STAT 420: Methods of Applied Statistics

Simple Linear Regression II

Ruoqing Zhu, Ph.D. <rqzhu@illinois.edu>

Course website: https://sites.google.com/site/teazrq/teaching/STAT420

University of Illinois at Urbana-Champaign

February 1, 2017

Simple Linear Regression

- From last lecture, we learned how to perform linear regression on one predictor and an intercept:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- The regression coefficients can be obtained through

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$
$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = r_{xy} \frac{s_y}{s_x}$$

- We can calculate these estimations without knowing the complete data, but only some key statistics.

What's next?

- Properties of the least squares estimator
- Evaluating the goodness-of-fit of the model
- Estimating the error variance

Simple Linear Regression

- Load the skin cancer dataset and perform a simple linear regression and answer the following questions:
  1). Predict the mortality rates, i.e., $\widehat{y}$, for states with latitude at 30 and 45.
  2). Calculate the sum of squared errors produced by this regression line
  3). Suppose another researcher claims that the regression line should be

  $$y = 389 - 6x$$

  which line fits the data better?
- Learn how to plot using R
- Source code provided in SIR II.r

Coefficient of determination: R^2

```
> cancer = read.table("skincancer.txt", header = TRUE)
> x = cancer$Lat
> y = cancer$Mort
> ybar = mean(y)
> xbar = mean(x)
> beta1 = sum((x - xbar)*(y - ybar))/sum( (x- xbar)^2)
> beta0 = ybar - beta1*xbar

# predicted values at 30 and 45
> beta0 + beta1*c(30, 45)
[1] 209.8603 120.1957

# sum of squared errors
> sum((y - beta0 - beta1*x)^2)
[1] 17173.07

# another fit
> sum((y - 389 + 6*x)^2)
[1] 17230.04
```

Properties of the least squares estimator

- The sum of errors (not squared) is 0. Why?

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \widehat{y}_i) = 0$$

- The predictor vector $\mathbf{x}$ is orthogonal to the error vector $\mathbf{e}$. Why?

$$\mathbf{x}^\mathsf{T}\mathbf{e} = \sum_{i=1}^{n} x_i e_i = 0$$

- The regression line passes through the centroid $(\bar{x}, \bar{y})$. Why?

Partitioning the variance

- We partition observed variation by expressing $y_i = \widehat{y}_i + e_i$.
- The total variance (corrected for mean) SST is defined as

$$\text{SST} = \sum_i (y_i - \bar{y})^2$$

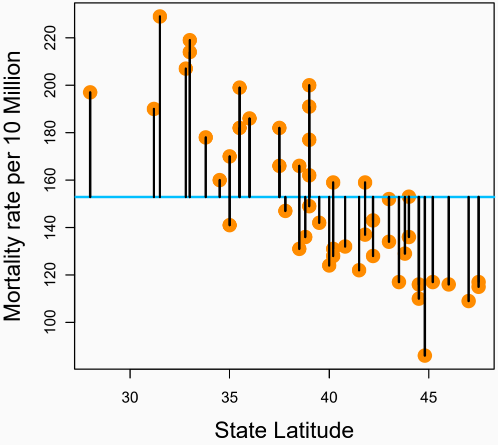
which can be partitioned into

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \widehat{y}_i + \widehat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} 2(y_i - \widehat{y}_i)(\widehat{y}_i - \bar{y}) \\
&= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 + 0 \\
&= \text{SS of Error} + \text{SS of Regression},
\end{aligned}$$

i.e., the sum of squares explained by the regression (SSR) and the sum of squared errors (SSE). This is the essential idea of the analysis of variance (ANOVA).

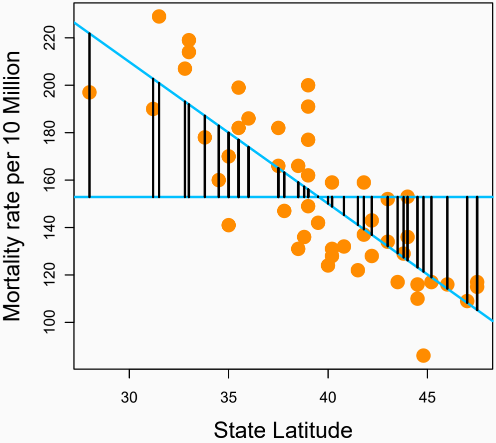
Partitioning the variance

SST

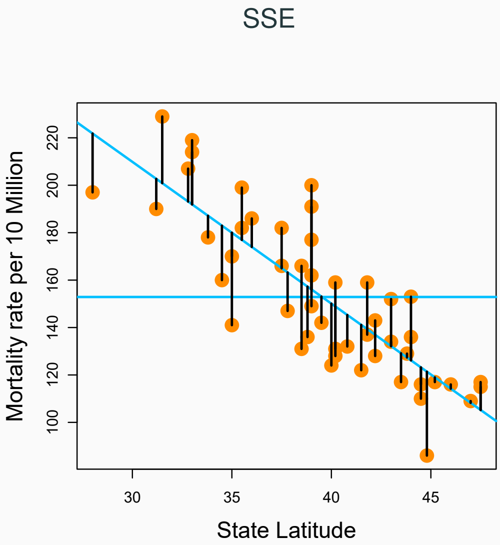


Partitioning the variance

SSR



Partitioning the variance



Coefficient of determination R^2

- SST is a fixed value for a given dataset. SSR and SSE changes depending on the regression coefficient.
- We can create a descriptive measure of linear association,

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

- $0 \leq R^2 \leq 1$
- Interpretation: R^2 tells us the percentage of variance explained by the regression line when the relationship is indeed linear. It represents the goodness-of-fit.
- Remark 1: R^2 does not necessarily tell us that we can make accurate predictions even when R^2 is large.
- Remark 2: Reasons for R^2 being close to 0?

Simple Linear Regression

Load the “cheddar” data using `library(faraway)` and `data(cheddar)`. Perform the regression model:

$$\text{Taste} = \beta_0 + \beta_1 \text{Lactic acid concentration}$$

Calculate:

- The parameter estimates of β_0 and β_1 using the least square method we introduced
- Calculate the residual sum of squares SSE
- Calculate the coefficient of determination R^2. R^2 is:

$$\text{A: } \leq 0.3; \quad \text{B: } \in (0.3, 0.5]; \quad \text{C: } \in (0.5, 0.7]; \quad \text{D: } > 0.7;$$

Coefficient of determination: R^2

- Fit a linear model to the skin cancer data and calculate the SST, SSE and R^2.

```
# Calculate the coefficient of determination
> n = nrow(cancer)
> SSR = var(beta0 + beta1*x)*(n-1)
> SST = var(y)*(n-1)
> SSR/SST
[1] 0.6798296
```

- Fit the model using the variable "*longitude*" instead, what is the R^2?

A: ≤ 0.3; B: $\in (0.3, 0.5]$; C: $\in (0.5, 0.7]$; D: > 0.7;

Estimating the error variance

- There is an other part of the model that we haven't analyzed: ϵ
- We approximate these error terms using $e_i = y_i - \hat{y}_i$
- Hence, to estimate the variance σ^2 of the error term ϵ.

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{\text{degrees of freedom}} = \frac{\text{SSE}}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n-2}$$

  We call this the mean square error (MSE).
- Why the degrees of freedom is $n-2$?

Source of Variation	DF	Sum of Squares
Regression	1	SSR = $\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2$
Residual error	n-2	SSE = $\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$
Total	n-1	SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$

- We have n observations, where is that last degree of freedom?

The "cats" data

Load the "cats" data using library(MASS) and data(cats). "MASS" is a pre-installed package in R. Calculate the following statistics: $\bar{x}$, $\bar{y}$, s_x, s_y, r_{xy}, $\sum_i^n x_i y_i$. Consider the regression model:

$$\text{Height} = \beta_0 + \beta_1 \text{Weight}$$

- Use 4 of these statistics to calculate both $\widehat{\beta}_0$ and $\widehat{\beta}_1$
- Use only 3 of these statistics to calculate $\widehat{\beta}_1$

The “cats” data

Load the “cats” data using `library(MASS)` and `data(cats)`. “`MASS`” is a pre-installed package in `R`. Calculate the following statistics: $\bar{x}$, $\bar{y}$, s_x, s_y, r_{xy}, $\sum_i^n x_i y_i$. Consider the regression model:

$$\text{Height} = \beta_0 + \beta_1 \text{Weight}$$

- Use 4 of these statistics to calculate both $\widehat{\beta}_0$ and $\widehat{\beta}_1$
- Use only 3 of these statistics to calculate $\widehat{\beta}_1$
- Calculate the coefficient of determination R^2.

The “cats” data

Load the “cats” data using library(MASS) and data(cats). “MASS” is a pre-installed package in R. Calculate the following statistics: $\bar{x}$, $\bar{y}$, s_x, s_y, r_{xy}, $\sum_i^n x_i y_i$. Consider the regression model:

$$\text{Height} = \beta_0 + \beta_1 \text{Weight}$$

- Use 4 of these statistics to calculate both $\widehat{\beta}_0$ and $\widehat{\beta}_1$
- Use only 3 of these statistics to calculate $\widehat{\beta}_1$
- Calculate the coefficient of determination R^2.
- Calculate the MSE $\widehat{\sigma}^2$, and it is

A: > 2; B: ≤ 2; C: 😱

The "cheddar" data

Load the "cheddar" data using `library(faraway)` and `data(cheddar)`. Use "taste" as the outcome variable and compare two models

$$\text{model 1:} \qquad \text{taste} = \beta_0 + \beta_1 \text{Acetic}$$
$$\text{model 2:} \qquad \text{taste} = \beta_0 + \beta_1 \text{Lactic}$$

- If we use the coefficient of determination as the criterion, which variable seems to fit the data better?

A: Acetic; B: Lactic; C: Equally well

The "cheddar" data

Load the "cheddar" data using library(faraway) and data(cheddar).
Use "taste" as the outcome variable and compare two models

$$\text{model 1:} \qquad \text{taste} = \beta_0 + \beta_1 \text{Acetic}$$
$$\text{model 2:} \qquad \text{taste} = \beta_0 + \beta_1 \text{Lactic}$$

- If we use the coefficient of determination as the criterion, which variable seems to fit the data better?

  A: Acetic; B: Lactic; C: Equally well

- Which model gives smaller estimation $\widehat{\sigma}^2$

  A: Acetic model; B: Lactic model; C: Equally well

Inference about the coefficients

- If we model the mortality rate using longitude instead of latitude, it gives a regression line

$$y = 182.7696 - 0.3287x.$$

  But, does this mean the regression line, especially the slope $\beta_1 = -0.3287$, is truly meaningful?
- How to test whether β_1 is zero or not?
- We will develop some new techniques of linear regression using matrix representation and derive the property of the parameter estimations.
- Let's first use a simulation study to analyze it.

A simulation study for hypothesis testing

- We want to test the hypothesis:

$$\mathrm{H}_0 : \beta_1 = 0 \quad \text{vs.} \quad \mathrm{H}_0 : \beta_1 \neq 0$$

- Similar to the hypothesis testing of means (such as t-test), we want to obtain a p-value (meaning?).
  - In the current data we obtained $\widehat{\beta}_1 = -0.3287$.
  - The idea of p-value is: if we repeatedly generate new datasets under the Null hypothesis, and calculate $\widehat{\beta}_1$, how many of them (proportion) have an estimation "more extreme" than -0.3287, i.e., < -0.3287 or > 0.3287
  - How to setup the simulation study?

A simulation study for hypothesis testing

- There could be many different ways to setup the simulation. Here I give one choice:
- We want to generate new dataset from the model:

$$Y = \beta_0 + 0X + \epsilon$$

- Take the longitude values in the skin cancer dataset as x values
- Use the mean of mortality rates as the true β_0, 152.8776
- Take ϵ i.i.d. from $\mathcal{N}(0, \sigma^2)$, where we use the MSE in our current data: 1116.829
- Generate 1000 independent datasets, with 49 observations in each dataset, and perform the regression. Record all estimations of β_1

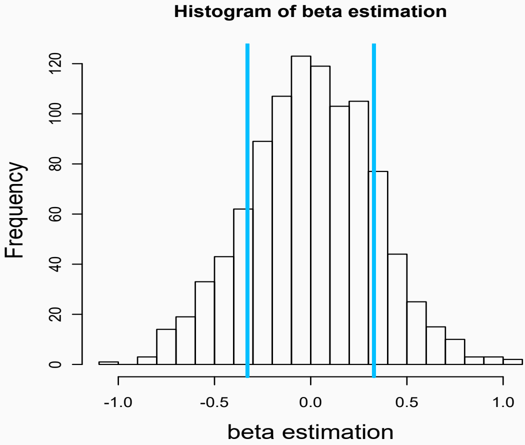
A simulation study for hypothesis testing

```
# Setup the underlying data generater
> x = cancer$Long
> beta0 = mean(cancer$Mort)
> n = nrow(cancer)
> sigma2 = sum(lm(Mort ~ Long, data= cancer)$residuals^2)/(n-2)

# Generate 10000 independent datasets and perform the regression
# Record the estimation $\widehat{\beta}_1$ from each dataset
> hatb = rep(NA, 10000)
> for (i in 1:10000)
+ {
+   y = beta0 + rnorm(n, mean = 0, sd = sqrt(sigma2))
+   hatb[i] = lm(y ~ x)$coef[2]
+ }

# Check how many of them have more extreme value than -0.3287
> mean( abs(hatb) > 0.3287)
[1] 0.314
```

A simulation study



A simulation study for hypothesis testing

```
# Compare it to the theoretical p-value obtained in the lm() function
> summary(lm(Mort ~ Long, data= cancer))

Call:
lm(formula = Mort ~ Long, data = cancer)

Residuals:
    Min      1Q  Median      3Q     Max
-63.898 -25.995  -5.952  21.856  78.444

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 182.7696    29.8893   6.115  1.8e-07 ***
Long         -0.3287     0.3245  -1.013    0.316
---
Signif. codes:  0 ?***?0.001 ?**?0.01 ?*?0.05 ?.?0.1 ? ?1

Residual standard error: 33.42 on 47 degrees of freedom
Multiple R-squared:  0.02137, Adjusted R-squared:  0.0005491
F-statistic: 1.026 on 1 and 47 DF,  p-value: 0.3162
```