회귀분석론 HW4

212STG18 예지혜

2.3

2.3.1 Explain why this graph and the graph in Problem 2.2 suggests that using log-scale is preferable if fitting simple linear regression is desired.

```
library(alr4)
```

```
## Warning: package 'alr4' was built under R version 3.6.3
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## Warning: package 'effects' was built under R version 3.6.3
```

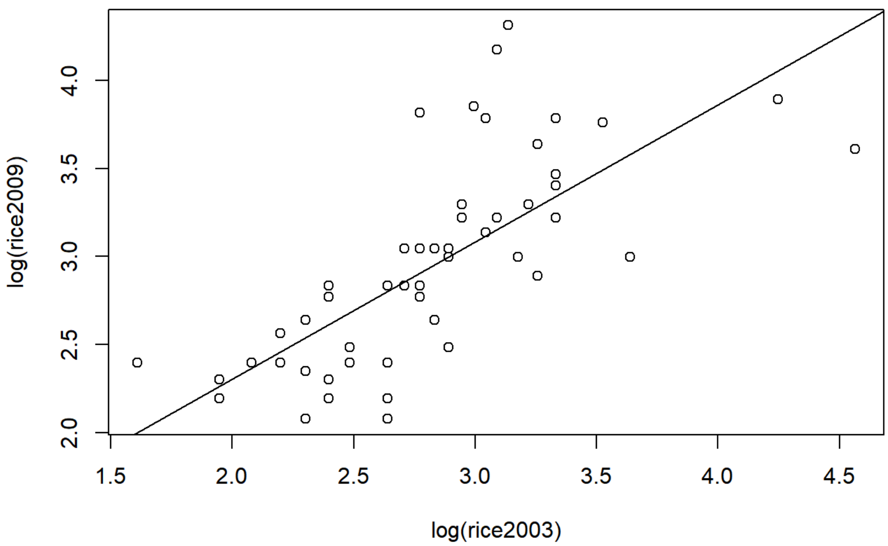
```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
head(UBSprices)
```

```
##           bigmac2009 bread2009 rice2009 bigmac2003 bread2003 rice2003
## Amsterdam         19        10       11         16         9        9
## Athens            30        13       27         21        12       19
## Auckland          19        19       13         19        19        9
## Bangkok           45        43       27         50        42       25
## Barcelona         21        17        8         22        19       10
## Berlin            19        10       17         16        10       16
```

```
lm1 <- lm(log(rice2009)~log(rice2003), UBSprices)
plot(log(rice2009)~log(rice2003), UBSprices)
abline(lm1)
```



로그 스케일을 적용한 그래프가 더 직선에 가깝고, 고르게 분포하는 것으로 보아 등분산에 가깝기 때문에 SLR에 더 적합하다. 이상치도 개선된 것을 확인할 수 있다.

2.3.2 E(log(y)|x) = β0 + β1*log(x). Give an interpretation of β0 and β1 in this setting, assuming β1 > 0.

```
summary(lm1)
```

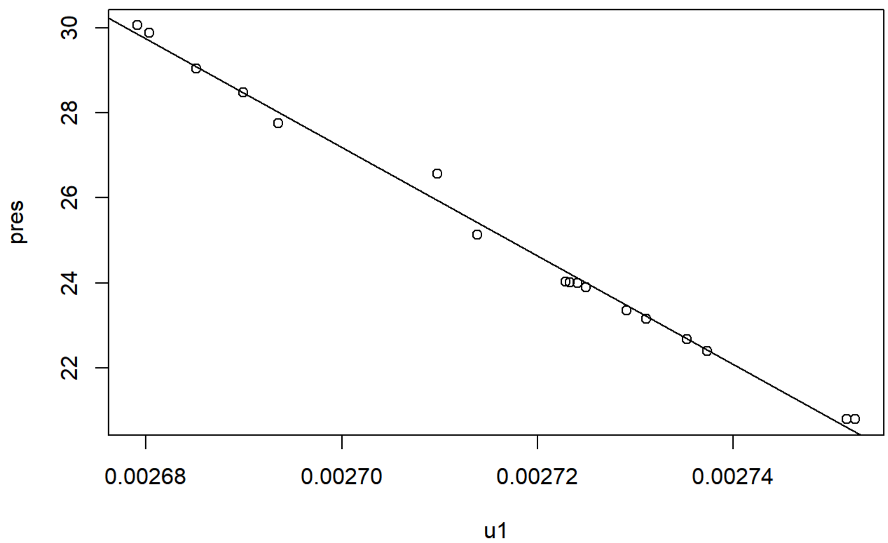
```
##
## Call:
## lm(formula = log(rice2009) ~ log(rice2003), data = UBSprices)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72251 -0.26950  0.00795  0.15346  1.12229
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.7470     0.2964   2.520   0.0148 *
## log(rice2003)   0.7787     0.1038   7.503 7.82e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4013 on 52 degrees of freedom
## Multiple R-squared:  0.5198, Adjusted R-squared:  0.5106
## F-statistic: 56.29 on 1 and 52 DF,  p-value: 7.819e-10
```

이 회귀식을 다시 표현하면 $E(y) = exp(\beta 0) * (x)^(\beta 1)$ 이다. 따라서 β1에 의해 y가 지수 성장인지, 선형인지, 점진적 성장을 하는지 결정된다. 또한 exp(β0)에 의해 그 기울기가 결정이 된다.
피팅 결과를 이 식에 적용해보면, $E(y) = exp(0.7470) * (x)^(0.7787)$ 이다. β1이 1보다 작으므로 선형 아래로 점진적 성장을 하며, 그 크기에 exp(0.7470)이라는 1보다 큰 값을 곱하면 E(y)값을 알 수 있다.

2.7

2.7.1 Draw the plot of pres versus u1, and verify that apart from case 12 the 17 points in Forbes's data fall close to a straight line. Explain why the apparent slope in this graph is negative when the slope in Figure 1.4a is positive.

```
forbe <- Forbes
forbe$u1 <- 1/((5/9)*forbe$bp+255.37)
plot(pres~u1, forbe)
abline(lm(pres~u1, forbe))
```



bp의 선형 변형에 역수를 취했기 때문에 그 관계또한 음의 관계로 나타난다.

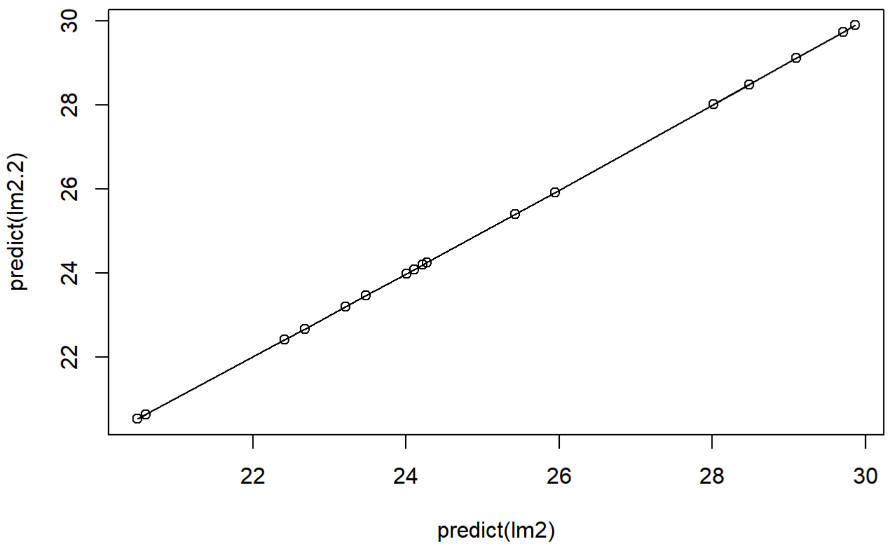
2.7.2 Compute the linear regression implied by (2.23), and summarize your results.

```
lm2 <- lm(pres ~ u1, forbe)
summary(lm2)
```

```
##
## Call:
## lm(formula = pres ~ u1, data = forbe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28216 -0.12643 -0.05569  0.17111  0.62569
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.723e+02  7.013e+00   53.08   <2e-16 ***
## u1          -1.278e+05  2.581e+03  -49.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2433 on 15 degrees of freedom
## Multiple R-squared:  0.9939, Adjusted R-squared:  0.9935
## F-statistic:  2451 on 1 and 15 DF,  p-value: < 2.2e-16
```

2.7.3 To compare these two mean functions, draw the plot of the fitted values from Forbes's mean function fit versus the fitted values from (2.23).

```
lm2.2 <- lm(pres ~ bp, forbe)
plot(predict(lm2), predict(lm2.2))
lines(predict(lm2), predict(lm2.2))
```



두 fitted value를 비교해보면 거의 동일하다. 둘 중 어떤 피팅이 낫다고 말하기 어렵다.

2.7.4

```
hooker <- Hooker
hooker$u1 <- 1/((5/9)*hooker$bp+255.37)
lm2.3 <- lm(pres~u1, hooker)
summary(lm2.3)
```

```
##
## Call:
## lm(formula = pres ~ u1, data = hooker)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6671 -0.2751 -0.1478  0.3161  0.9427
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.089e+02  5.559e+00   55.57   <2e-16 ***
## u1          -1.045e+05  2.011e+03  -51.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.405 on 29 degrees of freedom
## Multiple R-squared:  0.9894, Adjusted R-squared:  0.989
## F-statistic:  2701 on 1 and 29 DF,  p-value: < 2.2e-16
```

```
# 두 회귀식 비교
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

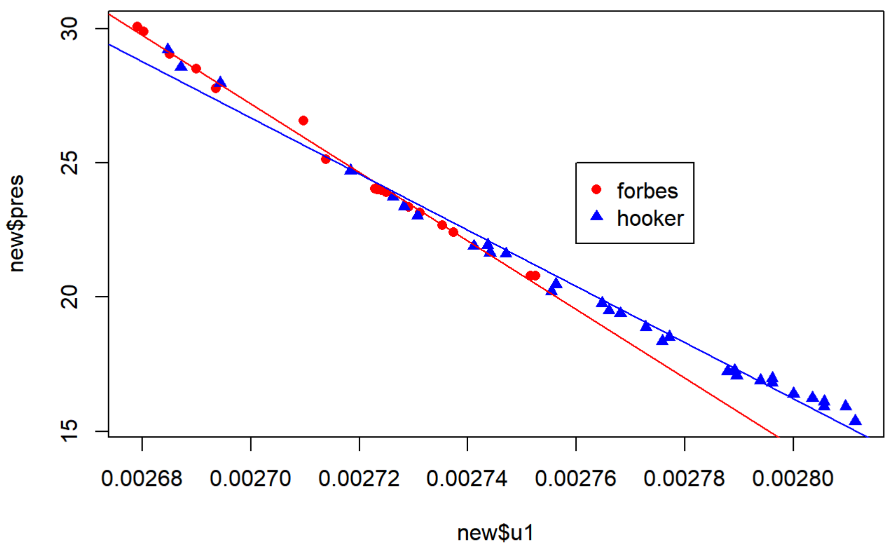
```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
new <- bind_rows(list(data_F = forbe, data_H = hooker), .id="data")

plot(new$u1, new$pres, col = ifelse(new$data=='data_F','red','blue'), pch = c(16, 17)[as.factor(new$data)])
legend(0.00276, 25,
       legend = c("forbes", "hooker"),
       col = c("red", "blue"),
       pch = c(16, 17))
abline(lm2, col = c('red'))
abline(lm2.3, col = c('blue'))
```



2.9외 일부 문제는 맨 뒷부분에.

2.17.2

```
lm.snake <- lm(Y~X-1, snake)
summary(lm.snake)
```

```
##
## Call:
## lm(formula = Y ~ X - 1, data = snake)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4207 -1.4924 -0.1935  1.6515  3.0771
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## X  0.52039    0.01318   39.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.7 on 16 degrees of freedom
## Multiple R-squared:  0.9898, Adjusted R-squared:  0.9892
## F-statistic:  1559 on 1 and 16 DF,  p-value: < 2.2e-16
```

```
df <- nrow(snake)-1
(sigma.square <- sum((snake$Y-lm.snake$fitted.values)^2)/df)
```

```
## [1] 2.889149
```

```
c(lower = lm.snake$coefficients - qt(1-0.05,df)*sqrt(vcov(lm.snake)),
  upper = lm.snake$coefficients + qt(1-0.05,df)*sqrt(vcov(lm.snake)))
```

```
##     lower     upper
## 0.4973811 0.5434069
```

```
(t.val <- (lm.snake$coefficients - 0.49)/sqrt(vcov(lm.snake)))
```

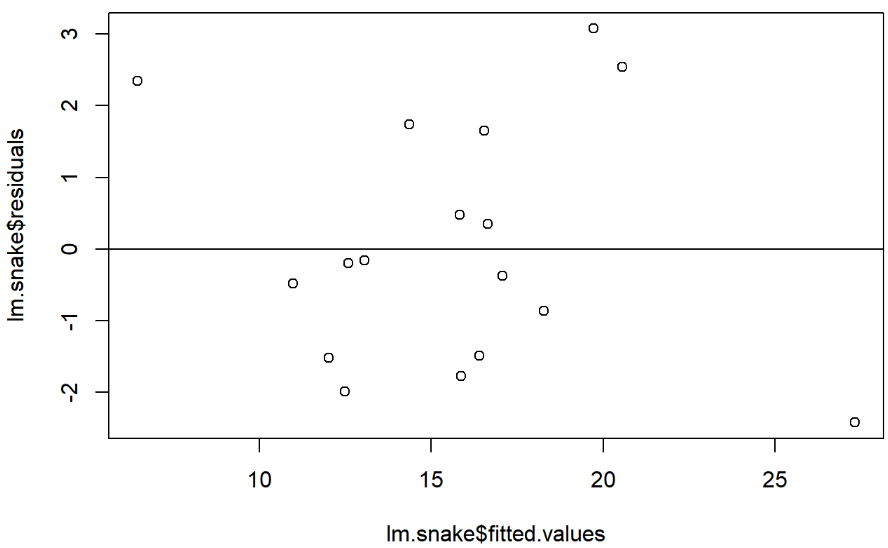
```
##          X
## X 2.305853
```

```
(p.val <- 1-pt(t.val, df))
```

```
##            X
## X 0.01742104
```

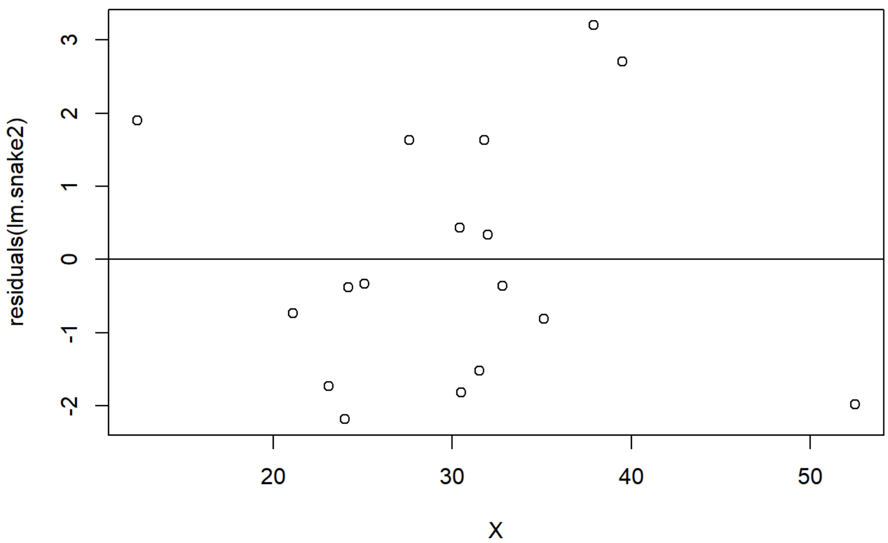
2.17.3

```
plot(lm.snake$fitted.values, lm.snake$residuals)
abline(h=0)
```



뚜렷한 패턴은 보이지 않으나 이상적인 그래프는 아니므로 단순 선형회귀와 비교해보자.

```
lm.snake2 <- lm(Y ~ X, snake)
plot(residuals(lm.snake2) ~ X, snake)
abline(h=0)
```



별 차이가 없다.

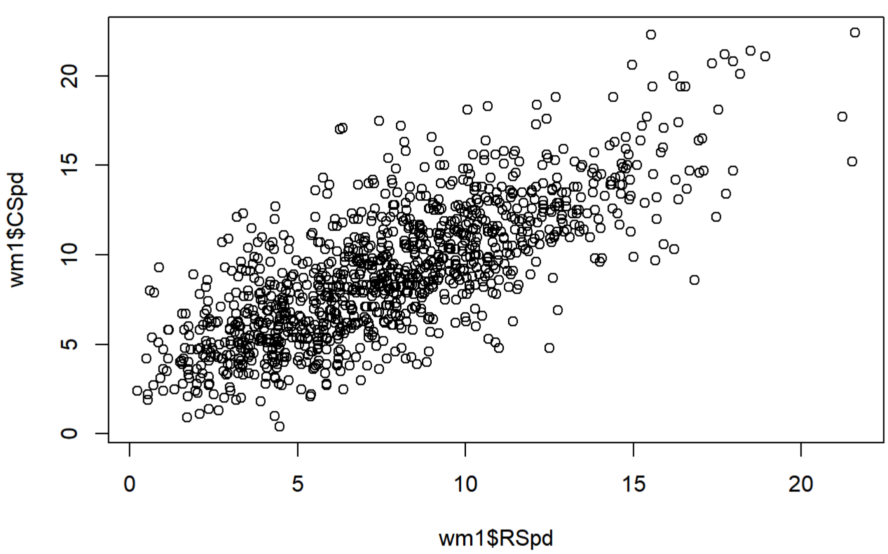
2.21

2.21.1 Draw the scatterplot of the response CSpd versus the predictor RSpd. Is the simple linear regression model plausible for these data?

```
head(wm1)
```

```
##         Date CSpd   RSpd
## 1  2002/1/1/0  6.9 5.9666
## 2  2002/1/1/6  7.1 7.2176
## 3 2002/1/1/12  7.8 7.9405
## 4 2002/1/1/18  6.9 6.0174
## 5  2002/1/2/0  5.5 6.1646
## 6  2002/1/2/6  3.1 1.7687
```

```
plot(wm1$RSpd, wm1$CSpd)
```



데이터가 중심에 몰려있긴 하지만 선형 관계가 보이므로 Simple linear regression도 좋을 것 같다.

2.21.2 Fit the simple regression of the response on the predictor, and present the appropriate regression summaries.

```
lm.21 <- lm(CSpd ~ RSpd, wm1)
summary(lm.21)
```

```
##
## Call:
## lm(formula = CSpd ~ RSpd, data = wm1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7877 -1.5864 -0.1994  1.4403  9.1738
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.14123    0.16958   18.52   <2e-16 ***
## RSpd         0.75573    0.01963   38.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.466 on 1114 degrees of freedom
## Multiple R-squared:  0.5709, Adjusted R-squared:  0.5705
## F-statistic:  1482 on 1 and 1114 DF,  p-value: < 2.2e-16
```

2.21.3 Obtain a 95% prediction interval for CSpd at a time when RSpd = 7.4285.

```
predict(lm.21, newdata=data.frame(RSpd=c(7.4285)), interval="prediction", level=.95)
```

```
##        fit      lwr      upr
## 1 8.755197 3.914023 13.59637
```

2.21.5

```
df = nrow(wm1)-2
fcov = (nrow(wm1)-1)*cov(wm1[-1])
SXX = fcov[2,2]
SXY = fcov[1,2]
SYY = fcov[1,1]
RSS = SYY - SXY^2/SXX
sigmahat2 = RSS/df
sqrt(sigmahat2)
```

```
## [1] 2.466234
```

```
m = 62039
se.mean = sqrt(sigmahat2*(1/m+1/nrow(wm1)+(7.4285-mean(wm1$RSpd))^2/SXX))
c(point = predict(lm.21, newdata=data.frame(RSpd=c(7.4285)), level=.95),
  lower = predict(lm.21, newdata=data.frame(RSpd=c(7.4285)), level=0.95)-qt(1-0.05/2, df)*se.mean,
  upper = predict(lm.21, newdata=data.frame(RSpd=c(7.4285)), level=0.95)+qt(1-0.05/2, df)*se.mean)
```

```
##  point.1  lower.1  upper.1
## 8.755197 8.608433 8.901962
```

2.9 Invariance

2.9.1)

I: $E(Y|X=x) = \beta_0+\beta_1 x$

II: $E(Y|Z=z) = \gamma_0 + \gamma_1 z = \gamma_0 + \gamma_1(ax+b)$

$\Rightarrow \beta_0+\beta_1 x = \gamma_0+\gamma_1 ax+\gamma_1 b$

$\therefore \beta_1 = \gamma_1 a,\ \beta_0 = \gamma_0+\gamma_1 b$

$\Rightarrow \gamma_1 = \beta_1/a$

$\gamma_0 = \beta_0-\gamma_1 b = \beta_0 - \frac{b}{a}\beta_1$

① $\therefore$ γ_1은 a값에, γ_0는 a값과 b값 모두에 영향받는다.

② $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum(y_i-\hat{y}_i)^2}{n-2}$

$\Rightarrow$ y값과 $\hat{y}$은 변화가 없으므로 $\hat{\sigma}^2$은 서로 같다.

③ $\hat{\beta}_1=0$ 에 대한 검정통계량 : $\frac{\hat{\beta}_1}{\hat{\sigma}\sqrt{\frac{1}{\sum(x_i-\bar{x})^2}}}$

$\hat{\gamma}_1=0$ 에 대한 검정통계량 :

$$\frac{\hat{\gamma}_1}{\hat{\sigma}\sqrt{\frac{1}{\sum(z_i-\bar{z})^2}}} = \frac{\frac{1}{a}\hat{\beta}_1}{\hat{\sigma}\sqrt{\frac{1}{\sum a^2(x_i-\bar{x})^2}}} = \frac{\cancel{\frac{1}{a}}\hat{\beta}_1}{\cancel{\frac{1}{a}}\hat{\sigma}\sqrt{\frac{1}{\sum(x_i-\bar{x})^2}}}$$

$=$ $\therefore$ 동일

④ $\hat{\beta}_0=0$ 검정통계량 : $\frac{\hat{\beta}_0}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}}}$

$\hat{\gamma}_0=0$ 검정통계량 : $\frac{\hat{\beta}_0-\frac{b}{a}\hat{\beta}_1}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(a\bar{x}+b)^2}{a^2\sum(x_i-\bar{x})^2}}}$

$\neq$ b값 때문에 다르게 나타난다

2.9.2)

$\mathrm{I}: E(Y|X=x) = \beta_0 + \beta_1 x$

$\mathrm{III}: E(V|X=x) = \delta_0 + \delta_1 x$

$\Rightarrow V = dY$ 이므로 $\delta_0 + \delta_1 x = d\beta_0 + d\beta_1 x$

$\therefore \beta_0 = \frac{1}{d}\delta_0, \quad \beta_1 = \frac{1}{d}\delta_1$

$\delta_0 = d\beta_0, \quad \delta_1 = d\beta_1$

① δ_0와 δ_1은 각각 β_0와 β_1에 d를 곱한 값이다.

② $\hat{\sigma}_Y^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}$

$\hat{\sigma}_V^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum(dY - d\hat{Y}_i)^2}{n-2} = \frac{d^2\sum(Y_i - \hat{Y}_i)^2}{n-2} = d^2\hat{\sigma}_Y^2$

분산은 d^2을 곱한 형태이다.

③ $\hat{\beta}_1 = 0$ 에 대한 검정통계량 : $\frac{\hat{\beta}_1}{\hat{\sigma}\sqrt{\frac{1}{\sum(x_i-\bar{x})^2}}}$

$\hat{\delta}_1 = 0$ 에 대한 검정통계량 :

$$\frac{\hat{\delta}_1}{\hat{\sigma}\sqrt{\frac{1}{\sum(x_i-\bar{x})^2}}} = \frac{d\hat{\beta}_1}{\hat{\sigma}\sqrt{\frac{1}{\sum(x_i-\bar{x})^2}}} \Rightarrow d\text{를 곱한 형태}$$

④ $\hat{\beta}_0 = 0$ 검정통계량 : $\frac{\hat{\beta}_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}}}$

$\hat{\delta}_0 = 0$ 검정통계량 : $\frac{d\hat{\beta}_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}}}$ $\Rightarrow$ d를 곱한형태.

2.17

$$E(y|x) = \beta_1 x, \quad \bar{Y} = \beta_1 \bar{X}$$

2.17.1)

minimize $Q = \Sigma(Y_i - \beta_1 X_i)^2$

$$\frac{\partial Q}{\partial \beta_1} = -2\Sigma(Y_i - \beta_1 X_i) X_i = 0$$ 이 되는 $\hat{\beta}_1$

$$\Sigma X_i(Y_i - \hat{\beta}_1 X_i) = 0$$

$$\therefore \hat{\beta}_1 = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

$$E(\hat{\beta}_1|X) = E\left(\frac{\Sigma x_i y_i}{\Sigma x_i^2} \,\Big|\, X\right) = \Sigma x_i E(y_i|X) / \Sigma x_i^2$$

$$= \Sigma x_i \cdot \beta_1 x_i / \Sigma x_i^2 = \beta_1 \Sigma x_i^2 / \Sigma x_i^2 = \beta_1$$: unbiased

$$Var(\hat{\beta}_1|X) = Var\left(\frac{\Sigma x_i y_i}{\Sigma x_i^2} \,\Big|\, X\right) = \Sigma x_i^2 Var(y_i|X) / (\Sigma x_i^2)^2$$

$$= \sigma^2 / \Sigma x_i^2$$

$$\hat{\sigma}^2 : RSS = \Sigma(y_i - \hat{\beta}_1 x_i)^2 = \Sigma(y_i^2 - 2\hat{\beta}_1 x_i y_i + \hat{\beta}_1^2 x_i^2)$$

$$= \Sigma y_i^2 - 2\hat{\beta}_1 \Sigma x_i y_i + \hat{\beta}_1^2 \Sigma x_i^2$$

$$= \Sigma y_i^2 - \frac{2(\Sigma x_i y_i)^2}{\Sigma x_i^2} + \frac{(\Sigma x_i y_i)^2}{(\Sigma x_i^2)^{\cancel{2}}} \cdot \cancel{\Sigma x_i^2}$$

$$= \Sigma y_i^2 - \frac{(\Sigma x_i y_i)^2}{\Sigma x_i^2}$$

$$\therefore \hat{\sigma}^2 = \left(\Sigma y_i^2 - \frac{(\Sigma x_i y_i)^2}{\Sigma x_i^2}\right) / (n-1)$$

이때, 파라미터가 β_1 1개이므로 자유도는 $n-1$이다.

2.21.4

(1) 피팅된 식을 $y = \hat{\beta_0} + \hat{\beta_1} x$ 라 하자.

$$E(\widetilde{y_*}) = \frac{1}{m}\sum_{i=1}^{m}(\hat{\beta_0} + \hat{\beta_1} x_{*i}) = \hat{\beta_0} + \hat{\beta_1}\cdot\frac{1}{m}\sum_{i=1}^{m} x_{*i}$$

$$= \hat{\beta_0} + \hat{\beta_1}\cdot\overline{x_*}$$

$\therefore$ 예측된 y값의 평균은 $x = \bar{x}$ 일때 예측된 값과 같다.

(2) $$Var(\widetilde{y_*} \mid x) = \underbrace{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(\overline{x_*} - \bar{x})^2}{S_{xx}}\right)}_{\text{적합된 파라미터에의한 분산}} + \hat{\sigma}^2$$

prediction 에 의한 분산

$$\therefore Var(\overline{\widetilde{y_*}} \mid x) = \underbrace{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(\overline{x_*} - \bar{x})^2}{S_{xx}}\right)}_{\text{동일}} + \frac{\hat{\sigma}^2}{m}$$

m개의 prediction 평균에 의한 분산

$$\therefore se(\overline{\widetilde{y_*}} \mid x) = \sqrt{\frac{\hat{\sigma}^2}{m} + \hat{\sigma}^2\left(\frac{1}{n} + \frac{(\overline{x_*} - \bar{x})^2}{S_{xx}}\right)}$$