APPLIED STATISTICS

Simple Linear Regression and Its Estimation

Dr Tao Zou

Research School of Finance, Actuarial Studies & Statistics
The Australian National University

Last Updated: Wed Aug 2 08:35:02 2017

Overview

- Introduction to Simple Linear Regression (SLR)
- SLR Model Assumptions
- Estimation of SLR Model

References

1. **F.L. Ramsey and D.W. Schafer** (2012)
   Chapter 7 of *The Statistical Sleuth*

2. The slides are made by **R Markdown**.
   http://rmarkdown.rstudio.com

Simple Linear Regression

Simple linear regression (SLR) is used to describe the **mean** of the **response**, as a function of a single **explanatory variable**.

For example: using a person's height (explanatory) to predict his/her weight (response), or using lean body mass (explanatory) to predict muscle strength (response).

What is a response variable?

Example: Old Faithful

Old Faithful is a cone geyser located in Yellowstone National Park in Wyoming, United States.

"oldfaithful.csv"

	A	B	C
1	DATE	INTERVAL	DURATION
2	1	78	4.40
3	1	74	3.90
4	1	68	4.00
5	1	76	4.00
6	1	80	3.50
7	1	84	4.10
8	1	50	2.30
9	1	93	4.70
10	1	55	1.70
11	1	76	4.90
12	1	58	1.70
13	1	74	4.60
14	1	75	3.40
15	2	80	4.30
16	2	56	1.70
17	2	80	3.90
18	2	69	3.70
19	2	57	3.10
20	2	90	4.00

DURATION (explanatory): Duration of Old Faithful Eruptions (mins).

INTERVAL (response): Interval until Subsequent Eruption (mins).

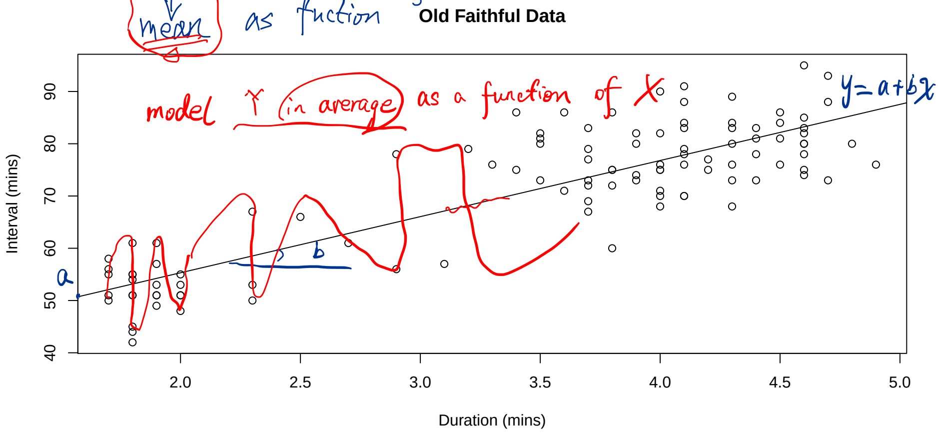
www.geyserstudy.org/geyser.aspx?pGeyserNo=OLDFAITHFUL

R Code

```
rm(list=ls())
setwd('~/Desktop/Research/AppliedStat2017/L1')
oldfaith<-read.table("oldfaithful.csv",header=T,sep=",")

plot(oldfaith$DURATION,oldfaith$INTERVAL,ylab="Interval (mins)",xlab="Duration (mins)",
     main="Old Faithful Data")
fit<-lm(oldfaith$INTERVAL~oldfaith$DURATION)
abline(fit)
```

mean as fuction of



√L3

Regression Terminology

The **regression** of the **response variable** on the **explanatory variable** is a **mathematical relationship between** the **mean** of the **response variable** and the **explanatory variable**.

In the Old Faithful example the **mean** of the **response variable** is modelled as a straight line **function** of the **explanatory variable**.

Notation: Let Y and X denote, respectively, the response variable and the explanatory variable.

- $\mu\{Y|X\}$, will represent **the regression of Y on X = the mean of Y as a function of X**.
- $\sigma\{Y|X\}$, will represent the standard deviation of Y as a function of X.

SLR and Interpretation

The SLR model specifies a particular form for $\mu\{Y|X\}$:

$$\mu\{Y|X\} = \beta_0 + \beta_1 X.$$

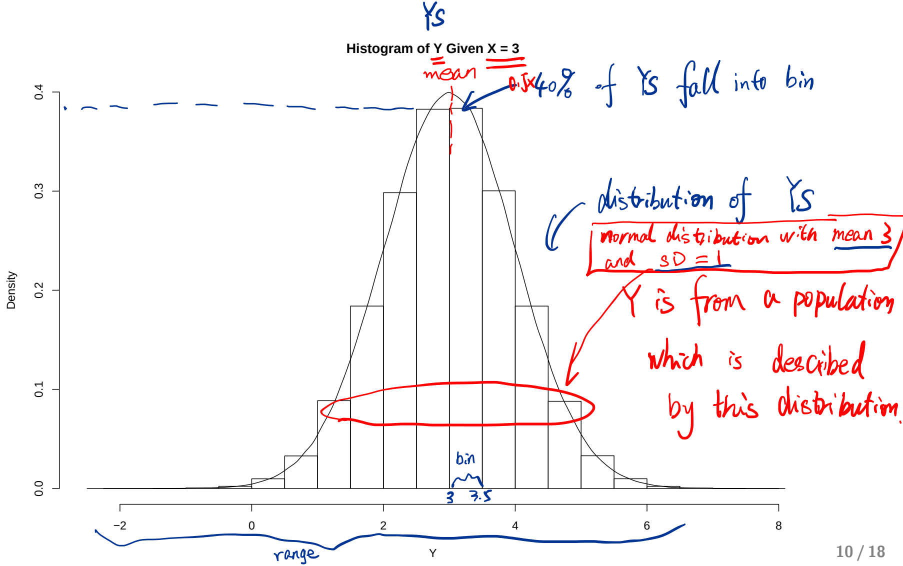
Two parameters (or called regression coefficients) are involved, where β_0 is the intercept and β_1 is the slope.

- β_0 is the mean of Y when X takes the value 0.
- β_1 is the increase in the mean of Y per one-unit increase in X.

Both β_0 and β_1 are unknown in the model.

SLR Model Assumptions

For each value of the explanatory variable ($X = x$), imagine there is a (sub)population of response values (realisations of Y).



SLR Model Assumptions (Con'd)

1. **Linearity**: The means of the populations fall on a straight-line function of the explanatory variable ($\mu\{Y|X\} = \beta_0 + \beta_1 X$).

2. **Normality**: There is a normally distributed population of responses for each value of the explanatory variable.

3. **Constant variance**: The population standard deviations are all equal: $\sigma\{Y|X\} = \sigma$.

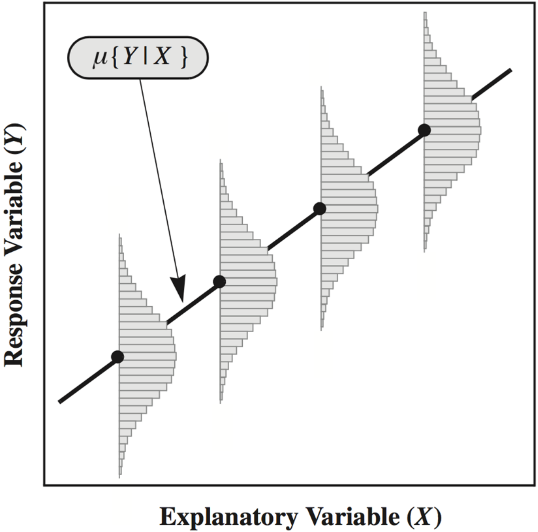
4. **Independence**: The selection of an observation from any of the populations is independent of the selection of any other observations. Briefly speaking, observations $(X_1, Y_1), \cdots, (X_n, Y_n)$ are independent, where n is called sample size.

can be described by

Remark: 2 & 3 ~~imply~~ $Y = \mu\{Y|X\} + \mathcal{E}$, where $\mathcal{E} \sim N(0, \sigma^2)$. It follows $Y \sim N(\mu\{Y|X\}, \sigma^2)$.

124

SLR Model Assumptions (Con'd)



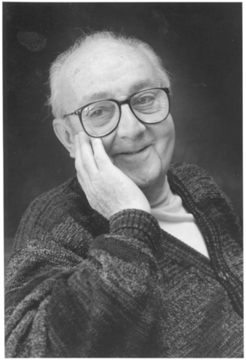
Picture taken from class text: "The Statistical Sleuth".

The Ideal Normal, SLR Model

Real data will not conform perfectly to these assumptions!

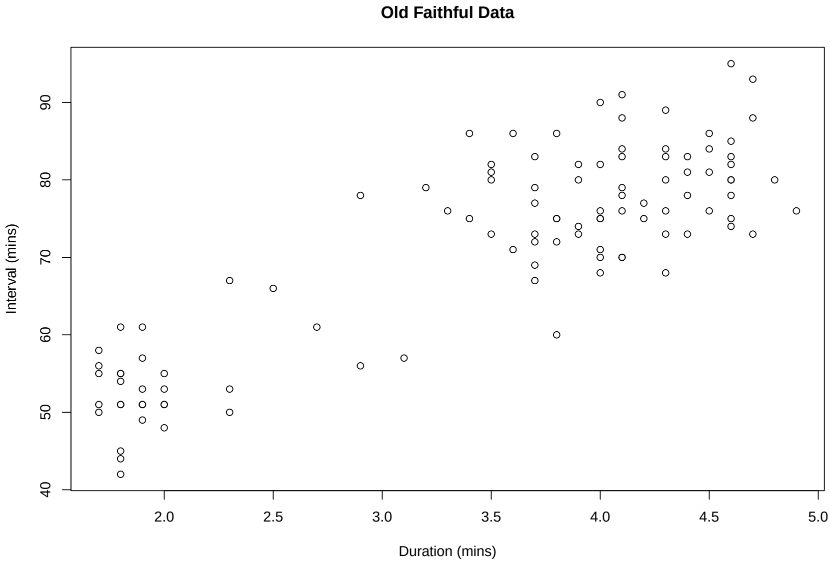
For example, $\mu\{Y|X\}$ is often not a straight line. However, $\mu\{Y|X\}$ can often be well approximated by a straight line.

We will talk later about the robustness of SLR to assumption violations.



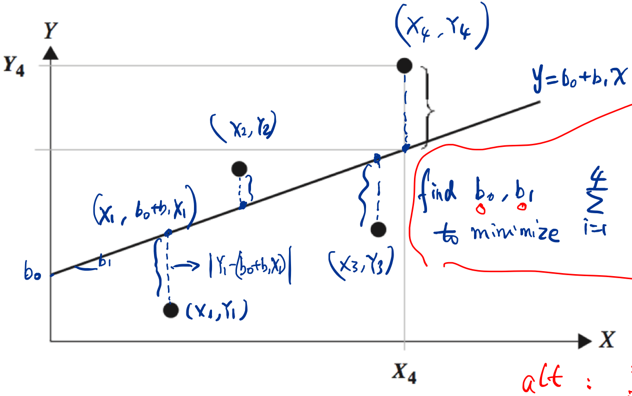
George E. P. Box (1919 - 2013)
"All models are wrong, but some are useful."

Estimation of SLR Parameters



Estimation of SLR Parameters (Con'd)

The method of least squares (LS) is used to obtain the "best fitting" straight line $\Rightarrow$ "best fitting" intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$, which are called the estimates of unkown β_0 and β_1, respectively.



$i=1,2,3,4$

find b_0, b_1 to minimize $\sum_{i=1}^{4} |Y_i - (b_0+b_1X_i)|^2$

abs

alt: $\sum_{i=1}^{4} |Y_i - (b_0+b_1X_i)|$

Picture taken from class text: "The Statistical Sleuth".

Estimation of SLR Parameters (Con'd)

Given the observations $(X_1, Y_1), \cdots, (X_n, Y_n)$, the LS estimates of β_1 and β_0 are chosen to minimise:

$$Q(b_1, b_0) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

The estimators of β_1 and β_0 are those values of b_1 and b_0, that minimise $Q(b_1, b_0)$.

The values of b_1 and b_0 that minimise $Q(b_1, b_0)$ are given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \text{ and } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

where $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$ and $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$.

Estimates are unbiased: $\mathrm{E}(\hat{\beta}_k) = \beta_k$, $k = 1, 0$.

How are these solutions obtained?

Fitting Values and Residuals

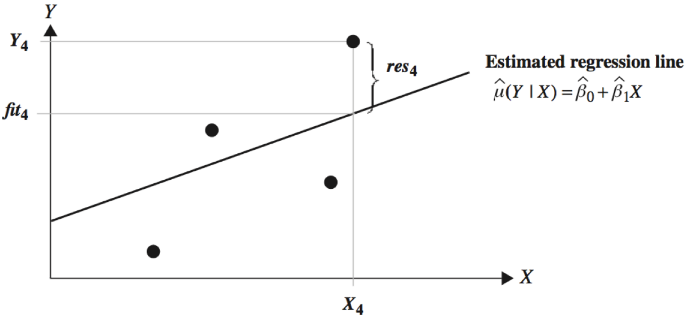
Using $\hat{\beta}_0$ and $\hat{\beta}_1$, the estimated mean function is given by:

$$\hat{\mu}\{Y|X\} = \hat{\beta}_0 + \hat{\beta}_1 X \text{ (plug-in idea)}.$$

- The estimated mean is called the fitted or predicted value:

$$\text{fit}_i = \hat{Y}_i = \hat{\mu}\{Y_i|X_i\} = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

- Residual: $\text{res}_i = \hat{\mathcal{E}}_i = Y_i - \hat{Y}_i$.



Picture taken from class text: "The Statistical Sleuth".

Example: Old Faithful (Con'd)

```r
names(fit)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```r
fit$coefficients
```

```
##      (Intercept) oldfaith$DURATION
##         33.82821          10.74097
```

```r
head(fit$fitted.values)
```

```
##        1        2        3        4        5        6
## 81.08848 75.71800 76.79209 76.79209 71.42161 77.86619
```

```r
head(fit$residuals)
```

```
##          1          2          3          4          5          6
## -3.0884837 -1.7179979 -8.7920941 -0.7920941  8.5783917  6.1338098
```

For Old Faithful (note the notation hat " ^ "):

$$\hat{\mu}\{\mathrm{INTERVAL}|\mathrm{DURATION}\} = 33.8 + 10.7 \times \mathrm{DURATION}.$$

Interpretation: If DURATION is increased by one-unit, the estimated mean of INTERVAL will increase 10.7 unit.