Lecture 3: Linear Regression and Linear Algebra Review

STAT GR5206 *Statistical Computing & Introduction to Data Science*

Gabriel Young
Columbia University

May 25, 2017

Last Time

- **Filtering**. Accessing elements of a structure based on some criteria. `v[v>5]`, `m[ m[,1]!=0, ]`.
- **Lists**. Elements can all be different types. Access like `l[[3]]`, `l$name`. Create with `list()`.
- `NA` **and** `NULL` **values**. `NA` is missing data and `NULL` doesn't exist.
- **Factors and Tables**. Factors is how `R` classifies categorical variables.
- **Dataframes**. Used for data that is organized with rows indicating cases and columns indicating variables.
- **Importing and Exporting Data in** `R`. Use `read.csv()` and `read.table()` depending on dataset type. The working directory.
- **Control Statements**. We studied iteration, `for` loops and `while` loops, and `if, else` statements.
- **Vectorized Operations**. To be used instead of iterations.

Multiple Linear Regression

Multiple Linear Regression

Example

A large national grocery retailer tracks productivity and costs of its facilities closely. Consider a data set obtained from a single distribution center for a one-year period. Each data point for each variable represents one week of activity. The variables included are number of cases shipped in thousands (X_1), the indirect costs of labor as a percentage of total costs (X_2), a qualitative predictor called holiday that is coded 1 if the week has a holiday and 0 otherwise (X_3), and total labor hours (Y).

Multiple Linear Regression

Suppose, as statisticians, we are asked to build a model to predict total labor hours in the future using this dataset.

What information would be useful to provide such a model?

- Is there a relationship between holidays and total labor hours? What about number of cases shipped? Indirect costs?
- How strong are these relationships?
- Is the relationship linear?

Multiple Linear Regression

Suppose, as statisticians, we are asked to build a model to predict total labor hours in the future using this dataset.

What information would be useful to provide such a model?

- Is there a relationship between holidays and total labor hours? What about number of cases shipped? Indirect costs?
- How strong are these relationships?
- Is the relationship linear?

Multiple linear regression can be used to answer each of these questions.

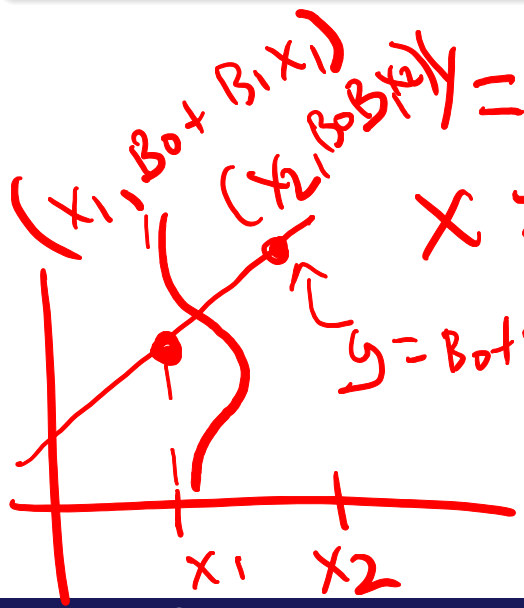
Multiple Linear Regression

Models a relationship between two or more **explanatory** variables and a **response** variable by fitting a linear equation to observed data.

General Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

with $\epsilon \sim N(0, \sigma^2)$. ~> homogenity of variance



Y = response, dependent

X = covariates, ind. variables, features, predicators

$Y = f(\underline{X}) + \epsilon$

$\underline{X} = X_1 \cdots X_p$

Multiple Linear Regression

Models a relationship between two or more **explanatory** variables and a **response** variable by fitting a linear equation to observed data.

General Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

with $\epsilon \sim N(0, \sigma^2)$.

Coefficient β_j quantifies the association between the predictor and the response.

Interpret β_j as the average effect on Y of one unit increase of X_j, *holding all other predictors fixed*.

Multiple Linear Regression

Matrix Formulation

Using a set of training observations (data): $(Y_i, X_{i1}, X_{i2}, \ldots, X_{ip})$ for $i = 1, 2, \ldots, n$, we want to estimate the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i,$$

with $\epsilon \sim N(0, \sigma^2)$. We can represent this with matrices as follows:

vector of 0s

bold vector bold bold

$$Y = X\beta + \epsilon,$$

multiple linear regression

$\epsilon \sim N(0, \sigma^2 \cdot I)$ index

where

$$Y = (Y_1, Y_2, \ldots, Y_n)^T \in \mathbb{R}^n, \quad X = \text{design matrix} \in \mathbb{R}^{n \times (p+1)}$$

$$\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{p+1}, \quad \epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T \in \mathbb{R}^n.$$

The Training Set

Note that we refer to the observations as the **training data** because we will use these training data observations to **train**, or teach, our method how to estimate the model.

The Training Set

Note that we refer to the observations as the **training data** because we will use these training data observations to **train**, or teach, our method how to estimate the model.

This is in contrast to the **test set** which is data that isn't used to estimate (or train) the model, but rather to test how well the model is at prediction.

Multiple Linear Regression

Example (Multiple Linear Regression Model)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

where,

- Total labor hours (Y).
- Number of cases shipped (in thousands) (X_1) .
- Indirect costs of labor as a percentage of total costs (X_2).
- Holiday (X_3) with

$$X_{i3} = \begin{cases} 1 & \text{holiday week,} \\ 0 & \text{otherwise.} \end{cases}$$

Multiple Linear Regression

Example

Case	Y	$X1$	$X2$	$X3$
1	4264	305.657	7.17	0
2	4496	328.476	6.20	0
3	4317	317.164	4.61	0
4	4292	366.745	7.02	0
5	4945	265.518	8.61	1
6	4325	301.995	6.88	0
⋮	⋮	⋮	⋮	⋮
48	4993	442.782	7.61	1
49	4309	322.303	7.39	0
50	4499	290.455	7.99	0
51	4186	411.750	7.83	0
52	4342	292.087	7.77	0

Multiple Linear Regression

Design Matrix

$$X = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

What is the dimension of the design matrix?

Example

need a placeholder for β_0

$$X = \begin{pmatrix} 1 & 305.657 & 7.17 & 0 \\ 1 & 328.476 & 6.20 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 411.750 & 7.83 & 0 \\ 1 & 292.087 & 7.77 & 0 \end{pmatrix}$$

Parameter Estimation

Using the training data, how do we estimate the parameters of the linear regression model? How do we find

$$\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T$$

which provide predictions

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p? \tag{1}$$

We say, how do we **fit** or how do we **train** the model?

Least Squares

Parameter Estimation

Using the training data, how do we estimate the parameters of the linear regression model? How do we find

$$\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T$$

Least Squares is equivalent to Max Likelihood b/c of normality

which provide predictions

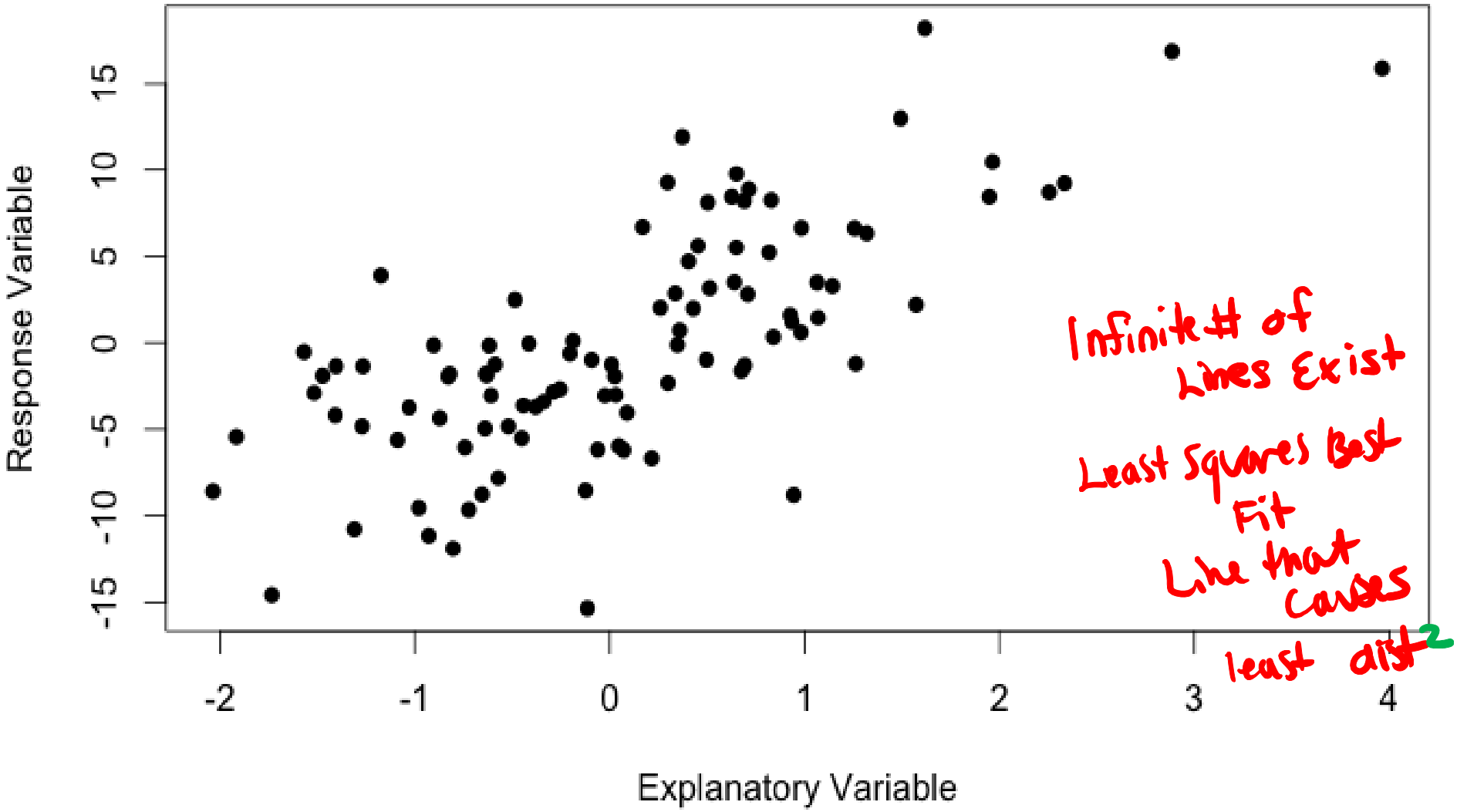
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p? \tag{1}$$

We say, how do we **fit** or how do we **train** the model?

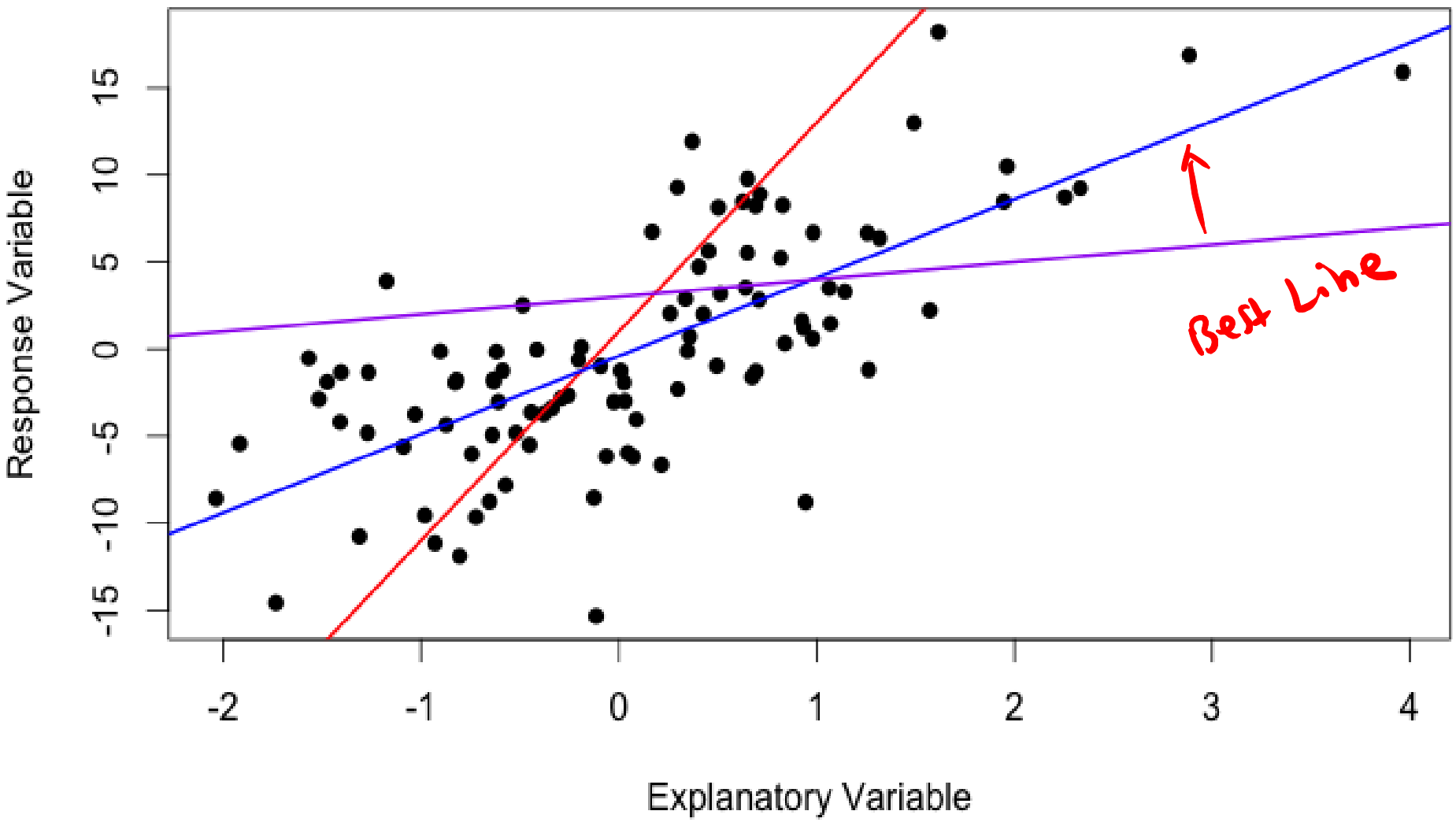
Least Squares Estimate

The **least squares line** is calculated from the training data by minimizing the sum of the squares of the vertical deviations from each data point to the line.

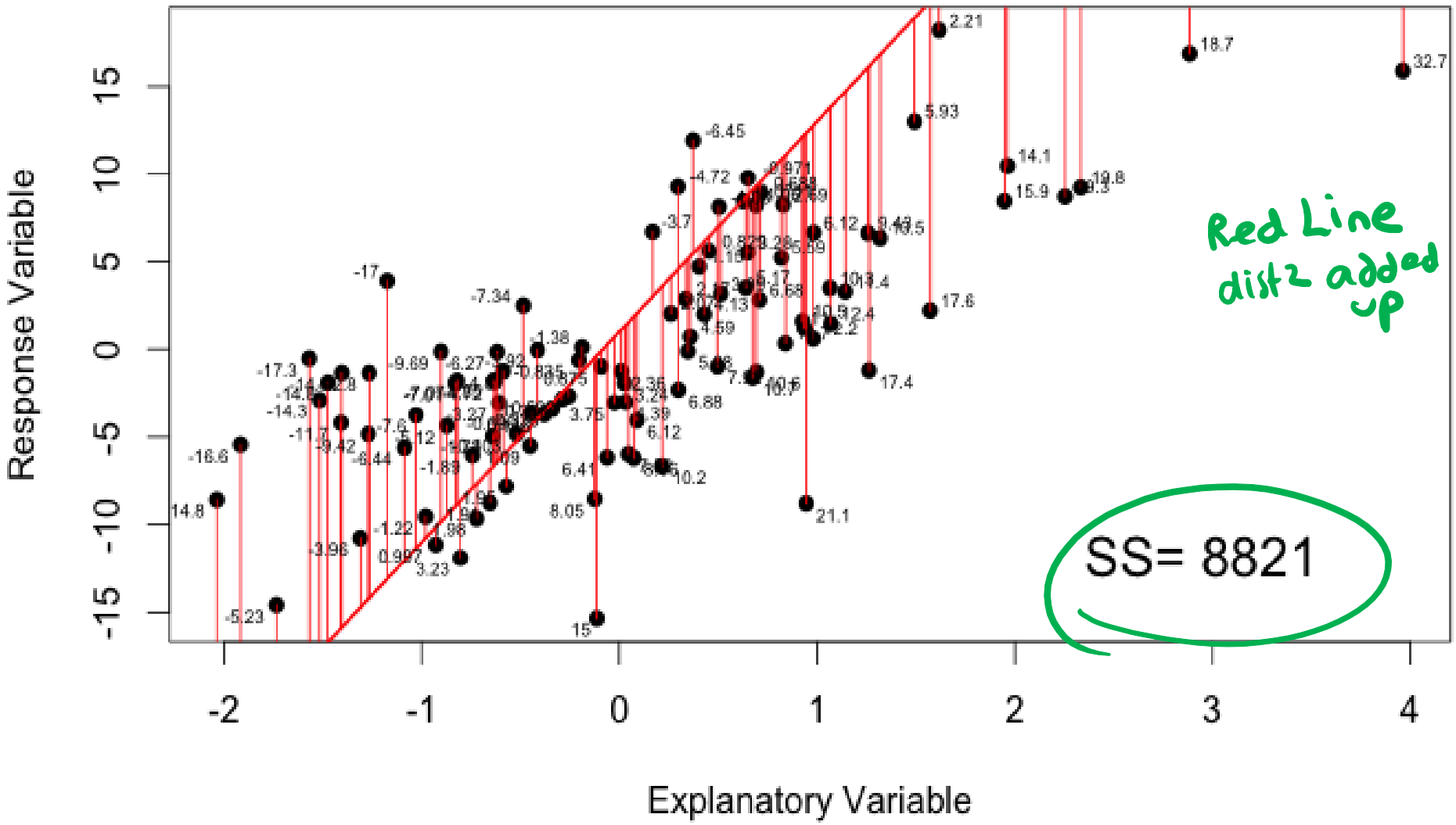
The Least Squares Line



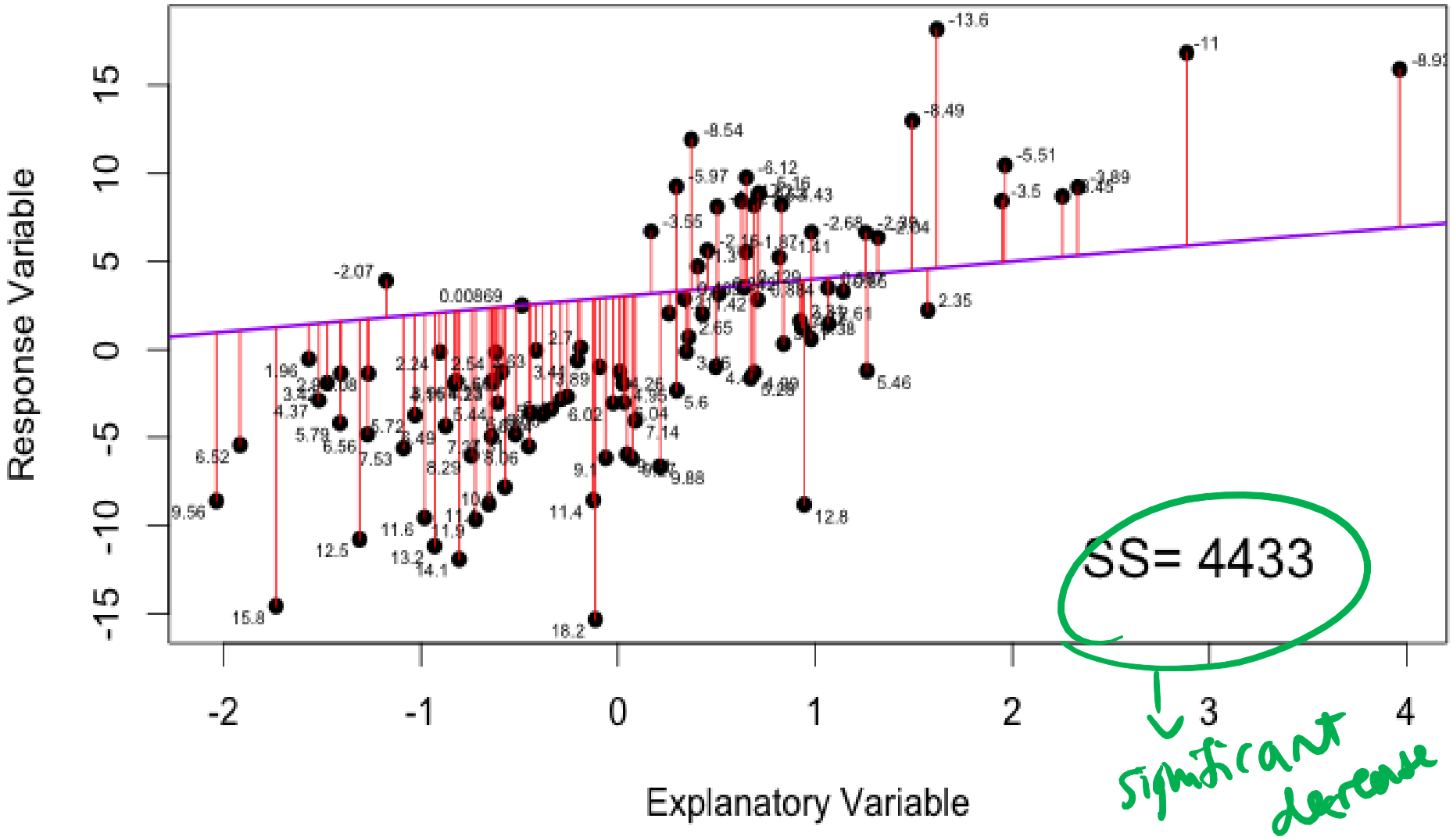
The Least Squares Line



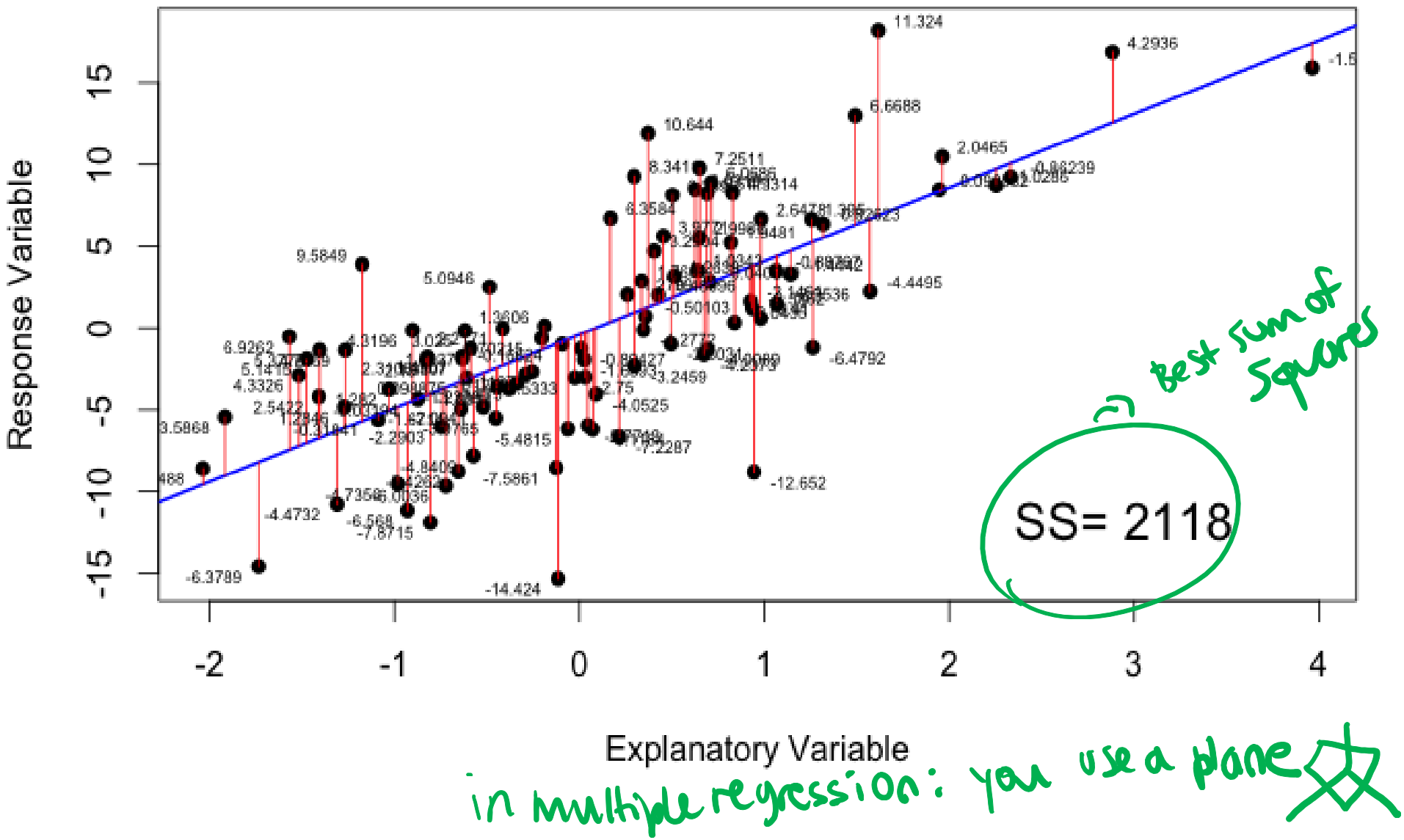
The Least Squares Line



The Least Squares Line



The Least Squares Line



Parameter Estimation

Least Squares Estimate

Define an objective function $Q(b)$ as follows.

$$Q(b_0, b_1, \ldots, b_p) := \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_p X_{ip}))^2$$
$$= \|Y - Xb\|^2,$$

dot product squared

which we minimize with respect to $b = (b_0, b_1, b_2, \ldots, b_p) \in \mathbb{R}^{p+1}$.

Every common problem is an optimization problem

Parameter Estimation

Theorem

If design matrix X has full column rank, then the global minimizer of

$$Q(b) = \|Y - Xb\|^2$$

with respect to $b = (b_0, b_1, \ldots, b_{p-1})^T$ *is*

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

↪ optimal regression function

↳ produces a vector of $\hat{\beta}$

To come:

- What do we mean by **full column rank**?
- Is there a **geometric interpretation** of this result?

$$\underline{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & & \\ 1 & x_{n1} & x_{n2} \end{bmatrix}$$

Bilodeau

$$(\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y} = \underline{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \leftarrow \qquad \underline{Y} = \begin{bmatrix} y_1 \\ y_n \end{bmatrix}$$

Parameter Estimation

Sketch of Theorem Proof for Least Squares

First note,

$$\begin{aligned} Q(b) = \|Y - Xb\|^2 &= (Y - Xb)^T (Y - Xb) \\ &= Y^T Y - Y^T Xb + (Xb)^T Y + (Xb)^T Xb \\ &= b^T X^T Xb - 2Y^T Xb + Y^T Y. \end{aligned}$$

Taking the first derivative of $Q(b)$ with respect to b and equating the derivative equal to zero gives

$$2X^T Xb - 2X^T Y = 0.$$

need full column rank b/c

won't be invertible

Solving the expression for b yields $b = (X^T X)^{-1} X^T Y$.

The second derivative is a positive definite matrix which implies that $Q(b)$ achieves its minimum at $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Parameter Estimation

Example

```
> Grocery <- read.table("Kutner_6_9.txt", header=T)
> head(Grocery)
```

```
     Y       X1   X2 X3
1 4264 305.657 7.17  0
2 4496 328.476 6.20  0
3 4317 317.164 4.61  0
4 4292 366.745 7.02  0
5 4945 265.518 8.61  1
6 4325 301.995 6.88  0
```

```
> # Construct design matrix
> X <- cbind(rep(1,52), Grocery$X1, Grocery$X2, Grocery$X3)
```

row of 1s

all the Xs

Parameter Estimation

Example

Least Square Estimator: $\hat{\beta} = (X^T X)^{-1} X^T Y$

```
> beta_hat <- solve((t(X) %*% X)) %*% t(X) %*% Grocery$Y
> round(t(beta_hat), 2)
```

```
        [,1] [,2]   [,3]   [,4]
[1,] 4149.89 0.79 -13.17 623.55
```

What is the estimated model?

Parameter Estimation

Example

Least Square Estimator: $\hat{\beta} = (X^T X)^{-1} X^T Y$

```
> beta_hat <- solve((t(X) %*% X)) %*% t(X) %*% Grocery$Y
> round(t(beta_hat), 2)
```

```
        [,1] [,2]   [,3]   [,4]
[1,] 4149.89 0.79 -13.17 623.55
```

What is the estimated model?

$$\hat{Y} = 4149.89 + 0.79X_1 - 13.17X_2 + 623.56X_3$$

$$\widehat{\text{Labor.Hours}} = 4149.89 + 0.79 \times \text{Cases.Shipped} - 13.17 \times \text{Indirect.Costs} + 623.56 \times \text{Holiday}.$$

Parameter Estimation

Estimated Model:

$$\widehat{\text{Labor.Hours}} = 4149.89 + 0.79 \times \text{Cases.Shipped} - 13.17 \times \text{Indirect.Costs} + 623.56 \times \text{Holiday}.$$

Example: Prediction

How many labor hours does out model predict for a **holiday** week with **350000 cases** shipped and indirect costs at **8.5 percent**?

Parameter Estimation

Estimated Model:

$$\widehat{\text{Labor.Hours}} = 4149.89 + 0.79 \times \text{Cases.Shipped} - 13.17 \times \text{Indirect.Costs} + 623.56 \times \text{Holiday}.$$

Example: Prediction

How many labor hours does out model predict for a **holiday** week with **350000 cases** shipped and indirect costs at **8.5 percent**?

$$\begin{aligned}\widehat{\text{Labor.Hours}} &= 4149.89 + 0.79(\mathbf{350000}) - 13.17(\mathbf{8.5}) + 623.56(\mathbf{1}) \\ &= 4938.01 \text{ hours}.\end{aligned}$$

Fitting Linear Models in R

`lm(formula,data)` is used to fit linear models.

linear model function

Example

```
> lm0 <- lm(Y ~ X1 + X2 + X3, data = Grocery)
> lm0
```

formula for regression

response

regressing Y

different features

```
Call:
lm(formula = Y ~ X1 + X2 + X3, data = Grocery)

Coefficients:
(Intercept)          X1          X2          X3
  4149.8872      0.7871    -13.1660    623.5545
```

estimation from least square formulation via matrix

Fitted Values and Residuals

- The i^{th} fitted value is denoted $\hat{Y}_i$ and defined by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_{p-1} X_{i,p-1}.$$

- Denote the fitted values by the $n \times 1$ vector

$$\hat{Y} = \left(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n\right)^T = X\hat{\beta}.$$

- The i^{th} residual denoted e_i is the difference between the actual response value and its corresponding fitted value: $e_i = Y_i - \hat{Y}_i$.
- Denote the residuals by the $n \times 1$ vector

$$e = (e_1, e_2, \ldots, e_n)^T = Y - \hat{Y}.$$

Fitted Values and Residuals

For an estimated linear model in R,

- Compute **residuals** with `residuals()`.
- Compute **fitted values** with `fitted()`.

Fitted Values and Residuals

For an estimated linear model in R,

- Compute **residuals** with `residuals()`.
- Compute **fitted values** with `fitted()`.

Example

```
> lm0 <- lm(Y ~ X1 + X2 + X3, data = Grocery)
> residuals(lm0)[1:5]
```

```
       1          2          3          4          5
-32.06348 169.20509 -21.82543 -54.11955  75.93372
```

```
> fitted(lm0)[1:5]
```

```
       1        2        3        4        5
4296.063 4326.795 4338.825 4346.120 4869.066
```

Model Summary

```
> summary(lm0)

Residuals:
    Min      1Q  Median      3Q     Max
-264.05 -110.73  -22.52   79.29  295.75

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4149.8872   195.5654  21.220  < 2e-16 ***
X1             0.7871     0.3646   2.159   0.0359 *
X2           -13.1660    23.0917  -0.570   0.5712
X3           623.5545    62.6409   9.954 2.94e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 143.3 on 48 degrees of freedom
Multiple R-squared:  0.6883,  Adjusted R-squared:  0.6689
F-statistic: 35.34 on 3 and 48 DF,  p-value: 3.316e-12
```

summary of linear Model

classical output

Linear Algebra Review

Linear Independence

Definition

A set of vectors $\{v_1, \ldots, v_p\}$, each in $\mathbb{R}^n$, is **linearly independent** if

$$c_1 v_1 + c_2 v_2 + \cdots + c_p v_p = 0$$

has only the trivial solution, i.e., $c_1 = c_2 = \cdots = c_p = 0$.

Any combination of vector = linearly dependent

Linear Independence

Definition

A set of vectors $\{v_1, \ldots, v_p\}$, each in $\mathbb{R}^n$, is **linearly independent** if

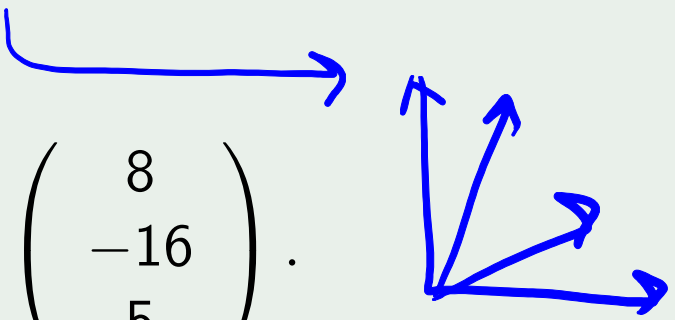
$$c_1 v_1 + c_2 v_2 + \cdots + c_p v_p = 0$$

has only the trivial solution, i.e., $c_1 = c_2 = \cdots = c_p = 0$.

Example

Vectors v_1 and v_2 are **linearly independent** since if $c_1 v_1 + c_2 v_2 = 0$, then it must be the case $c_1 = c_2 = 0$.

$$v_1 = \begin{pmatrix} 3 \\ -2 \\ 7 \end{pmatrix}, \; v_2 = \begin{pmatrix} 8 \\ -16 \\ 5 \end{pmatrix}.$$

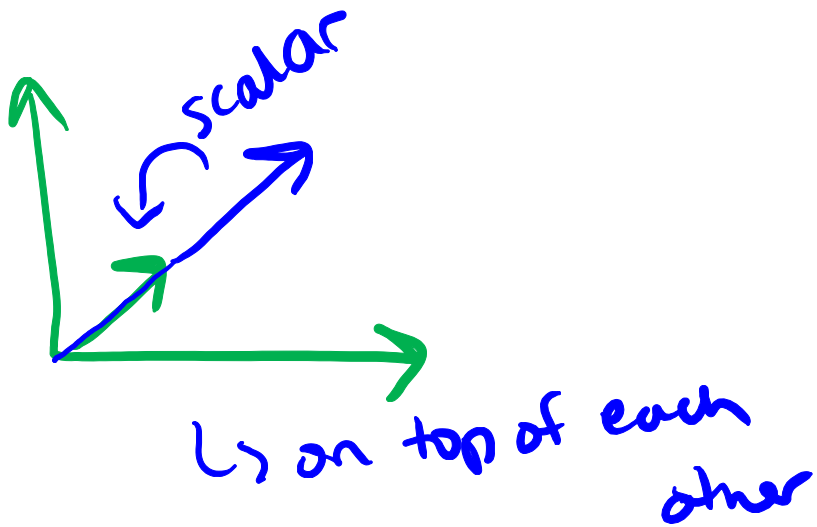


Linear Dependence

Definition

A set of vectors $\{v_1, \ldots, v_p\}$, each in $\mathbb{R}^n$, is **linearly dependent** if there exists scalars $c_1, \ldots, c_p$, not all zero, such that

$$c_1 v_1 + c_2 v_2 + \cdots + c_p v_p = 0.$$



Linear Dependence

Definition

A set of vectors $\{v_1, \ldots, v_p\}$, each in $\mathbb{R}^n$, is **linearly dependent** if there exists scalars $c_1, \ldots, c_p$, not all zero, such that

$$c_1 v_1 + c_2 v_2 + \cdots + c_p v_p = 0.$$

Example

Vectors v_1 and v_2 are **linearly dependent** since $2v_1 + v_2 = 0$.

$$v_1 = \begin{pmatrix} 3 \\ -2 \\ 7 \end{pmatrix}, \; v_2 = \begin{pmatrix} -6 \\ 4 \\ -14 \end{pmatrix}.$$

Note $-2 \times v_1 = v_2$.

Span

Definition

The **span** of a set of vectors $\{v_1, \ldots, v_p\}$, each in $\mathbb{R}^n$, is the collection of all vectors that can be written in the form

$$c_1 v_1 + c_2 v_2 + \cdots + c_p v_p$$

where $c_1, \ldots, c_p$ are scalars. Denote this set by $\text{Span}\{v_1, \ldots, v_p\}$.

Span

Definition

The **span** of a set of vectors $\{v_1, \ldots, v_p\}$, each in $\mathbb{R}^n$, is the collection of all vectors that can be written in the form

$$c_1 v_1 + c_2 v_2 + \cdots + c_p v_p$$

where $c_1, \ldots, c_p$ are scalars. Denote this set by $\text{Span}\{v_1, \ldots, v_p\}$.

- $\text{Span}\{v_1, \ldots, v_p\}$ is the set of all linear combinations of $v_1, \ldots, v_p$.
- Say $\text{Span}\{v_1, \ldots, v_p\}$ is the **subspace** of $\mathbb{R}^n$ generated by $v_1, \ldots, v_p$.
- A vector b is in $\text{Span}\{v_1, \ldots, v_p\}$ if a solution exists to the vector equation

$$x_1 v_1 + x_2 v_2 + \cdots + x_p v_p = b.$$

Example

Consider vectors

$$v_1 = \begin{pmatrix} 3 \\ -2 \\ 7 \end{pmatrix}, \; v_2 = \begin{pmatrix} -2 \\ 12 \\ 9 \end{pmatrix}, \; b = \begin{pmatrix} 8 \\ -16 \\ 5 \end{pmatrix}.$$

Is the vector b an element of Span$\{v_1, v_2\}$?

Example

Consider vectors

$$v_1 = \begin{pmatrix} 3 \\ -2 \\ 7 \end{pmatrix}, \; v_2 = \begin{pmatrix} -2 \\ 12 \\ 9 \end{pmatrix}, \; b = \begin{pmatrix} 8 \\ -16 \\ 5 \end{pmatrix}.$$

Is the vector b an element of $\text{Span}\{v_1, v_2\}$?

To check if $b \in \text{Span}\{v_1, v_2\}$ we must find scalars c_1, c_2 such that

$$c_1 v_1 + c_2 v_2 = b.$$

Notice,

$$2v_1 - v_2 = 2\begin{pmatrix} 3 \\ -2 \\ 7 \end{pmatrix} - \begin{pmatrix} -2 \\ 12 \\ 9 \end{pmatrix} = \begin{pmatrix} 8 \\ -16 \\ 5 \end{pmatrix}. \; \surd$$

Hence $b \in \text{Span}\{v_1, v_2\}$.

Column Space

Definition

The **column space** of a matrix A is the set $\mathcal{C}(A)$ of all linear combinations of the columns of A. If $A = [v_1, \ldots, v_n]$, then $\mathcal{C}(A) = \text{Span}\{v_1, \ldots, v_n\}$.

Column Space

Definition

The **column space** of a matrix A is the set $\mathcal{C}(A)$ of all linear combinations of the columns of A. If $A = [v_1, \ldots, v_n]$, then $\mathcal{C}(A) = \text{Span}\{v_1, \ldots, v_n\}$.

Example

$$A = \begin{pmatrix} v_1 & v_2 \end{pmatrix} = \begin{pmatrix} 3 & -2 \\ -2 & 12 \\ 7 & 9 \end{pmatrix}, \quad b = \begin{pmatrix} 8 \\ -16 \\ 5 \end{pmatrix}.$$

b is in $\mathcal{C}(A)$ because

$$2v_1 - v_2 = 2\begin{pmatrix} 3 \\ -2 \\ 7 \end{pmatrix} - \begin{pmatrix} -2 \\ 12 \\ 9 \end{pmatrix} = \begin{pmatrix} 8 \\ -16 \\ 5 \end{pmatrix}.$$

Rank

Definition

The **rank** of a matrix A, denoted rank(A), is the number of linearly independent columns of A.

Linearly dependent

columns of measurement in ft & inches

Rank

Definition

The **rank** of a matrix A, denoted rank(A), is the number of linearly independent columns of A.

Example

$$A = \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix} = \begin{pmatrix} 3 & -2 & 8 \\ -2 & 12 & -16 \\ 7 & 9 & 5 \end{pmatrix}.$$

rank$(A) = 2$ because v_1 and v_2 are linearly independent, but

$$0 = 2v_1 - v_2 - v_3 = 2\begin{pmatrix} 3 \\ -2 \\ 7 \end{pmatrix} - \begin{pmatrix} -2 \\ 12 \\ 9 \end{pmatrix} - \begin{pmatrix} 8 \\ -16 \\ 5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

combination =

Inverse

Theorem

The following statements are equivalent if A is a $p \times p$ matrix

- *A is invertible.*
- $\mathcal{C}(A) = \mathbb{R}^p$
- $rank(A) = p$

if you don't have full column rank you can't invert matrix!

Theorem

If X is a $n \times p$ matrix with $rank(X) = p$, then $rank(X^T X) = p$

full rank = no linear combination in matrix

matrix packages to get this

Inverse

Theorem

The following statements are equivalent if A is a $p \times p$ matrix

- *A is invertible.*
- $\mathcal{C}(A) = \mathbb{R}^p$
- $rank(A) = p$

Theorem

If X is a $n \times p$ matrix with $rank(X) = p$, then $rank(X^T X) = p$

Why do we care about the above result?

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

non full rank can't have

In the Grocery retailer example consider introducing another variable *not holiday*.

Example

For $i = 1, 2, \ldots, 52$ weeks,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

$$X_{i3} = \begin{cases} 1 & \text{holiday week} \\ 0 & \text{otherwise} \end{cases} \qquad X_{i4} = \begin{cases} 1 & \text{not holiday week} \\ 0 & \text{otherwise} \end{cases}$$

For week i,

- Total labor hours (Y_i)
- Number of cases shipped (X_{i1})
- Indirect costs of the total labor hours as a percentage (X_{i2})
- Holiday (X_{i3})
- Not holiday (X_{i4})

if added up it'll end up being all 1s
- not invertible
- linear dependence

Multiple Linear Regression

Example

$$X = \begin{pmatrix} 1 & 305657 & 7.17 & 0 & 1 \\ 1 & 328476 & 6.20 & 0 & 1 \\ 1 & 317164 & 4.61 & 0 & 1 \\ 1 & 366745 & 7.02 & 0 & 1 \\ 1 & 265518 & 8.61 & 1 & 0 \\ 1 & 301995 & 6.88 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 442782 & 7.61 & 1 & 0 \\ 1 & 322303 & 7.39 & 0 & 1 \\ 1 & 290455 & 7.99 & 0 & 1 \\ 1 & 411750 & 7.83 & 0 & 1 \\ 1 & 292087 & 7.77 & 0 & 1 \end{pmatrix}$$

X_1 $\quad$ $X_4 + X_5 = X_1$

rank 4 despite 5 columns bc rank deficient

(2) (1) (3) (4)

Notice: $col_1 = col_4 + col_5 \longrightarrow \text{rank}(X) = 4, \text{rank}(X^T X) = 4.$

The Projection Approach to Linear Regression

Some Definitions

Definition

1. The **squared norm** of a vector $X \in \mathbb{R}^n$ is

$$\|X\|^2 = \sum_{i=1}^{n} X_i^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

2. The **distance** between vectors $X, Y \in \mathbb{R}^n$ is

$$\|X - Y\| = \|Y - X\|$$

3. The **inner product** of $X, Y \in \mathbb{R}^n$ is

$$<X, Y> = X^T Y = \sum_{i=1}^{n} X_i Y_i = X_1 Y_1 + X_2 Y_2 + \cdots + X_n Y_n.$$

4. Vectors $X, Y \in \mathbb{R}^n$ are **orthogonal** or perpendicular if $<X, Y> = 0$.

↳ best case of linear independence

Projections

Definition

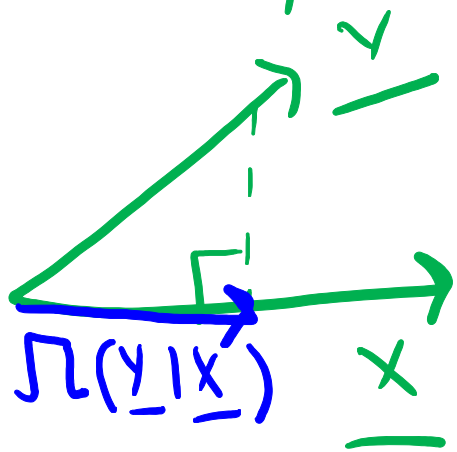
Let $X, Y \in \mathbb{R}^n$. The **projection** of Y onto X, written $\Pi(Y|X)$, is defined to be

$$\Pi(Y|X) = \frac{<X, Y>}{\|X\|^2} X.$$

dot product of $X \cdot Y$ → scalar

norm2 → scalar

projection = vector b/c there's a vector on the right hand side



Projections

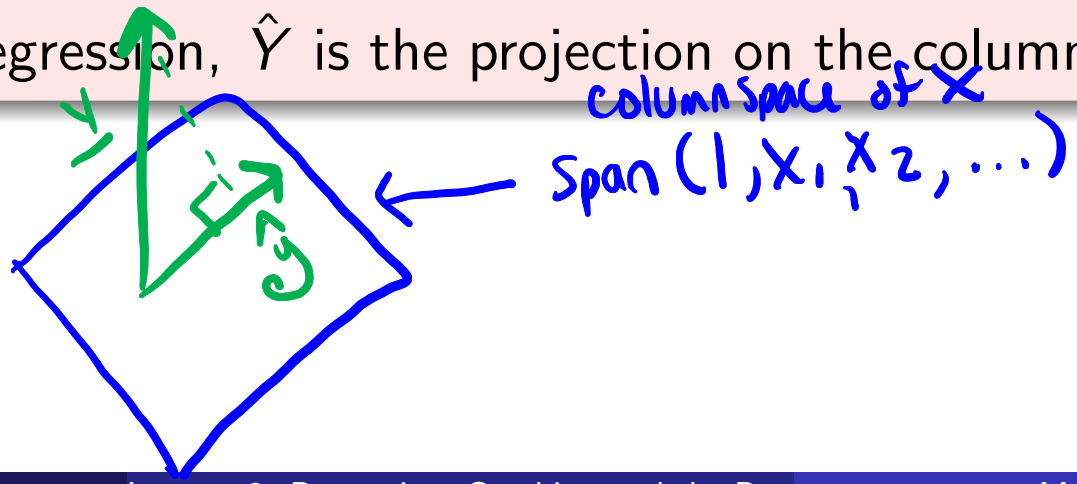
Definition

Let $X, Y \in \mathbb{R}^n$. The **projection** of Y onto X, written $\Pi(Y|X)$, is defined to be

$$\Pi(Y|X) = \frac{<X, Y>}{\|X\|^2} X.$$

Important Idea

In multiple linear regression, $\hat{Y}$ is the projection on the column space of X.



Projections

Example

Consider the least squares estimator for the regression through the origin model: $Y = \beta X_1 + \epsilon$. Then, from the Theorem

$$\hat{\beta} = (X_1^T X_1)^{-1} X_1^T Y = \frac{<X_1, Y>}{\|X_1\|^2}.$$

$y_i = \beta x_i + \epsilon$

$\underline{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ $\underline{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

$\underline{X}^T \underline{X} = \sum x_i^2$

$(\underline{X}^T \underline{X})^{-1} X^T Y = (\sum x_i^{-2})^{-1} \sum x_i y_i$

regression through origin w/ 1 covariant

Projections

Example

Consider the least squares estimator for the regression through the origin model: $Y = \beta X_1 + \epsilon$. Then, from the Theorem

$$\hat{\beta} = (X_1^T X_1)^{-1} X_1^T Y = \frac{<X_1, Y>}{\|X_1\|^2}.$$

Then we find predicted values

$$\hat{Y} = \hat{\beta} X_1 = \frac{<X_1, Y>}{\|X_1\|^2} X_1 = \Pi(Y|X_1),$$

by the *projection formula*.

Optional Reading

- Chapter 3 (3.1, 3.2, 3.6) from *An Introduction to Statistical Learning*.
- Chapter 1 (Vectors and Vector Spaces) found here from G. Donald Allen's Linear Algebra course (class website) at Texas A & M.