06

데이터분석방법론(1)

Simple Linear Regression

통계·데이터과학과 장영재 교수

한국방송통신대학교 대학원

학습목차

01

Introduction

1. Initial Data Analysis

- This is a critical step that should always be performed. It looks simple but it is vital.
- Numerical summaries - means, sds, five-number summaries, correlations.
- Graphical summaries
  - One variable - Boxplots, histograms etc.
  - Two variables - scatterplots.
  - Many variables - interactive graphics.
- Look for outliers, data-entry errors and skewed or unusual distributions. Are the data distributed as you expect?
- Getting data into a form suitable for analysis by cleaning out mistakes and aberrations is often time consuming. It often takes more time than the data analysis itself.

2. When to use Regression Analysis

- Regression analyses have several possible objectives including

1. Prediction of future observations.
2. Assessment of the effect of, or relationship between, explanatory variables on the response.
3. A general description of data structure.

- Extensions exist to handle multivariate responses, binary responses (logistic regression analysis) and count responses (poisson regression).

3. History

- Regression-type problems were first considered in the 18th century concerning navigation using astronomy.
- Legendre developed the method of least squares in 1805. Gauss claimed to have developed the method a few years earlier and showed that the least squares was the optimal solution when the errors are normally distributed in 1809. The methodology was used almost exclusively in the physical sciences until later in the 19th century.
- Francis Galton coined the term regression to mediocrity in 1875 in reference to the simple regression equation
- Galton used this equation to explain the phenomenon that sons of tall fathers tend to be tall but not as tall as their fathers while sons of short fathers tend to be short but not as short as their fathers. This effect is called the regression effect.

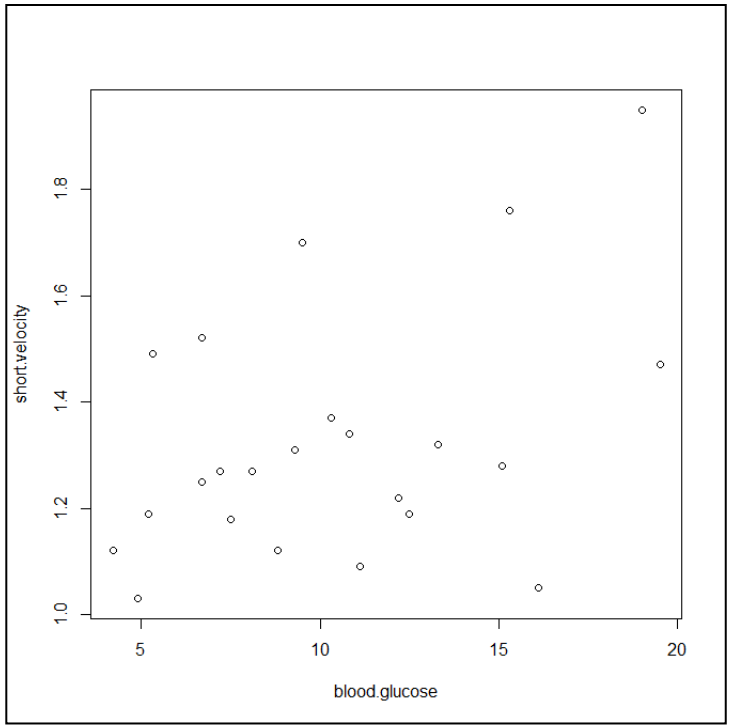
02

Model and Inference

1. Relation between two variables

- We consider situations where you want to describe the relation between two variables using linear regression analysis.

```
> attach(thuesen)
> head(thuesen)
  blood.glucose short.velocity
1          15.3           1.76
2          10.8           1.34
3           8.1           1.27
4          19.5           1.47
5           7.2           1.27
6           5.3           1.49
> plot(blood.glucose,  short.velocity)
```



2. Model

- Linear regression model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

, ε_i are assumed independent and $N(0, \sigma^2)$

- β : the increase per unit change in x.
- The method of least squares : Find the values of α and β that minimize the sum of squared residuals

$$SS_{res} = \sum_i (y_i - (\alpha + \beta x_i))^2$$

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

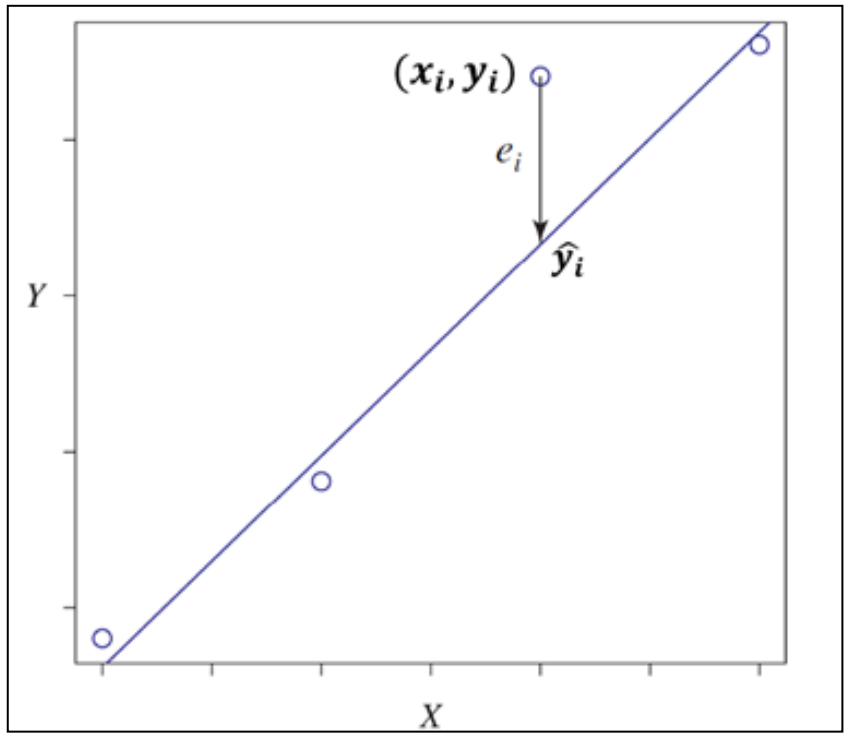
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

- The residual variance is estimated as SSres/(n − 2)

2. Model

- **Fitted linear regression model**

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$



Residual

$$e_i = y_i - \hat{y}_i$$

3. Decomposition of the total variation

- The sum of squares (SST), which represents the total variation, can be divided into the variation explained by the regression equation (SSR) and the sum of square errors (SSE), which is the sum of squares of the unexplained residuals.

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$$

$$n-1 = 1 + n-2$$

4. Estimation

- Under the assumption of Normal distribution of errors, $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$
- Estimation of β

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$E(\hat{\beta}) = \beta,$$

$$Var(\hat{\beta}) = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 Var(Y_i)\right)}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

4. Estimation

- Under the assumption of Normal distribution of errors, $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$
- Estimation of α

$$E(\hat{\alpha}) = E(\bar{Y} - \hat{\beta}\bar{x}) = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha,$$

$$Var(\hat{\alpha}) = Var(\bar{Y} - \hat{\beta}\bar{x}) = Var(\bar{Y}) + (\bar{x})^2 Var(\hat{\beta}) - 2\bar{x}Cov(\bar{Y}, \hat{\beta})$$

$$= \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + 0 = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha} \sim N\left(\alpha, \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\sigma^2\right)$$

5. Hypothesis test

- To test the null hypothesis that $\beta = 0$

  검정통계량 $t_0 = \frac{\hat{\beta}}{s.e.(\hat{\beta})} \sim t\,(n-2)$

- 분산분석표 (Analysis of Variance Table)

요인	제곱합	자유도	평균제곱	F값
회귀	SSR	1	MSR	MSR/MSE
잔차	SSE	n-2	MSE	
합	SST	n-1		

- 결정계수 $R^2 = SSR/SST$
- 검정 $H_0 : \beta = 0 \; vs \; H_1 : \beta \neq 0$
- 검정통계량 $F_0 = MSR/MSE \sim F\,(1, n-2)$

6. Example of Simple linear regression

```
> attach(thuesen)
> head(thuesen,3)
  blood.glucose short.velocity
1          15.3           1.76
2          10.8           1.34
3           8.1           1.27
> thu.reg <- lm(short.velocity~blood.glucose)
> thu.reg

Call:
lm(formula = short.velocity ~ blood.glucose)

Coefficients:
  (Intercept)  blood.glucose
      1.09781        0.02196
```

- The best-fitting straight line is seen to be

$$\text{short.velocity} = 1.098 + 0.0220 \times \text{blood.glucose}$$

6. Example of Simple linear regression

```
> summary(thu.reg)

Call:
lm(formula = short.velocity ~ blood.glucose)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40141 -0.14760 -0.02202  0.03001  0.43490

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.09781    0.11748   9.345 6.26e-09 ***
blood.glucose 0.02196    0.01045   2.101   0.0479 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2167 on 21 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1737,    Adjusted R-squared:  0.1343
F-statistic: 4.414 on 1 and 21 DF,  p-value: 0.0479
```

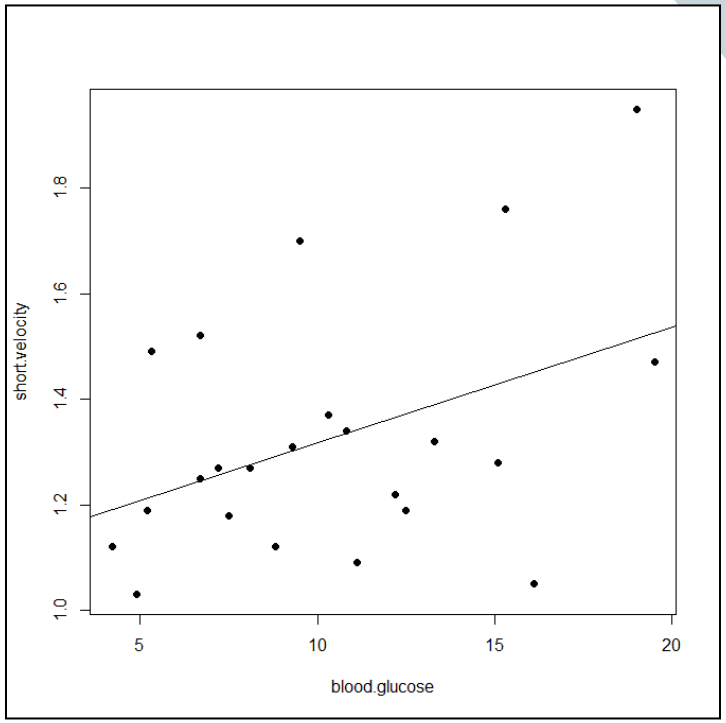
6. Example of Simple linear regression (Estimation)

```
> anova(thu.reg)
Analysis of Variance Table

Response: short.velocity
              Df  Sum Sq  Mean Sq F value Pr(>F)
blood.glucose  1 0.20727 0.207269   4.414 0.0479 *
Residuals     21 0.98610 0.046957
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
0.05 ‘.’ 0.1 ‘
```

```
> plot(blood.glucose,short.velocity, pch=19)
> abline(thu.reg)
```

```
> ## abline(1.1, 0.022)
```



6. Example of Simple linear regression

- Fitted value (추정값) $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$
- Residual (잔차) $e_i = y_i - \hat{y}_i$

lm.velo <- lm(short.velocity~blood.glucose)

```
> fitted(lm.velo)
      1       2       3       4       5       6       7       8
1.433841 1.335010 1.275711 1.526084 1.255945 1.214216 1.302066 1.341599
      9      10      11      12      13      14      15      17
1.262534 1.365758 1.244964 1.212020 1.515103 1.429449 1.244964 1.190057
     18      19      20      21      22      23      24
1.324029 1.372346 1.451411 1.389916 1.205431 1.291085 1.306459
> resid(lm.velo)
           1            2            3            4            5            6
 0.326158532  0.004989882 -0.005711308 -0.056084062  0.014054962  0.275783754
           7            8            9           10           11           12
 0.007933665 -0.251598875 -0.082533795 -0.145757649  0.005036223 -0.022019994
          13           14           15           17           18           19
 0.434897199 -0.149448964  0.275036223 -0.070057471  0.045971143 -0.182346406
          20           21           22           23           24
-0.401411486 -0.069916424 -0.175431237 -0.171085074  0.393541161
```

6. Example of Simple linear regression (deleting NA's)

- To put the fitted line on the plot

```
> plot(blood.glucose,short.velocity)
> lines(blood.glucose,fitted(lm.velo))
```

Error in xy.coords(x, y) : 'x' and 'y' lengths differ

→Missing cases

```
> lines(blood.glucose[!is.na(short.velocity)],fitted(lm.velo))
```

: The technique works but becomes clumsy if there are missing values in several variables:

...blood.glucose[!is.na(short.velocity) & !is.na(blood.glucose)]...

- You can use the na.exclude method for NA handling. This can be set either as an argument to lm or as an option;

```
> options(na.action=na.exclude)
> lm.velo <- lm(short.velocity~blood.glucose)
```

6. Example of Simple linear regression

```
> options(na.action=na.exclude)
> lm.velo <- lm(short.velocity~blood.glucose)
> fitted(lm.velo)
       1        2        3        4        5        6        7        8
1.433841 1.335010 1.275711 1.526084 1.255945 1.214216 1.302066 1.341599
       9       10       11       12       13       14       15       16
1.262534 1.365758 1.244964 1.212020 1.515103 1.429449 1.244964       NA
      17       18       19       20       21       22       23       24
1.190057 1.324029 1.372346 1.451411 1.389916 1.205431 1.291085 1.306459
> resid(lm.velo)
           1            2            3            4            5            6
 0.326158532  0.004989882 -0.005711308 -0.056084062  0.014054962  0.275783754
           7            8            9           10           11           12
 0.007933665 -0.251598875 -0.082533795 -0.145757649  0.005036223 -0.022019994
          13           14           15           16           17           18
 0.434897199 -0.149448964  0.275036223           NA -0.070057471  0.045971143
          19           20           21           22           23           24
-0.182346406 -0.401411486 -0.069916424 -0.175431237 -0.171085074  0.393541161
>
```

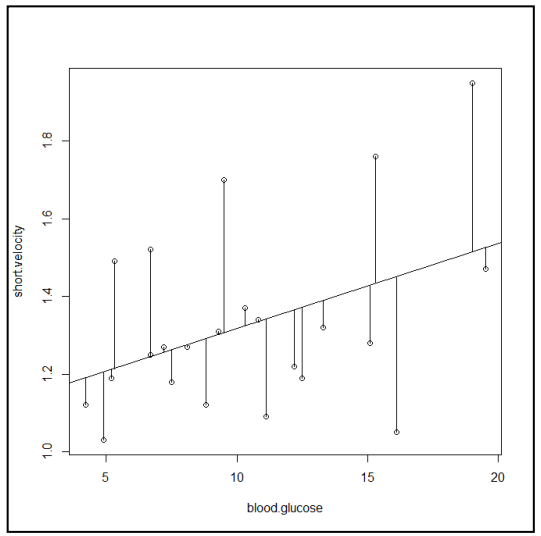
6. Example of Simple linear regression (Diagnosis)

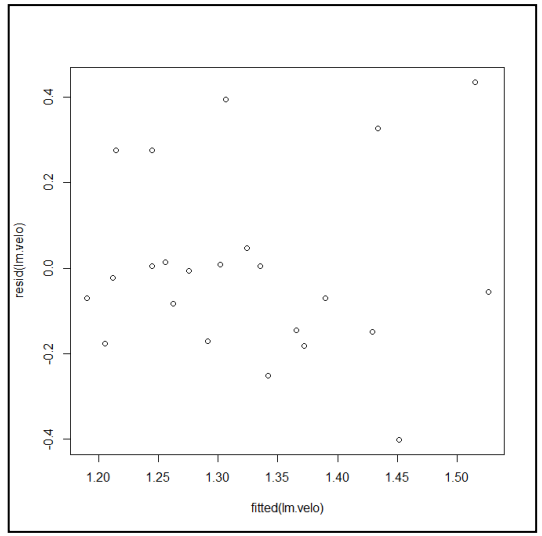
- To create a plot where residuals are displayed by connecting observations to corresponding points on the fitted line

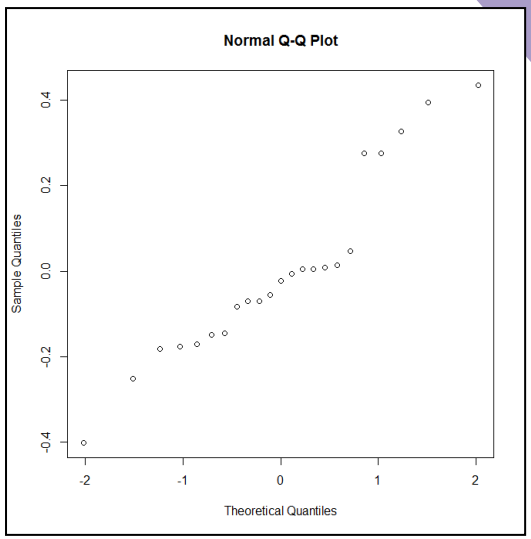
```
> plot(blood.glucose,short.velocity)
> abline(lm.velo)
> segments(blood.glucose,fitted(lm.velo),  blood.glucose,short.velocity)
```

```
> plot(fitted(lm.velo),resid(lm.velo))
```

```
> qqnorm(resid(lm.velo))
```







7. Prediction and confidence bands

- Fitted lines are often presented with uncertainty bands around them.
  There are two kinds of bands, often referred to as the “narrow” and “wide” limits.
- The narrow bands, confidence bands, reflect the uncertainty about the line itself, like the SEM expresses the precision with which a mean is known.
- The wide bands, prediction bands, include the uncertainty about future observations.

```
> predict(lm.velo)
       1        2        3        4        5        6        7        8
1.433841 1.335010 1.275711 1.526084 1.255945 1.214216 1.302066 1.341599
       ...
```

```
> predict(lm.velo,int="c")
        fit      lwr      upr
1  1.433841 1.291371 1.576312
2  1.335010 1.240589 1.429431
3  1.275711 1.169536 1.381887
```

```
> predict(lm.velo,int="p")
        fit       lwr      upr
1  1.433841 0.9612137 1.906469
2  1.335010 0.8745815 1.795439
3  1.275711 0.8127292 1.738693
```

7. Prediction and confidence bands

- Fitted lines are often presented with uncertainty bands around them.
  There are two kinds of bands, often referred to as the “narrow” and “wide” limits.
- The narrow bands, confidence bands, reflect the uncertainty about the line itself,
  like the SEM expresses the precision with which a mean is known.
- The wide bands, prediction bands, include the uncertainty about future observations.

At $X = x_0$, mean of Y is $\mu_{Y|X} = \alpha + \beta x_0$, so its point estimation is $\hat{Y}_0 = \hat{\alpha} + \hat{\beta} X_0$.
$(1-\alpha) \times 100\%$ CI of $\mu_{Y|X}$: $[\hat{y}_0 \pm t_{n-2,\, a/2} \cdot SE(\hat{y}_0)]$

$$Var(\hat{y}_0) = Var(\hat{\alpha} + \hat{\beta} x_0) = Var(\hat{\alpha}) + (x_0)^2 Var(\hat{\beta}) + 2x_0 Cov(\hat{a}, \hat{\beta})$$

$$Cov(\hat{\alpha}, \hat{\beta}) = Cov\left(\sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_j - \bar{x})^2} \right) Y_i, \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

$$Cov(Y_i, Y_i) = \sigma^2$$

7. Prediction and confidence bands

- Fitted lines are often presented with uncertainty bands around them.
  There are two kinds of bands, often referred to as the “narrow” and “wide” limits.
- The narrow bands, confidence bands, reflect the uncertainty about the line itself,
  like the SEM expresses the precision with which a mean is known.
- The wide bands, prediction bands, include the uncertainty about future observations.

$(1-\alpha)\times 100\%$ CI of $\mu_{Y|X}$:

$$\hat{y}_0 \pm t_{n-2,\,\alpha/2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

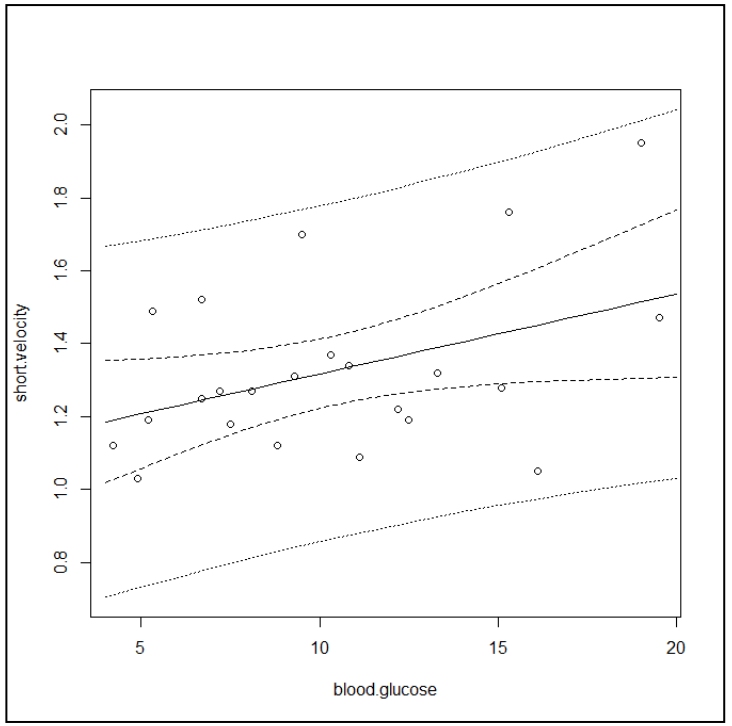
$(1-\alpha)\times 100\%$ PI of $\mu_{Y|X}$:

$$\hat{y}_0 \pm t_{n-2,\,\alpha/2} \cdot s \cdot \sqrt{1+\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

7. Prediction and confidence bands

- The best way to add prediction and confidence intervals to a scatterplot is to use the matlines function, which plots the columns of a matrix against a vector.

```
> pred.frame <-   data.frame(blood.glucose=4:20)
> pp <- predict(lm.velo, int="p", newdata=pred.frame)
> pc <- predict(lm.velo, int="c", newdata=pred.frame)
> plot(blood.glucose,short.velocity,
+         ylim=range(short.velocity, pp, na.rm=T))
> pred.gluc <- pred.frame$blood.glucose
> matlines(pred.gluc, pc, lty=c(1,2,2), col="black")
> matlines(pred.gluc, pp, lty=c(1,3,3), col="black")
```



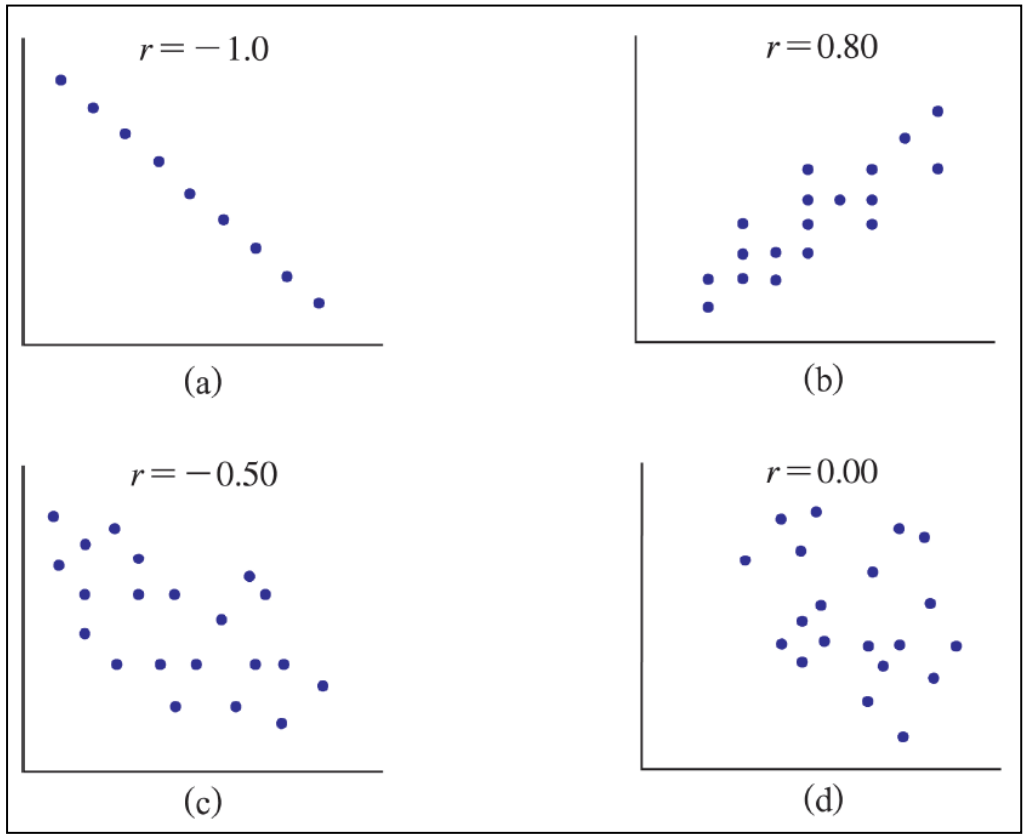
03

Correlation

1. Concept

- **Correlation coefficient :**

  A measure of the strength of a linear relationship between two continuous variables



2. Pearson Correlation coefficient

- Pearson Correlation coefficient :

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

```
> # in case of missing values
> cor(blood.glucose,short.velocity)
[1] NA
> cor(blood.glucose,short.velocity,use="complete.obs")
[1] 0.4167546
> cor(thuesen,use="complete.obs")
                blood.glucose short.velocity
blood.glucose       1.0000000      0.4167546
short.velocity      0.4167546      1.0000000

> cor.test(blood.glucose,short.velocity)
        Pearson's product-moment correlation
data:  blood.glucose and short.velocity
t = 2.101, df = 21, p-value = 0.0479
```

3. Nonparametric correlations

- Spearman's rank correlation coefficient : Nonparametric correlation
- Kendall's τ : based on counting the number of concordant and discordant pairs.

```
> cor.test(blood.glucose,short.velocity,method="spearman")
        Spearman's rank correlation rho
data:  blood.glucose and short.velocity
S = 1380.364, p-value = 0.1392
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.318002
> cor.test(blood.glucose,short.velocity,method="kendall")
        Kendall's rank correlation tau
data:  blood.glucose and short.velocity
z = 1.5604, p-value = 0.1187
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.2350616
```

07

다음시간 안내

Multiple regression