Lecture 3: Linear Regression, Linear Algebra Review and Bootstrapping

STAT GR5206 *Statistical Computing & Introduction to Data Science*

Linxi Liu
Columbia University

September 21, 2018

Last Time

- `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.

Class Notes

Check Yourself (Warm Up)

Iris

- Use the built-in `iris` dataset:
- This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

Check Yourself (Warm Up)

Tasks: Filtering Data

Use the built-in `iris` dataset:

- How many of the iris are species `versicolor` and have a petal width of less than or equal to 1.2?
- What is the mean petal length of the `setosa` species iris? (Do this with filtering and with `tapply()`.)
- Make a table of iris species for only those iris with sepal width greater than or equal to 3.0.
- Use the `ifelse()` command to create a new variable `Versicolor` that's an indicator variable (1 if the iris species is versicolor and 0 otherwise). Use the `table()` function to check your result.

Code Example

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

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:

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

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

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

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?

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

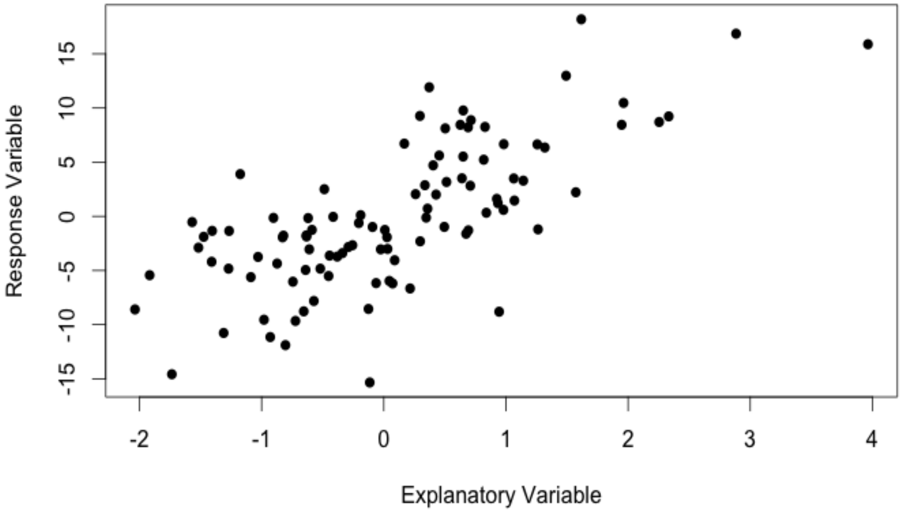
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p? \tag{1}$$

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

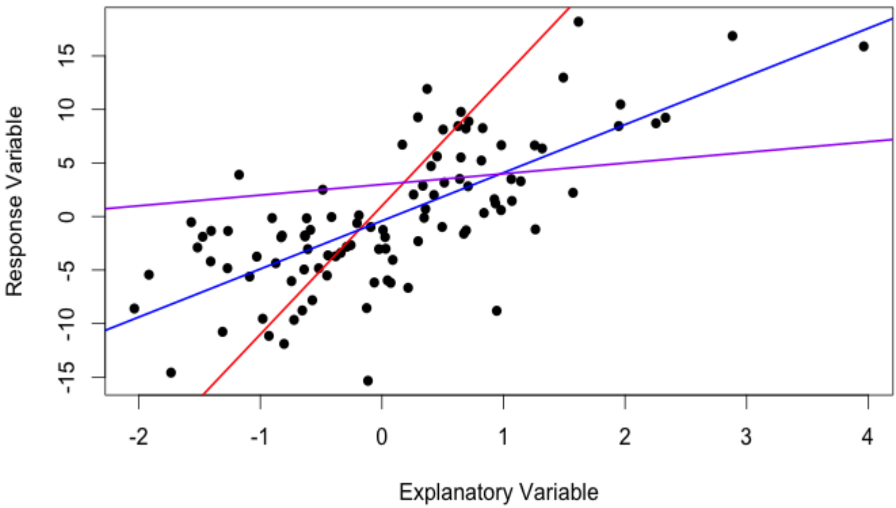
Least Squares Estimate

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.

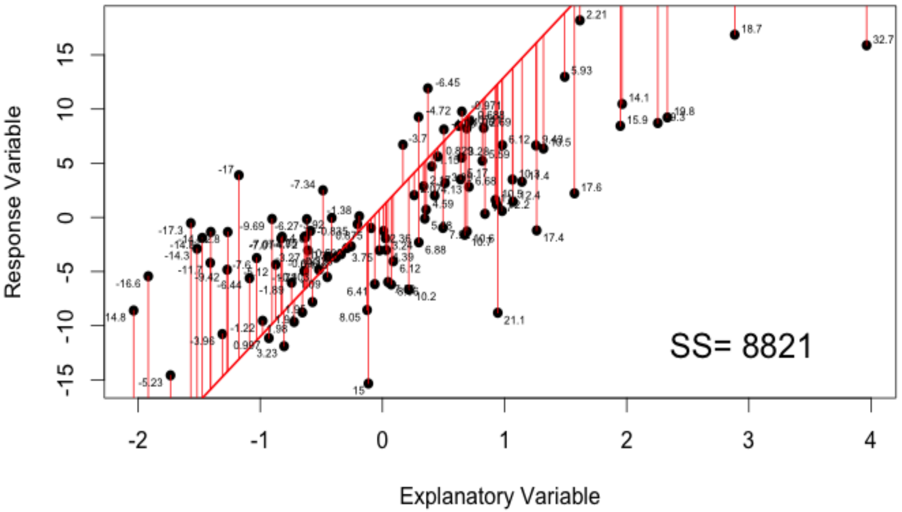
The Least Squares Line



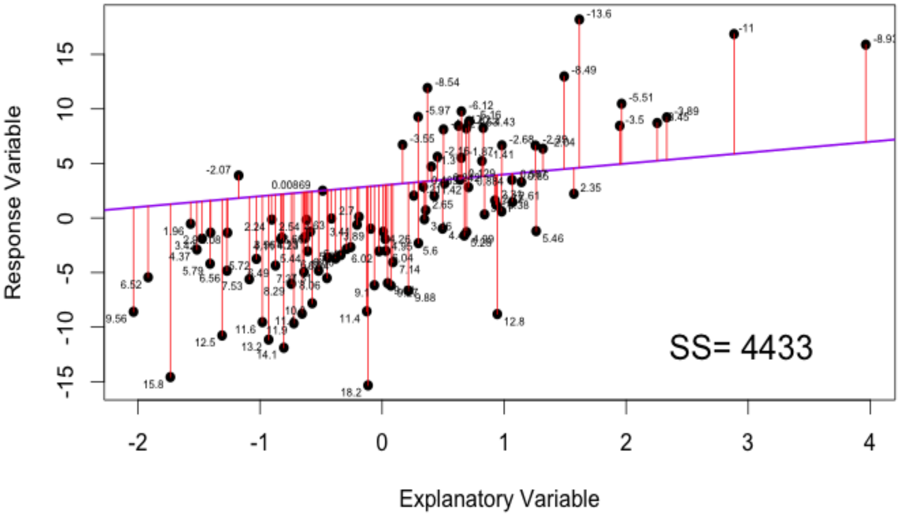
The Least Squares Line



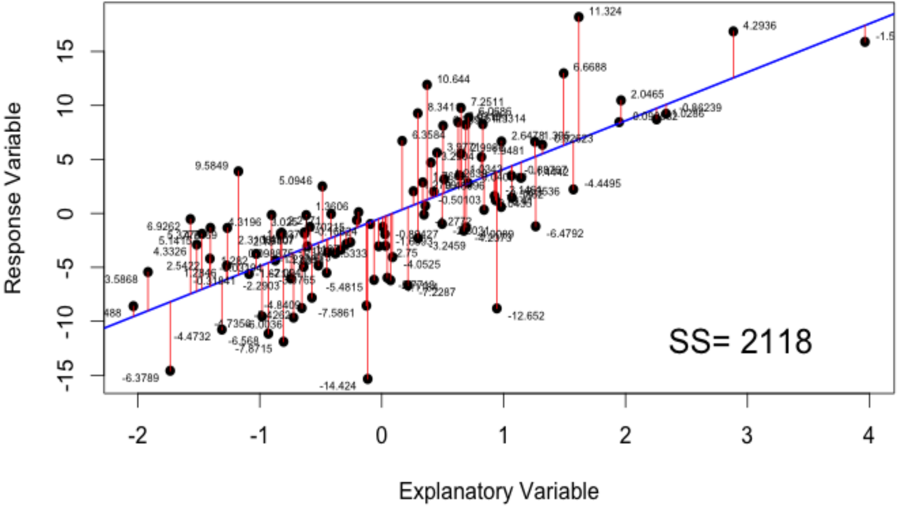
The Least Squares Line



The Least Squares Line



The Least Squares Line



Parameter Estimation

Least Squares Estimate

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

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

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

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

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

To come:

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

Parameter Estimation

Sketch of Theorem Proof

First note,

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

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

$$2X^TXb - 2X^TY = 0.$$

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

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

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)
```

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?

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

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

Parameter Estimation

Estimated Model:

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

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:

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

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.

Example

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

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

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

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
```

Linear Algebra Review

Rank

Definition

The **rank** of a matrix A, denoted $\text{rank}(A)$, is the number of linearly independent columns of A.

Rank

Definition

The **rank** of a matrix A, denoted $\text{rank}(A)$, is the number of linearly independent columns of A.

Example

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

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

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

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 $rank(X^T X) = p$

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?

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

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})

Multiple Linear Regression

Example

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

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

The Bootstrap

The Bootstrap Principle

- If we could repeat an experiment over and over again, we could actually find a very good approximation to the sampling distribution.
- Grocery example: If I had 1000 years of data, run the regression model on each year to see how estimates change.

The Bootstrap Principle

- If we could repeat an experiment over and over again, we could actually find a very good approximation to the sampling distribution.
- Grocery example: If I had 1000 years of data, run the regression model on each year to see how estimates change.
- Often too expensive or time-consuming.
- Bradley Efron's Idea (1979): Use computers to **simulate** replication.
- Instead of repeatedly obtaining new, independent datasets from the *population*, we repeatedly obtain datasets from the *sample* itself, the original dataset.

Bootstrap Methods

"Pull yourself up by your bootstraps!"

Bootstrap Methods

To get a bootstrap estimate,

1. Resample from the original data n times *with replacement* (note an original data observation could be in the new sample more than once),
2. Use the new dataset to compute a bootstrap estimate,
3. Repeat this to create B new datasets, and B new estimates.

Bootstrap Methods

Formally, you have original data $(x_i)_{i=1}^n$ and you are interested in estimating a population parameter Θ from the data. Label the estimate $\hat{\Theta}$.

Procedure

1. For $b = 1, \ldots, B$,
   - Create a new dataset $\mathcal{B}_b = (x_i^{(b)})_{i=1}^n$ by sampling from original dataset *with replacement*.
   - Use the new dataset to find an estimate $\hat{\Theta}^{(b)}$.

Bootstrap Methods

Formally, you have original data $(x_i)_{i=1}^n$ and you are interested in estimating a population parameter Θ from the data. Label the estimate $\hat{\Theta}$.

Procedure

1. For $b = 1, \ldots, B$,
    - Create a new dataset $\mathcal{B}_b = (x_i^{(b)})_{i=1}^n$ by sampling from original dataset *with replacement*.
    - Use the new dataset to find an estimate $\hat{\Theta}^{(b)}$.
2. The collection $(\hat{\Theta}^{(b)} - \hat{\Theta})_{b=1}^B$ estimates the sampling distribution of $\hat{\Theta} - \Theta$.

Example: Gaussian Random Variables

You sample $n = 100$ data points, $x_1, \ldots, x_{100} \sim N(\mu, 1)$. (Recall, Lab 1.)

```
> n <- 100
> vec <- rnorm(n, mean = mu)
> head(vec)
```

```
[1] -0.6264538  0.1836433 -0.8356286  1.5952808  0.3295078
[6] -0.8204684
```

- What's a good estimator for μ?

Example: Gaussian Random Variables

You sample $n = 100$ data points, $x_1, \ldots, x_{100} \sim N(\mu, 1)$. (Recall, Lab 1.)

```
> n <- 100
> vec <- rnorm(n, mean = mu)
> head(vec)
```

```
[1] -0.6264538  0.1836433 -0.8356286  1.5952808  0.3295078
[6] -0.8204684
```

- What's a good estimator for μ?

```
> mean(vec)
```

```
[1] 0.1088874
```

Set $\hat{\mu} = 0.11$. Recall, $\hat{\mu} \sim N(\mu, 1/100)$.

Example: Gaussian Random Variables

You sample $n = 100$ data points, $x_1, \ldots, x_{100} \sim N(\mu, 1)$. (Recall, Lab 1.)

```
> n <- 100
> vec <- rnorm(n, mean = mu)
> head(vec)
```

```
[1] -0.6264538  0.1836433 -0.8356286  1.5952808  0.3295078
[6] -0.8204684
```

- What's a good estimator for μ?

```
> mean(vec)
```

```
[1] 0.1088874
```

Set $\hat{\mu} = 0.11$. Recall, $\hat{\mu} \sim N(\mu, 1/100)$.

- How can we estimate $Var(\hat{\mu})$?

Example: Gaussian Random Variables

We'll use the bootstrap to estimate the variance! For $b = 1 : B$,

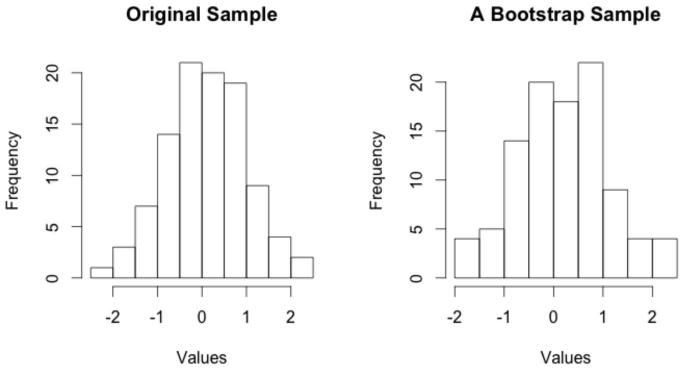
- Resample $x_1, \ldots, x_{100}$ *with replacement* to get $x_1^{(b)}, \ldots, x_{100}^{(b)}$.
- Compute $\hat{\mu}^{(b)} = \frac{1}{100} \sum_{i=1}^{100} x_i^{(b)}$.

```
> B <- 1000
> estimates <- vector(length = B)
> for (b in 1:B) {
+   new_sample <- sample(vec, size = n, replace = TRUE)
+   estimates[b] <- mean(new_sample)
+ }
> head(estimates)
```

```
[1] 0.12250487 0.10894538 0.21117547 0.05405239 0.16694190
[6] 0.13804749
```

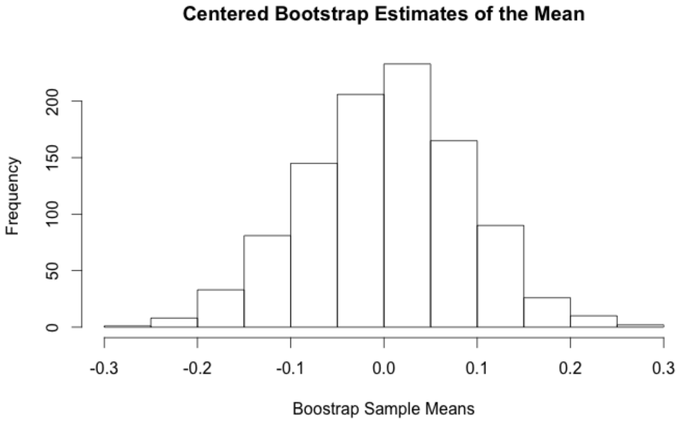
Example: Gaussian Random Variables

A histogram of the original sample and a histogram of a single resampled bootstrap sample.



Example: Gaussian Random Variables

The **Bootstrap Distribution of the Statistic**. Recall $(\hat{\mu}^{(b)} - \hat{\mu})_{b=1}^{B}$ approximates the sampling distribution of $\hat{\mu} - \mu$.



Example: Gaussian Random Variables

We'll use the bootstrap to estimate the variance!

Estimating the Variance

```
> var(estimates)
```

```
[1] 0.007380355
```

True variance: $Var(\hat{\mu}) = \frac{\sigma^2}{n} = \frac{1}{100} = 0.01$.

Regular Bootstrap Intervals

Regular Bootstrap Interval Formula:

- The 95% bootstrap interval is (L, U), where

$$L = 2\hat{\Theta} - \hat{\Theta}^*_{0.975} \quad \text{and} \quad U = 2\hat{\Theta} - \hat{\Theta}^*_{0.025}$$

- Note that $\hat{\Theta}^*_p$ is the p^{th} percentile of $\hat{\Theta}_{b=1}^{B}$

Regular Bootstrap Intervals

Regular Bootstrap Interval Formula:

- The 95% bootstrap interval is (L, U), where

$$L = 2\hat{\Theta} - \hat{\Theta}^*_{0.975} \quad \text{and} \quad U = 2\hat{\Theta} - \hat{\Theta}^*_{0.025}$$

- Note that $\hat{\Theta}^*_p$ is the p^{th} percentile of $\hat{\Theta}^B_{b=1}$

```
> L <- 2*mean(vec)-quantile(estimates,.975);L
```

```
      97.5%
-0.05218752
```

```
> U <- 2*mean(vec)-quantile(estimates,.025);U
```

```
     2.5%
0.2797513
```

Regular Bootstrap Intervals (Derivation)

Regular Bootstrap Intervals (Derivation)

Percentile Based Bootstrap Intervals

Percentile Based Bootstrap Formula:

- The 95% percentile bootstrap interval is (L, U), where

$$L = \hat{\Theta}^*_{0.025} \quad \text{and} \quad U = \hat{\Theta}^*_{0.975}$$

- Note that $\hat{\Theta}^*_p$ is the p^{th} percentile of $\hat{\Theta}_{b=1}^{B}$

Percentile Based Bootstrap Intervals

Percentile Based Bootstrap Formula:

- The 95% percentile bootstrap interval is (L, U), where

$$L = \hat{\Theta}^*_{0.025} \quad \text{and} \quad U = \hat{\Theta}^*_{0.975}$$

- Note that $\hat{\Theta}^*_p$ is the p^{th} percentile of $\hat{\Theta}^B_{b=1}$

```
> L <- quantile(estimates,.025);L
```

```
      2.5%
-0.0619766
```

```
> U <- quantile(estimates,.975);U
```

```
    97.5%
0.2699623
```

Nonparametric Testing

Bootstrapping Summary

Bootstrapping is very flexible!

- Bootstrapping gives you a distribution over estimators.
- This can be used to:
  - Approximate more complicated metrics (medians, quantiles, etc.).
  - Approximate distributional properties.
  - Create confidence intervals.
- By resampling $(x_i, y_i)_{i=1}^n$ pairs, we could create bootstrap estimators for linear model regression parameters.

Intervals

- Intervals can be used as a nonparametric hypothesis testing procedure.
- Both the regular and percentile based intervals are common techniques.
- The percentile based bootstrap interval is more intuitive to construct.

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.
- Chapter 6 (The Bootstrap) in Advanced Data Analysis from an Elementary Point of View.