경영데이터분석기초

- SPSS, Excel을 활용한 통계분석 -

유 진 호

jhyoo@smu.ac.kr

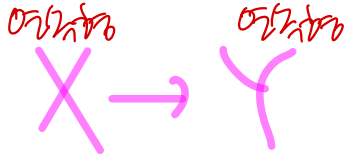
회귀분석(Regression Analysis)

회귀분석(regression analysis)은 변수 x(원인)가 변수 y(결과)에 주는 영향을 알아보기 위한 분석방법이다. 변수 x(원인)를 독립변수(independent variable) 또는 설명변수라 하고, 변수 y(결과)를 종속변수(dependent variable)라 한다.

독립 변수가 1개인 경우를 단순회귀분석(simple regression)이라 하며, 2개 이상인 경우를 다중회귀분석(multiple regression)이라 한다.

ex. X_1 : 강수량 X_2 : 야구장 → Y

회귀분석을 이용하면 변수 x(원인)가 변수 y(결과)에 주는 영향의 정도를 수치화할 수 있기 때문에 예측 등에 활용이 가능하다.



회귀분석에서 종속변수는 간격척도 혹은 비율척도로 측정된 계량적 자료이어야 하며, 독립변수도 마찬가지로 간격척도 혹은 비율척도로 측정된 계량적 자료여야 한다. 하지만 경우에 따라 명목척도 혹은 서열척도로 측정된 자료가 사용될 수 있으며, 이 경우 독립변수를 더미변수(dummy variable)라고 한다.

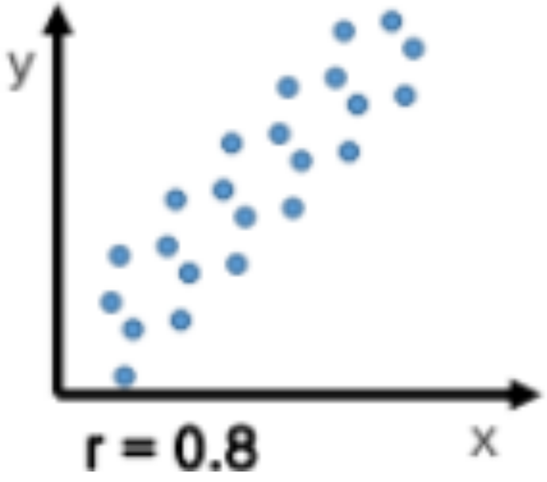
상관분석에서 ···

상관계수

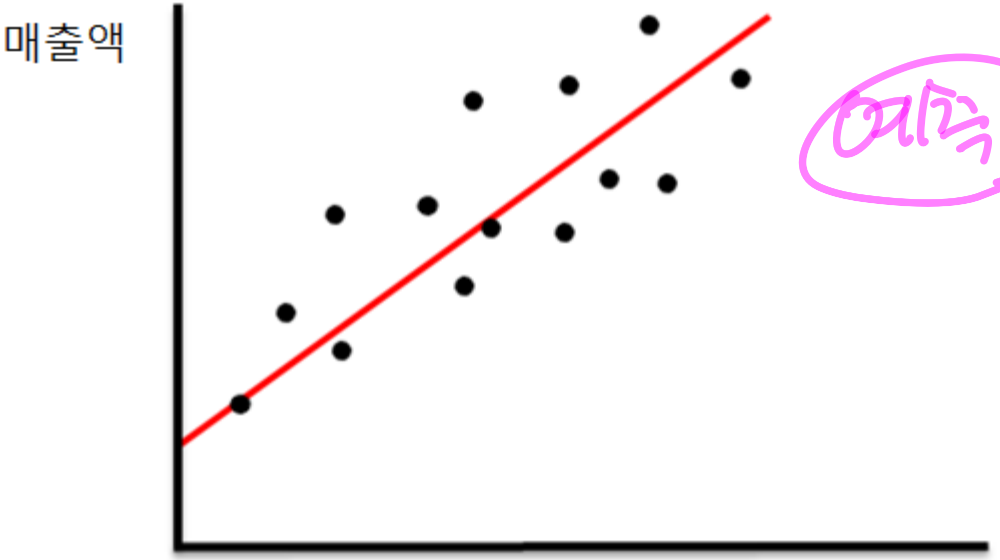
예측을 위해서는 회귀분석 반드시! 이해.

회귀직선

매출



광고비



예측

광고비

bank.

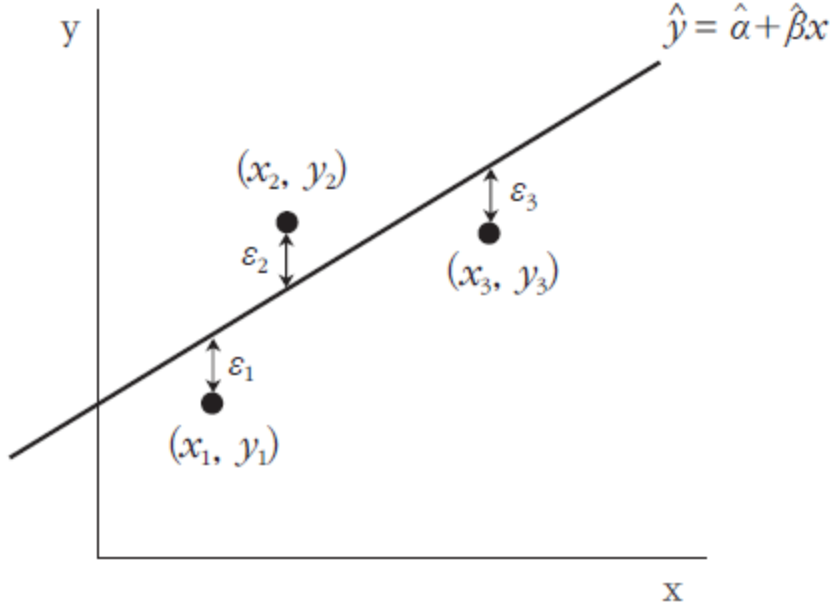
n=4924

변수 x(원인)와 변수 y(결과)의 산점도(scatter plot)이 아래와 같다고 할 때, 회귀분석은 관측값과 예측값의 차이를 최소화하는 직선의 기울기와 절편을 찾는 것을 말한다. 수식으로 표현하면 아래와 같다.

$\hat{\beta}$ $\hat{\alpha}$

$$\min \Sigma \varepsilon_i^{\,2} = \min \Sigma (y_i - \hat{y})^2$$
$$= \min \Sigma (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

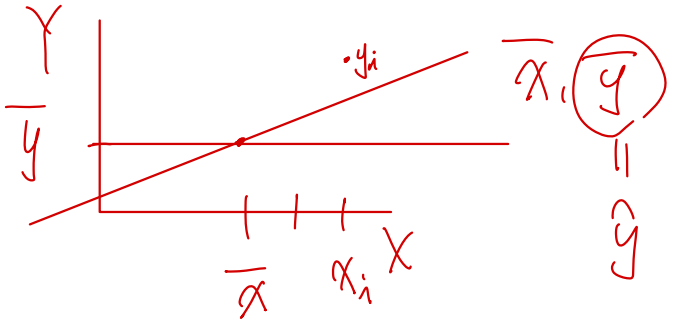
즉, 회귀분석은 관측값(y_i)과 예측값($\hat{y}_i$)의 차이를 최소화하는 직선의 기울기(β)와 절편(α)를 찾는 것이다.



최소제곱법(Method of Least Squares, 최소자승법)
단순회귀분석에서 최소제곱법을 이용해 회귀계수를 계산하면,

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

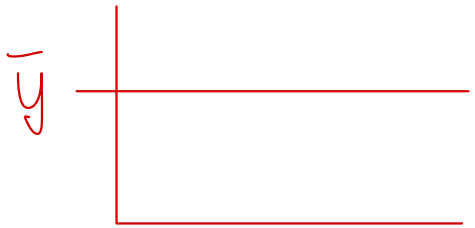
$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$



만약 Xi = $\bar{x}$ 이면?

$\hat{\beta} = 0$

$\hat{\alpha} = \bar{y}$

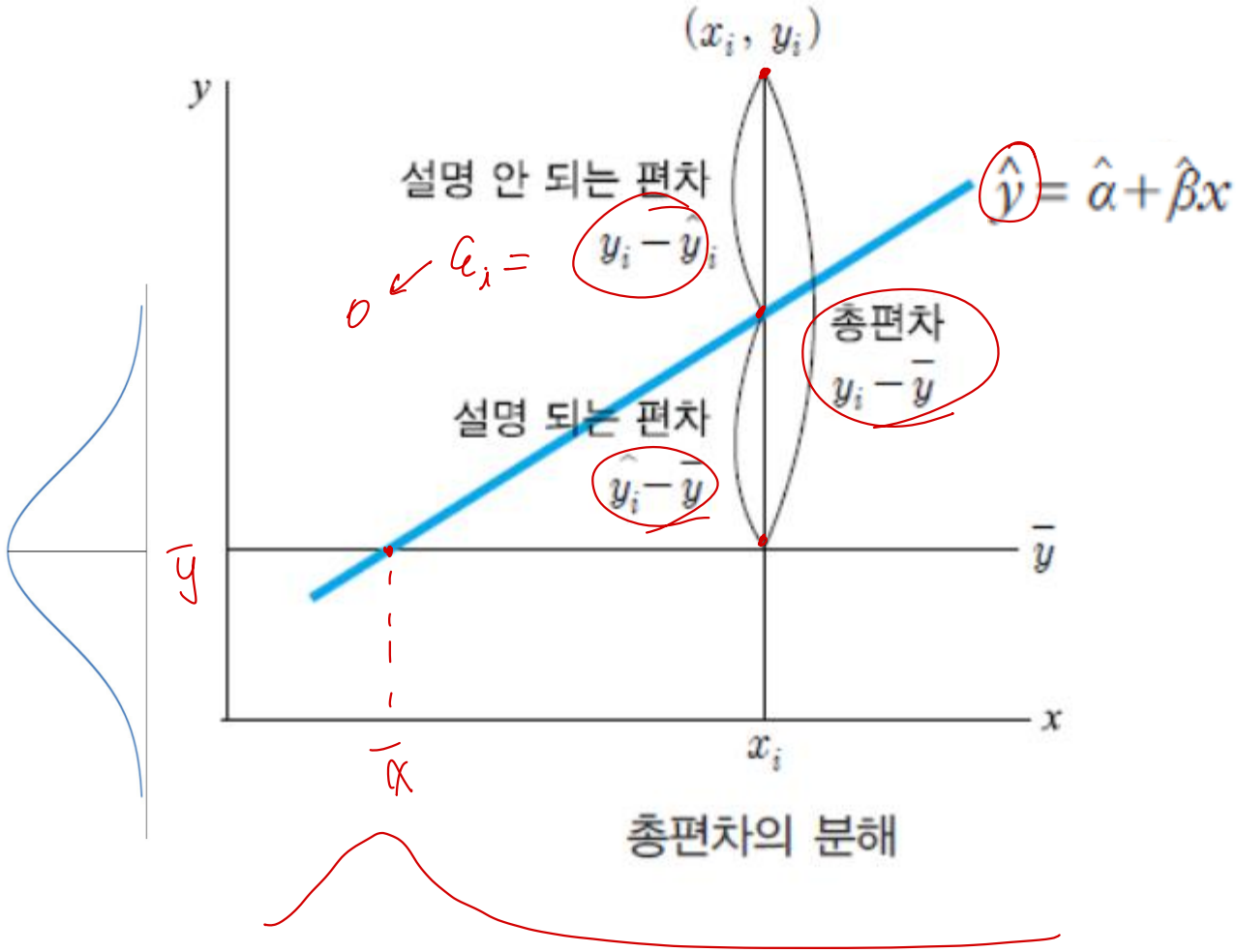


$\hat{y} = \hat{\alpha} + \hat{\beta}x = \hat{\alpha} = \bar{y}$

$\hat{y} = ?$

의미는?

회귀선의 설명력을 평가 => 결정계수



<총편차>의 제곱합

= <설명되는 편차>의 제곱합

+ <설명 안되는 편차>의 제곱합

① 총제곱합(total sum of squares)

$$\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \tag{12-26}$$

② 회귀직선에 의하여 설명되는 회귀제곱합(regression sum of squares)

$$\mathrm{SSR} = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2 \tag{12-27}$$

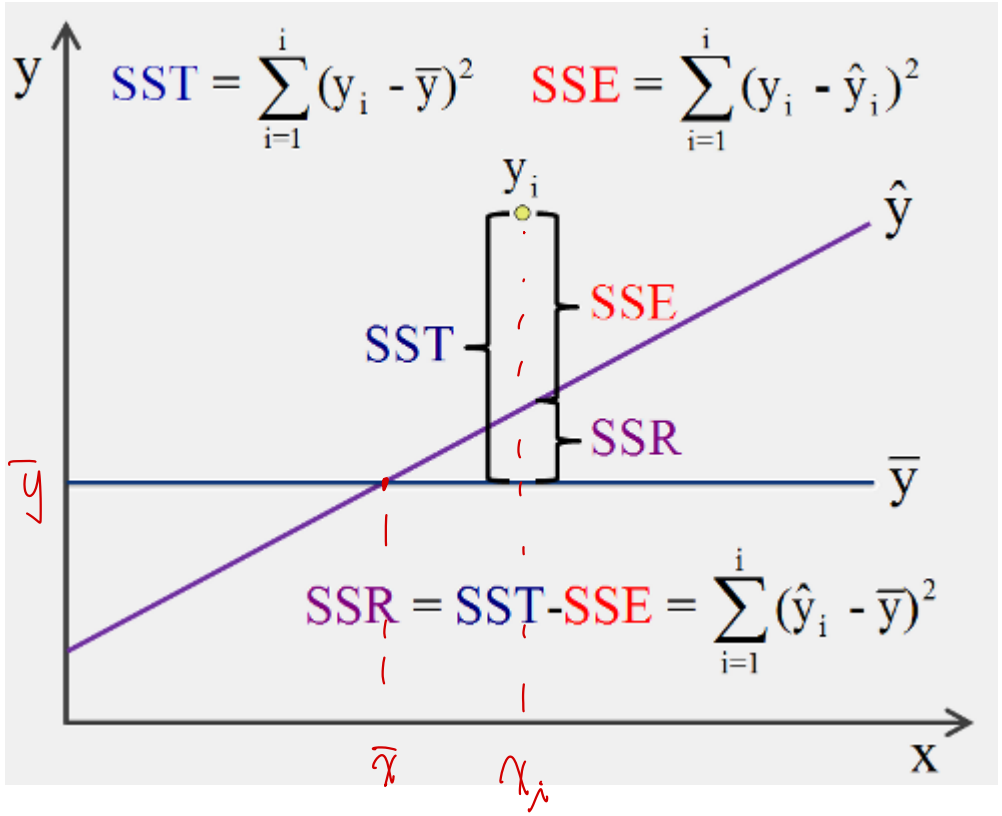
③ 회귀직선에 의하여 설명이 안되는 잔차제곱합(error sum of squares)

$$\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{12-28}$$

즉,

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE} \tag{12-29}$$

회귀선의 설명력을 평가 => 결정계수



$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{SSE}$$

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

회귀선의 설명력을 평가 => 결정계수

결정계수는 총변동 중에서 회귀선에 의하여 설명되는 비율을 의미한다.

결정계수의 범위는 0≦결정계수≦1의 값을 지닌다. 모든 관찰값과 회귀식이 일치한다면 결정계수가 1이 되어 독립변수와 종속변수간에 100%의 상관관계가 있다고 할 수 있다. 즉, 결정계수의 값이 1에 가까울수록 회귀선은 표본을 설명하는데 유용하다.

[수정된 결정계수] 회귀분석이 단계적으로 전개될 때 자유도를 고려하여 조정된 결정계수값으로서, 일반적으로 모집단의 결정계수를 추정할 때 더 자주 사용된다. 표본의 수가 충분히 큰 경우에는 위의 결정계수값과 동일하다.

종속변수는 독립변수에 의해 100% 설명된다.

X → Y

3) 설명력

모형 요약

모형	R	R제곱	수정된 R제곱	추정값의 표준오차
1	.084[a]	.007	.007	2999.379

a. 예측값 : (상수), age

1) 적합성

분산분석[a]

		제곱합	자유도	평균 제곱	F	유의확률
1	회귀모형	287649607.9	1	287649607.9	31.974	.000[b]
	잔차	40654156696	4519	8996272.781		
	합계	40941806304	4520			

a. 종속변수 : balance

b. 예측값 : (상수), age

2) 회귀식

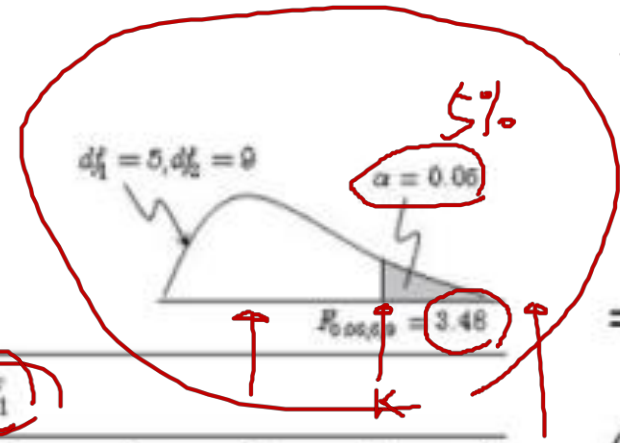
계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	440.651	179.303		2.458	.014
	age	23.852	4.218	.084	5.655	.000

a. 종속변수 : balance

$y = 440.651 + 23.852$

F분포표



=fdist(x,df1,df2)

(예)

=fdist(3.48,5,9)

➔ 0.05

분모의 자유도 df_2	(분자의 자유도) df_1 1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54

Y

N

0.05 > 0.000

$$F = \frac{MSR}{MSE} \sim F(1, n-2)$$

- **회귀평균제곱(MSR : Regression Mean Square)** : 회귀제곱합(SSR)을 자신의 자유도로 나눈 값
- **잔차평균제곱(MSE : Error Mean Square)** : 잔차제곱합(SSE)을 자신의 자유도로 나눈 값

회귀직선이 적합한 모형이라는 것을 의미

[표 12-3] 단순회귀분석의 분산분석표

요인	제곱합(SS)	자유도	제곱평균(MS)	검정통계량	기각역
회귀	SSR	1	$\mathrm{MSR}=\mathrm{SSR}$	$\frac{\mathrm{MSR}}{\mathrm{MSE}}$	$F_\alpha(1,\, n-2)$
잔차	SSE	$n-2$	$\mathrm{MSE}=\frac{\mathrm{SSE}}{n-2}$		
계	SST	$n-1$			

결과해석하기

- 상관분석
  - Age와 balance의 상관계수는 ###, 유의확률은 ###이므로 유의수준 0.05에서 상관성이 존재하다. 다만, 상관계수값이 작으므로 아주 작은 상관관계가 존재한다고 할 수 있다.

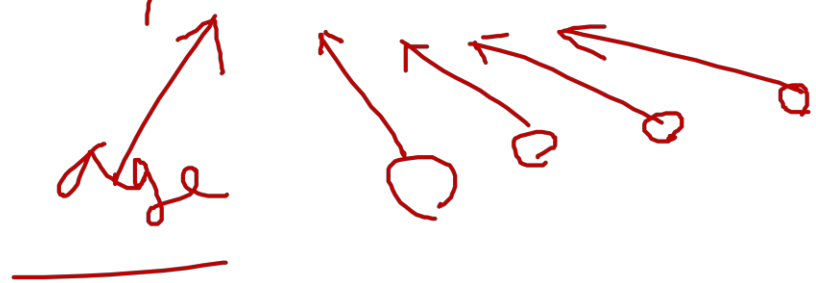
- 교차분석(CrossTab)
  - 카이제곱(X^2) 통계량값이 ###, 유의확률은 ###이므로 유의수준 0.05에서 [Y 집단]별로 [결혼상태]는 서로 다르다(즉, 서로 연관성이 있다)고 할 수 있다.

- 평균차이분석(Means)
  - F통계량 값이 866.5, 유의확률은 0.000 이므로 유의수준 0.05에서 [Y 집단]별로 [duration]은 통계적으로 유의한 차이가 있다.

- 회귀분석
  - 분산분석표에 따라 F 통계량값이 ###, 유의확률은 ###이므로 유의수준 0.05에서 회귀모형이 통계적으로 유의하다고 할 수 있다.
  - t 통계량값이 ###, 유의확률은 ###이므로 유의수준 0.05에서 회귀계수는 0이 아니다. 즉, [나이]는 [balance]에 영향을 준다(나이에 따라 balance는 달라진다)고 할 수 있다. 다만,...

회귀식이 통계적으로 유의한지를 검정하는 분산분석표에서, 통계량의 유의확률이 0.000으로서 0.05보다 작다. 즉, 이 회귀식은 통계적으로 매우 유의하다고 할 수 있다.

[age=23.852, 유의확률 0.000] age의 회귀계수는 23.852이고, 이 회귀계수의 통계적 유의성을 검정하는 t값은 5.655로 유의확률이 0.000<0.05이므로, 이 회귀계수는 통계적으로 유의하다고 볼 수 있다. 다만, 결정계수 값이 0.007로 매우 작기 때문에 balance에 영향을 주는 요인으로서의 설명력은 매우 낮다고 할 수 있다.



9주차.숙제.회귀분석

1) SPSS로 회귀분석 실시하기(age -> balance). 그 결과를 *.xls 파일로 export(내보내기)하여 통계 결과에 대한 해석을 달기

2) 엑셀에서 (age -> balance) 회귀분석 계산한 값과 SPSS 결과값(캡쳐해서) 비교하기 (excel 자료 제출하기).