text{and} \quad \lambda &= (\sigma^{-1}\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{I}-\mathbf{H})(\sigma^{-1}\mathbf{X}\boldsymbol{\beta}) \\
&= \frac{1}{\sigma^2}\boldsymbol{\beta}^\top\mathbf{X}^\top[\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]\mathbf{X}\boldsymbol{\beta} \\
&= \frac{1}{\sigma^2}\boldsymbol{\beta}^\top[\mathbf{X}^\top\mathbf{X}-\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}]\boldsymbol{\beta}
\end{aligned}
$$

Distribution of SSE

Theorem. $\frac{\mathsf{SSE}}{\sigma^2} \sim \chi^2_{n-p}$

Proof.

$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \quad \Rightarrow \quad \sigma^{-1}\mathbf{Y} \sim \mathcal{N}_n(\sigma^{-1}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$. Hence,

$$
\begin{aligned}
\frac{\mathsf{SSE}}{\sigma^2} &= (\sigma^{-1}\mathbf{Y})^\top(\mathbf{I}-\mathbf{H})(\sigma^{-1}\mathbf{Y}) \sim \chi^2_{\mathrm{rank}[\mathbf{I}-\mathbf{H}]}(\lambda), \\
\text{where} \quad \mathrm{rank}[\mathbf{I}-\mathbf{H}] &= \mathrm{trace}[\mathbf{I}-\mathbf{H}] = \mathrm{trace}[\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top] \\
&= \mathrm{trace}\,\mathbf{I} - \mathrm{trace}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top] \\
&= n - \mathrm{trace}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}] = n - p, \\
\text{and} \quad \lambda &= (\sigma^{-1}\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{I}-\mathbf{H})(\sigma^{-1}\mathbf{X}\boldsymbol{\beta}) \\
&= \frac{1}{\sigma^2}\boldsymbol{\beta}^\top\mathbf{X}^\top[\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]\mathbf{X}\boldsymbol{\beta} \\
&= \frac{1}{\sigma^2}\boldsymbol{\beta}^\top[\mathbf{X}^\top\mathbf{X} - \mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}]\boldsymbol{\beta} = 0
\end{aligned}
$$

Some properties of the hat matrix

Lemma. $\mathbf{H1} = \mathbf{1}$

Proof.

$\square$

Remark.

Some properties of the hat matrix

Lemma. $\mathbf{H}\mathbf{1} = \mathbf{1}$

Proof.

$$\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X} = \mathbf{X}$$

$\square$

Remark.

Some properties of the hat matrix

Lemma. $\mathbf{H1} = \mathbf{1}$

Proof.

$$\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X} = \mathbf{X}$$

$$\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix} = \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}$$

$\square$

Remark.

Some properties of the hat matrix

Lemma. $\mathbf{H1} = \mathbf{1}$

Proof.

$$
\begin{aligned}
\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X} &= \mathbf{X} \\
\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix} &= \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix} \\
\begin{bmatrix}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{1} & \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\tilde{\mathbf{X}}\end{bmatrix} &= \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}
\end{aligned}
$$

□

Remark.

Some properties of the hat matrix

Lemma. $\mathbf{H1} = \mathbf{1}$

Proof.

$$\begin{aligned}
\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X} &= \mathbf{X} \\
\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix} &= \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix} \\
\begin{bmatrix}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{1} & \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\tilde{\mathbf{X}}\end{bmatrix} &= \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}
\end{aligned}$$

Hence, $\mathbf{H1} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{1} = \mathbf{1}$. □

Remark.

Some properties of the hat matrix

Lemma. $\mathbf{H}\mathbf{1} = \mathbf{1}$

Proof.

$$\begin{aligned}
\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X} &= \mathbf{X} \\
\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix} &= \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix} \\
\begin{bmatrix}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{1} & \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\tilde{\mathbf{X}}\end{bmatrix} &= \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}
\end{aligned}$$

Hence, $\mathbf{H}\mathbf{1} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{1} = \mathbf{1}$. $\square$

Remark. $\mathbf{H}\tilde{\mathbf{X}} = \tilde{\mathbf{X}}$

Some properties of the hat matrix

Lemma. $\mathbf{H}\mathbf{1} = \mathbf{1}$

Proof.

$$\begin{aligned}
\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X} &= \mathbf{X} \\
\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix} &= \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix} \\
\begin{bmatrix}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{1} & \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\tilde{\mathbf{X}}\end{bmatrix} &= \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}
\end{aligned}$$

Hence, $\mathbf{H}\mathbf{1} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{1} = \mathbf{1}$. $\square$

Remark. $\mathbf{H}\tilde{\mathbf{X}} = \tilde{\mathbf{X}}$, $\mathbf{1}^\top\mathbf{H} = \mathbf{1}^\top$

Some properties of the hat matrix

Lemma. $\mathbf{H}\mathbf{1} = \mathbf{1}$

Proof.

$$\begin{aligned}
\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X} &= \mathbf{X} \\
\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix} &= \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix} \\
\begin{bmatrix}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{1} & \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\tilde{\mathbf{X}}\end{bmatrix} &= \begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}
\end{aligned}$$

Hence, $\mathbf{H}\mathbf{1} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{1} = \mathbf{1}$. $\square$

Remark. $\mathbf{H}\tilde{\mathbf{X}} = \tilde{\mathbf{X}}$, $\mathbf{1}^\top\mathbf{H} = \mathbf{1}^\top$, $\tilde{\mathbf{X}}^\top\mathbf{H} = \tilde{\mathbf{X}}^\top$

Distribution of SSR

Theorem. $\frac{\mathrm{SSR}}{\sigma^2} \sim \chi^2_{p-1}(\lambda)$

Distribution of SSR

Theorem. $\frac{\text{SSR}}{\sigma^2} \sim \chi^2_{p-1}(\lambda)$ with $\lambda = \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}} \tilde{\boldsymbol{\beta}}$,

Distribution of SSR

Theorem. $\frac{\mathrm{SSR}}{\sigma^2} \sim \chi^2_{p-1}(\lambda)$ with $\lambda = \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}} \tilde{\boldsymbol{\beta}}$, where

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix},$$

Distribution of SSR

Theorem. $\frac{\text{SSR}}{\sigma^2} \sim \chi^2_{p-1}(\lambda)$ with $\lambda = \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}}\tilde{\boldsymbol{\beta}}$, where

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \bar{\mathbf{X}} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{bmatrix}$$

Distribution of SSR

Theorem. $\frac{\text{SSR}}{\sigma^2} \sim \chi^2_{p-1}(\lambda)$ with $\lambda = \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}} \tilde{\boldsymbol{\beta}}$, where

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \bar{\mathbf{X}} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{bmatrix}$$

and $\bar{x}_j := \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ is the average value of the j^{th} predictor.

Proof.

Distribution of SSR

Theorem. $\frac{\text{SSR}}{\sigma^2} \sim \chi^2_{p-1}(\lambda)$ with $\lambda = \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}}\tilde{\boldsymbol{\beta}}$, where

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \bar{\mathbf{X}} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{bmatrix}$$

and $\bar{x}_j := \frac{1}{n}\sum_{i=1}^n x_{ij}$ is the average value of the j^{th} predictor.

Proof.
The proof is analogous to the case of $k = 1$.

Distribution of SSR

Theorem. $\frac{\text{SSR}}{\sigma^2} \sim \chi^2_{p-1}(\lambda)$ with $\lambda = \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}} \tilde{\boldsymbol{\beta}}$, where

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \bar{\mathbf{X}} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{bmatrix}$$

and $\bar{x}_j := \frac{1}{n}\sum_{i=1}^n x_{ij}$ is the average value of the j^{th} predictor.

Proof.
The proof is analogous to the case of $k = 1$. Let's focus on λ.

Distribution of SSR

Theorem. $\frac{\text{SSR}}{\sigma^2} \sim \chi^2_{p-1}(\lambda)$ with $\lambda = \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}}\tilde{\boldsymbol{\beta}}$, where

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \bar{\mathbf{X}} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{bmatrix}$$

and $\bar{x}_j := \frac{1}{n}\sum_{i=1}^n x_{ij}$ is the average value of the j^{th} predictor.

Proof.
The proof is analogous to the case of $k = 1$. Let's focus on λ.

Let $\tilde{\mathbf{X}} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}$,

Distribution of SSR

Theorem. $\frac{\text{SSR}}{\sigma^2} \sim \chi^2_{p-1}(\lambda)$ with $\lambda = \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}}\tilde{\boldsymbol{\beta}}$, where

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \bar{\mathbf{X}} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{bmatrix}$$

and $\bar{x}_j := \frac{1}{n}\sum_{i=1}^n x_{ij}$ is the average value of the j^{th} predictor.

Proof.
The proof is analogous to the case of $k = 1$. Let's focus on λ.

Let $\tilde{\mathbf{X}} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}$, then $\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{1} & \tilde{\mathbf{X}} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \tilde{\boldsymbol{\beta}} \end{bmatrix} = \beta_0 \mathbf{1} + \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}$

Distribution of SSR

Proof (cont'd).

Since $\frac{\mathrm{SSR}}{\sigma^2} = \left(\frac{1}{\sigma}\mathbf{Y}\right)^\top \left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\left(\frac{1}{\sigma}\mathbf{Y}\right)$ and $\frac{1}{\sigma}\mathbf{Y} \sim \mathcal{N}_n\left(\frac{1}{\sigma}\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\right)$,

Distribution of SSR

Proof (cont'd).

Since $\frac{\text{SSR}}{\sigma^2} = (\frac{1}{\sigma}\mathbf{Y})^\top \left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right) (\frac{1}{\sigma}\mathbf{Y})$ and $\frac{1}{\sigma}\mathbf{Y} \sim \mathcal{N}_n(\frac{1}{\sigma}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$,

$$\lambda = \frac{1}{\sigma^2}\boldsymbol{\beta}^\top\mathbf{X}^\top[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top]\mathbf{X}\boldsymbol{\beta}$$

Distribution of SSR

Proof (cont'd).

Since $\frac{\text{SSR}}{\sigma^2} = (\frac{1}{\sigma}\mathbf{Y})^\top \left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right) \left(\frac{1}{\sigma}\mathbf{Y}\right)$ and $\frac{1}{\sigma}\mathbf{Y} \sim \mathcal{N}_n(\frac{1}{\sigma}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$,

$$\begin{aligned}
\lambda &= \frac{1}{\sigma^2}\boldsymbol{\beta}^\top\mathbf{X}^\top[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top]\mathbf{X}\boldsymbol{\beta} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}^\top \\ \tilde{\mathbf{X}}^\top\end{bmatrix}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top]\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix}
\end{aligned}$$

Distribution of SSR

Proof (cont'd).

Since $\frac{\text{SSR}}{\sigma^2} = (\frac{1}{\sigma}\mathbf{Y})^\top \left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\left(\frac{1}{\sigma}\mathbf{Y}\right)$ and $\frac{1}{\sigma}\mathbf{Y} \sim \mathcal{N}_n(\frac{1}{\sigma}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$,

$$\begin{aligned}
\lambda &= \frac{1}{\sigma^2}\boldsymbol{\beta}^\top\mathbf{X}^\top[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top]\mathbf{X}\boldsymbol{\beta} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}^\top \\ \tilde{\mathbf{X}}^\top\end{bmatrix}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top]\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top \\ \tilde{\mathbf{X}}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \tilde{\mathbf{X}}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix}
\end{aligned}$$

Distribution of SSR

Proof (cont'd).

Since $\frac{\text{SSR}}{\sigma^2} = (\frac{1}{\sigma}\mathbf{Y})^\top \left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right) \left(\frac{1}{\sigma}\mathbf{Y}\right)$ and $\frac{1}{\sigma}\mathbf{Y} \sim \mathcal{N}_n(\frac{1}{\sigma}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$,

$$
\begin{aligned}
\lambda &= \frac{1}{\sigma^2}\boldsymbol{\beta}^\top\mathbf{X}^\top[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top]\mathbf{X}\boldsymbol{\beta} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}^\top \\ \tilde{\mathbf{X}}^\top\end{bmatrix}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top]\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top \\ \tilde{\mathbf{X}}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \tilde{\mathbf{X}}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}^\top - \mathbf{1}^\top \\ \tilde{\mathbf{X}}^\top - \tilde{\mathbf{X}}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix}
\end{aligned}
$$

Distribution of SSR

Proof (cont'd).

Since $\frac{\text{SSR}}{\sigma^2} = \left(\frac{1}{\sigma}\mathbf{Y}\right)^\top \left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\left(\frac{1}{\sigma}\mathbf{Y}\right)$ and $\frac{1}{\sigma}\mathbf{Y} \sim \mathcal{N}_n\left(\frac{1}{\sigma}\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\right)$,

$$
\begin{aligned}
\lambda &= \frac{1}{\sigma^2}\boldsymbol{\beta}^\top\mathbf{X}^\top[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top]\mathbf{X}\boldsymbol{\beta} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}^\top \\ \tilde{\mathbf{X}}^\top\end{bmatrix}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top]\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top \\ \tilde{\mathbf{X}}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \tilde{\mathbf{X}}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}^\top - \mathbf{1}^\top \\ \tilde{\mathbf{X}}^\top - \tilde{\mathbf{X}}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}0 & \mathbf{0}^\top \\ \mathbf{0} & \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} - \tilde{\mathbf{X}}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix}
\end{aligned}
$$

Distribution of SSR

Proof (cont'd).

Since $\frac{\text{SSR}}{\sigma^2} = (\frac{1}{\sigma}\mathbf{Y})^\top \left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\left(\frac{1}{\sigma}\mathbf{Y}\right)$ and $\frac{1}{\sigma}\mathbf{Y} \sim \mathcal{N}_n(\frac{1}{\sigma}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$,

$$\begin{aligned}
\lambda &= \frac{1}{\sigma^2}\boldsymbol{\beta}^\top\mathbf{X}^\top[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top]\mathbf{X}\boldsymbol{\beta} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}^\top \\ \tilde{\mathbf{X}}^\top\end{bmatrix}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top]\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top \\ \tilde{\mathbf{X}}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \tilde{\mathbf{X}}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}^\top - \mathbf{1}^\top \\ \tilde{\mathbf{X}}^\top - \tilde{\mathbf{X}}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}0 & \mathbf{0}^\top \\ \mathbf{0} & \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} - \tilde{\mathbf{X}}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix} \\
&= \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top[\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} - \tilde{\mathbf{X}}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\tilde{\mathbf{X}}]\tilde{\boldsymbol{\beta}}
\end{aligned}$$

Distribution of SSR

Proof (cont'd).

Since $\frac{\text{SSR}}{\sigma^2} = \left(\frac{1}{\sigma}\mathbf{Y}\right)^\top \left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\left(\frac{1}{\sigma}\mathbf{Y}\right)$ and $\frac{1}{\sigma}\mathbf{Y} \sim \mathcal{N}_n(\frac{1}{\sigma}\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$,

$$
\begin{aligned}
\lambda &= \frac{1}{\sigma^2}\boldsymbol{\beta}^\top\mathbf{X}^\top[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top]\mathbf{X}\boldsymbol{\beta} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}^\top \\ \tilde{\mathbf{X}}^\top\end{bmatrix}[\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top]\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \mathbf{1}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top \\ \tilde{\mathbf{X}}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - \tilde{\mathbf{X}}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1}^\top - \mathbf{1}^\top \\ \tilde{\mathbf{X}}^\top - \tilde{\mathbf{X}}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\end{bmatrix}\begin{bmatrix}\mathbf{1} & \tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix}\beta_0 & \tilde{\boldsymbol{\beta}}^\top\end{bmatrix}\begin{bmatrix}0 & \mathbf{0}^\top \\ \mathbf{0} & \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} - \tilde{\mathbf{X}}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\tilde{\mathbf{X}}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \tilde{\boldsymbol{\beta}}\end{bmatrix} \\
&= \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top[\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} - \tilde{\mathbf{X}}^\top\mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top\tilde{\mathbf{X}}]\tilde{\boldsymbol{\beta}} = \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top\bar{\mathbf{X}}^\top\bar{\mathbf{X}}\tilde{\boldsymbol{\beta}}
\end{aligned}
$$

Means of SSE and SSR

Theorem. Let $\mathsf{MSE} := \frac{\mathsf{SSE}}{n-p}$, then $\mathrm{E}[\mathsf{MSE}] = \sigma^2$.

Proof.

Means of SSE and SSR

Theorem. Let $\text{MSE} := \frac{\text{SSE}}{n-p}$, then $\mathrm{E}[\text{MSE}] = \sigma^2$.

Proof.
$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p}$

Means of SSE and SSR

Theorem. Let $\text{MSE} := \frac{\text{SSE}}{n-p}$, then $\mathrm{E}[\text{MSE}] = \sigma^2$.

Proof.

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p} \quad \Rightarrow \quad \mathrm{E}\left[\frac{\text{SSE}}{\sigma^2}\right] = n - p$$

Means of SSE and SSR

Theorem. Let $\mathrm{MSE} := \frac{\mathrm{SSE}}{n-p}$, then $\mathrm{E}[\mathrm{MSE}] = \sigma^2$.

Proof.

$$\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2_{n-p} \quad \Rightarrow \quad \mathrm{E}\left[\frac{\mathrm{SSE}}{\sigma^2}\right] = n - p \quad \Rightarrow \quad \mathrm{E}\left[\frac{\mathrm{SSE}}{n-p}\right] = \sigma^2 \qquad \square$$

Means of SSE and SSR

Theorem. Let MSE $:= \frac{\text{SSE}}{n-p}$, then $\mathrm{E}[\text{MSE}] = \sigma^2$.

Proof.

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p} \quad \Rightarrow \quad \mathrm{E}\left[\frac{\text{SSE}}{\sigma^2}\right] = n-p \quad \Rightarrow \quad \mathrm{E}\left[\frac{\text{SSE}}{n-p}\right] = \sigma^2 \qquad \square$$

Theorem. Let MSR $:= \frac{\text{SSR}}{p-1}$, then $\mathrm{E}[\text{MSR}] = \sigma^2 + \frac{1}{p-1}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}} \tilde{\boldsymbol{\beta}}$

Proof.

Means of SSE and SSR

Theorem. Let $\mathsf{MSE} := \frac{\mathsf{SSE}}{n-p}$, then $\mathrm{E}[\mathsf{MSE}] = \sigma^2$.

Proof.
$$\frac{\mathsf{SSE}}{\sigma^2} \sim \chi^2_{n-p} \quad \Rightarrow \quad \mathrm{E}\left[\frac{\mathsf{SSE}}{\sigma^2}\right] = n-p \quad \Rightarrow \quad \mathrm{E}\left[\frac{\mathsf{SSE}}{n-p}\right] = \sigma^2 \qquad \square$$

Theorem. Let $\mathsf{MSR} := \frac{\mathsf{SSR}}{p-1}$, then $\mathrm{E}[\mathsf{MSR}] = \sigma^2 + \frac{1}{p-1}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}} \tilde{\boldsymbol{\beta}}$

Proof.

$\square$

Means of SSE and SSR

Theorem. Let $\text{MSE} := \frac{\text{SSE}}{n-p}$, then $\mathrm{E}[\text{MSE}] = \sigma^2$.

Proof.
$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p} \quad \Rightarrow \quad \mathrm{E}\left[\frac{\text{SSE}}{\sigma^2}\right] = n-p \quad \Rightarrow \quad \mathrm{E}\left[\frac{\text{SSE}}{n-p}\right] = \sigma^2 \qquad \square$$

Theorem. Let $\text{MSR} := \frac{\text{SSR}}{p-1}$, then $\mathrm{E}[\text{MSR}] = \sigma^2 + \frac{1}{p-1}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}} \tilde{\boldsymbol{\beta}}$

Proof.
$$\frac{\text{SSR}}{\sigma^2} \sim \chi^2_{p-1}(\lambda) \text{ with } \lambda = \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}} \tilde{\boldsymbol{\beta}}$$

$\square$

Means of SSE and SSR

Theorem. Let $\text{MSE} := \frac{\text{SSE}}{n-p}$, then $\mathrm{E}[\text{MSE}] = \sigma^2$.

Proof.

$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p} \quad \Rightarrow \quad \mathrm{E}\left[\frac{\text{SSE}}{\sigma^2}\right] = n - p \quad \Rightarrow \quad \mathrm{E}\left[\frac{\text{SSE}}{n-p}\right] = \sigma^2$ $\square$

Theorem. Let $\text{MSR} := \frac{\text{SSR}}{p-1}$, then $\mathrm{E}[\text{MSR}] = \sigma^2 + \frac{1}{p-1}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}}\tilde{\boldsymbol{\beta}}$

Proof.

$$\frac{\text{SSR}}{\sigma^2} \sim \chi^2_{p-1}(\lambda) \text{ with } \lambda = \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}}\tilde{\boldsymbol{\beta}}$$

$$\Rightarrow \quad \mathrm{E}\left[\frac{\text{SSR}}{\sigma^2}\right] = p - 1 + \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}}\tilde{\boldsymbol{\beta}}$$

$\square$

Means of SSE and SSR

Theorem. Let $\mathrm{MSE} := \frac{\mathrm{SSE}}{n-p}$, then $\mathrm{E}[\mathrm{MSE}] = \sigma^2$.

Proof.
$$\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2_{n-p} \quad \Rightarrow \quad \mathrm{E}\left[\frac{\mathrm{SSE}}{\sigma^2}\right] = n-p \quad \Rightarrow \quad \mathrm{E}\left[\frac{\mathrm{SSE}}{n-p}\right] = \sigma^2 \qquad \square$$

Theorem. Let $\mathrm{MSR} := \frac{\mathrm{SSR}}{p-1}$, then $\mathrm{E}[\mathrm{MSR}] = \sigma^2 + \frac{1}{p-1}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}} \tilde{\boldsymbol{\beta}}$

Proof.
$$\begin{aligned}
&\frac{\mathrm{SSR}}{\sigma^2} \sim \chi^2_{p-1}(\lambda) \text{ with } \lambda = \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}} \tilde{\boldsymbol{\beta}} \\
\Rightarrow \quad & \mathrm{E}\left[\frac{\mathrm{SSR}}{\sigma^2}\right] = p-1 + \frac{1}{\sigma^2}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}} \tilde{\boldsymbol{\beta}} \\
\Rightarrow \quad & \mathrm{E}\left[\mathrm{MSR}\right] = \sigma^2 + \frac{1}{p-1}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}} \tilde{\boldsymbol{\beta}}
\end{aligned}$$

$\square$

Analysis of variance

In multiple regression we can decompose the total sum of squares into the SSR and SSE pieces.

Analysis of variance

In multiple regression we can decompose the total sum of squares into the SSR and SSE pieces. The table is now

Source	SS	df	MS	$\mathrm{E}[MS]$
Regression	$\mathrm{SSR} = \sum(\hat{Y}_i - \bar{Y})^2$	$p-1$	$\frac{\mathrm{SSR}}{p-1}$	$\sigma^2 + \mathrm{QF}$
Error	$\mathrm{SSE} = \sum(Y_i - \hat{Y})^2$	$n-p$	$\frac{\mathrm{SSE}}{n-p}$	σ^2
Total	$\mathrm{SST} = \sum(Y_i - \bar{Y})^2$	$n-1$		

where $p = k + 1$.

Analysis of variance

In multiple regression we can decompose the total sum of squares into the SSR and SSE pieces. The table is now

Source	SS	df	MS	$\mathrm{E}[MS]$
Regression	$\mathrm{SSR} = \sum(\hat{Y}_i - \bar{Y})^2$	$p-1$	$\frac{\mathrm{SSR}}{p-1}$	$\sigma^2 + \mathrm{QF}$
Error	$\mathrm{SSE} = \sum(Y_i - \hat{Y})^2$	$n-p$	$\frac{\mathrm{SSE}}{n-p}$	σ^2
Total	$\mathrm{SST} = \sum(Y_i - \bar{Y})^2$	$n-1$		

where $p = k + 1$.

Here, QF stands for "quadratic form" and is given by

$$\mathrm{QF} = \frac{1}{p-1}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}}\tilde{\boldsymbol{\beta}} = \frac{1}{k}\sum_{j=1}^{k}\sum_{s=1}^{k}\beta_j\beta_s\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{is}-\bar{x}_s) \geq 0.$$

Analysis of variance

In multiple regression we can decompose the total sum of squares into the SSR and SSE pieces. The table is now

Source	SS	df	MS	$\mathrm{E}[MS]$
Regression	$\mathrm{SSR} = \sum(\hat{Y}_i - \bar{Y})^2$	$p-1$	$\frac{\mathrm{SSR}}{p-1}$	$\sigma^2 + \mathrm{QF}$
Error	$\mathrm{SSE} = \sum(Y_i - \hat{Y})^2$	$n-p$	$\frac{\mathrm{SSE}}{n-p}$	σ^2
Total	$\mathrm{SST} = \sum(Y_i - \bar{Y})^2$	$n-1$		

where $p = k + 1$.

Here, QF stands for "quadratic form" and is given by

$$\mathrm{QF} = \frac{1}{p-1}\tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}}\tilde{\boldsymbol{\beta}} = \frac{1}{k}\sum_{j=1}^{k}\sum_{s=1}^{k}\beta_j\beta_s\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{is} - \bar{x}_s) \geq 0.$$

Note that $\mathrm{QF} = 0 \quad \Leftrightarrow \quad \beta_1 = \beta_2 = \ldots = \beta_k = 0.$

Overall F-test for a regression relationship (p. 226)

In multiple regression, our F-test based on $F^* = \frac{\text{MSR}}{\text{MSE}}$ tests whether the *entire set* of predictors $x_1, \ldots, x_k$ explains a significant amount of variation in Y.

Overall F-test for a regression relationship (p. 226)

In multiple regression, our F-test based on $F^* = \frac{\text{MSR}}{\text{MSE}}$ tests whether the *entire set* of predictors $x_1, \ldots, x_k$ explains a significant amount of variation in Y.

If MSR $\approx$ MSE, there's no evidence that *any* of the predictors are useful. If MSR $\gg$ MSE, then *some* of them are useful.

Overall F-test for a regression relationship (p. 226)

In multiple regression, our F-test based on $F^* = \frac{\text{MSR}}{\text{MSE}}$ tests whether the *entire set* of predictors $x_1, \ldots, x_k$ explains a significant amount of variation in Y.

If MSR $\approx$ MSE, there's no evidence that *any* of the predictors are useful. If MSR $\gg$ MSE, then *some* of them are useful.

Formally, we test

$$\mathrm{H}_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

Overall F-test for a regression relationship (p. 226)

In multiple regression, our F-test based on $F^* = \frac{\text{MSR}}{\text{MSE}}$ tests whether the *entire set* of predictors $x_1, \ldots, x_k$ explains a significant amount of variation in Y.

If MSR $\approx$ MSE, there's no evidence that *any* of the predictors are useful. If MSR $\gg$ MSE, then *some* of them are useful.

Formally, we test

$$\mathrm{H}_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

versus

$$\mathrm{H}_a : \text{ at least one } \beta_j \neq 0.$$

Overall F-test for a regression relationship (p. 226)

In multiple regression, our F-test based on $F^* = \frac{\text{MSR}}{\text{MSE}}$ tests whether the *entire set* of predictors $x_1, \ldots, x_k$ explains a significant amount of variation in Y.

If MSR $\approx$ MSE, there's no evidence that *any* of the predictors are useful. If MSR $\gg$ MSE, then *some* of them are useful.

Formally, we test

$$\text{H}_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

versus

$$\text{H}_a : \text{ at least one } \beta_j \neq 0.$$

If $F^* > F_{p-1,n-p,1-\alpha}$, we reject H_0 and conclude that *something* is going on, there is *some* relationship between one or more of the $x_1, \ldots, x_k$ and Y. R provides a p-value for this test.

R^2 in case of multiple regression

The **coefficient of multiple determination** is

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}.$$

measures the proportion of sample variation in Y explained by its *linear* relationship with the predictors $x_1, \dots, x_k$. As before, $0 \leq R^2 \leq 1$.

R^2 in case of multiple regression

The **coefficient of multiple determination** is

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}.$$

measures the proportion of sample variation in Y explained by its *linear* relationship with the predictors $x_1, \ldots, x_k$. As before, $0 \leq R^2 \leq 1$.

When we add a predictor to the model R^2 can only increase. (Why?)

R^2 in case of multiple regression

The **coefficient of multiple determination** is

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}.$$

measures the proportion of sample variation in Y explained by its *linear* relationship with the predictors $x_1, \ldots, x_k$. As before, $0 \leq R^2 \leq 1$.

When we add a predictor to the model R^2 can only increase. (Why?)

The **adjusted** R^2

$$R_a^2 = 1 - \frac{\mathrm{SSE}/(n-p)}{\mathrm{SST}/(n-1)}$$

accounts for the number of predictors in the model. It may decrease when we add useless predictors to the model.

Dwayne Studio Regression Output

https://github.com/zh3nis/MATH440/blob/main/chp06/mlr.R

```
> m = lm(sales ~ children + income, data=dwayne)
> summary(m)
...
Multiple R-squared:  0.9167,Adjusted R-squared:  0.9075
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10

> anova(m)
Analysis of Variance Table

Response: sales
          Df  Sum Sq Mean Sq  F value    Pr(>F)
children   1 23371.8 23371.8 192.8962 4.64e-11 ***
income     1   643.5   643.5   5.3108  0.03332 *
Residuals 18  2180.9   121.2
```

Conclusions?

Dwayne Studio Regression Output

https://github.com/zh3nis/MATH440/blob/main/chp06/mlr.R

```
> m = lm(sales ~ children + income, data=dwayne)
> summary(m)
...
Multiple R-squared:  0.9167,Adjusted R-squared:  0.9075
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10

> anova(m)
Analysis of Variance Table

Response: sales
          Df  Sum Sq Mean Sq  F value   Pr(>F)
children   1 23371.8 23371.8 192.8962 4.64e-11 ***
income     1   643.5   643.5   5.3108  0.03332 *
Residuals 18  2180.9   121.2
```

Conclusions?
We reject $H_0 : \beta_1 = \beta_2 = 0$ at any reasonable significance level α.
About 92% of the total variability in the data is explained by the linear regression model.

Inference about individual regression parameters

The overall F-test concerns the *entire set* of predictors $x_1, \ldots, x_k$.

Inference about individual regression parameters

The overall F-test concerns the *entire set* of predictors $x_1, \ldots, x_k$.

If the F-test is significant, we will want to determine *which* of the individual predictors contribute significantly to the model.

Inference about individual regression parameters

The overall F-test concerns the *entire set* of predictors $x_1, \ldots, x_k$.

If the F-test is significant, we will want to determine *which* of the individual predictors contribute significantly to the model.

We will talk about this shortly, but the main methods are forward selection, backwards elimination, stepwise procedures, C_p, R_a^2, LASSO, etc.

Mean and covariance matrix of a vector

Recall: If $\mathbf{Y}$ is a random vector, then its expected value is also a vector

$$\mathrm{E}[\mathbf{Y}] = \begin{bmatrix} \mathrm{E}[Y_1] \\ \mathrm{E}[Y_2] \\ \vdots \\ \mathrm{E}[Y_n] \end{bmatrix}$$

The random vector $\mathbf{Y}$ also has a covariance matrix

$$\mathrm{Cov}[\mathbf{Y}] = \begin{bmatrix} \mathrm{Cov}[Y_1, Y_1] & \mathrm{Cov}[Y_1, Y_2] & \cdots & \mathrm{Cov}[Y_1, Y_n] \\ \mathrm{Cov}[Y_2, Y_1] & \mathrm{Cov}[Y_2, Y_2] & \cdots & \mathrm{Cov}[Y_2, Y_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[Y_n, Y_1] & \mathrm{Cov}[Y_n, Y_2] & \cdots & \mathrm{Cov}[Y_n, Y_n]. \end{bmatrix}$$

Multivariate normal

The **multivariate normal** density is given by

$$f(\mathbf{y}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\{-0.5(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\},$$

where $\mathbf{y} \in \mathbb{R}^n$. We write

$$\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Then $\mathrm{E}[\mathbf{Y}] = \boldsymbol{\mu}$ and $\mathrm{Cov}[\mathbf{Y}] = \boldsymbol{\Sigma}$.

For the general linear model,

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n).$$

Error vector

Note that along the diagonal of $\mathrm{Cov}[\mathbf{Y}]$, $\mathrm{Cov}[Y_i, Y_i] = \mathrm{Var}[Y_i]$.

For the general linear model,

$$\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$\mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma^2 \mathbf{I}_n = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}.$$

Inference for $\boldsymbol{\beta}$

$$\mathrm{Cov}[\mathbf{Y}] = \mathrm{Cov}[\underbrace{\mathbf{X}\boldsymbol{\beta}}_{\text{fixed}} + \underbrace{\boldsymbol{\epsilon}}_{\text{random}}] = \mathrm{Cov}[\boldsymbol{\epsilon}]$$

Fact: If $\mathbf{A}$ is a constant matrix, $\mathbf{a}$ is a constant vector, and $\mathbf{Y}$ is any random vector, then

$$\mathrm{E}[\mathbf{AY} + \mathbf{a}] = \mathbf{A}\mathrm{E}[\mathbf{Y}] + \mathbf{a},$$

$$\mathrm{Cov}[\mathbf{AY} + \mathbf{a}] = \mathbf{A}\mathrm{Cov}[\mathbf{Y}]\mathbf{A}^\top.$$

Back to the general linear model

For $\hat{\mathbf{Y}} = \mathbf{HY}$,

Back to the general linear model

For $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$,

$$\mathrm{E}[\hat{\mathbf{Y}}] = \mathbf{H}\mathrm{E}[\mathbf{Y}] = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}.$$

Back to the general linear model

For $\hat{\mathbf{Y}} = \mathbf{HY}$,

$$\mathrm{E}[\hat{\mathbf{Y}}] = \mathbf{H}\mathrm{E}[\mathbf{Y}] = \mathbf{HX}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}.$$

$$\mathrm{Cov}[\hat{\mathbf{Y}}] = \mathbf{H}\mathrm{Cov}[\mathbf{Y}]\mathbf{H}' = \sigma^2\mathbf{H},$$

Back to the general linear model

For $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$,

$$\mathrm{E}[\hat{\mathbf{Y}}] = \mathbf{H}\mathrm{E}[\mathbf{Y}] = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}.$$

$$\mathrm{Cov}[\hat{\mathbf{Y}}] = \mathbf{H}\mathrm{Cov}[\mathbf{Y}]\mathbf{H}' = \sigma^2\mathbf{H},$$

since $\mathbf{H}\mathbf{H}' = \mathbf{H}$ (property of a *projection matrix*).

Back to the general linear model

For $\hat{\mathbf{Y}} = \mathbf{HY}$,

$$\mathrm{E}[\hat{\mathbf{Y}}] = \mathbf{H}\mathrm{E}[\mathbf{Y}] = \mathbf{HX}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}.$$

$$\mathrm{Cov}[\hat{\mathbf{Y}}] = \mathbf{H}\mathrm{Cov}[\mathbf{Y}]\mathbf{H}' = \sigma^2\mathbf{H},$$

since $\mathbf{HH}' = \mathbf{H}$ (property of a *projection matrix*).
For $\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$,

Back to the general linear model

For $\hat{\mathbf{Y}} = \mathbf{HY}$,

$$\mathrm{E}[\hat{\mathbf{Y}}] = \mathbf{H}\mathrm{E}[\mathbf{Y}] = \mathbf{HX}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}.$$

$$\mathrm{Cov}[\hat{\mathbf{Y}}] = \mathbf{H}\mathrm{Cov}[\mathbf{Y}]\mathbf{H}' = \sigma^2\mathbf{H},$$

since $\mathbf{HH}' = \mathbf{H}$ (property of a *projection matrix*).
For $\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$,

$$\mathrm{E}[\mathbf{e}] = (\mathbf{I}_n - \mathbf{H})\mathrm{E}[\mathbf{Y}] = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{HX}\boldsymbol{\beta} = \mathbf{0},$$

Back to the general linear model

For $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$,

$$\mathrm{E}[\hat{\mathbf{Y}}] = \mathbf{H}\mathrm{E}[\mathbf{Y}] = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}.$$

$$\mathrm{Cov}[\hat{\mathbf{Y}}] = \mathbf{H}\mathrm{Cov}[\mathbf{Y}]\mathbf{H}' = \sigma^2\mathbf{H},$$

since $\mathbf{H}\mathbf{H}' = \mathbf{H}$ (property of a *projection matrix*).
For $\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$,

$$\mathrm{E}[\mathbf{e}] = (\mathbf{I}_n - \mathbf{H})\mathrm{E}[\mathbf{Y}] = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{0},$$

as $\mathbf{H}\mathbf{X} = \mathbf{X}$ (projection matrix again).

Back to the general linear model

For $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$,

$$\mathrm{E}[\hat{\mathbf{Y}}] = \mathbf{H}\mathrm{E}[\mathbf{Y}] = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}.$$

$$\mathrm{Cov}[\hat{\mathbf{Y}}] = \mathbf{H}\mathrm{Cov}[\mathbf{Y}]\mathbf{H}' = \sigma^2\mathbf{H},$$

since $\mathbf{H}\mathbf{H}' = \mathbf{H}$ (property of a *projection matrix*).
For $\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$,

$$\mathrm{E}[\mathbf{e}] = (\mathbf{I}_n - \mathbf{H})\mathrm{E}[\mathbf{Y}] = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{0},$$

as $\mathbf{H}\mathbf{X} = \mathbf{X}$ (projection matrix again).

$$\mathrm{Cov}[\mathbf{e}] = (\mathbf{I}_n - \mathbf{H})\mathrm{Cov}[\mathbf{Y}](\mathbf{I}_n - \mathbf{H})' = \sigma^2(\mathbf{I}_n - \mathbf{H}).$$

(Why?)

Mean and variance of $\mathbf{b}$

Finally, $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is *unbiased*

$$\mathrm{E}[\mathbf{b}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}[\mathbf{Y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

and has covariance matrix

$$\begin{aligned}\mathrm{Cov}[\mathbf{b}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Cov}[\mathbf{Y}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

Hence,

$$\mathbf{b} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

Table of regression effects

From the previous slide, the jth estimated coefficient β_j,

$$\mathrm{Var}[b_j] = \sigma^2 c_{jj},$$

where c_{jj} is the jth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.

As usually, we *estimate* the standard deviation of b_j by $\mathrm{s}[b_j] = \sqrt{\mathrm{MSE} \cdot c_{jj}}$ yielding

$$\frac{b_j - \beta_j}{\mathrm{s}[b_j]} \sim t_{n-p}$$

Note: R gives each $\mathrm{s}[b_j]$ as well as b_j, $t_j^* = b_j/\mathrm{s}[b_j]$, and a p-value for testing each $\mathrm{H}_0 : \beta_j = 0$.

Dwayne output

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.8571    60.0170  -1.147   0.2663
children      1.4546     0.2118   6.868    2e-06 ***
income        9.3655     4.0640   2.305   0.0333 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.01 on 18 degrees of freedom
Multiple R-squared:  0.9167,Adjusted R-squared:  0.9075
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10
```

We reject $\mathrm{H}_0 : \beta_1 = 0$ at the $\alpha = 0.01$ level and $\mathrm{H}_0 : \beta_2 = 0$ at the $\alpha = 0.05$ level.

Individual tests of $\mathrm{H}_0 : \beta_j = 0$

Note: A test of $\mathrm{H}_0 : \beta_j = 0$ versus $\mathrm{H}_a : \beta_j = 0$ – available in the table of regression coefficients – is a test of whether predictor x_j is necessary in a model *with the other remaining predictors included*.

For the Dwayne Studio Data:

Individual tests of $\mathrm{H}_0 : \beta_j = 0$

Note: A test of $\mathrm{H}_0 : \beta_j = 0$ versus $\mathrm{H}_a : \beta_j = 0$ – available in the table of regression coefficients – is a test of whether predictor x_j is necessary in a model *with the other remaining predictors included*.

For the Dwayne Studio Data:

- The R summary gives us $F^* = \mathrm{MSR}/\mathrm{MSE} = 99.10$ with associated p-value < 0.0001 (it is actually 2×10^{-10}!). We strongly reject $\mathrm{H}_0 : \beta_1 = \beta_2 = 0$.

Individual tests of $\mathrm{H}_0 : \beta_j = 0$

Note: A test of $\mathrm{H}_0 : \beta_j = 0$ versus $\mathrm{H}_a : \beta_j = 0$ – available in the table of regression coefficients – is a test of whether predictor x_j is necessary in a model *with the other remaining predictors included*.

For the Dwayne Studio Data:

- The R summary gives us $F^* = \mathrm{MSR}/\mathrm{MSE} = 99.10$ with associated p-value < 0.0001 (it is actually 2×10^{-10}!). We strongly reject $\mathrm{H}_0 : \beta_1 = \beta_2 = 0$.
- 95% CI's are $(1.01, 1.90)$ for β_1 and $(0.83, 17.90)$ for β_2.

Individual tests of $\mathrm{H}_0 : \beta_j = 0$

Note: A test of $\mathrm{H}_0 : \beta_j = 0$ versus $\mathrm{H}_a : \beta_j = 0$ – available in the table of regression coefficients – is a test of whether predictor x_j is necessary in a model *with the other remaining predictors included*.

For the Dwayne Studio Data:

- The R summary gives us $F^* = \mathrm{MSR}/\mathrm{MSE} = 99.10$ with associated p-value < 0.0001 (it is actually 2×10^{-10}!). We strongly reject $\mathrm{H}_0 : \beta_1 = \beta_2 = 0$.
- 95% CI's are $(1.01, 1.90)$ for β_1 and $(0.83, 17.90)$ for β_2.
- For example, we are 95% confident that mean sales increases by \$1010 to \$1900 for every 1000 increase in kids 16 and under, holding income constant.

Individual tests of $\mathrm{H}_0 : \beta_j = 0$

Note: A test of $\mathrm{H}_0 : \beta_j = 0$ versus $\mathrm{H}_a : \beta_j = 0$ – available in the table of regression coefficients – is a test of whether predictor x_j is necessary in a model *with the other remaining predictors included*.

For the Dwayne Studio Data:

- The R summary gives us $F^* = \mathrm{MSR}/\mathrm{MSE} = 99.10$ with associated p-value < 0.0001 (it is actually 2×10^{-10}!). We strongly reject $\mathrm{H}_0 : \beta_1 = \beta_2 = 0$.
- 95% CI's are $(1.01, 1.90)$ for β_1 and $(0.83, 17.90)$ for β_2.
- For example, we are 95% confident that mean sales increases by \$1010 to \$1900 for every 1000 increase in kids 16 and under, holding income constant.
- For $\mathrm{H}_0 : \beta_1 = 0$ we get $p < 0.0001$; for $\mathrm{H}_0 : \beta_2 = 0$ we get $p = 0.03$. Are people under 16 (x_1) and income (x_2) important in the model?

CI for mean response and PI for new response

Let's construct a CI for the mean response corresponding to a set of values

$$\mathbf{x}_{\text{new}} = \begin{bmatrix} 1 \\ X_{\text{new},1} \\ X_{\text{new},2} \\ \vdots \\ X_{\text{new},k} \end{bmatrix}$$

CI for mean response and PI for new response

Let's construct a CI for the mean response corresponding to a set of values

$$\mathbf{x}_{\text{new}} = \begin{bmatrix} 1 \\ x_{\text{new},1} \\ x_{\text{new},2} \\ \vdots \\ x_{\text{new},k} \end{bmatrix}$$

We want to make inferences about

$$\mathrm{E}[Y_{\text{new}}] = \mathbf{x}_{\text{new}}^{\top} \boldsymbol{\beta} = \beta_0 + \beta_1 x_{\text{new},1} + \ldots + \beta_k x_{\text{new},k}.$$

CI for mean response and PI for new response

- A point estimate is $\hat{Y}_{\text{new}} = \widehat{\mathrm{E}[Y_{\text{new}}]} = \mathbf{x}_{\text{new}}\mathbf{b}$.

CI for mean response and PI for new response

- A point estimate is $\hat{Y}_{\text{new}} = \widehat{\mathrm{E}[Y_{\text{new}}]} = \mathbf{x}_{\text{new}}\mathbf{b}$.
- Then $\mathrm{E}[\hat{Y}_{\text{new}}] = \mathrm{E}[\mathbf{x}_{\text{new}}\mathbf{b}] = \mathbf{x}_{\text{new}}\mathrm{E}[\mathbf{b}] = \mathbf{x}_{\text{new}}\boldsymbol{\beta}$.

CI for mean response and PI for new response

- A point estimate is $\hat{Y}_{\text{new}} = \widehat{\mathrm{E}[Y_{\text{new}}]} = \mathbf{x}_{\text{new}}\mathbf{b}$.
- Then $\mathrm{E}[\hat{Y}_{\text{new}}] = \mathrm{E}[\mathbf{x}_{\text{new}}\mathbf{b}] = \mathbf{x}_{\text{new}}\mathrm{E}[\mathbf{b}] = \mathbf{x}_{\text{new}}\boldsymbol{\beta}$.
- Also, $\mathrm{Var}[\hat{Y}_{\text{new}}] = \mathrm{Cov}[\mathbf{x}'_{\text{new}}\mathbf{b}] = \mathbf{x}'_{\text{new}}\mathrm{Cov}[\mathbf{b}]\mathbf{x}_{\text{new}} = \sigma^2\mathbf{x}'_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{\text{new}}$.

CI for mean response and PI for new response

- A point estimate is $\hat{Y}_{\text{new}} = \widehat{\mathrm{E}[Y_{\text{new}}]} = \mathbf{x}_{\text{new}}\mathbf{b}$.
- Then $\mathrm{E}[\hat{Y}_{\text{new}}] = \mathrm{E}[\mathbf{x}_{\text{new}}\mathbf{b}] = \mathbf{x}_{\text{new}}\mathrm{E}[\mathbf{b}] = \mathbf{x}_{\text{new}}\boldsymbol{\beta}$.
- Also, $\mathrm{Var}[\hat{Y}_{\text{new}}] = \mathrm{Cov}[\mathbf{x}'_{\text{new}}\mathbf{b}] = \mathbf{x}'_{\text{new}}\mathrm{Cov}[\mathbf{b}]\mathbf{x}_{\text{new}} = \sigma^2\mathbf{x}'_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{\text{new}}$.

So

CI for mean response and PI for new response

- A point estimate is $\hat{Y}_{\text{new}} = \widehat{\mathrm{E}[Y_{\text{new}}]} = \mathbf{x}_{\text{new}}\mathbf{b}$.
- Then $\mathrm{E}[\hat{Y}_{\text{new}}] = \mathrm{E}[\mathbf{x}_{\text{new}}\mathbf{b}] = \mathbf{x}_{\text{new}}\mathrm{E}[\mathbf{b}] = \mathbf{x}_{\text{new}}\boldsymbol{\beta}$.
- Also, $\mathrm{Var}[\hat{Y}_{\text{new}}] = \mathrm{Cov}[\mathbf{x}'_{\text{new}}\mathbf{b}] = \mathbf{x}'_{\text{new}}\mathrm{Cov}[\mathbf{b}]\mathbf{x}_{\text{new}} = \sigma^2\mathbf{x}'_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{\text{new}}$.

So

- An *approximate* $100 \cdot (1-\alpha)\%$ CI for $\mathrm{E}[Y_{\text{new}}]$ is

$$\hat{Y}_{\text{new}} \pm t_{n-p,1-\alpha/2}\sqrt{\mathrm{MSE} \cdot \mathbf{x}_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{\text{new}}},$$

CI for mean response and PI for new response

- A point estimate is $\hat{Y}_{\text{new}} = \widehat{\mathrm{E}[Y_{\text{new}}]} = \mathbf{x}_{\text{new}}\mathbf{b}$.
- Then $\mathrm{E}[\hat{Y}_{\text{new}}] = \mathrm{E}[\mathbf{x}_{\text{new}}\mathbf{b}] = \mathbf{x}_{\text{new}}\mathrm{E}[\mathbf{b}] = \mathbf{x}_{\text{new}}\boldsymbol{\beta}$.
- Also, $\mathrm{Var}[\hat{Y}_{\text{new}}] = \mathrm{Cov}[\mathbf{x}'_{\text{new}}\mathbf{b}] = \mathbf{x}'_{\text{new}}\mathrm{Cov}[\mathbf{b}]\mathbf{x}_{\text{new}} = \sigma^2\mathbf{x}'_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{\text{new}}$.

So

- An *approximate* $100 \cdot (1-\alpha)\%$ CI for $\mathrm{E}[Y_{\text{new}}]$ is

$$\hat{Y}_{\text{new}} \pm t_{n-p,1-\alpha/2}\sqrt{\mathrm{MSE} \cdot \mathbf{x}_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{\text{new}}},$$

- An *approximate* $100 \cdot (1-\alpha)\%$ prediction interval for a new response $Y_{\text{new}} = \mathbf{x}'_{\text{new}}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{\text{new}}$ is

$$\hat{Y}_{\text{new}} \pm t_{n-p,1-\alpha/2}\sqrt{\mathrm{MSE} \cdot [1 + \mathbf{x}'_{\text{new}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{\text{new}}]},$$

Dwayne Studios

https://github.com/zh3nis/MATH440/blob/main/chp06/dwayne_ci_pi.R

Say we want to estimate mean sales in cities with $x_1 = 65.4$ thousand children and per capita disposable income of $x_2 = 17.6$ thousand dollars.

```
> new_x = data.frame(children=65.4, income=17.6)
> predict.lm(m, new_x, interval="confidence", level=0.95)
       fit      lwr      upr
1 191.1039 185.2911 196.9168
> predict.lm(m, new_x, interval="prediction", level=0.95)
       fit      lwr     upr
1 191.1039 167.2589 214.949
```

Checking model assumptions

The general linear model assumes the following:

1. A linear relationship between $\mathrm{E}[Y]$ and associated predictors $x_1, \ldots, x_k$.
2. The errors have constant variance.
3. The errors are normally distributed.
4. The errors are independent.

Checking model assumptions

The general linear model assumes the following:

1. A linear relationship between $\mathrm{E}[Y]$ and associated predictors $x_1, \ldots, x_k$.
2. The errors have constant variance.
3. The errors are normally distributed.
4. The errors are independent.

We estimate the unknown $\epsilon_1, \ldots, \epsilon_n$ with the residuals $e_1, \ldots, e_n$.

Checking model assumptions

The general linear model assumes the following:

1. A linear relationship between $\mathrm{E}[Y]$ and associated predictors $x_1, \ldots, x_k$.
2. The errors have constant variance.
3. The errors are normally distributed.
4. The errors are independent.

We estimate the unknown $\epsilon_1, \ldots, \epsilon_n$ with the residuals $e_1, \ldots, e_n$. Assumptions can be checked informally using plots and formally using tests.

Linearity b/w mean response and predictors

- Scatterplots of $\{(x_{ij}, Y_i)\}_{i=1}^{n}$ for each predictor $j = 1, \ldots, k$. Look for "nonlinear" patterns. These are *marginal* relationships, and do not get at the simultaneous relationship among variables.

Linearity b/w mean response and predictors

- Scatterplots of $\{(x_{ij}, Y_i)\}_{i=1}^n$ for each predictor $j = 1, \ldots, k$. Look for "nonlinear" patterns. These are *marginal* relationships, and do not get at the simultaneous relationship among variables.
- Look at residuals versus each predictor $\{(x_{ij}, e_i)\}_{i=1}^n$, *and* residuals versus fitted values $\{(\hat{Y}_i, e_i)\}_{i=1}^n$.

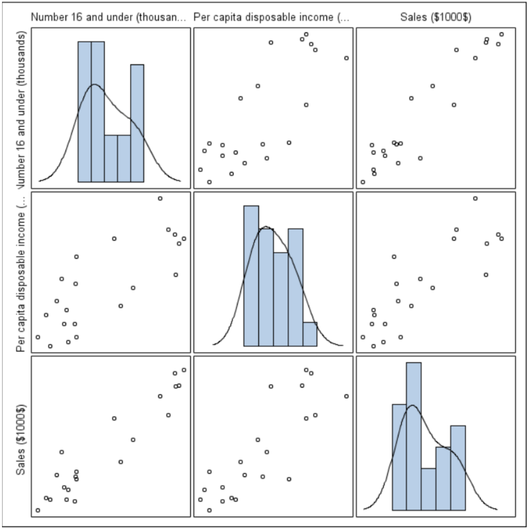
Linearity b/w mean response and predictors

- Scatterplots of $\{(x_{ij}, Y_i)\}_{i=1}^n$ for each predictor $j = 1, \ldots, k$. Look for "nonlinear" patterns. These are *marginal* relationships, and do not get at the simultaneous relationship among variables.
- Look at residuals versus each predictor $\{(x_{ij}, e_i)\}_{i=1}^n$, *and* residuals versus fitted values $\{(\hat{Y}_i, e_i)\}_{i=1}^n$.
- Look for non-random (especially curved) pattern in the residual plots, indicating violation of linear mean.

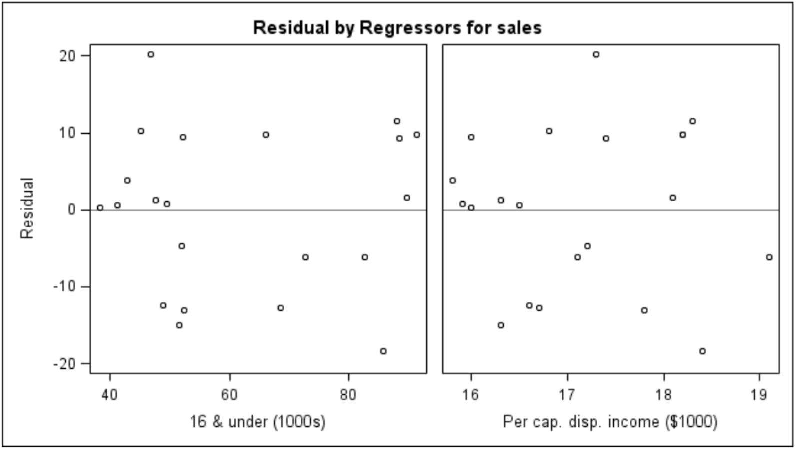
Linearity b/w mean response and predictors

- Scatterplots of $\{(x_{ij}, Y_i)\}_{i=1}^n$ for each predictor $j = 1, \ldots, k$. Look for "nonlinear" patterns. These are *marginal* relationships, and do not get at the simultaneous relationship among variables.
- Look at residuals versus each predictor $\{(x_{ij}, e_i)\}_{i=1}^n$, *and* residuals versus fitted values $\{(\hat{Y}_i, e_i)\}_{i=1}^n$.
- Look for non-random (especially curved) pattern in the residual plots, indicating violation of linear mean.
- Remedies: (i) choose different functional form of model, (ii) transformation of one or more predictor variables.

Scatterplot matrix



Residuals vs predictors

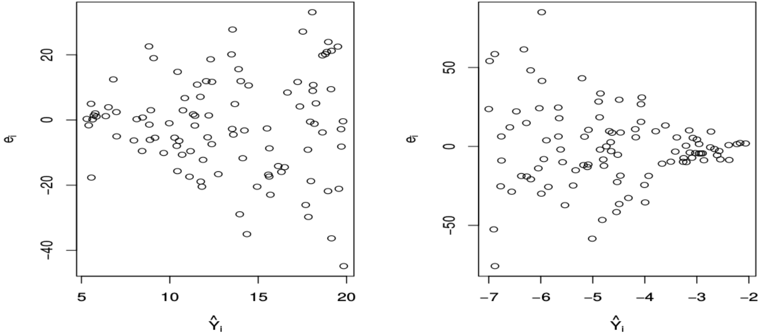


Constant variance

- Often the most worrisome assumption

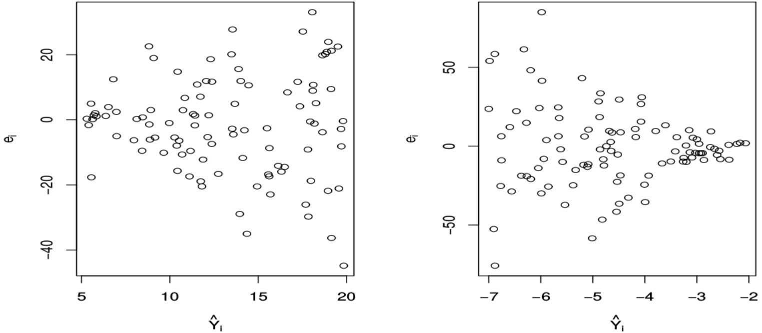
Constant variance

- Often the most worrisome assumption
- Violation indicated by "megaphone shape" in residual plot:



Constant variance

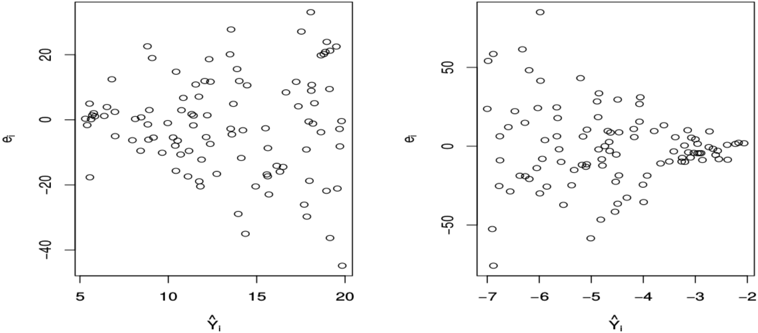
- Often the most worrisome assumption
- Violation indicated by "megaphone shape" in residual plot:



- Easy remedy: transform the response, e.g. $Y^* = \log(Y)$ or $Y^* = \sqrt{Y}$.

Constant variance

- Often the most worrisome assumption
- Violation indicated by "megaphone shape" in residual plot:



- Easy remedy: transform the response, e.g. $Y^* = \log(Y)$ or $Y^* = \sqrt{Y}$.
- *A more advanced method: weighted least squares (Chapter 11).

Constant variance

- Breusch-Pagan test (pp. 118–119): tests whether the log error variance increase or decrease linearly with the predictor(s).

Constant variance

- Breusch-Pagan test (pp. 118–119): tests whether the log error variance increase or decrease linearly with the predictor(s). Let $Y_i \sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma_i^2)$, set

$$\ln \sigma_i^2 = \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_k x_{ik}$$

Constant variance

- Breusch-Pagan test (pp. 118–119): tests whether the log error variance increase or decrease linearly with the predictor(s). Let $Y_i \sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma_i^2)$, set

$$\ln \sigma_i^2 = \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_k x_{ik}$$

and test

$$\mathrm{H}_0 : \alpha_1 = \ldots = \alpha_k = 0 \qquad \Leftrightarrow \qquad \ln \sigma_i^2 = \alpha_0,$$

Constant variance

- Breusch-Pagan test (pp. 118–119): tests whether the log error variance increase or decrease linearly with the predictor(s). Let $Y_i \sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma_i^2)$, set

$$\ln \sigma_i^2 = \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_k x_{ik}$$

and test

$$\mathrm{H}_0 : \alpha_1 = \ldots = \alpha_k = 0 \qquad \Leftrightarrow \qquad \ln \sigma_i^2 = \alpha_0,$$

Requires large samples & assumes normal errors.

Constant variance

- Breusch-Pagan test (pp. 118–119): tests whether the log error variance increase or decrease linearly with the predictor(s). Let $Y_i \sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma_i^2)$, set

$$\ln \sigma_i^2 = \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_k x_{ik}$$

and test

$$\mathrm{H}_0 : \alpha_1 = \ldots = \alpha_k = 0 \qquad \Leftrightarrow \qquad \ln \sigma_i^2 = \alpha_0,$$

Requires large samples & assumes normal errors.

- Brown-Forsythe (Levene) test (pp. 116–117): Robust to non-normal errors. Requires user to break data into groups and test for constancy error variance across groups (not natural for continuous data).

Constant variance

- Breusch-Pagan test (pp. 118–119): tests whether the log error variance increase or decrease linearly with the predictor(s). Let $Y_i \sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma_i^2)$, set

$$\ln \sigma_i^2 = \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_k x_{ik}$$

and test

$$\mathrm{H}_0 : \alpha_1 = \ldots = \alpha_k = 0 \qquad \Leftrightarrow \qquad \ln \sigma_i^2 = \alpha_0,$$

Requires large samples & assumes normal errors.

- Brown-Forsythe (Levene) test (pp. 116–117): Robust to non-normal errors. Requires user to break data into groups and test for constancy error variance across groups (not natural for continuous data).
- Graphical methods have advantage of checking for *general violations*, not just violation of a specific type.

Breusch Pagan test in R

```
> library(lmtest)
> dwayne = read.table("path/to/CH06FI05.txt", header=FALSE)
> colnames(dwayne) = c("children", "income", "sales")
> m = lm(sales ~ children + income, data=dwayne)
> bptest(m)

studentized Breusch-Pagan test

data:  m
BP = 1.949, df = 2, p-value = 0.3774
```

Breusch Pagan test in R

```
> library(lmtest)
> dwayne = read.table("path/to/CH06FI05.txt", header=FALSE)
> colnames(dwayne) = c("children", "income", "sales")
> m = lm(sales ~ children + income, data=dwayne)
> bptest(m)

studentized Breusch-Pagan test

data:  m
BP = 1.949, df = 2, p-value = 0.3774
```

With p-value = .3774 we do not reject $\mathrm{H}_0 : \sigma_i = \sigma$ at $\alpha = 0.05$, no evidence of non-constant variance.

Normality of errors

Diagnostics include...

- Q-Q plot of $e_1, \ldots, e_n$.

Normality of errors

Diagnostics include...

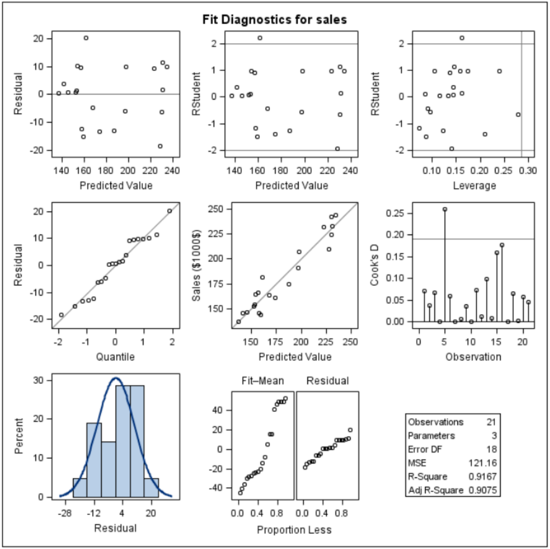
- Q-Q plot of $e_1, \dots, e_n$.
- Formal test for normality: Shapiro-Wilk (Section 3.5), essentially based on the correlation coefficient r for expected versus observed in normal Q-Q plot.

Normality of errors

Diagnostics include...

- Q-Q plot of $e_1, \ldots, e_n$.
- Formal test for normality: Shapiro-Wilk (Section 3.5), essentially based on the correlation coefficient r for expected versus observed in normal Q-Q plot.
- Remedy: transformation of Y or any of $x_1, \ldots, x_k$, *nonparametric methods (e.g. additive models), *robust regression (least sum of absolute distances), *median regression.

Standard diagnostics



Test for normal residuals in Portrait data

```
                    Tests for Normality

Test                  --Statistic---    -----p Value------
Shapiro-Wilk          W      0.954073    Pr < W       0.4056
Kolmogorov-Smirnov    D      0.147126    Pr > D       >0.1500
Cramer-von Mises      W-Sq   0.066901    Pr > W-Sq    >0.2500
Anderson-Darling      A-Sq   0.432299    Pr > A-Sq    >0.2500
```

We do not reject $\mathrm{H}_0 : \epsilon_1, \ldots, \epsilon_n$ are normally distributed

Test for normal residuals in Portrait data

```
                   Tests for Normality

Test                  --Statistic---    -----p Value------
Shapiro-Wilk          W     0.954073    Pr < W      0.4056
Kolmogorov-Smirnov    D     0.147126    Pr > D     >0.1500
Cramer-von Mises      W-Sq  0.066901    Pr > W-Sq  >0.2500
Anderson-Darling      A-Sq  0.432299    Pr > A-Sq  >0.2500
```

We do not reject $\mathrm{H}_0 : \epsilon_1, \ldots, \epsilon_n$ are normally distributed

The Anderson-Darling test looks primarily for evidence of non-normal data in the tails of a distribution; the Shapiro-Wilk emphasizes lack of symmetry in the distribution; i.e. less emphasis placed on the tails.

Comments

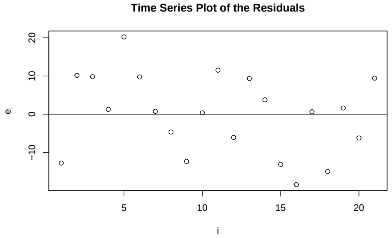
- With large sample sizes, the normality assumption is not critical *unless you are predicting new observations*.

Comments

- With large sample sizes, the normality assumption is not critical *unless you are predicting new observations*.
- The formal test will not tell you the *type* of departure from normality (e.g. bimodal, skew, heavy or light tails, etc.).
- Q-Q plots help answer these questions (if the mean is specified correctly).

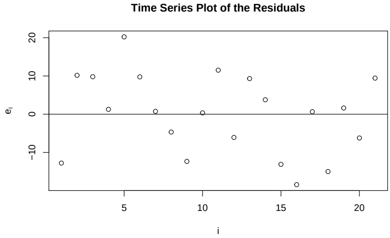
Independence

As in the case of SLR, plot of e_i vs i:



Independence

As in the case of SLR, plot of e_i vs i:



and Runs test of independence:

```
> res = m$residuals
> library(lawstat)
> runs.test(res)

Runs Test - Two sided

data:  res
Standardized Runs Statistic = -0.66258, p-value = 0.5076
```

Transformations of variables

- Some violations of our model assumptions may be fixed by transforming one or more predictors $x_1, \ldots, x_k$ or Y.

Transformations of variables

- Some violations of our model assumptions may be fixed by transforming one or more predictors $x_1, \ldots, x_k$ or Y.
- If the only problem is a nonlinear relationship between Y and the predictors, i.e. constant variance seems okay, a transformation of one or more of the $x_1, \ldots, x_k$ is preferred.

Transformations of variables

- Some violations of our model assumptions may be fixed by transforming one or more predictors $x_1, \ldots, x_k$ or Y.
- If the only problem is a nonlinear relationship between Y and the predictors, i.e. constant variance seems okay, a transformation of one or more of the $x_1, \ldots, x_k$ is preferred.
- If non-constant variance appears in one or more plots of Y versus the predictors, a transformation in Y can help...or make it worse!

Transformations of variables

- Some violations of our model assumptions may be fixed by transforming one or more predictors $x_1, \ldots, x_k$ or Y.
- If the only problem is a nonlinear relationship between Y and the predictors, i.e. constant variance seems okay, a transformation of one or more of the $x_1, \ldots, x_k$ is preferred.
- If non-constant variance appears in one or more plots of Y versus the predictors, a transformation in Y can help. . . or make it worse!
- *Data analysis is an art*. The best way to learn how to analyze data is to analyze data.

Transformations of variables

- Some violations of our model assumptions may be fixed by transforming one or more predictors $x_1, \ldots, x_k$ or Y.
- If the only problem is a nonlinear relationship between Y and the predictors, i.e. constant variance seems okay, a transformation of one or more of the $x_1, \ldots, x_k$ is preferred.
- If non-constant variance appears in one or more plots of Y versus the predictors, a transformation in Y can help...or make it worse!
- *Data analysis is an art*. The best way to learn how to analyze data is to analyze data.
- A nonlinear relationship *could* show itself in the scatterplot matrix of Y_i versus x_{ij} for $j = 1, \ldots, k$, or the residuals e_i versus x_{ij}.

Transformations of variables

- Some violations of our model assumptions may be fixed by transforming one or more predictors $x_1, \ldots, x_k$ or Y.
- If the only problem is a nonlinear relationship between Y and the predictors, i.e. constant variance seems okay, a transformation of one or more of the $x_1, \ldots, x_k$ is preferred.
- If non-constant variance appears in one or more plots of Y versus the predictors, a transformation in Y can help…or make it worse!
- *Data analysis is an art*. The best way to learn how to analyze data is to analyze data.
- A nonlinear relationship *could* show itself in the scatterplot matrix of Y_i versus x_{ij} for $j = 1, \ldots, k$, or the residuals e_i versus x_{ij}.
- The chosen transformation should roughly mimic the relationship seen in the plot.

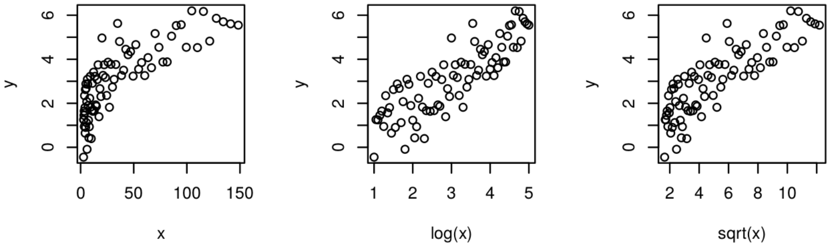
Transformations for $x_{i1}, \ldots, x_{ik}$

Examples of transformations for predictors are:

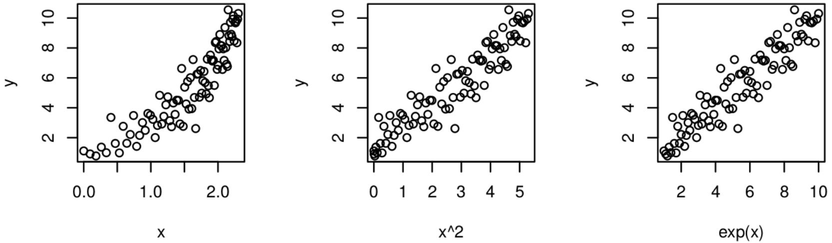
- $x^* = \log(x)$
- $x^* = \sqrt{x}$
- $x^* = 1/x$
- $x^* = \exp(x)$ or $x^* = \exp(-x)$

We will examine *marginal* relationships and transformation "fixes".

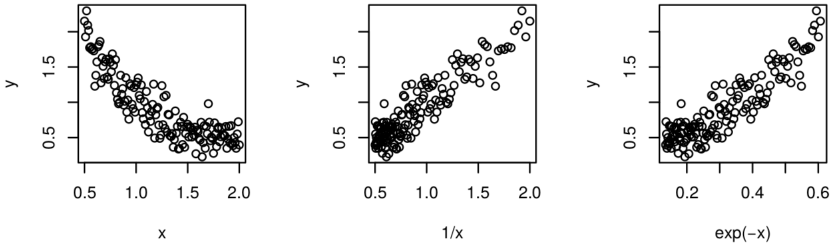
Example 1: transforming a predictor



Example 2: transforming a predictor



Example 3: transforming a predictor



Transforming a response

If there is evidence of nonconstant error variance, a transformation of Y can often fix things. Examples include:

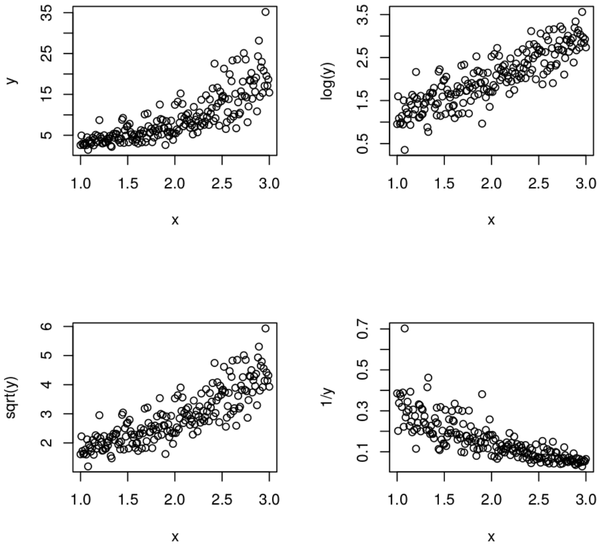
- $Y^* = \log(Y)$
- $Y^* = \sqrt{Y}$
- $Y^* = 1/Y$

See Figure 3.15, page 132.

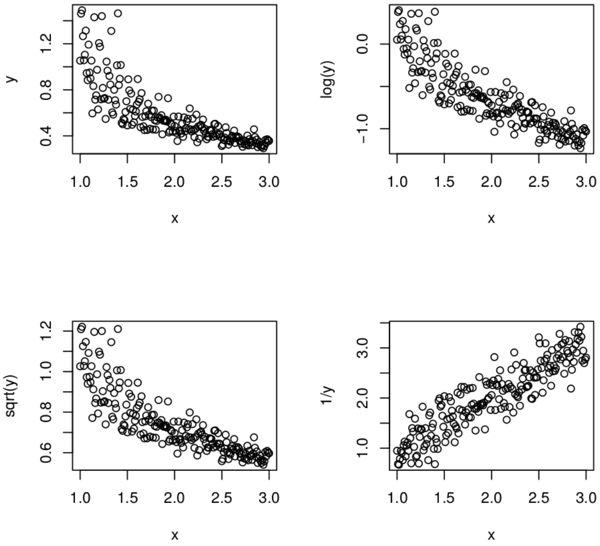
All of these are included in the Box-Cox family of transformations.

For some data, a transformation in Y may be followed by one or more transformations in the $x_{i1}, \ldots, x_{ik}$.

Example 4: transforming the response



Example 5: transforming the response



Box-Cox transformations

Box-Cox transformations are of the type

$$Y^* = \frac{Y^\lambda - 1}{\lambda}$$

where λ is estimated from the data, typically $-3 \leq \lambda \leq 3$. These include

$$\begin{aligned}
\lambda = 2 & \qquad Y^* = (Y^2 - 1)/2 \sim Y^2 \\
\lambda = 1 & \qquad Y^* = Y - 1 \sim Y \\
\lambda = 0 & \qquad Y^* = \log(Y) \\
\lambda = -1 & \qquad Y^* = 1 - 1/Y \sim 1/Y \\
\lambda = -2 & \qquad Y^* = 1/2 - 1/(2Y^2) \sim 1/Y^2
\end{aligned}$$

R will help you to pick λ automatically.