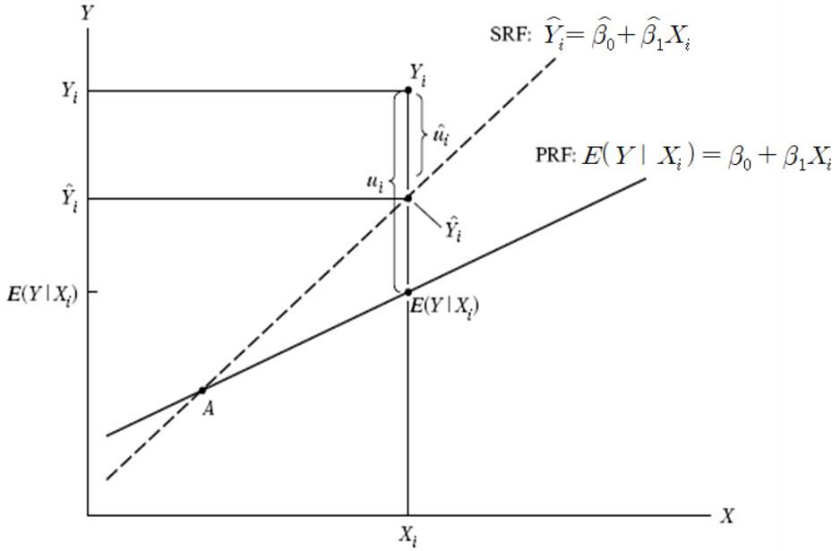
4주차 1차시 : 단순회귀분석(예측)

1.예측

1.예측

- 모형설정→모형추정→가설검정→예측이라는 단순회귀분석의 절차에 따라 주어진 독립변수의 값에 대한 종속변수의 값을 구하는 예측(forecasting, prediction)이라고 함
- 모집단회귀선 위의 한 점을 예측하는 평균예측(mean prediction)과 독립변수의 값에 대응하는 개별 Y의 값을 예측하는 개별예측(individual prediction)이 있음
- 예측에는 주어진 X에 대해 하나의 Y 값을 구하는 점예측(point forecasting)과 점예측에 대한 신뢰구간을 구하는 구간예측(interval forecasting)이 있음
- 구간예측을 위해서는 실제치와 예측치의 차이인 예측오차의 분산을 알아야 함
- 일반적으로 개별예측에서 점예측치 및 예측구간대를 구함



- $Y_i = \beta_0 + \beta_1 X_i + u_i$에서 독립변수 $X_i = X_0$일 때 평균예측과 개별예측은 각각 다음과 같음
  (평균예측) $E(Y_0|X_0) = \beta_0 + \beta_1 X_0$
  (개별예측) $Y_0 = \beta_0 + \beta_1 X_0 + u_0$

- 평균예측과 개별예측의 점 예측치(이는 추정량이 됨)는 모두 동일하며 다음과 같음
  (점예측치) $\widehat{Y_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$

- 예측오차는 평균예측(또는 개별예측)과 점 예측치(추정량)의 차이 이므로 각각 다음과 같음
  (평균예측오차) $E(Y_0|X_0) - \widehat{Y_0} = \beta_0 + \beta_1 X_0 - (\hat{\beta}_0 + \hat{\beta}_1 X_0) = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) X_0$
  (개별예측오차) $Y_0 - \widehat{Y_0} = \beta_0 + \beta_1 X_0 + u_0 - (\hat{\beta}_0 + \hat{\beta}_1 X_0) = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) X_0 + u_0$

- 따라서 예측오차의 분산은 각각 다음과 같음
  (평균예측오차분산) $\mathrm{var}((E(Y_0|X_0) - \widehat{Y_0}) = \mathrm{var}(\hat{\beta}_0) + X_0^2 \mathrm{var}(\hat{\beta}_1) + 2X_0 cov(\hat{\beta}_0, \hat{\beta}_1)$
  $= \sigma_u^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2} \right) + X_0^2 \sigma_u^2 \frac{1}{\sum x_i^2} + 2X_0 \frac{-\bar{X}\sigma_u^2}{\sum x_i^2} = \sigma_u^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n x_i^2} \right)$
  (개별예측오차분산) $\mathrm{var}(Y_0 - \widehat{Y_0}) = \mathrm{var}(\hat{\beta}_0) + X_0^2 \mathrm{var}(\hat{\beta}_1) + 2X_0 cov(\hat{\beta}_0, \hat{\beta}_1) + \sigma_u^2$
  $= \sigma_u^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2} \right) + X_0^2 \sigma_u^2 \frac{1}{\sum x_i^2} + 2X_0 \frac{-\bar{X}\sigma_u^2}{\sum x_i^2} + \sigma_u^2 = \sigma_u^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n x_i^2} \right)$

(점예측)) $\hat{Y_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$ ($X = X_0$일때)

(평균예측의 예측오차 분산)

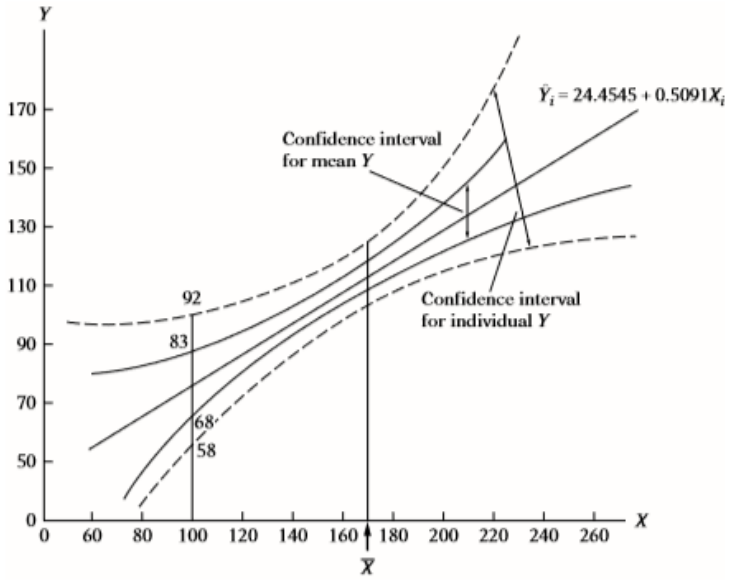
$$\sigma_\epsilon^2 = \sigma_u^2 (\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2})$$

(개별예측의 예측오차 분산)

$$\sigma_\epsilon^2 = \sigma_u^2 (1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2})$$

(예측구간)

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{(n-2, \frac{\alpha}{2})} \sigma_\epsilon$$



(예측오차 분산에 대한 해석)

①표본의 크기(n)가 커질수록 예측오차의 분산이 작아진다. 즉, 관측자료가 많을수록 좋은 예측을 할 수 있다

②교란항의 분산이 커질수록 예측오차의 분산이 커진다. 즉, 원래 회귀모형에서 불확실성이 커서 교란항의 분산이 크면 예측이 어려울 수밖에 없다

③독립변수의 표본평균으로부터 멀어질수록 예측오차도 커진다. 즉, 예측에 주어진 독립변수의 값이 평균으로부터 멀어질수록 표본회귀함수의 예측력은 크게 감소한다

(예제-계속)

(점예측) 추정 회귀식 $\hat{Y} = 0.4 + 1.4\,X$에서
홍보비 지출액이 7(천만 원)일 때 즉, $X_0 = 7$일 때
연간매출액의 개별예측치 및 평균예측치는
$\hat{Y} = 0.4 + (1.4)(7)$ =10.2억 원

(예측오차 분산)

$\hat{\sigma}_u^2$= 1.4667, 독립변수의 평균 $\bar{X} = 4$, 표본의 크기 $n = 5$, $\sum_{i=1}^{n} x_i^2 = 10$

-평균예측오차분산 : $\sigma_\epsilon^2 = \sigma_u^2\left(\frac{1}{n} + \frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right) = (1.4667)(\frac{1}{5} + \frac{(7-4)^2}{10}) = 1.61$

-개별예측오차분산 : $\sigma_\epsilon^2 = \sigma_u^2\left(1 + \frac{1}{n} + \frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right) = (1.4667)(1 + \frac{1}{5} + \frac{(7-4)^2}{10}) = 3.08$

(95% 예측구간)

자유도 \ p	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.923
3	0.1638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032

점예측치가 10.2이고, 예측오차의 분산이 각각 3.08, 1.61이며, $t_{3,0.025}$=3.182

-평균예측구간 : $\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{(n-2,\frac{\alpha}{2})}\,\sigma_\epsilon = 10.2 \pm (3.182)\sqrt{1.61} = [6.15, 14.24]$

-개별예측구간 : $\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{(n-2,\frac{\alpha}{2})}\,\sigma_\epsilon = 10.2 \pm (3.182)\sqrt{3.08} = [4.61, 15.78]$

```
> x0<-7
> (yhat<-beta0+beta1*x0)
[1] 10.2
>
> (sigesq_ind<-sigusq*(1+(1/n)+((x0-mx)^2/sumdxsq)))
[1] 3.08
> (sige_ind<-sqrt(sigesq_ind))
[1] 1.754993
> (yhat_ind_lb<-(yhat-(-tc)*sige_ind))
[1] 4.614829
> (yhat_ind_ub<-(yhat+(-tc)*sige_ind))
[1] 15.78517
>
>
> (sigesq_mean<-sigusq*((1/n)+((x0-mx)^2/sumdxsq)))
[1] 1.613333
> (sige_mean<-sqrt(sigesq_mean))
[1] 1.270171
> (yhat_mean_lb<-(yhat-(-tc)*sige_mean))
[1] 6.15775
> (yhat_mean_ub<-(yhat+(-tc)*sige_mean))
[1] 14.24225
```

```
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
        0.4          1.4

> ols<-lm(y~x)
> summary(ols)

Call:
lm(formula = y ~ x)

Residuals:
         1          2          3          4          5
 8.000e-01 -6.000e-01 -4.441e-16 -1.400e+00  1.200e+00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.400      1.625   0.246   0.8214
x              1.400      0.383   3.656   0.0354 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.211 on 3 degrees of freedom
Multiple R-squared:  0.8167,    Adjusted R-squared:  0.7556
F-statistic: 13.36 on 1 and 3 DF,  p-value: 0.03535

> confint(ols)
                2.5 %   97.5 %
(Intercept) -4.7708632 5.570863
x            0.1812159 2.618784
```

$\widehat{\beta_0}$

$\widehat{\beta_1}$

```
> predict(lm(y~x))
  1   2   3   4   5
3.2 4.6 6.0 7.4 8.8

> new<-data.frame(x = 7)

> predict(lm(y~x), new, se.fit = TRUE)
$fit
   1
10.2

$se.fit
[1] 1.270171

$df
[1] 3

$residual.scale
[1] 1.21106
```

점예측

평균예측오차의 표준오차

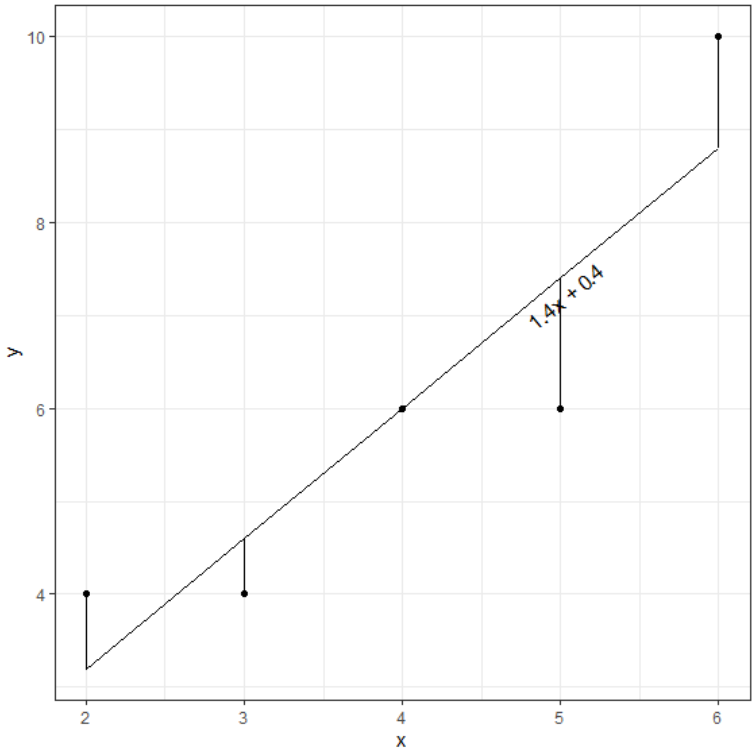
교란항의 표준오차

```
> pred.w.plim <- predict(lm(y ~ x), new, interval = "prediction")

> pred.w.plim
   fit      lwr      upr
1 10.2 4.614829 15.78517
```

개별예측

```
> pred.w.clim<-predict(lm(y ~ x), new, se.fit=T, interval = "confidence")

> pred.w.clim
$fit
   fit     lwr      upr
1 10.2 6.15775 14.24225

$se.fit
[1] 1.270171

$df
[1] 3

$residual.scale
[1] 1.21106
```

평균예측



평균예측오차분산=$1.270171^2 = 1.61$

개별예측오차분산=$1.61 + 1.21106^2 = 3.08$

```
> ggPredict(ols, xpos=0.74, vjust=1.5,show.error=T)
```