STAT302 Methods of Data Analysis 1
Module 2: Multiple Linear Regression Models and Basics

Austin Brown

September 15, 2024

Previous lecture review

Beyond basic statistics: simple linear regression

- **Population model:** $Y|X = x \sim N(\beta_0 + \beta_1 x, \sigma^2)$
- **Parameters:** $\beta_0, \beta_1, \sigma^2$
- **Sample:** Independent pairs $Y_i|X = x_i \; i = 1, \ldots, n$
- **Data:** Pairs $(y_1, x_1), \ldots, (y_n, x_n)$

Putting this all together: the **simple linear regression model** is

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

for $i = 1, \ldots, n$ where

- $e_i \sim N(0, \sigma^2)$ are i.i.d.
- x_i are known numbers
- $\beta_0, \beta_1, \sigma^2$ are unknown parameters

Estimation

- Given data $(y_1, x_1), \ldots, (y_n, x_n)$, we can find the unique minimizers of the residual sum of squares b_0^*, b_1^* and these have explicit solutions depending on the data.

$$b_1^* = \frac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$
$$b_0^* = \overline{y} - b_1^* \overline{x}$$

Simple Least Squares Estimators

Assume $(x_1, Y_1), \ldots, (x_1, Y_n)$ is a random sample from the population following the simple linear regression model. Then $\hat{\beta}_0, \hat{\beta}_1$ are the (random) simple least squares estimators for the unknown parameters β_0, β_1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$

Remarks

This concept is important. Estimates by minimizing something (not necessarily residual sum of squares) appears throughout statistics, machine learning, etc.

- Maximum likelihood
- Neural networks, Gradient boosting, etc.

Lecture 1: Multiple linear regression model

Learning goals

- Define the multiple linear regression model
- Understand the components / terminology of multiple linear regression models

Working example: iris data

Measurements (cm) of the variables sepal length and width and petal length and width for 50 flowers from each of 3 species of iris (setosa, versicolor, virginica).

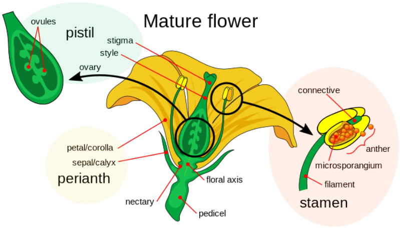


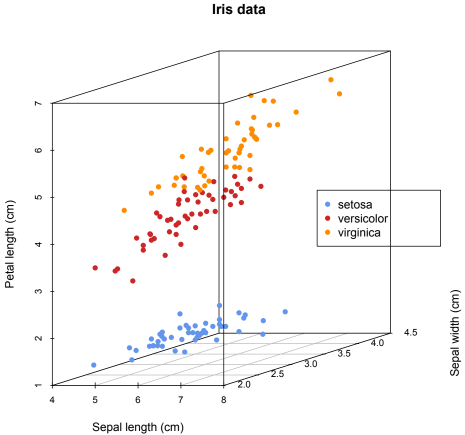
Figure: Picture by Mariana Ruiz

```
data(iris)
```

Working example: iris data

Our goal: Study the linear relationship between petal length and sepal length, width, and species. We are interested in **predicting** petal length.

Plot the iris data



Simple linear regression in matrix form

$Y_i = \beta_0 + \beta_1 x_i + e_i$, $e_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, n$. In vector form:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} (1, x_1)\boldsymbol{\beta} + e_1 \\ \vdots \\ (1, x_n)\boldsymbol{\beta} + e_n \end{pmatrix}$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad \boldsymbol{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

The simple linear regression model can be written:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

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

Multiple linear regression

- **Response**: Y is a random univariate dependent variable
- **Vector of predictors** $\boldsymbol{x} = (1, x_1, \ldots, x_p)^T$ fixed, $p+1$ dimension vector of intercept and p explanatory/predictor variables.

The multivariate regression population model

Population model specification:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + e.$$

- $\boldsymbol{\beta}$: parameter vector of dimension $p+1$
- σ^2: error variance parameter
- $e \sim N(0, \sigma^2)$
- $E(Y|X = \boldsymbol{x})$
  $= E(Y|X_1 = x_1, \ldots, X_p = x_p)$
  $= \beta_0 + \beta_1 x_1 + \cdots + x_p \beta_p$
- $\text{Var}(Y|X = \boldsymbol{x}) = \text{Var}(Y|X_1 = x_1, \ldots, X_p = x_p) = \sigma^2$

Then $Y_1|X = \boldsymbol{x}_1, \ldots, Y_n|X = \boldsymbol{x}_n$ is a random sample from the population.

The multivariate regression model

The **multiple linear regression model** is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + e_i \\ &= \boldsymbol{x_i}^T \boldsymbol{\beta} + e_i \end{aligned}$$

for $i = 1, \ldots, n$ where

- $\boldsymbol{\beta}$: parameter vector of dimension $p + 1$
- σ^2: error variance parameter
- $e_i \sim N(0, \sigma^2)$ i.i.d. errors

The multivariate regression model

Properties from the definition:

- $E(Y_i|X = \boldsymbol{x_i})$
  $= E(Y_i|X_1 = x_{i,1}, \dots, X_p = x_{i,p})$
  $= \beta_0 + \beta_1 x_{i,1} + \cdots + x_{i,p}\beta_p$
- On average, the response is linear in β
- $\mathsf{Var}(Y_i|X = \boldsymbol{x}_i) = \mathsf{Var}(Y_i|X_1 = x_{i,1}, \dots, X_p = x_{i,p}) = \sigma^2$

Matrix form of the multivariate regression model

The **multivariate linear regression model** can be written:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}.$$

where

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x_1}^T \\ \vdots \\ \boldsymbol{x_n}^T \end{pmatrix} = \begin{pmatrix} 1, x_{1,1}, \ldots, x_{1,p} \\ \vdots \\ 1, x_{n,1}, \ldots, x_{n,p} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \boldsymbol{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \sim N(\mathbf{0}_n, \sigma^2 I_n)$$

Poll question

Question: How many predictor variables are in this model?

Fit the regression model

We will estimate the parameters with least squares estimator (random) $\hat{\beta}$ by minimizing the RSS. We will learn the details next lecture.

Linear part of the model

On average, the response is linear in this sense:

$$\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}.$$

This is linear in $\boldsymbol{\beta}$.

The entries of $\boldsymbol{x}$ may be complex:

- Continuous/discrete predictors
- Dummy variables for categorical predictors
- polynomial orders of predictors x^2, x^3, etc.
- Transformations of predictors $log(x), sin(x)$
- Interactions or cross products between predictors $x_1 x_2$, etc.

Fit the regression model

The **fitted regression** is

$$\begin{aligned}\hat{E}(Y_i|X=\boldsymbol{x_i}) &= \hat{Y}_i \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_p x_{i,p} \\ &= \boldsymbol{x_i}^T \hat{\boldsymbol{\beta}}.\end{aligned}$$

- **Estimates / Predicts** the response on average given covariates $\boldsymbol{x_i}$

Combine these predictions:

$$\hat{E}(\boldsymbol{Y}|\boldsymbol{X}) = \hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} \boldsymbol{x_1}^T\hat{\boldsymbol{\beta}} \\ \vdots \\ \boldsymbol{x_n}^T\hat{\boldsymbol{\beta}} \end{pmatrix}.$$

Lecture 1: Activity

Fit the iris data

```r
# Fit iris with R
fitted_model = lm(Petal.Length ~ Sepal.Length + Sepal.
    Width, data = iris)

# What does this do?
beta_hat = coef(fitted_model)
```

Fit the iris data

```
# What does this do?
Xbeta_hat = fitted(fitted_model)

# What does this do?
X = model.matrix(fitted_model)

# What does this do?
Xbeta_hat = X %*% beta_hat
```

Visualize the fit to the iris data

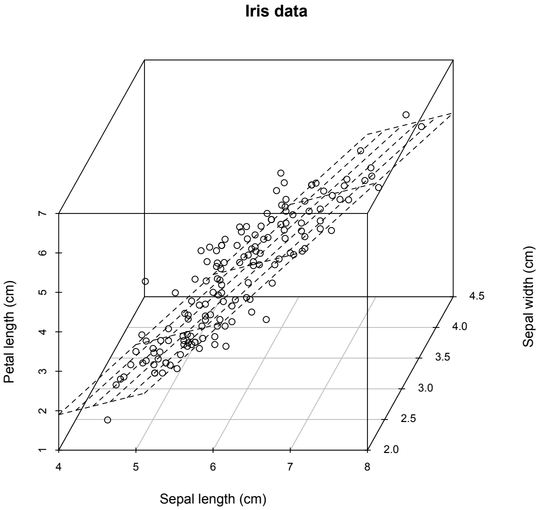


Figure: Linear regression for the iris data

Lecture 2: Estimation

Learning goals

- Understand least squares estimation
- Carry out the least squares procedure

Least squares components

- Data: $\boldsymbol{X}, \boldsymbol{y}$
- Matrix:

$$\boldsymbol{X} = \begin{pmatrix} 1, x_{1,1}, \ldots, x_{1,p} \\ \vdots \\ 1, x_{n,1}, \ldots, x_{n,p} \end{pmatrix}$$

- vector:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

- vector: unique minimizer to the residual sum of squares $\boldsymbol{b}^*$ depending on the data.

Residual

Choose vector $\boldsymbol{b}$ of dimension $p+1$ (not the regression parameter here).

- Residual using this choice $\boldsymbol{b}$: $y_i - \boldsymbol{x}_i^T \boldsymbol{b}$

We generally say the (computed) **residual** is

$$y_i - \hat{y}_i = y_i - \boldsymbol{x}_i^T \boldsymbol{b}^*$$

using least squares estimate $\boldsymbol{b}^*$. Vector of all residuals

$$\boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - X\boldsymbol{b}^*.$$

Residual sum of squares

- Residual sum of squares (RSS) using a chosen $\boldsymbol{b}$:

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x_i}^T\boldsymbol{b})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$$

(not the unknown regression parameter here)

Least squares estimates

Find $\boldsymbol{b}^*$ that minimizes RSS over all possible choices of b (i.e. minimize over all possible unkown parameters):

$$\min_{\boldsymbol{b}} \sum_{i=1}^{n} (y_i - \boldsymbol{x_i}^T \boldsymbol{b})^2.$$

Least squares estimates

(This slide is for understanding and not required on an exam)

$$\partial_{b_l}\partial_{b_k}\sum_{i=1}^{n}(y_i - \boldsymbol{x_i}^T\boldsymbol{b})^2$$

The Hessian for all $\boldsymbol{b}$:

$$2\boldsymbol{X}^T\boldsymbol{X}$$

Least squares estimates

If all partial derivatives of the RSS at $\boldsymbol{b}^*$ are 0, then $\boldsymbol{b}^*$ is the unique a minimum if $\boldsymbol{X}^T\boldsymbol{X}$ is invertible (positive-definite).

Least squares procedure

1. Take all partial derivatives of the RSS with respect to the vector $\boldsymbol{b}$ and set to these to 0.
2. Rearrange the equations to solve for $\boldsymbol{b}^*$.
3. $\boldsymbol{b}^*$ will be the unique minimizer if $X^T X$ is invertible (use a computer to check in practice).

Least squares procedure

$$\partial_{b_0} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_{\boldsymbol{i}}^T \boldsymbol{b})^2 =$$

$$\vdots$$

$$\partial_{b_p} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_{\boldsymbol{i}}^T \boldsymbol{b})^2 =$$

Least squares procedure

$$\begin{pmatrix} 1 & \cdots & 1 \\ x_{1,1} & \cdots & x_{n,1} \\ \vdots & \cdots & \vdots \\ x_{1,p} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} \boldsymbol{x_1^T b} - y_1 \\ \vdots \\ \boldsymbol{x_n^T b} - y_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

Putting in matrix form:

Least squares procedure

Solve:

$$\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{b} - \boldsymbol{y}) = 0$$

Least squares procedure

If $\boldsymbol{X}^T\boldsymbol{X}$ is invertible,

$$\boldsymbol{b}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

is the unique minimizer of the RSS.

Remark: R will improve the stability and doesn't generally use this form exactly as it looks here.

Least squares estimator

Now assume $(\boldsymbol{x}_1, Y_1), \cdots, (\boldsymbol{x}_n, Y_n)$ are from the linear regression model. If $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, the (random) least squares estimator for the unknown parameter vector $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

Lecture 3: Discussion

- What assumptions did we make on the data for least squares estimation?

Lecture 2: Application example

Example

TABLE 7.1 Data for Example 7.2

Observation Number	y	x_1	x_2
1	2	0	2
2	3	2	6
3	2	2	7
4	7	2	5
5	6	4	9
6	8	4	8
7	10	4	7
8	7	6	10
9	8	6	11
10	12	6	9
11	11	8	15
12	14	8	13

Figure: Data set from Rencher's textbook

Example

You are given,

$$\sum y_i = 90, \sum x_{i,1} = 52, \qquad \sum x_{i,2} = 102, \sum x_{i,1}x_{i,2} = 536$$
$$\sum x_{i,1}^2 = 296, \sum x_{i,2}^2 = 1004, \quad \sum x_{i,1}y_i = 482, \sum x_{i,2}y_i = 872$$

$$(X^T X)^{-1} = \begin{pmatrix} 0.9747634 & 0.2429022 & -0.22870662 \\ 0.2429022 & 0.1620662 & -0.11119874 \\ -0.2287066 & -0.1111987 & 0.08359621 \end{pmatrix}$$

Compute the $X^T X$ and $\hat{\boldsymbol{\beta}}$.

Example

Use the form of the entries and plug in the values:

$$
\begin{aligned}
X^T X &= \begin{pmatrix} n & \sum x_{i,1} & \sum x_{i,2} \\ \sum x_{i,1} & \sum x_{i,1}^2 & \sum x_{i,1} x_{i,2} \\ \sum x_{i,2} & \sum x_{i,2} x_{i,1} & \sum x_{i,2}^2 \end{pmatrix} \\
&=
\end{aligned}
$$

Example

Use the representation:

$$X^T y = \begin{pmatrix} \sum y_i \\ \sum x_{i,1} y_i \\ \sum x_{i,2} y_i \end{pmatrix}$$

Perform the matrix multiplication:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \begin{pmatrix} 0.9747634 & 0.2429022 & -0.22870662 \\ 0.2429022 & 0.1620662 & -0.11119874 \\ -0.2287066 & -0.1111987 & 0.08359621 \end{pmatrix} \begin{pmatrix} 90 \\ 482 \\ 872 \end{pmatrix} \\ &= \end{aligned}$$

Example

```r
# Load data
y = c(2, 3, 2, 7, 6, 8, 10, 7, 8, 12, 11, 14)
x1 = c(0, 2, 2, 2, 4, 4, 4, 6, 6, 6, 8, 8)
x2 = c(2, 6, 7, 5, 9, 8, 7, 10, 11, 9, 15, 13)

# Construct X matrix (design matrix)
ones = rep(1, 12)
X = matrix(c(ones, x1, x2), ncol = 3)
```

Example

```r
# Compute the least squares solution
XtX = t(X) %*% X
Xty = t(X) %*% y
XtX_inv = solve(XtX) # solve() inverts the matrix X^T X

beta_hat = XtX_inv %*% Xty
```

Example

```r
# Compare with lm
fit = lm(y ~ x1 + x2)
```

Compute the RSS with R:

```r
RSS = sum( fit$residuals^2 )
```

Lecture 3: Estimate σ^2, Interpretation and applications

Learning goals

- Estimate σ^2
- Understand categorical predictors
- Interpret regression coefficients

Lecture 3: Estimate σ^2

Estimate σ^2

- (Random) Residual: $\hat{e}_i = Y_i - \hat{Y}_i = Y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$
- (Random) RSS: $\sum_{i=1}^{n} \hat{e_i}^2$

Estimate σ^2

Under the regression model assumption,

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-(p+1)}\sum_{i=1}^{n}(y_i - \boldsymbol{x_i}^T\hat{\boldsymbol{\beta}})^2 \\ &= \frac{1}{n-(p+1)}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\end{aligned}$$

is used for an estimator/estimate for σ^2.

- Remember the denominator: number of samples minus "size of $\hat{\boldsymbol{\beta}}$"
- [Rencher and Schaalje, 2008] uses the notation: s^2

Poll question

For the iris data of 150 flowers, the formula for the lm function is

$$Petal.Length \sim Sepal.Length + Sepal.Width$$

and the (computed) RSS is 61.43675.

Question: Compute the estimate to σ^2.

Lecture 3: Categorical predictors

Categorical predictors

Consider predictor data of the form $(A, .1), (A, -.1), (C, .5), (B, .2)$. The fitted model using dummy variables:

$$\hat{E}(Y_i|X = \boldsymbol{x}_i) = \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 I(x_i = B) + \hat{\beta}_2 I(x_i = C)$$

Create two dummy variables one if predictor category is B and 0 otherwise and another one if predictor category is C and 0 otherwise.

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Poll question

Question: We have 100 observations with a response, a continuous predictor, and a categorical predictor with 10 categories / levels. What are the dimensions of $\boldsymbol{X}$?

Question: We have a discrete predictor that takes values from $1-100$. Should we treat this as continuous or categorical?

Example: Categorical predictors iris data

```
lm(Petal.Length ~ Species, data=iris)
```

Example: Categorical predictors for iris data

Let's go through what R does here the long way to understand.

Lecture 3: Interpretation of regression coefficients

Working example: iris data

Measurements (cm) of the variables sepal length and width and petal length and width for 50 flowers from each of 3 species of iris (setosa, versicolor, virginica).

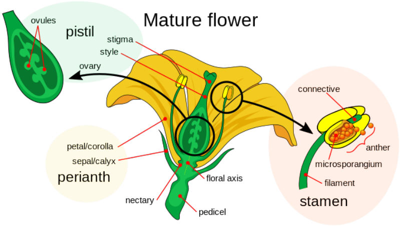


Figure: Picture by Mariana Ruiz

```
data(iris)
```

Interpretation of coefficients

$$\hat{E}(Y|X = \boldsymbol{x}) = \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

- $\hat{\beta}_0$: Estimated the average of $Y|X = \boldsymbol{x}$ with all predictors 0.
- $\hat{\beta}_1$: Estimated average change $Y|X = \boldsymbol{x}$ for one unit increase in x_1 with all other predictors held fixed.
- $\cdots$
- $\hat{\beta}_p$: Estimated average change $Y|X = \boldsymbol{x}$ for one unit increase in x_p with all other predictors held fixed.

Example: iris data

The fitted regression:

$$\hat{Petal}_L = -2.525 + 1.776 * Sepal_L - 1.339 * Sepal_W$$

```
lm(Petal.Length ~ Sepal.Length + Sepal.Width, data =
    iris)
```

Question: Interpret $\hat{\beta}_0$ in the context of the problem. Interpret $\hat{\beta}_1$ in the context of the problem. Interpret $\hat{\beta}_2$ in the context of the problem.

Fit the iris data

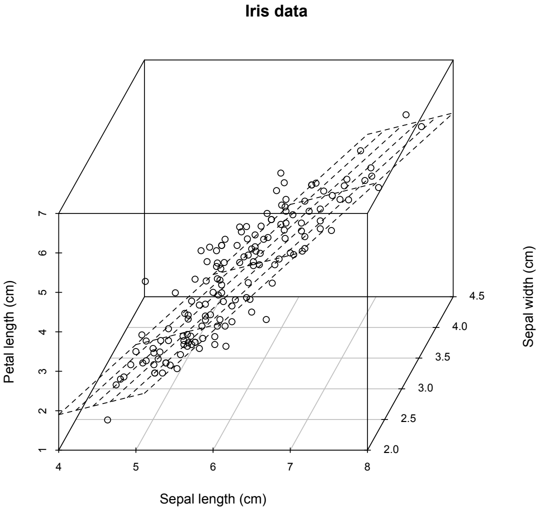
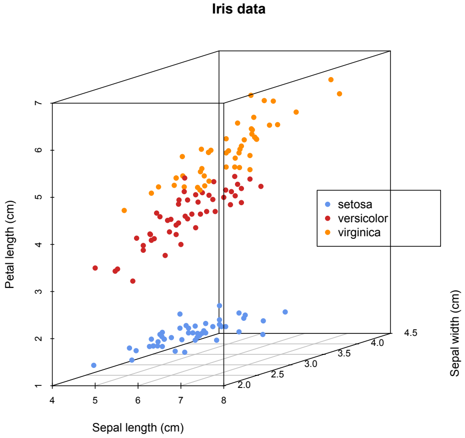


Figure: Linear regression for iris data

Visualization of species in iris data



Question: Does our current model take into account different species?

Interpretation of indicator coefficients with no interactions

We have a categorical predictor x with levels A, B, C and a continuous predictor x_3:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 I(\text{x is category B}) + \hat{\beta}_2 I(\text{x is category C}) + \hat{\beta}_3 x_3 + \cdots$$

Here each category has a different intercept but the same slope.

- $\hat{\beta}_0$: Estimated the average of $Y|X = \boldsymbol{x}$ for category A and all other predictors 0
- $\hat{\beta}_1$: Intercept of category B minus Intercept of category A.
- $\hat{\beta}_1$: Estimated average change of the response from changing to category B from the category A

Example: iris data

The fitted regression:

$$\hat{Petal_L} = -1.63 + 2.17I(\text{versicolor}) + 3.05I(\text{virginica}) + 0.65Sepal_L - 0.04Sepal_W$$

```r
lm(Petal.Length ~ Species + Sepal.Length + Sepal.Width,
    data = iris)
```

Question: How do we interpret the coefficient for the species virginica? What are the dimensions of the matrix $\boldsymbol{X}$?

Example: iris data

Given species is **setosa**,

$$\hat{Petal_L} = -1.63 + 0.65 Sepal_L - 0.04 Sepal_W$$

Given species is **versicolor**,

$$\hat{Petal_L} = (-1.63 + 2.17) + 0.65 Sepal_L - 0.04 Sepal_W$$

- $\hat{\beta}_0 + \hat{\beta}_1$ [Intercept for setosa] plus [intercept for versicolor] minus [intercept for setosa]

Given species is **virginica**,

$$\hat{Petal_L} = (-1.63 + 3.05) + 0.65 Sepal_L - 0.04 Sepal_W$$

- $\hat{\beta}_0 + \hat{\beta}_2$ [Intercept for setosa] plus [intercept for virginica] minus [intercept for setosa]

Fit the iris data

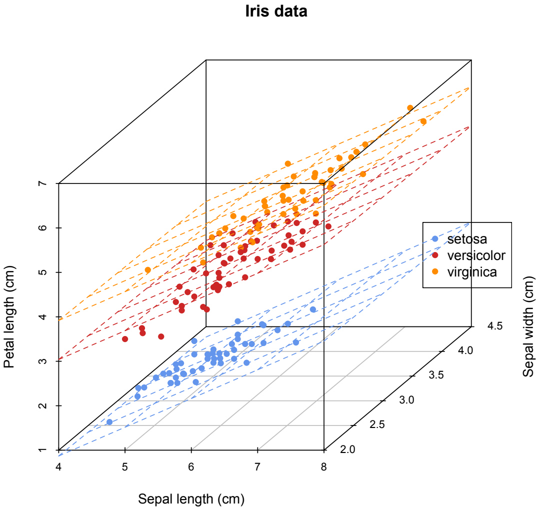


Figure: Linear regression for iris data

Iris data discussion

- Do all the species have the same slopes for sepal length/width?
- Is this reasonable? Can linear regression express this?

Module takeaways

Module takeaways

1. What is similar/different between simple and multiple linear regression? (e.g. estimation, formulae, interpretation, notation, etc.)
2. How do we estimate/compute the coefficients of our model?
3. What is the correct way to interpret the estimated coefficients?
4. How does the interpretation change when we use indicator variables and interaction terms?

References I

Alvin C Rencher and G Bruce Schaalje. *Linear models in statistics*. John Wiley & Sons, 2008.