APPLIED STATISTICS

Inferential Tools for Simple Linear Regression

Dr Tao Zou

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

Last Updated: Fri Aug 4 09:00:59 2017

Overview

- Sampling Distribution of Estimation
- Standard Error of Estimation
- Hypothesis Testing
- Confidence Intervals and Prediction Intervals

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

Distinguish Parameters and Estimation

SLR model $\mu\{Y|X\} = \beta_0 + \beta_1 X$ and **real data** $(X_1, Y_1), \cdots, (X_n, Y_n)$.

Parameters	**Estimation** (notation hat " ˆ ")
β_1	$\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$ (denoted by $\hat{\beta}_1$)
β_0	$\bar{Y} - \hat{\beta}_1\bar{X}$ (denoted by $\hat{\beta}_0$)
unknown for **real data**	can be computed based on real data
the value is unique	can be different for different datasets

Hence, if we have another sample/dataset, e.g.,
$(X_{n+1}, Y_{n+1}), \cdots, (X_{n+n}, Y_{n+n})$, we will obtain different realisations of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Sampling Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

The distributions of the realisations are the sampling distributions.

We consider the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ given the values of the explanatory variables.

It can be shown mathematically that the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are both **normal**.

Sampling Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ (Con'd)

Sampling distribution of $\hat{\beta}_1$:

$\hat{\beta}_1$ is normal distributed;

$$\text{Mean}=\text{E}(\hat{\beta}_1) = \beta_1;$$

$$\text{Spread}=\text{SD}(\hat{\beta}_1)=\sigma\sqrt{\frac{1}{(n-1)s_X^2}}.$$

Sampling distribution of $\hat{\beta}_0$:

$\hat{\beta}_0$ is normal distributed;

$$\text{Mean}=\text{E}(\hat{\beta}_0) = \beta_0;$$

$$\text{Spread}=\text{SD}(\hat{\beta}_0)=\sigma\sqrt{\frac{1}{n}+\frac{\bar{X}^2}{(n-1)s_X^2}}.$$

Here, $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Knowing the sampling distributions allows us to make inferences about β_0 and β_1.

Remark: SLR model assumptions 1 & 2 & 3 can be described by $Y = \beta_0 + \beta_1 X + \mathcal{E}$, where $\mathcal{E} \sim N(0, \sigma^2)$. It follows $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$.

Example: Simulation for SLR $\mu(Y|X) = \beta_0 + \beta_1 X$.

1. Set the **unknown** parameters $\beta_0 = 2$ and $\beta_1 = 1$ by yourselves.

```
rm(list=ls())
beta0=2;beta1=1
```

2. Randomly generate $R = 1000$ **repeated samples** of $\{Y_i, X_i\}_{i=1}^n$. For each sample, obtain the **statistics** required in your analysis. Here we consider the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ as the statistics.

```
n=100
R=1000
hatbeta0=rep(0,R)          #space to store the different realisations of estimations
hatbeta1=rep(0,R)
set.seed(1)
for(r in 1:R) {
  X=1:n                    #our X values
  errors=rnorm(n)          #generate a set of errors, which implies sigma=1
  Y=beta0+beta1*X+errors   #generate a set of response values
  SLRfit=lm(Y~X)           #fit the SLR
  hatbeta0[r]=SLRfit$coef[1]   #get the estimation
  hatbeta1[r]=SLRfit$coef[2]
}
```

3. The sampling distribution of the statistics can be described by the $R = 1000$ different statistics values from $R = 1000$ **repeated samples**.

"set.seed()"

returns random numbers for $N(0,1)$

```
set.seed(1)
head(rnorm(n))
```

1 2 3 4 5 6

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

```
set.seed(2)
head(rnorm(n))
```

```
## [1] -0.89691455  0.18484918  1.58784533 -1.13037567 -0.08025176  0.13242028
```

```
set.seed(1)
head(rnorm(n))
```

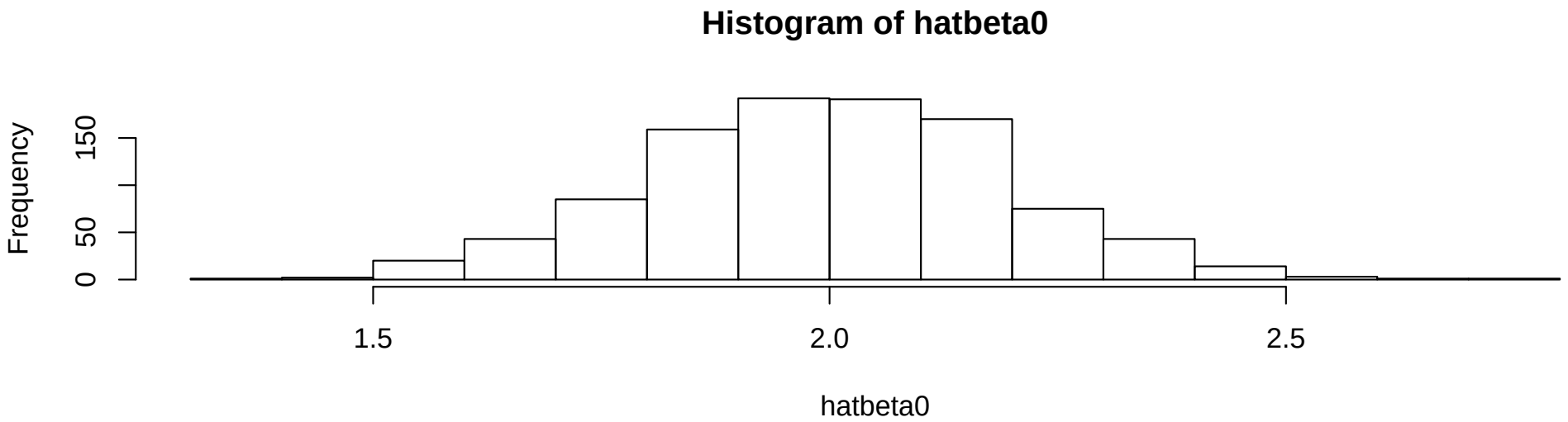
```
## [1] -0.6264538  0.1836433 -0.8356286  1.5952808  0.3295078 -0.8204684
```

seed=1	seed=2	···
-0.6264538	0.89691455	···
0.1836433	0.18484918	···
-0.8356286	1.58784533	···
1.5952808	-1.13037567	···
0.3295078	-0.08025176	···
-0.8204684	0.13242028	···
⋮	⋮	⋮

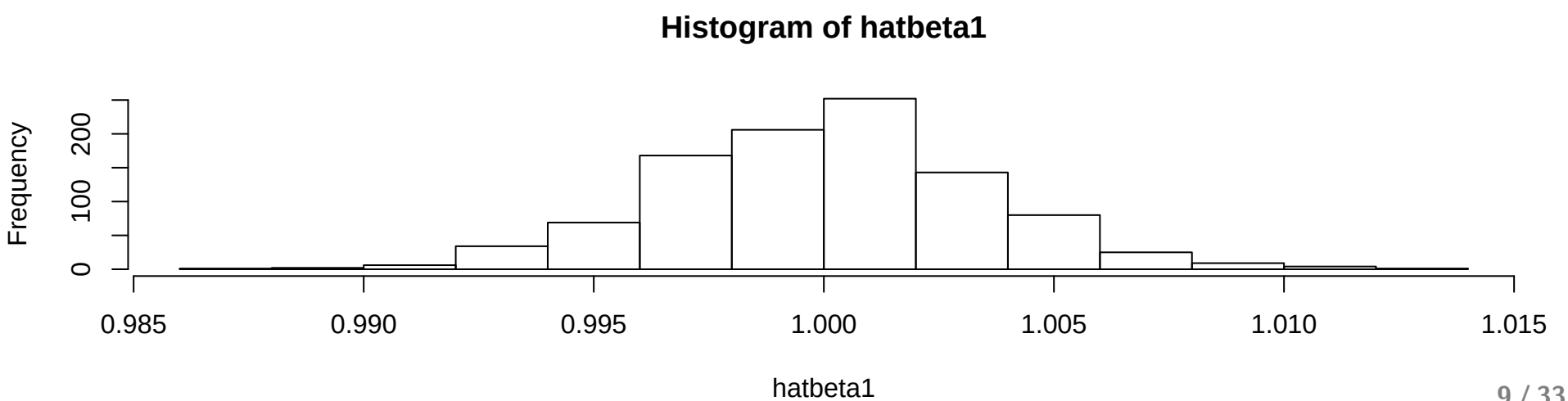
Histogram

The sampling distribution of the least squares estimates can be approximated by

```
hist(hatbeta0)
```




```
hist(hatbeta1)
```



Mean

The mean and variance/standard deviation of the statistics are both determined by the sampling dstribution. Hence, those can also be approximated by the $R = 1000$ different statistics values from $R = 1000$ **repeated samples**.

```
mean(hatbeta0)
```

```
## [1] 1.997998
```

is an approximate of $\mathrm{E}\hat{\beta}_0 = \beta_0 = 2$.

```
mean(hatbeta1)
```

```
## [1] 0.9999952
```

is an approximate of $\mathrm{E}\hat{\beta}_1 = \beta_1 = 1$.

Variance/Standard Deviation

```
sd(hatbeta1)
```

```
## [1] 0.003462471
```

is an approximate of

$$\mathrm{SD}(\hat{\beta}_1) = \sigma\sqrt{\frac{1}{(n-1)s_X^2}}$$

```
sigma=1
sigma*sqrt(1/((n-1)*(sd(X))^2))
```

```
## [1] 0.003464275
```

Advantage of simulation: no need of knowing the formulas for sampling distribution, mean, variance/standard deviation and etc. But $R = 1000$ **repeated samples** are required.

// End of Simulation

Standard Errors of $\hat{\beta}_0$ and $\hat{\beta}_1$

Standard deviation of $\hat{\beta}_1$:

$$\mathrm{SD}(\hat{\beta}_1)=\sigma\sqrt{\frac{1}{(n-1)s_X^2}}.$$

Standard deviation of $\hat{\beta}_0$:

$$\mathrm{SD}(\hat{\beta}_0)=\sigma\sqrt{\frac{1}{n}+\frac{\bar{X}^2}{(n-1)s_X^2}}.$$

However, for a real dataset, σ is unknown. But we can estimate it by

$$\hat{\sigma}=\sqrt{\frac{\sum_{i=1}^n \mathrm{res}_i^2}{n-2}},$$

where $n-2$ is called the number of degrees of freedom. Here we can understand it as a number such that $\mathrm{E}(\hat{\sigma}^2)=\sigma^2$.

Standard error of $\hat{\beta}_1$:

$$\mathrm{SE}(\hat{\beta}_1)=\hat{\sigma}\sqrt{\frac{1}{(n-1)s_X^2}}.$$

Standard error of $\hat{\beta}_0$:

$$\mathrm{SE}(\hat{\beta}_0)=\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{X}^2}{(n-1)s_X^2}}.$$

Plug-in idea.

Practical Sampling Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

Another form of the sampling distribution for $\hat{\beta}_0$ or $\hat{\beta}_1$ is

$$\frac{\hat{\beta}_k - \beta_k}{\mathrm{SD}(\hat{\beta}_k)} \sim N(0,1), \text{ for } k = 0, 1.$$

However, since σ is unknown, $\mathrm{SD}(\hat{\beta}_k)$ is unknown and this form is not practically useful.

Under the condition of the normal SLR model, it can be shown mathematically that

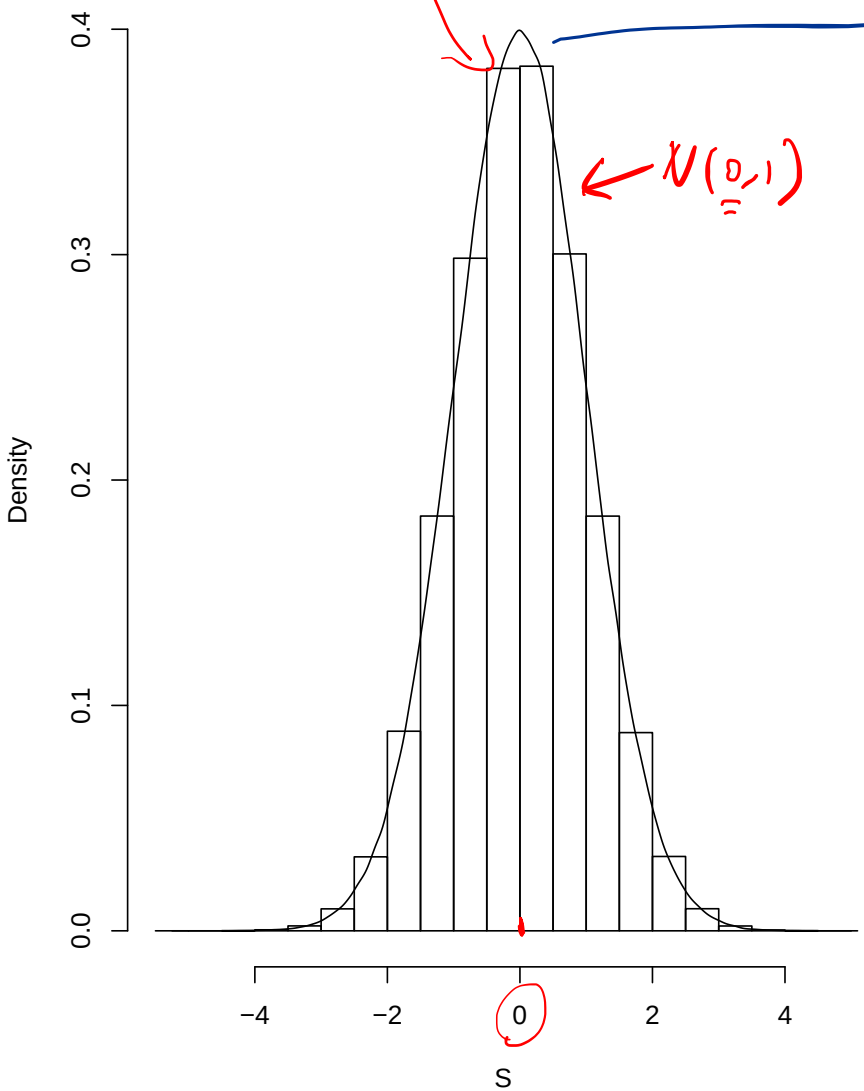
plug-in idea

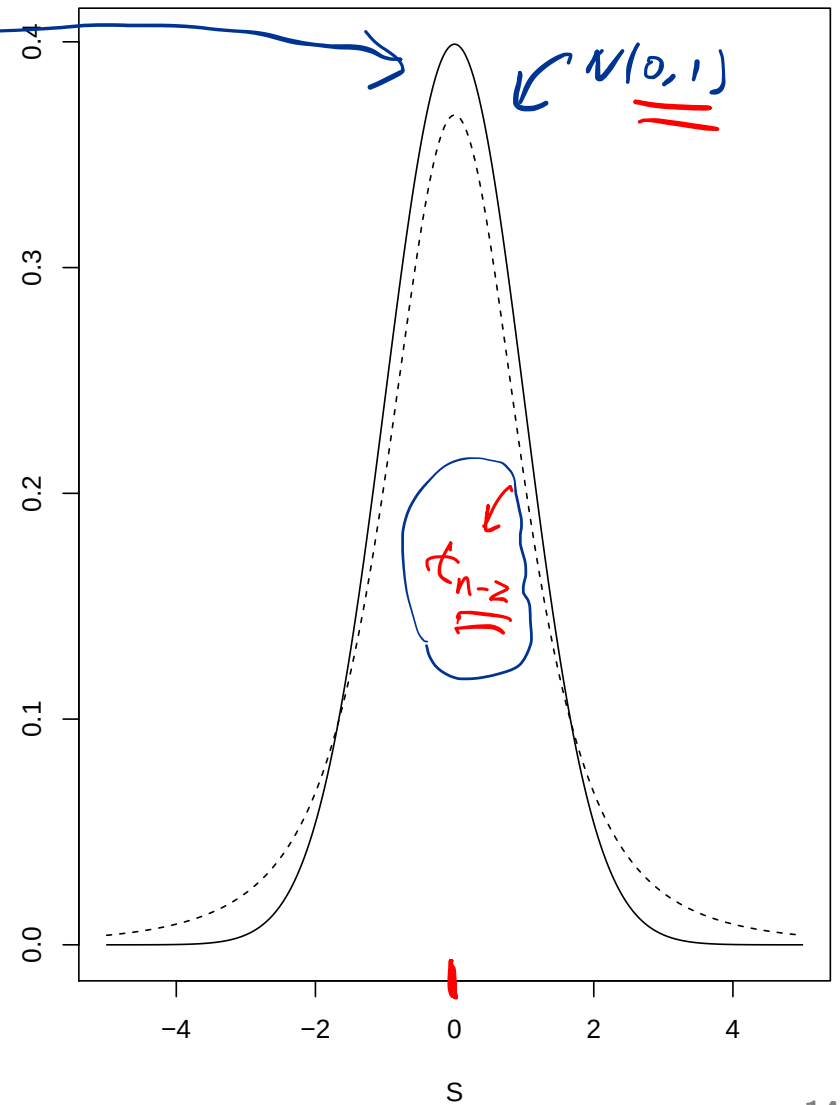
$$\frac{\hat{\beta}_k - \beta_k}{\mathrm{SE}(\hat{\beta}_k)} \sim t_{n-2}, \text{ for } k = 0, 1.$$

Here, $\mathrm{SE}(\hat{\beta}_k)$ is known.

Hypothesis Testing and confidence intervals for β_0 and β_1 can be based on these practical sampling distributions.

Normal Distribution and t Distribution





Hypothesis Testing for β_1: t-Test

$$\mathrm{H}_0 : \beta_1 = 0 \leftrightarrow \mathrm{H}_a : \beta_1 \neq 0.$$

$$\text{Test Statistic} = TS = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}.$$

The idea of hypothesis testing:

Intuitively, we should compare $\hat{\beta}_1$ and 0. However, even if $\beta_1 = 0$ (under H_0), $\hat{\beta}_1$ is numerically different from 0.

```
beta0=2
beta1=0
set.seed(1)
n=100
X=1:n                                  #our X values
errors=rnorm(n)                          #generate a set of errors, which implies sigma=1
Y=beta0+beta1*X+errors                 #generate a set of response values
SLRfit=lm(Y~X)                         #fit the SLR
SLRfit$coef                            #get the estimation
```

```
##   (Intercept)                X
##  2.1316657322 -0.0004510567
```

This is due to the variability of $\hat{\beta}_1$ ($\mathrm{SE}(\hat{\beta}_1)$).

Idea of Hypothesis Testing (Con'd)

But if $\beta_1 \neq 0$ (under H_a), the estimation $\hat{\beta}_1$ is supposed to be far away from 0, namly $|\hat{\beta}_1 - 0|$ should be very large.

However, the variability of $\hat{\beta}_1$ should be taken into consideration. Accordingly, we use $TS = (\hat{\beta}_1 - 0)/\mathrm{SE}(\hat{\beta}_1)$ as a measure to reject H_0.

If $|TS|$ is too large, then intuitively $\beta_1 = 0$ is possibly not true and we should reject H_0.

But how to quantify "too large"?

Idea of Hypothesis Testing (Con'd)

If $\beta_1 = 0$ (under H_0), recall the practical sampling distribution of $\hat{\beta}_1$, and we have

$$TS = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)} \sim t_{n-2}.$$

The test statistic (TS) is then compared to the t distribution.

Due to the nature of the t distribution, most values of the t distribution concentrate in the middle around 0.

Hence, if TS falls into the two tails of the t distribution, namely $|TS|$ is too large, then it is unlikely that H_0 is true.

Idea of Hypothesis Testing (Con'd)

Accordingly, the t distribution quantifies "too large". Based on the t_{n-2} distribution and TS, statisticians develop a formula to compute a measure called p-value, namely

$$p\text{-value} = 2 \times \mathrm{P}(T > |TS|), \text{ where } T \sim t_{n-2}.$$

p-value $<$ predetermined significance level α (usually 0.05).
$\Rightarrow$
TS falls into the two tails of the t distribution.
$\Rightarrow$
$|TS|$ is too large.
$\Rightarrow$
Reject H_0.

short for page 21

The test is significant.

In summary, p-value $< \alpha \Rightarrow$ reject H_0; p-value $\geq \alpha \Rightarrow$ not reject H_0.

If $\mathrm{H}_0 : \beta_1 = 0$ is rejected, the sample data suggest that the mean of the response is linearly related to the explanatory variable.

Hypothesis Testing for β_0: t-Test

$$\mathrm{H}_0 : \beta_0 = 0 \leftrightarrow \mathrm{H}_a : \beta_0 \neq 0.$$

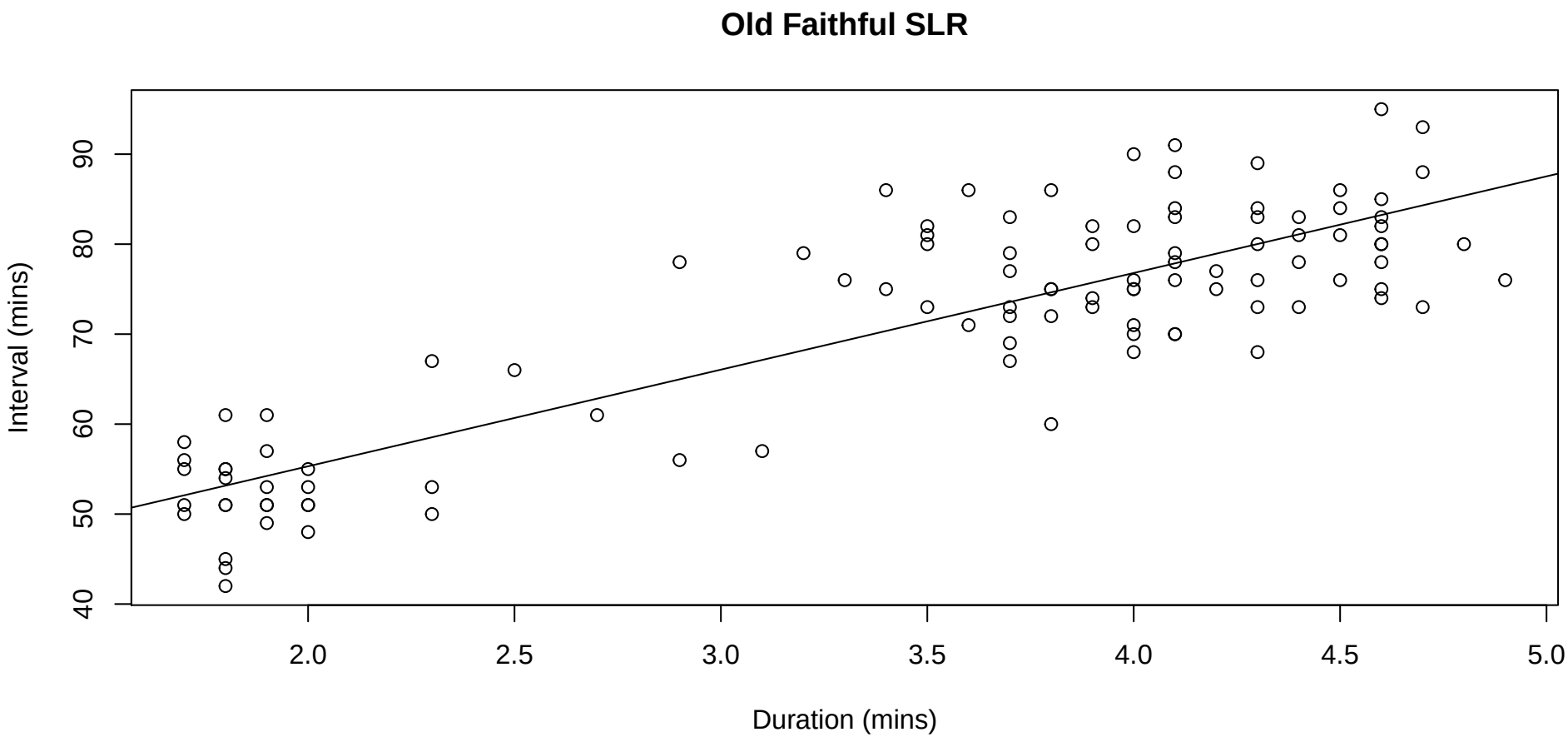
Test statistic is

$$TS = \frac{\hat{\beta}_0 - 0}{\mathrm{SE}(\hat{\beta}_0)},$$

which should be compared to the t_{n-2} distribution.

Example: Old Faithful (Con'd)

```
rm(list=ls())
#reading in the data
oldfaith=read.table("oldfaithful.csv",header=T,sep=",")
#fitting the SLR
oldfaith.reg=lm(oldfaith$INTERVAL~oldfaith$DURATION)
#Plotting the data
plot(oldfaith$DUR,oldfaith$INT,xlab="Duration (mins)", ylab="Interval (mins)", main="Old Faithful SLR")
#adding the fitted SLR
abline(oldfaith.reg)
```



R Code

```
#information about the SLR
summary(oldfaith.reg)
```

```
##
## Call:
## lm(formula = oldfaith$INTERVAL ~ oldfaith$DURATION)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.644  -4.440  -1.088   4.467  15.652
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        33.8282     2.2618   14.96   <2e-16 ***
## oldfaith$DURATION  10.7410     0.6263   17.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.683 on 105 degrees of freedom
## Multiple R-squared:  0.7369, Adjusted R-squared:  0.7344
## F-statistic: 294.1 on 1 and 105 DF,  p-value: < 2.2e-16
```

$H_0: \beta_0 = 0$

$H_0: \beta_1 = 0 \leftrightarrow H_a: \beta_1 \neq 0$

$\hat{\beta}_1$ $SE(\hat{\beta}_1)$ $TS = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$ P-value

$SE(\varepsilon) = \hat{\sigma}$

See page 31

- $\hat{\beta}_1 = 10.74$, $\text{SE}(\hat{\beta}_1) = 0.63$, $\hat{\sigma} = 6.68$.
- Test Statistic$= (10.74 - 0)/0.63 = 17.15$.
- p-value corresponding to this TS is 0.
- Reject H_0. Conclude $\beta_1 \neq 0$.
- Hence, β_1 is (statistically) significantly different from 0.

$H_0: \beta_1 = 0$ ✗

$\hat{\beta}_1$

We perform HT to obtain

Example: Simulation

```r
beta0=2
beta1=0.5
set.seed(1)
n=5
X=1:n                          #our X values
errors=rnorm(n)                  #generate a set of errors, which implies sigma=1
Y=beta0+beta1*X+errors         #generate a set of response values
SLRfit=lm(Y~X)                 #fit the SLR
summary(SLRfit)                #get the estimation and hypothesis testing
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##        1        2        3        4        5
## -0.09101  0.38673 -0.96490  1.13365 -0.46447
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.1322     0.9745   1.162    0.329
## X             0.8324     0.2938   2.833    0.066 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9291 on 3 degrees of freedom
## Multiple R-squared:  0.7279, Adjusted R-squared:  0.6372
## F-statistic: 8.026 on 1 and 3 DF,  p-value: 0.06603
```

p-value

Conclusion: We **do not have enough evidence to reject** $\mathrm{H}_0 : \beta_1 = 0$ at $\alpha = 5\%$. Is it correct for us to accept H_0 and conclude $\beta_1 = 0$?

Regression and Causality

The existence of a statistical relation between the response Y and the explanatory variable X (i.e. reject H_0 and conclude $\beta_1 \neq 0$) does not necessarily imply that the response depends causally on the explanatory variable.

For example, data on reading ability (response Y, or called dependent variable) and shoe size (explanatory variable X, or called independent variable) for a sample of young children aged 4-8 will show a positive relationship ($\beta_1 \neq 0$).

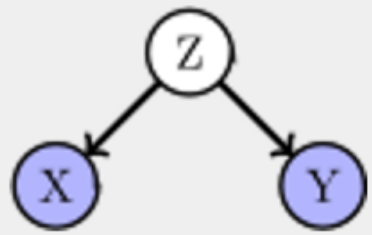
This relation does not imply, however, that an increase in shoe size causes an increase in reading ability.

Regression and Causality (Con'd)

There are other variables (confounders) such as age (Z) that affect both reading ability (Y) and shoe size (X).

Confounders (Confounding)

From Wikipedia



In statistics, a confounding variable is an extraneous variable in a statistical model that correlates (directly or inversely) with both the dependent variable and the independent variable.
A spurious relationship is a perceived relationship between an independent variable and a dependent variable that has been estimated incorrectly because the estimate fails to account for a confounding factor. The incorrect estimation suffers from omitted-variable bias.

Solution: potential confounding variables should be included in the regression model $\Rightarrow$ multiple linear regression.

Confidence Intervals (CI) for β_0 and β_1

Recall the practical sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\frac{\hat{\beta}_k - \beta_k}{\mathrm{SE}(\hat{\beta}_k)} \sim t_{n-2}, \text{ for } k = 0, 1.$$

Using this information, a $(1-\alpha)$ CI for β_k is

$$\left[\hat{\beta}_k - t_{n-2,\alpha/2} \times \mathrm{SE}(\hat{\beta}_k), \hat{\beta}_k + t_{n-2,\alpha/2} \times \mathrm{SE}(\hat{\beta}_k)\right],$$

where $t_{n-2,\alpha/2}$ is the $1-\alpha/2$ quantile of t_{n-2}, namely

$$\mathrm{P}(T \leq t_{n-2,\alpha/2}) = 1-\alpha/2 \text{ or } \mathrm{P}(T > t_{n-2,\alpha/2}) = \alpha/2$$

for $T \sim t_{n-2}$.

Later we use

$$\hat{\beta}_k \mp t_{n-2,\alpha/2} \times \mathrm{SE}(\hat{\beta}_k)$$

to simplify the notaion of CI.

Confidence Interval for *Mean of Response*

According to the SLR model, at some specified value $X = x_0$, the mean of response is

$$\mu\{Y|X = x_0\} = \beta_0 + \beta_1 x_0.$$

The estimate of the mean response at $X = x_0$ is

$$\hat{\mu}\{Y|X = x_0\} = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Similarly as the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$, we can obtain

$$\mathrm{SE}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{(n-1)s_X^2}}.$$

Confidence Interval for *Mean of Response* (Con'd)

Similarly as practical sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$, we can obtain

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)}{\mathrm{SE}(\hat{\beta}_0 + \hat{\beta}_1 x_0)} \sim t_{n-2}.$$

A $(1-\alpha)$ CI for $\mu\{Y|X = x_0\} = \beta_0 + \beta_1 x_0$ is

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \mp t_{n-2,\alpha/2} \times \mathrm{SE}(\hat{\beta}_0 + \hat{\beta}_1 x_0),$$

where $t_{n-2,\alpha/2}$ is the $1-\alpha/2$ quantile of t_{n-2}, namely

$$\mathrm{P}(T \leq t_{n-2,\alpha/2}) = 1 - \alpha/2 \text{ or } \mathrm{P}(T > t_{n-2,\alpha/2}) = \alpha/2$$

for $T \sim t_{n-2}$.

Example: Old Faithful (Con'd)

For eruptions that last 3 minutes, find a 95% confidence interval for the mean interval until the subsequent eruption.

```
#CI for mean of response
Y=oldfaith$INTERVAL
X=oldfaith$DURATION
fit<-lm(Y~X)
x0=data.frame(X=3)
predict(fit,x0,interval='confidence',level=0.95)
```

```
##        fit      lwr      upr
## 1 66.05112 64.64817 67.45408
```

Please try to use the formula given in the notes to verify this result.

The CI gives a range of values for the mean of response that are consistent with the data. Using a 95% CI is standard.

We can say that the CI is for the mean of response with **confidence coefficient** 95%, or the mean of response lies in the CI with **confidence** 95%.

Prediction Interval (PI) for *a Future Response*

The best **single estimation** for the mean of response at some specified value $X = x_0$, i.e. $\mu\{Y|X = x_0\} = \beta_0 + \beta_1 x_0$, is $\hat{\beta}_0 + \hat{\beta}_1 x_0$.

CI is the **interval estimation** for **the mean of response at some specified value** $X = x_0$, namely $\beta_0 + \beta_1 x_0$.

Consider **at a new observed explanatory variable** X_{new}, we are interested in predict **the future response** Y_{new}.

Under SLR,

$$Y_{\text{new}} = \beta_0 + \beta_1 X_{\text{new}} + \mathcal{E}_{\text{new}}.$$

Hence, the best **single prediction** for the future response Y_{new} at X_{new} is

$$\text{Prediction}(Y_{\text{new}}|X_{\text{new}}) = \hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}}.$$

PI is the **interval prediction** for **the future response** Y_{new} at X_{new}.

Prediction Interval (PI) for *a Future Response* (Con'd)

Confidence Interval

Target: $\beta_0 + \beta_1 x_0$; **Single estimation**: $\hat{\beta}_0 + \hat{\beta}_1 x_0$.

$$\frac{\text{Estimation} - \text{Target}}{\text{SE}(\text{Estimation} - \text{Target})} = \frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)}{\text{SE}\{\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)\}} \sim t_{n-2}.$$

It is worth noting that $\text{SE}\{\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0)\} = \text{SE}(\hat{\beta}_0 + \hat{\beta}_1 x_0)$.

Prediction Interval

Target: Y_{new}; **Single prediction**: $\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}}$.

$$\frac{\text{Prediction} - \text{Target}}{\text{SE}(\text{Prediction} - \text{Target})} = \frac{(\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}}) - Y_{\text{new}}}{\text{SE}\{(\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}}) - Y_{\text{new}}\}} \sim t_{n-2}.$$

What is $\text{SE}\{(\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}}) - Y_{\text{new}}\}$?

Prediction Interval (PI) for *a Future Response* (Con'd)

$$\begin{aligned} & (\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}}) - Y_{\text{new}} \\ = \; & (\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}}) - (\beta_0 + \beta_1 X_{\text{new}} + \mathcal{E}_{\text{new}}) \\ = \; & \{\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}} - (\beta_0 + \beta_1 X_{\text{new}})\} + (0 - \mathcal{E}_{\text{new}}). \end{aligned}$$

So intuitively,

$$\begin{aligned} & \text{SE}^2\{(\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}}) - Y_{\text{new}}\} \\ = \; & \text{SE}^2\{\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}} - (\beta_0 + \beta_1 X_{\text{new}})\} + \text{SE}^2(0 - \mathcal{E}_{\text{new}}) \\ = \; & \text{SE}^2(\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}}) + \text{SE}^2(\mathcal{E}_{\text{new}}), \end{aligned}$$

where $\text{SE}(\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}})$ is given in the CI, and $\text{SD}(\mathcal{E}_{\text{new}}) = \sigma$ implies $\text{SE}(\mathcal{E}_{\text{new}}) = \hat{\sigma}$.

Hence,

$$\text{SE}\{(\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}}) - Y_{\text{new}}\} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_{\text{new}} - \bar{X})^2}{(n-1)s_X^2}}.$$

Prediction Interval (PI) for *a Future Response* (Con'd)

$$\frac{\text{Prediction} - \text{Target}}{\text{SE}(\text{Prediction} - \text{Target})} = \frac{(\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}}) - Y_{\text{new}}}{\text{SE}\{(\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}}) - Y_{\text{new}}\}} \sim t_{n-2}.$$

A $(1-\alpha)$ PI for Y_{new} at X_{new} is

$$(\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}}) \mp t_{n-2,\alpha/2} \times \text{SE}\{(\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}}) - Y_{\text{new}}\},$$

where $t_{n-2,\alpha/2}$ is the $1-\alpha/2$ quantile of t_{n-2}, namely

$$\text{P}(T \leq t_{n-2,\alpha/2}) = 1 - \alpha/2 \text{ or } \text{P}(T > t_{n-2,\alpha/2}) = \alpha/2$$

for $T \sim t_{n-2}$.

Example: Old Faithful (Con'd)

An eruption just lasted for 3 minutes. What will be the 95% prediction for the interval until the subsequent eruption? (contrast this with the above CI example.)

```
#PI for future response
Xnew=data.frame(X=3)
predict(fit,Xnew,interval='prediction',level=0.95)
```

```
##        fit      lwr      upr
## 1 66.05112 52.72668 79.37557
```

The PI gives a range of plausible values for the interval until the next eruption.

A 95% PI succeeds in capturing the future value in 95% of its applications.