1강. 단순선형회귀(Simple Linear Regression)

■ 주요용어

용어	해설
단순선형회귀 (simple linear regression)	두 연속 변수(variable) 사이의 관계를 해석하기 위한 통계적 방법
최소제곱법(least squares method)	예측 오차(prediction error)의 제곱합을 최소화하는 방식으로 모델의 계수 (coefficient) 또는 파라미터(parameter) b_0, b_1을 결정하는 방법

■ 정리하기

1. 단순선형회귀(simple linear regression)를 적용하기 위한 조건은 특정 예측변수(predictor) 값에서의 잔차(residual error)들이 독립(Independence, 독립성)이어야 하고, 평균이 0이고(Linearity, 선형성) 분산이 일정(Equal Variances, 등분산성)한 정규분포(normal distribution)을 따르는 것(Normality, 정규성)이다. 정리하면 다음과 같다.

1) Linearity, 선형성

2) Independence, 독립성

3) Normality, 정규성

4) Equal Variances, 등분산성

2. 데이터에 가장 잘 들어맞는 직선을 찾기 위해 잔차(residual error)들의 제곱의 합을 최소화하는 방법을 사용한다.

■ 연습문제

1. 단순선형회귀(simple linear regression)를 적용할 수 있기 위한 조건이 아닌 것은?

(x: 예측변수(predictor variable), y: 반응변수(response variable))

(1) x와, 특정한 x에서의 y의 평균이 선형적인 관계를 가져야 한다.

(2) 특정한 x에서의 잔차(residual error)들이 서로 독립이어야 한다.

(3) 특정한 x에서의 잔차(residual error)들이 평균이 1인 정규분포(normal distribution)를 따라야 한다.

(4) 특정한 x에서의 잔차(residual error)들이 일정한 분산(variance)을 가져야 한다.

정답 : (3)

해설 특정한 x에서의 잔차(residual error)들은 평균이 0인 정규분포(normal distribution)를 따라야 합니다.

2. 단순선형회귀(simple linear regression)에서 최소제곱법(least squares)을 이용하여 기울기(slope)를 구하는 식에 해당하는 것은?

(x_i: i번째 샘플의 예측변수(predictor variable), y_i: i번째 샘플의 반응변수(response variable), $\bar{x}$: x 전체의 평균, $\bar{y}$: y 전체의 평균, n: 샘플 수, $\hat{y}_i$: i번째 샘플에 대한 예측(prediction))

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

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

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

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

정답: (1)

해설: 단순선형회귀(simple linear regression)에서 최소제곱법(least squares)을 이용하면 기울기(slope) $b_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$, 절편(intercept) $b_0=\bar{y}-b_1\bar{x}$와 같이 계산하게 됩니다.

2강. SLR Model Evaluation

■ 주요용어

용어	해설
평균제곱오차 (MSE)	σ^2(y의 분산(variance)에 대한 추정값(estimate), $\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$
결정계수 (coefficient of determination, r^2)	반응변수(response variable)의 변화가 예측변수(predictor variable)에 의해 설명되는 정도를 나타내는 값

■ 정리하기

1. σ^2(y의 분산(variance))을 알 수 없기 때문에 평균제곱오차(MSE)로 σ^2를 추정한다.

2. r^2은 반응변수(response variable)의 변화를 예측변수(predictor variable)로 설명할 수 있는 정도를 나타낸다.

3. t-test를 이용하여 x와 y 사이에 선형 관계가 존재하는지를 테스트할 수 있다.

■ 연습문제

1. 단순선형회귀(simple linear regression)에서 평균제곱오차(MSE)를 계산하는 다음 식의 〈 〉 자리에 들어갈 식으로 알맞은 것은? (단, n은 샘플 수)

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\langle \quad \rangle}$$

(1) n

(2) $n-1$

(3) $n-2$

(4) $n+1$

정답: (3)

2. r^2(coefficient of determination)에 대한 설명으로 틀린 것은?

(1) r^2는 0과 1 사이의 값을 가진다.

(2) r^2은 반응변수(response variable)의 변화를 예측변수(predictor variable)로 설명할 수 있는 정도를 나타낸다.

(3) r^2 값이 0이면 반응변수(response variable)는 예측변수(predictor variable)와 관계가 없다.

(4) r^2 값은 소수의 데이터 포인트들에 의해 크게 변화할 수 있다.

정답: (3)

해설: r^2 값이 0이면 반응변수(response variable)는 예측변수(predictor variable)와 선형적인 관계가 없지만, 비선형적인 관계를 가질 수도 있기 때문에 관계가 없다고 말할 수는 없습니다.

3. 단순선형회귀(simple linear regression)의 기울기(slope)인 β_1 의 $100(1-\alpha)$ 퍼센트 신뢰구간(confidence interval)의 크기는 다음과 같이 정의된다. 신뢰구간(confidence interval)의 크기에 대한 설명으로 옳지 않은 것은? (n: 샘플 수)

$$2 \times t_{\left(\frac{\alpha}{2}, n-2\right)} \times \left(\frac{\sqrt{MSE}}{\sqrt{\sum(x_i - \bar{x})^2}}\right)$$

(1) 샘플 수가 많아질수록 신뢰구간(confidence interval)의 크기는 감소한다.

(2) α가 커질수록 신뢰구간(confidence interval)의 크기는 증가한다.

(3) 예측변수(predictor variable)인 x의 값이 넓게 퍼져있을수록 신뢰구간(confidence interval)의 크기는 감소한다.

(4) 평균제곱오차(MSE)가 작아질수록 신뢰구간(confidence interval)의 크기는 감소한다.

정답: (2)

해설: α가 커질수록 신뢰수준(confidence level)은 감소하므로 신뢰구간(confidence interval)의 크기는 감소하게 됩니다.

3강. SLR Estimation & Prediction

■ 주요용어

용어	해설
평균반응에 대한 신뢰 구간 (confidence interval for mean response)	특정 x에 대한(즉, x를 고정시키고 생각한) y 평균 μ_Y의 신뢰구간(confidence interval)
예측구간 (prediction interval)	새로운 y 관측치(instance)에 대한 예측 범위

■ 정리하기

1. 평균반응(mean response) μ_Y에 대한 신뢰구간(confidence interval)과, 새로운 y 관측치(instance)에 대한 예측구간(prediction interval)은 서로 다른 의미를 가진다.

2. 예측구간(prediction interval)의 크기는 평균반응(mean response)에 대한 신뢰구간(confidence interval)의 크기보다 크다.

3. 오차항(error term)이 정규분포(normal distribution)를 따를 때 신뢰구간(confidence interval)과 예측구간(prediction interval)에 관한 공식을 사용할 수 있다.

■ 연습문제

1. population parameter인 σ^2(y의 분산(variance))을 알기 어렵기 때문에 σ^2의 추정값(estimate)으로 사용되는 값은?

(1) r^2

(2) 회귀제곱합(SSR)

(3) 잔차제곱합(SSE)

(4) 평균제곱오차(MSE)

정답: (4)

해설: $\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$과 같이 계산되는 평균제곱오차(MSE)는 y_i가 $\hat{y}_i$로부터 얼마나 떨어져 있는지의 제곱의 평균이기 때문에 σ^2의 추정값(estimate)으로 사용되는 것이 자연스럽다고 할 수 있습니다.

2. 단순선형회귀(simple linear regression)에서 평균반응(mean response) μ_Y의 $100(1-\alpha)$ 퍼센트 신뢰구간(confidence interval)은 다음과 같이 정의된다. 신뢰구간(confidence interval)의 크기에 대한 설명으로 옳지 않은 것은? (n: 샘플 수)

$$\hat{y}_h \pm t_{\left(\frac{\alpha}{2},n-2\right)} \times \sqrt{MSE \times \left(\frac{1}{n}+\frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

(1) 샘플 수가 많아질수록 신뢰구간(confidence interval)의 크기는 감소한다.

(2) α가 커질수록 신뢰구간(confidence interval)의 크기는 감소한다.

(3) x_h가 $\bar{x}$에 가까울수록 신뢰구간(confidence interval)의 크기는 증가한다.

(4) 평균제곱오차(MSE)가 작아질수록 신뢰구간(confidence interval)의 크기는 감소한다.

정답: (3)

해설: x_h가 $\bar{x}$에 가까울수록 $(x_h-\bar{x})^2$의 크기가 감소하므로 신뢰구간(confidence interval)의 크기는 감소하게 됩니다. 즉, $\bar{x}$에 가까운 x 값에 대해서 예측할수록 예측이 더 정확해지는 셈입니다.

3. 평균반응(mean response) μ_Y에 대한 신뢰구간(confidence interval)과, 새로운 y 관측치(instance)에 대한 예측구간(prediction interval)의 차이에 대한 설명으로 옳지 않은 것은?

(1) 새로운 y 관측치(instance)에 대한 예측구간(prediction interval)은 μ_Y에 대한 추정과 σ^2에 대한 추정에 기반한다.

(2) σ^2에 대한 추정에 평균제곱오차(MSE)를 사용한다.

(3) 예측구간(prediction interval)의 크기는 평균반응(mean response)에 대한 신뢰구간(confidence interval)의 크기보다 작다.

(4) 예측구간(prediction interval)에 대한 공식은 평균반응(mean response)에 대한 신뢰구간(confidence interval)의 공식보다 오차항(error term)의 정규분포(normal distribution)에 대한 가정에 더 강하게 의존한다.

정답: (3)

해설: 예측구간(prediction interval)은 은 μ_Y에 대한 추정과 σ^2에 대한 추정 두 가지에 기반하기 때문에 그 크기가 μ_Y에 대한 추정인 신뢰구간(confidence interval)의 크기보다 커지게 됩니다.

4강. SLR Model Assumptions

■ 주요용어

용어	해설
잔차 대 적합치 그림(residuals vs. fits plot)	데이터 포인트의 적합치(fitted value) 를 x축으로, 잔차(residual) 값을 y 축으로 가지는 산점도(scatter plot)
잔차 대 예측변수 그림(residuals vs. predictor plot)	데이터 포인트의 예측변수 값(predictor value)을 x축으로, 잔차(residual) 값을 y 축으로 가지는 산점도(scatter plot)
잔차 대 순서 그림(residuals vs. order plot)	데이터가 수집된 순서를 x축으로, 잔차(residual) 값을 y 축으로 가지는 산점도(scatter plot)

■ 정리하기

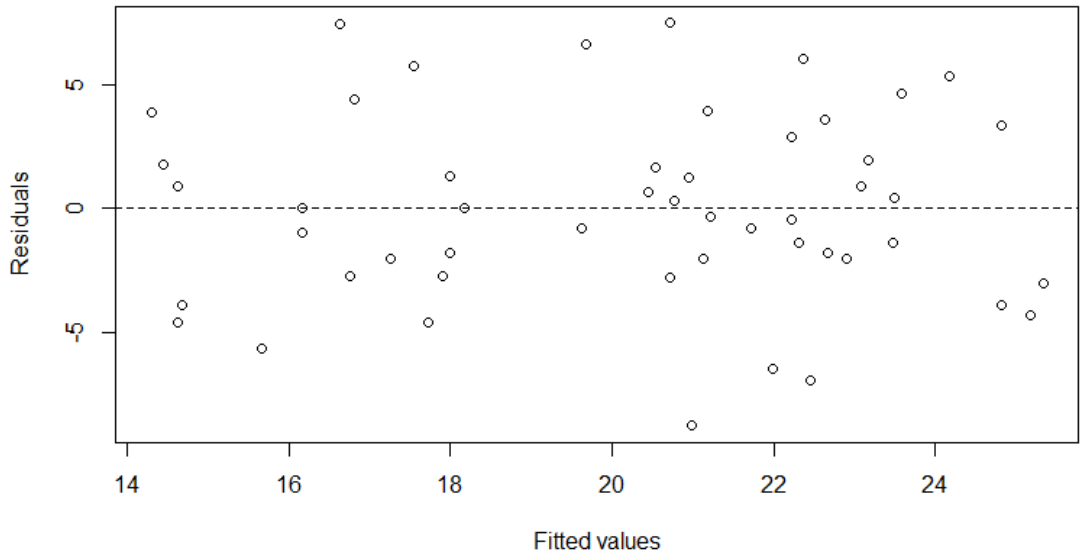
1. 잔차 대 적합치 그림(residuals vs. fits plot)을 통해 단순선형회귀(simple linear regression)의 선형성(Linearity) 가정을 확인할 수 있다.

2. 잔차 대 적합치 그림(residuals vs. fits plot)을 통해 단순선형회귀(simple linear regression)의 등분산성(Equal Variances) 가정을 확인할 수 있다.

3. 잔차 대 순서 그림(residuals vs. order plot)을 통해 단순선형회귀(simple linear regression)의 독립성(Independence) 가정을 확인할 수 있다.

■ 연습문제

1. 다음 잔차 대 적합치 그림(residuals vs. fits plot)을 보고 해석한 결과로 알 수 없는 것은?



(1) 여러 적합치(fitted value)들에서 비슷한 분산(variance)을 보인다.

(2) 잔차(residual) 값이 0 주위에서 특정한 패턴 없이 분포하고 있다.

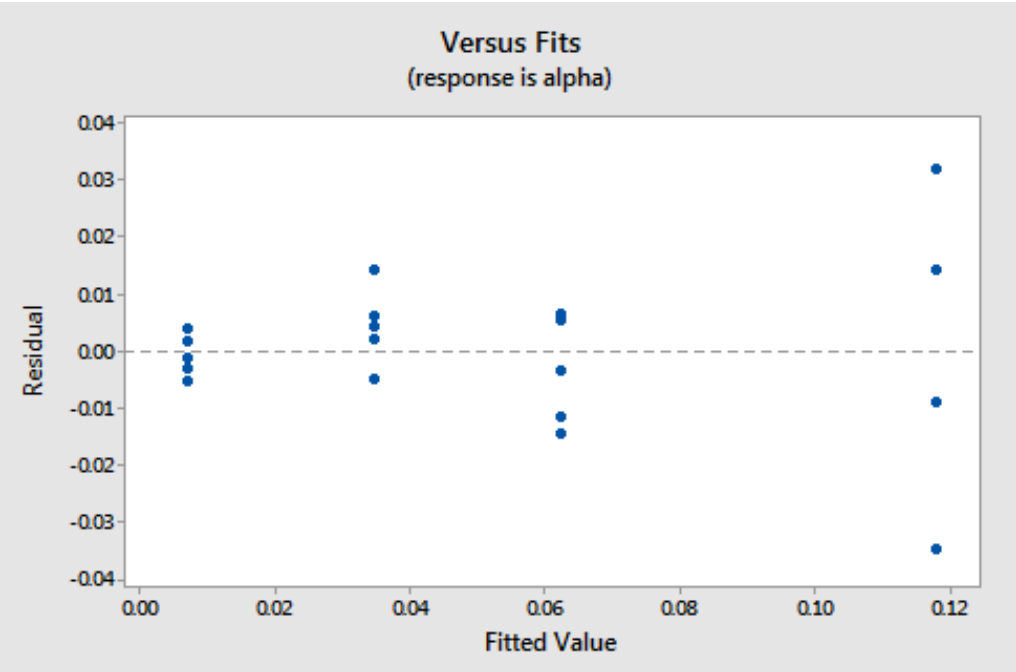
(3) 이 그림으로 미루어 봤을 때 예측변수(predictor variable)와 반응변수(response variable) 사이에 선형 관계(linear relationship)가 성립한다.

(4) 잔차(residual) 값들이 독립적(independent)이다.

정답: (4)

해설: 잔차 대 적합치 그림(residuals vs. fits plot)으로는 잔차(residual) 값들이 독립적인지를 판단할 수 없습니다. 해당 판단을 위해서는 잔차 대 순서 그림(residuals vs. order plot)이 필요합니다.

2. 다음 잔차 대 적합치 그림(residuals vs. fits plot)을 해석했을 때, 이 데이터가 위반하고 있다고 판단 가능한 단순선형회귀(simple linear regression)의 가정은?



(1) 선형성(Linearity)

(2) 독립성(Independence)

(3) 정규성(Normality)

(4) 등분산성(Equal Variances)

정답: (4)

해설: 적합치(fitted value)가 커짐에 따라 분산(variance)이 커지는 양상을 보여주기 때문에 등분산성(Equal Variances) 가정에 맞지 않는다는 것을 알 수 있습니다.

3. 다음 중 단순선형회귀(simple linear regression)의 가정을 확인하는 방법으로 적합하지 않은 것은?

(1) 잔차 대 적합치 그림(residuals vs. fits plot)을 통해 단순선형회귀(simple linear regression)의 선형성(Linearity) 가정을 확인할 수 있다.

2. 잔차 대 적합치 그림(residuals vs. fits plot)을 통해 단순선형회귀(simple linear regression)의 등분산성(Equal Variances) 가정을 확인할 수 있다.

3. 잔차 대 순서 그림(residuals vs. order plot)을 통해 단순선형회귀(simple linear regression)의 정규성(Normality) 가정을 확인할 수 있다.

4. 잔차 대 순서 그림(residuals vs. order plot)을 통해 단순선형회귀(simple linear regression)의 독립성(Independence) 가정을 확인할 수 있다.

정답: (3)

해설: 잔차 대 순서 그림(residuals vs. order plot)을 통해 확인할 수 있는 것은 단순선형회귀(simple linear regression)의 독립성(Independence) 가정입니다.

5강. SLR Model Assumptions II

■ 주요용어

용어	해설
잔차의 정규 확률 그림(normal probability plot of residuals)	잔차(residual)의 이상적 백분위수(theoretical percentile)를 x축으로, 표본 백분위수(theoretical percentile)를 y축으로 가지는 그림
히스토그램 (histogram)	값들의 구간을 x축에 표기하고, 구간에 속하는 값들의 빈도를 높이로 가지는 막대(bar) 형태의 그래프

■ 정리하기

1. 잔차의 정규 확률 그림(normal probability plot of residuals)은 잔차(residual)의 이상적 백분위수(theoretical percentile)를 x축으로, 표본 백분위수(theoretical percentile)를 y축으로 가지는 그림이다.

2. 정규 확률 그림(normal proabability plot)으로 단순선형회귀(simple linear regression)의 가정 중 정규성(Normality) 가정을 확인할 수 있다.

3. 정규 확률 그림(normal proabability plot)이 히스토그램(histogram)보다 잔차(residual)가 정규분포(normal distribution)를 따르는지에 대해 확인하기가 더 쉽다.

■ 연습문제

1. 잔차의 정규 확률 그림(normal probability plot of residuals)에 대한 설명으로 적절하지 않은 것은?

(1) 이 그림으로 단순선형회귀(simple linear regression)의 가정 중 선형성(Linearity) 여부를 확인할 수 있다.

(2) 잔차(residual)의 이상적 백분위수(theoretical percentile)를 x축으로, 표본 백분위수(theoretical percentile)를 y축으로 가지는 그림이다.

(3) 잔차(residual)가 정규분포(normal distribution)를 따른다면 그림이 선형적(linear)인 형태를 가진다.

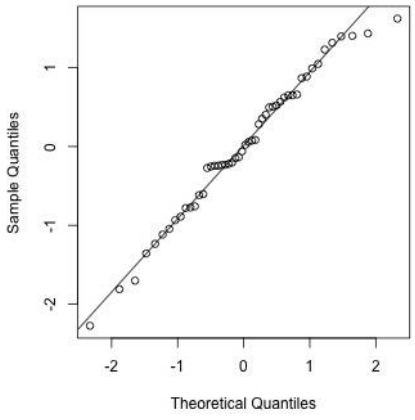
(4) 표준화된 잔차(standardized residual)의 경우, 중앙값(median)에 해당하는 잔차(residual)의 이상적 백분위수(theoretical percentile) 값은 0이다.

정답: (1)

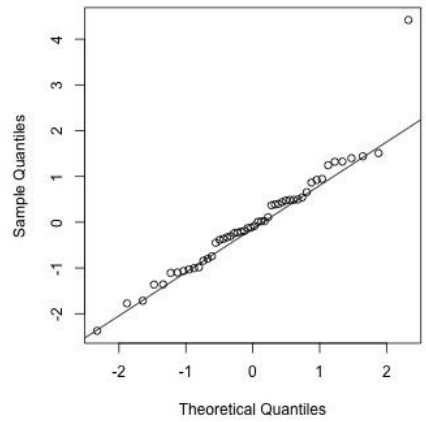
해설: 잔차의 정규 확률 그림(normal probability plot of residuals)을 이용하면 잔차(residual)가 정규분포(normal distribution)를 따르는지에 대해 확인할 수 있습니다.

2. 다음 잔차의 정규 확률 그림(normal probability plot of residuals) 중, 잔차(residual)의 분포가 두꺼운 꼬리(heavy-tail)를 가지는 경우는?

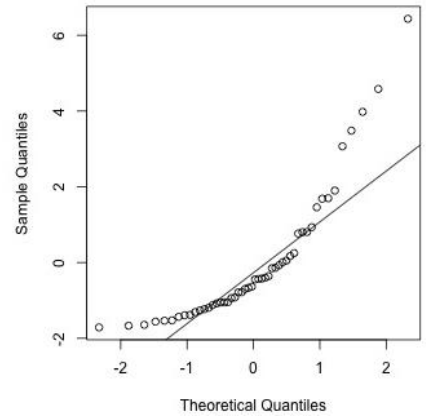
(1)



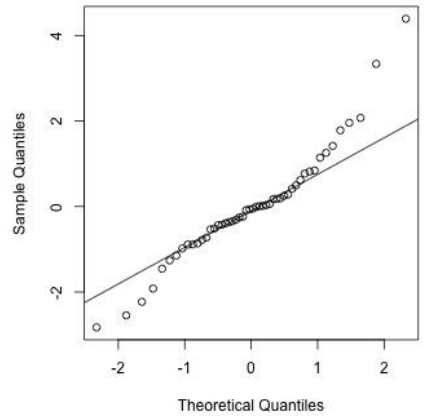
(2)



(3)



(4)



정답: (4)

해설: (4)의 경우에는 정규분포(normal distribution)를 나타내는 직선보다 더 먼 곳까지 데이터들이 분포되어 있기 때문에 꼬리(tail)가 두꺼워지는 형태가 됩니다.

3. 잔차(residual)에 대한 히스토그램(histogram)을 그렸을 때와, 정규 확률 그림(normal proabability plot)을 그렸을 때에 대한 설명으로 적절하지 않은 것은?

(1) 히스토그램(histogram)이 종 형태(bell-shape)를 가지면 대체로 잔차(residual)가 정규분포(normal distribution)를 따른다고 볼 수 있다.

(2) 히스토그램(histogram)은 구간 경계(breakpoint)의 선택과 상관없이 동일한 형태를 가진다.

(3) 정규 확률 그림(normal proabability plot)이 히스토그램(histogram)보다 잔차(residual)가 정규분포(normal distribution)를 따르는지에 대해 확인하기가 더 쉽다.

(3) 정규 확률 그림(normal proabability plot)에서 점들이 직선 형태를 이루고 있으면 잔차(residual)가 정규분포(normal distribution)를 따른다고 볼 수 있다.

정답: (2)

해설: 히스토그램(histogram)은 구간 경계(breakpoint)의 선택에 따라 그 형태가 크게 달라질 수 있습니다.

6강. 다중선형회귀(Multiple Linear Regression)

■ 주요용어

용어	해설
산점도 행렬 (scatter plot matrix)	모든 변수 쌍(pair)에 대한 산점도(scatter plot)를 행렬 형태로 배치한 것
adjusted R^2	다중 선형 회귀(multiple linear regression) 모델을 만드는 데 사용된 파라미터의 수를 고려한 R^2 값
최소제곱추정 (least squares estimation)	오차(error)의 제곱의 합을 최소로 만드는 방식으로 파라미터를 추정하는 것

■ 정리하기

- 산점도 행렬(scatter plot matrix)을 통해서 변수들간의 marginal relationship을 파악할 수 있다.

- 단순선형회귀(simple linear regression)의 가정과 개념들은 다중선형회귀(multiple linear regression)에서도 유사하게 성립한다.

- 기울기(slope)에 해당하는 계수(coefficient)는 다른 예측변수(predictor) 값들이 고정되었을 때, 계수(coefficient)와 연관된 예측변수(predictor) 값이 1 증가함에 따라 반응변수(response)의 평균이 변화하는 양을 의미한다.

- 최소제곱추정(least squares estimation)은 오차(error)들의 제곱의 합을 최소로 만드는 파라미터를 찾는 방식이다.

- 예측변수(predictor variable) 사이의 상관관계(correlation)가 크면, 선형회귀(linear regression) 모델의 파라미터 추정이 안정적으로 이루어지기 어렵다.

■ 연습문제

1. R^2 및 adjusted R^2 값에 대한 다음 설명 중 옳지 않은 것은?

(1) 모델의 파라미터(parameter) 수가 증가하면 R^2 값은 그대로이거나 증가한다.

(2) R^2 값은 반응변수(response)의 평균의 변화량을 예측변수(predictor)들의 변화로 설명할 수 있는 정도를 나타낸다.

(3) 모델의 파라미터(parameter) 수가 증가했는데 R^2 값이 그대로이면 adjusted R^2 값은 증가한다.

(4) 샘플 수가 파라미터(parameter)의 수에 비해 매우 커지면 adjusted R^2 값은 R^2 값과 유사해진다.

정답: (3)

해설: adjusted R^2는 파라미터(parameter)의 수가 늘어남에 따라 값에 페널티를 주는 방식으로 정의되기 때문에, 모델의 파라미터(parameter) 수가 증가했는데 R^2 값이 그대로이면 adjusted R^2 값은 감소하게 됩니다.

2. 다중선형회귀(multiple linear regression)에서 평균제곱오차(MSE)를 계산하는 다음 식의 < > 자리에 들어갈 식으로 옳은 것은? (단, n은 샘플 수이고 p는 파라미터(parameter)의 수)

$$MSE = \frac{SSE}{< >}$$

(1) n

(2) $n-1$

(3) $n-2$

(4) $n-p$

정답: (4)

해설: 샘플의 수를 나타내는 n에서 모델에 사용된 파라미터의 수만큼 자유도(degree of freedom)가 감소하게 됩니다. 따라서 p 개의 파라미터(parameter)가 사용되는 다중선형회귀(multiple linear regression) 모델에서는 $n-p$가 답이 됩니다.

3. 다음은 다중선형회귀(multiple linear regression)의 파라미터(parameter)들을 성분으로 가지는 벡터 $\mathbf{b}$를 최소제곱추정(least squares estimation)으로 구한 식을 나타낸다. 이와 관련된 설명으로 옳지 않은 것은? (단, X는 계획행렬(design matrix), $\mathbf{y}$ 는 반응변수(response variable) 값들을 성분으로 가지는 벡터이다.)

$$\mathbf{b} = (X^T X)^{-1} X^T \mathbf{y}$$

(1) X^{-1}가 존재하지 않으면 $\mathbf{b}$가 정의되지 않는다.

(2) 하나의 예측변수(predictor)가 다른 예측변수(predictor)들의 선형결합(linear combination)으로 나타날 수 있으면 X^{-1}가 존재하지 않는다.

(3) 데이터를 수집할 때 예측변수(predictor)들 간에 중복되는 정보가 많을수록 최소제곱추정(least squares estimation)으로 $\mathbf{b}$를 추정할 때 계산이 안정적으로 이루어진다.

(4) 예측변수(predictor)들 사이의 상관관계(correlation)가 높으면 $\mathbf{b}$의 계산이 불안정할 수 있다.

정답: (3)

해설: 예측변수(predictor)들 간에 중복되는 정보가 많다는 것은 예측변수(predictor)들 사이의 상관관계(correlation)가 높거나 하나의 예측변수(predictor)를 다른 예측변수(predictor) 값들로부터 계산할 수 있다는 뜻이 됩니다. 따라서 예측변수(predictor)들 간에 중복되는 정보가 많은 경우에는 $\mathbf{b}$의 계산이 불안정하거나 또는 일부 계수(coefficient)가 정의되지 않게 됩니다.

7강. MLR Model Evaluation

■ 주요용어

용어	해설
완전모델(full model)	모든 예측변수(predictor) 또는 파라미터를 포함하는 모델
축소모델(reduced model)	완전모델(full model)에 비해 일부 파라미터를 포함하지 않는, 귀무가설(null hypothesis)이 기술하는 모델

■ 정리하기

- 완전모델(full model)은 모든 예측변수(predictor) 또는 파라미터(parameter)를 포함하는 모델을 말한다.

- 축소모델(reduced model)은 완전모델(full model)에 비해 일부 파라미터(parameter)를 포함하지 않는, 귀무가설(null hypothesis)이 기술하는 모델을 말한다.

- general linear F-statistic은 $F^* = \left(\frac{SSE(R)-SSE(F)}{df_R-df_F}\right) \div \left(\frac{SSE(F)}{df_F}\right)$와 같이 정의된다.

- general linear F-statistic 값이 커질수록 축소모델(reduced model)보다 완전모델(full model)을 선호하게 된다.

- general linear F-statistic을 이용해 가설 검정(hypothesis test)을 수행함으로써 연구 질문에 대한 답을 얻을 수 있다.

■ 연습문제

1. 다중선형회귀(multiple linear regression) 모델을 생성시 예측변수(predictor variable) x_1, x_2, x_3가 있고, 각각의 예측변수(predictor variable)에 대응되는 계수(coefficient)를 $\beta_1, \beta_2, \beta_3$라 하자. 예측변수(predictor variable) 중 반응변수(response variable)와 연관이 있는 변수가 하나라도 있는지를 검정(test) 하고자 할 때, 귀무가설(null hypothesis)을 기술한 것으로 적절한 것은?

(1) $\beta_1 = \beta_2 = \beta_3 = 0$

(2) $\beta_1 = \beta_2 = \beta_3 = 1$

(3) $\beta_1 = 0$

(4) $\beta_2 = \beta_3 = 0$

정답: (1)

해설: 예측변수(predictor variable) 중 반응변수(response variable)와 연관이 있는 변수가 하나도 없다는 것이 귀무가설(null hypothesis)이 되므로 $\beta_1 = \beta_2 = \beta_3 = 0$이 됩니다.

2. 다중선형회귀(multiple linear regression) 모델을 생성시 예측변수(predictor variable) x_1, x_2, x_3가 있고, 각각의 예측변수(predictor variable)에 대응되는 계수(coefficient)를 $\beta_1, \beta_2, \beta_3$라 하자. x_1이 고정된 상태에서 x_2 또는 x_3가 반응변수(response variable)와 연관이 있는지 검정(test) 하고자 할 때, 귀무가설(null hypothesis)을 기술한 것으로 적절한 것은?

(1) $\beta_1 = \beta_2 = \beta_3 = 0$

(2) $\beta_2 = \beta_3 = 1$

(3) $\beta_1 = 0$

(4) $\beta_2 = \beta_3 = 0$

정답: (4)

해설: 이 경우 귀무가설(null hypothesis)은 x_2와 x_3 모두 연관이 없다는 것이므로 $\beta_2 = \beta_3 = 0$이 됩니다.

3. general linear F-statistic에 대한 다음 설명 중 옳지 않은 것은?

(1) $F^* = \left(\frac{SSE(R) - SSE(F)}{df_R - df_F}\right) \div \left(\frac{SSE(F)}{df_F}\right)$와 같이 정의된다.

(2) F-statistic 값이 커질수록 축소모델(reduced model)을 완전모델(full model) 보다 선호하게 된다.

(3) F-statistic 값이 커질수록 유의확률(p-value)은 작아진다.

(4) 축소모델(reduced model)은 귀무가설(null hypothesis)이 기술하는 모델을 말한다.

정답: (2)

해설: F-statistic 값이 커질수록 완전모델(full model)을 더 선호하게 됩니다.

8강. MLR Model Evaluation II

■ 주요용어

용어	해설
sequential sums of squares	모델에 예측변수(predictor variable)들을 추가했을 때 생기는 SSR의 증가량 또는 SSE의 감소량
overall F-test	모든 기울기 파라미터(slope parameter)가 0인지를 한꺼번에 확인하는 테스트

■ 정리하기

- 모든 기울기(slope) 파라미터(parameter)가 0인지를 테스트하는 overall F-test는 R의 summary 함수의 F-test 결과를 이용해 수행할 수 있다.

- 하나의 기울기(slope) 파라미터(parameter)가 0인지에 대한 테스트는 R의 anova 함수의 F-test 결과를 이용하거나 summary 함수의 t-test 결과를 이용해 수행할 수 있다.

- 파라미터(parameter)의 부분집합(subset)이 0인지에 대한 테스트는 general linear F-statistic을 직접 계산하여 수행한다.

■ 연습문제

1. 다음과 같은 anova 함수의 출력 결과가 있을 때, SSE(X3, X2, Area) - SSE(X2, Area)를 계산한 값으로 맞는 것은?

```
> anova(model.1)
Analysis of Variance Table

Response: Inf.
          Df  Sum Sq Mean Sq F value    Pr(>F)
Area       1 0.62492 0.62492 32.1115 4.504e-06 ***
X2         1 0.31453 0.31453 16.1622  0.000398 ***
X3         1 0.01981 0.01981  1.0181  0.321602
Residuals 28 0.54491 0.01946
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(1) 0.54491

(2) 0.31453

(3) 0.33434

(4) 0.01981

정답: (4)

해설: SSE(X3, X2, Area) - SSE(X2, Area) = SSR(X3 | X2, Area)이므로 X3의 Sum Sq 열의 값인 0.01981이 됩니다.

2. 다음과 같은 anova 함수의 출력 결과가 있을 때, SSR(X2, X3 | Area)를 계산한 값으로 맞는 것은?

```
> anova(model.1)
Analysis of Variance Table

Response: Inf.
          Df  Sum Sq Mean Sq F value    Pr(>F)
Area       1 0.62492 0.62492 32.1115 4.504e-06 ***
X2         1 0.31453 0.31453 16.1622  0.000398 ***
X3         1 0.01981 0.01981  1.0181  0.321602
Residuals 28 0.54491 0.01946
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(1) 0.54491

(2) 0.31453

(3) 0.33434

(4) 0.01981

정답: (3)

해설: Area가 기존에 모델에 포함되어 있었을 때 X2와 X3를 추가하면 증가하는 회귀제곱합(SSR) 값이므로 X2와 X3의 sequential sum of squares 값인 0.31453과 0.01981을 더해준 0.33434가 됩니다.

3. 다음과 같은 anova 함수의 출력 결과가 있을 때, model.1의 평균제곱오차(MSE) 값으로 맞는 것은?

```
> anova(model.1)
Analysis of Variance Table

Response: Inf.
          Df  Sum Sq Mean Sq F value    Pr(>F)
Area       1 0.62492 0.62492 32.1115 4.504e-06 ***
X2         1 0.31453 0.31453 16.1622  0.000398 ***
X3         1 0.01981 0.01981  1.0181  0.321602
Residuals 28 0.54491 0.01946
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(1) 0.54491

(2) 0.01946

(3) 0.01981

(4) 0.31453

정답: (2)

해설: model.1의 평균제곱오차(MSE) 값은 0.54491 / 28 = 0.01946 이 됩니다.

9강. MLR Estimation, Prediction & Model Assumptions

■ 주요용어

용어	해설
Shapiro-Wilk test	주어진 데이터가 정규분포(normal distribution)에 따라 분포되어 있는지를 확인하는 테스트의 일종

■ 정리하기

- 반응변수(response)와 관련된 구간(interval)의 너비(width)는 샘플 사이즈가 커질수록, 주어진 예측변수(predictor)들이 예측변수(predictor) 마다의 평균에 가까울 수록, 평균제곱오차(MSE)가 작을수록, 신뢰수준(confidence level)이 낮을수록 작아진다.

- 다중선형회귀(multiple linear regression)에서도 잔차 그림(residual plot)들을 이용하여 선형회귀(linear regression)의 가정(assumption)들을 만족하는지의 여부를 확인할 수 있다.

- Shapiro-Wilk test를 통해 잔차(residual)의 정규성(normality)에 대한 가설 검정(hypothesis test)을 수행할 수 있다.

- F-test를 이용해 잔차(residual)의 등분산성(equal variance) 가정에 대한 가설 검정(hypothesis test)을 수행할 수 있다.

■ 연습문제

1. Shapiro-Wilk test에 대한 다음 설명 중 옳지 않은 것은?

(1) 주어진 데이터가 정규분포(normal distribution)에 따라 분포되어 있는지의 여부를 확인하기 위한 테스트이다.

(2) 주어진 데이터가 정규분포(normal distribution)에 가까울수록 통계량(statistic) 값이 1에 가까워진다.

(3) 유의확률(p-value)이 0에 가까운 값이면 데이터가 정규분포(normal distribution)에 따라 분포되어 있다고 결론내릴 수 있다.

(4) 비슷한 종류의 테스트로 Anderson–Darling test나 Kolmogorov–Smirnov test가 있다.

정답: (3)

해설: Shapiro-Wilk test의 경우에는 귀무가설(null hypothesis)이 데이터가 정규분포(normal

distribution)에 따라 분포되어 있다는 것이기 때문에 유의확률(p-value)이 0에 가까운 값이면 귀무가설(null hypothesis)이 기각되어, 데이터가 정규분포(normal distribution)에서 벗어난다는 결론을 내리게 됩니다.

2. 다음 중 잔차(residual)들의 정규 확률 그림(normal probability plot)을 그리기 위한 R 코드로 적절한 것은?

(1) plot(x=fitted(model), y=residuals(model),
    xlab="Fitted values", ylab="Residuals",
    panel.last = abline(h=0, lty=2))

(2) hist(residuals(model), main="")

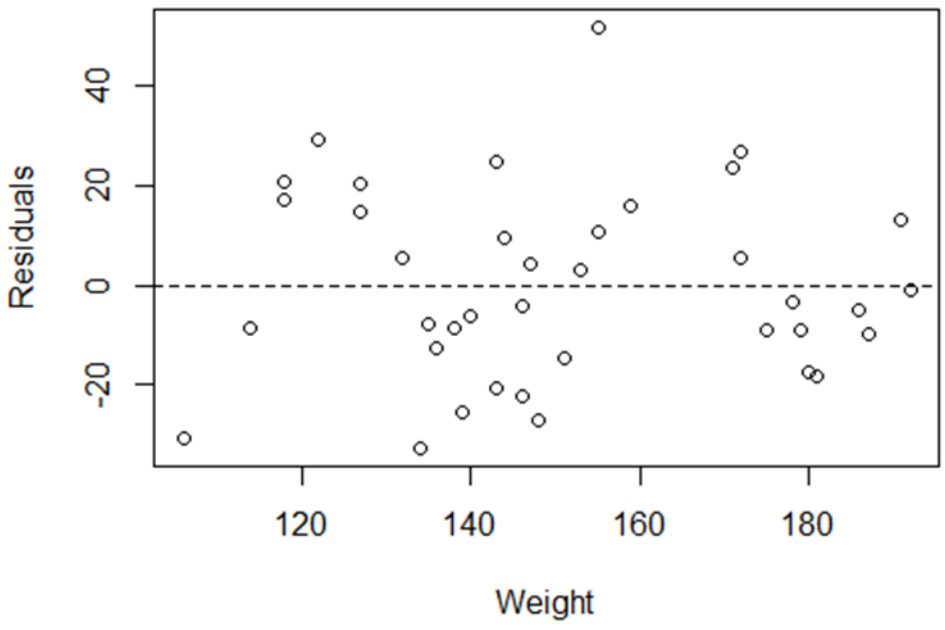
(3) qqnorm(residuals(model), main="", datax=TRUE)

(4) shapiro.test(residuals(model))

정답: (3)

해설: 정규 확률 그림(normal probability plot)을 그리기 위해서는 qqnorm 함수를 사용합니다.

3. 다음 그림은 주어진 샘플에 대해 다중선형회귀(multiple linear regression) model의 잔차(residual)를 y 값으로, 그리고 모델에 포함되지 않은 예측변수(predictor)인 Weight 값을 x 값으로 그린 그림이다. 이 그림을 해석한 것으로 옳지 않은 것은?



(1) 잔차(residual) 값들이 0을 중심으로 흩어져 있다.

(2) 잔차(residual)들의 분산(variance)은 Weight 값과 상관 없이 대체로 일정해 보인다.

(3) Weight와 잔차(residual) 사이에 선형 관계(linear relationship)는 없어 보인다.

(4) Weight 예측변수(predictor)를 모델에 추가하면 잔차(residual) 감소에 도움이 될 것으로 보인다.

정답: (4)

해설: Weight와 잔차(residual) 사이에 선형 관계(linear relationship) 또는 비선형적 관계(non-linear relationship)가 보이지 않기 때문에, 모델에 추가했을 때 잔차(residual) 감소에 도움이 될 것으로 생각할만한 부분이 보이지 않습니다.

10강. 범주형 예측변수(Categorical Predictors)

■ 주요용어

용어	해설
범주형 예측변수(categorical predictor)	연속적인 값이 아니라, 몇 개의 범주(category) 중 하나의 값을 가지는 예측변수(predictor)
이항변수(binary variable) 또는 지시변수(indicator variable)	두 개의 가능한 값을 가지는 변수

■ 정리하기

- c개의 가능한 값을 가지는 범주형 예측변수(categorical predictor)는 모델에서 $c-1$개의 지시변수(indicator variable)을 이용하여 표현된다.

- 반응변수(response)를 예측변수(predictor) 각각의 함수들의 합으로 나타낼 수 있을 때 이 모델에 additive effect를 가진다고 하고, 그렇지 않은 경우 interaction effect를 가진다고 한다.

- interaction effect가 있는 경우 interaction term을 모델에 추가함으로써 보다 정확한 모델을 만들 수 있다.

■ 연습문제

1. 가능한 값이 3개인 범주형 예측변수(categorical predictor)를 선형회귀(linear regression) 모델에서 표현하기 위해 사용해야 하는 지시변수(indicator variable)의 수는?

(1) 1

(2) 2

(3) 3

(4) 4

정답: (2)

해설: c개의 가능한 값을 가지는 범주형 예측변수(categorical predictor)는 모델에서 $c-1$개의 지시변수(indicator variable)를 이용하여 표현됩니다.

2. 지시변수(indicator variable)의 값으로 (1, 0)과 같은 코딩(coding) 방식을 사용할 때와 (1, -1)과 같은 코딩 방식을 사용할 때의 차이와 관련된 설명으로 옳지 않은 것은?

(1) 예측변수(predictor)들의 유의성(significance)에 대한 가설 검정(hypothesis test) 결과는 어떤 코딩 방식을 사용하는지와 상관 없이 동일하다.

(2) 모델의 예측(prediction) 결과는 어떤 코딩 방식을 사용하는지와 상관 없이 동일하다.

(3) 어떤 코딩 방식을 사용하든 지시변수(indicator variable)의 계수(coefficient) 값은 동일하다.

(4) 다른 코딩 방식을 사용하게 되면 모델의 절편(intercept) 값은 달라지게 된다.

정답: (3)

해설: 다른 코딩 방식을 사용하면 지시변수(indicator variable)의 계수(coefficient) 값이 달라지게 됩니다.

3. 어떤 선형회귀(linear regression) 모델에 대해 다음과 같은 관계가 표현하는 성질은? (단, μ_Y는 반응변수(response)의 평균, f는 함수, x는 예측변수(predictor)를 나타낸다.)

$$\mu_Y = f_1(x_1) + f_2(x_2) + \cdots + f_{p-1}(x_{p-1})$$

(1) additive effect

(2) interaction effect

(3) 다중공선성(multicollinearity)

(4) 과적합(overfitting)

정답: (1)

해설: 주어진 수식과 같은 관계가 성립할 때 additive effect를 가진다고 하고, 수식이 성립하지 않으면 interaction effect를 가진다고 합니다.

11강. 데이터 변환(Data Transformations)

■ 주요용어

용어	해설
로그 변환(log transformation)	변수에 로그를 씌우는 변환
역수 변환(reciprocal transformation)	변수의 역수를 취해주는 변환

■ 정리하기

- 잔차(residual)와 관련된 행동(behavior)들은 문제가 없는데, 비선형성(non-linearity)이 문제가 될 때 예측변수(predictor)들을 변환할 수 있다.

- 잔차(residual)와 관련된 행동(behavior)이 문제가 있을 때 y를 변환할 수 있다.

- 로그 변환(log transformation)이 많이 사용되는 이유는 예측변수(predictor)와 반응변수(response)의 관계가 지수(exponential) 함수나 다항(polynomial) 함수와 같이 비선형(non-linear)적인 관계일 때 이런 관계를 선형(linear)적으로 바꿔주는 특성이 있기 때문이다.

■ 연습문제

1. 다음 문장 중 ( ) 안에 들어갈 단어로 적절한 것은?

Transforming predictor values is appropriate when ( ) is the only problem.

(1) 비선형성(non-linearity)

(2) 독립성(independence)

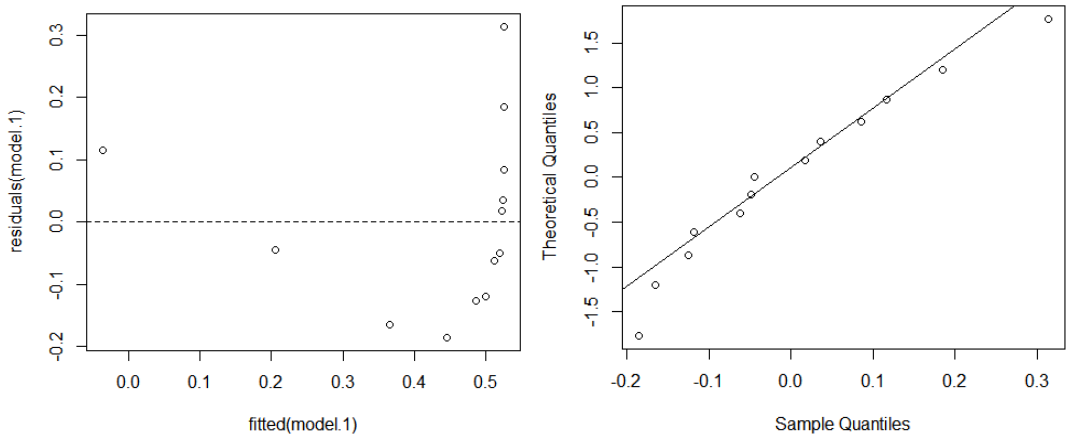
(3) 정규성(normality)

(4) 등분산성(equal variance)

정답: (1)

해설: 예측변수(predictor)들을 변환하는 것은 잔차(residual)와 관련된 행동(behavior)들은 대체로 문제가 없는데, 비선형성(non-linearity)이 문제가 될 때 효과를 볼 수 있습니다.

2. 다음은 어떤 선형회귀(linear regression) 모델의 잔차(residual)를 가지고 잔차 대 적합치 그림(residuals vs. fits plot)과 정규 확률 그림(normal probability plot)을 그린 것이다. 이 그림들로부터 판단할 수 있는 모델의 가정이 아닌 것은?



(1) 선형성(linearity)

(2) 독립성(independence)

(3) 정규성(normality)

(4) 등분산성(equal variances)

정답: (2)

해설: 잔차(residual)들의 독립성(independence)은 주어진 그림들로는 판단할 수 없습니다.

3. 다음 문장 중 ( ) 안에 들어갈 단어들의 쌍으로 적절한 것은?

Transforming the y values should be considered when there are ( ) or ( ) are the problems with the model.

(1) correlated residuals, non-normality

(2) correlated residuals, unequal variance

(3) non-normality, unequal variance

(4) multicollinearity, unequal variance

정답: (3)

해설: non-normality, unequal variance와 같이 잔차(residual)와 관련된 행동(behavior)이 문제가 있을 때 y를 변환하는 것이 도움이 될 수 있습니다.

12강. Model Building

■ 주요용어

용어	해설
과적합(overfitting)	데이터에 비해 상대적으로 모델의 파라미터가 많아서, 모델이 주어진 데이터에 너무 정확히 적합하려고 시도하다보니 오차(error)까지도 모델링하는 현상
일반화 (generalization)	모델이 보지 않은 데이터에 대해서 예측을 잘 수행할 수 있는 능력
교차 검증(cross validation)	주어진 데이터를 여러 분할의 훈련 집합(training set)과 검증 집합(validation set)으로 나누어서 훈련 집합(training set)으로 모델을 학습시킨 후, 검증 집합(validation set)으로 모델의 일반화(generalization) 능력을 측정하는 것을 반복하는 평가 방식

■ 정리하기

- 변수 선택(variable selection)을 위해 stepwise selection, forward selection, backward selection, best subset selection 등의 절차(procedure)를 사용할 수 있다.

- 모델에 많은 예측변수(predictor)들이 포함되어 파라미터(parameter)의 수가 많아지면 과적합(overfitting)과 같은 문제가 발생할 수 있다.

- 모델의 평가 시에 과적합(overfitting) 문제를 방지하기 위해 파라미터(parameter)의 수를 같이 고려하는 지표(metric)를 사용하거나 또는 cross validation과 같은 방법을 사용할 수 있다.

■ 연습문제

1. 다음 모델 평가(evaluation)를 위한 지표(metric) 중 파라미터(parameter)의 수를 고려하지 않는 지표(metric)는?

(1) AIC

(2) BIC

(3) APC

(4) r^2

정답: (4)

해설: (1)~(3)의 지표(metric)들은 파라미터(parameter)의 수인 p가 커질수록 값이 커지게 됩니다.

2. 다음 오차(error) 관련 지표(metric) 중 계산시에 파라미터(parameter)의 수가 커질수록 불이익(penalty)이 주어지는 지표(metric)는?

(1) 평균제곱오차(MSE)

(2) RMSE

(3) MAE

(4) adjusted R^2

정답: (4)

해설: adjusted R^2=$1-\left(\frac{n-1}{n-p}\right)(1-R^2)$와 같이 정의되므로 파라미터(parameter)의 수 p가 커질수록 불이익(penalty)이 주어지게 됩니다.

3. 과적합(overfitting)과 관련된 다음 설명 중 옳지 않은 것은?

(1) 주어진 데이터에 비해 모델의 파라미터(parameter) 수가 너무 많을 때 과적합(overfitting)이 일어날 수 있다.

(2) 과적합(overfitting)은 모델이 데이터의 전반적인 경향을 나타내는 것을 넘어서 오차(error)까지도 정확하게 표현하려고 할 때 나타난다.

(3) 과적합(overfitting)이 일어나면 모델의 일반화(generalization) 능력이 떨어질 수 있다.

(4) 더 많은 예측변수(predictor)가 모델에 포함될수록 과적합(overfitting)이 일어날 가능성이 낮아진다.

정답: (4)

해설: 모델에 포함되는 예측변수(predictor)의 수가 많아지면 예측변수(predictor)에 대응되는 계수(coefficient), 즉 파라미터(parameter)의 수도 많아지기 때문에 과적합(overfitting)이 발생할 가능성이 높아지게 됩니다.

13강. Influential Points

■ 주요용어

용어	해설
이상치(outlier)	y 값이 전반적인 트렌드에서 크게 벗어나는 데이터 포인트
high leverage point	예측변수(predictor) 값들이 극단적인 값을 가지는 데이터 포인트
influential data point	모델 생성 결과에 크게 영향을 주는 데이터 포인트

■ 정리하기

- leverage 값을 이용해 high leverage point를 식별할 수 있다.

- externally studentized residual을 사용하여 이상치(outlier)를 식별할 수 있다.

- DFFITS나 Cook's distance를 사용해 영향력 있는 관측값(influential point) 여부를 판단할 수 있다.

■ 연습문제

1. 다음 중 high leverage point를 식별하기 위해 이용할 수 있는 것은?

(1) Hat matrix

(2) studentized residual

(3) deleted residual

(4) externally studentized residual

정답: (1)

해설: Hat matrix의 주대각선상의 값인 leverage 값을 이용하면 high leverage를 갖는 point들을 식별할 수 있습니다.

2. 다음 중 이상치(outlier)를 식별하기 위해 이용되는 지표(metric)는?

(1) leverage

(2) DFFITS

(3) Cook's distance

(4) externally studentized residual

정답: (4)

해설: 잔차(residual) 값이 큰 데이터 포인트는 해당 포인트의 실제(actual) y 값이 모델의 트렌드와 차이가 많이 난다는 뜻이기 때문에 이상치(outlier) 라고 생각할 수 있습니다. 따라서 externally studentized residual 을 이용해 이상치(outlier)를 식별할 수 있습니다.

3. 다음 중 영향력 있는 관측값(influential point) 여부를 식별하기 위해 이용되는 지표(metric)는?

(1) leverage

(2) deleted residual

(3) Cook's distance

(4) externally studentized residual

정답: (3)

해설: Cook's distance는 $D_i = \frac{(y_i - \hat{y}_i)^2}{p \times MSE}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right)$과 같이 정의되어, 데이터 포인트의 잔차(residual)와 leverage가 클수록 값이 커지는 식으로 되어 있습니다. 따라서 어떤 포인트가 high leverage point 이고 이상치(outlier)에 가까우면 영향력 있는 관측값(influential point)으로 간주하게 됩니다.

14강. 다중공선성(Multicollinearity)

■ 주요용어

용어	해설
다중공선성 (multicollinearity)	둘 이상의 예측변수(predictor) 사이에서 상관관계(correlation)가 일정 이상으로 강하게 나타나는 문제
구조적 다중공선성(structural multicollinearity)	기존 예측변수(predictor)들을 사용해서 새로운 예측변수(predictor)를 만들어 낼 때 생기는 다중공선성(multicollinearity)
데이터 기반 다중공선성(data-based multicollinearity)	데이터 수집 과정을 통제할 수 없기 때문에 생기는, 주어진 데이터에 원래부터 존재하는 다중공선성(multicollinearity)

■ 정리하기

- 둘 이상의 예측변수(predictor) 사이에서 상관관계(correlation)가 일정 이상으로 강하게 나타날 때 다중공선성(multicollinearity)이 존재한다고 한다.

- 다중공선성(multicollinearity)이 존재하는 경우, 모델의 해석과 가설 검정(hypothesis test)에 어려움이 생긴다.

- VIF 값을 이용해서 다중공선성(multicollinearity)을 확인하고, 상관 행렬(correlation matrix)을 이용해서 상관관계가 높은 예측변수(predictor)들을 제거할 수 있다.

■ 연습문제

1. 다음 중 예측변수(predictor)들 사이에 다중공선성(multicollinearity)이 강하게 존재할 때 나타나는 효과가 아닌 것은?

(1) 어떤 예측변수(predictor)들이 모델에 포함되느냐에 따라 예측변수(predictor)의 유의성(significance)에 대한 가설 검정(hypothesis test)의 결과가 달라질 수 있다.

(2) 기존에 모델에 포함되어 있는 예측변수(predictor)들에 따라 새로운 예측변수(predictor)를 모델에 추가할 때 감소하는 오차(error)의 양이 달라지게 된다.

(3) 하나의 예측변수(predictor) 값을 변화시키면 다른 예측변수(predictor) 값이 동시에 변화하게 되기 때문에 계수(coefficient)의 의미를 해석하기가 어려워진다.

(4) 다중공선성(multicollinearity)이 존재하는 예측변수(predictor)들을 단독으로 포함하여 모델링할 때보다 한꺼번에 포함했을 때, 계수(coefficient) 추정값(estimate)의 표준 오차(standard error)가 감소한다.

정답: (4)

해설: 이 경우에는 계수(coefficient)의 표준 오차(standard error)가 커지게 됩니다.

2. 다음 다중공선성(multicollinearity)에 대한 설명 중 옳지 않은 것은?

(1) 둘 이상의 예측변수(predictor) 사이에서 상관관계(correlation)가 일정 이상으로 강하게 나타날 때 다중공선성(multicollinearity)이 존재한다고 한다.

(2) 다중공선성(multicollinearity)이 존재하는 예측변수(predictor)들을 포함하여 모델을 만들어도, 모델을 생성하는데 사용된 예측변수(predictor)들의 범위(scope) 안에서 이루어지는 예측(prediction)은 정확도에 크게 문제가 없다.

(3) 다중공선성(multicollinearity)이 존재할 때, pairwise correlation이 큰 예측변수(predictor) 쌍들을 찾아 그 중 일부를 모델에서 제거해나가는 방식으로 다중공선성(multicollinearity)을 줄일 수 있다.

(4) 예측변수(predictor)의 VIF 값이 작을수록 해당 예측변수(predictor)가 다른 예측변수(predictor)들과 더 강한 상관관계(correlation)를 가진다고 할 수 있다.

정답: (4)

해설: 예측변수(predictor)의 VIF 값이 클 때 해당 예측변수(predictor)가 다른 예측변수(predictor)들과 더 강한 상관관계(correlation)를 가진다고 할 수 있습니다.

3. 선형회귀(linear regression) 모델을 생성하고 summary 함수를 호출한 결과가 다음과 같을 때, Weight 변수에 대한 VIF(variance inflation factor) 값을 계산한 것으로 맞는 것은?

```
> model.2 <- lm(Weight ~ Age + BSA + Dur + Pulse + Stress)
> summary(model.2)

Call:
lm(formula = Weight ~ Age + BSA + Dur + Pulse + Stress)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7697 -1.0120  0.1960  0.6955  2.7035

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.674438   9.464742   2.079  0.05651 .
Age         -0.144643   0.206491  -0.700  0.49510
BSA         21.421654   3.464586   6.183 2.38e-05 ***
Dur          0.008696   0.205134   0.042  0.96678
Pulse        0.557697   0.159853   3.489  0.00361 **
Stress      -0.022997   0.013079  -1.758  0.10052
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.725 on 14 degrees of freedom
Multiple R-squared:  0.8812,    Adjusted R-squared:  0.8388
F-statistic: 20.77 on 5 and 14 DF,  p-value: 5.046e-06
```

(1) $\frac{1}{1.725-1} = 1.379$

(2) $\frac{1}{1-(5.046e-06)} = 1.000$

(3) $\frac{1}{1-0.8812} = 8.417$

(4) $\frac{1}{1-0.8388} = 6.203$

정답: (3)

해설: $VIF_k = \frac{1}{1-R_k^2}$ 와 같이 정의되므로(R_k^2: R^2 value obtained by regressing the kth predictor on the remaining predictors), (3)과 같이 계산됩니다.

15강. 일반화 선형 모델(Generalized Linear Model)

■ 주요용어

용어	해설
연결 함수(link function)	일반화 선형 모델(generalized linear model)에서 반응 변수(response variable)와 선형 예측자(linear predictor) $X\beta$ 사이의 관계를 정의하기 위해 사용되는 함수
승산(odds)	확률값이 π일 때 $\frac{\pi}{1-\pi}$와 같이 정의되는 값
logit function	확률값 π를 받아 $\log\left(\frac{\pi}{1-\pi}\right)$와 같이 로그 승산(log odds)으로 변환하는 함수
가능도(likelihood)	주어진 파라미터(parameter)들에 대해 주어진 데이터를 관측하게 될 확률

■ 정리하기

- 반응변수(response variable)가 정규분포(normal distribution)를 따르지 않는 경우에도 적절한 link function을 적용함으로써 선형 모델(linear model)을 확장할 수 있다.

- 반응변수(response)가 이항 변수(binary variable)의 형태를 가질 경우에는 logistic regression model을 적용할 수 있다.

- logistic regression에서는 sum of squares 대신 deviance가 사용된다.

■ 연습문제

1. 다음은 logistic regression model에 대해 anova 함수를 호출한 결과이다. 이 모델의 R^2 값을 계산한 결과로 맞는 것은?

```
> anova(model.2, test="Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: REMISS

Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev Pr(>Chi)

NULL                    26     34.372

LI    1   8.2988        25     26.073 0.003967 **

TEMP  1   1.2564        24     24.817 0.262338
```

(1) $1-\frac{24.817}{34.372}=0.278$

(2) $1-\frac{1.2564}{8.2988}=0.849$

(3) $\frac{24.817}{34.372}=0.722$

(4) $\frac{1.2564}{8.2988}=0.151$

정답: (1)

해설: $R^2=1-\frac{l(\hat{\beta})}{l(\hat{\beta}^{(0)})}$ 와 같이 정의되는데, Resid. Dev 값은 각각 $24.817: -2\times l(\hat{\beta}), 34.372: -2\times l(\hat{\beta}^{(0)})$ 를 나타내기 때문에, (1)과 같은 방식으로 R^2 값을 계산할 수 있습니다.

2. 다음은 logistic regression model에 대해 summary 함수를 호출한 결과이다. LI 예측변수(predictor)에 대한 Wald test statistic 값으로 맞는 것은?

```
> summary(model.2)


Call:

glm(formula = REMISS ~ LI + TEMP, family = "binomial")


Deviance Residuals:

    Min       1Q   Median       3Q      Max

-1.7173  -0.6430  -0.2613   0.7549   1.7268


Coefficients:

            Estimate Std. Error z value Pr(>¦z¦)

(Intercept)   44.932     46.389   0.969   0.3327

LI             3.260      1.338   2.437   0.0148 *

TEMP         -49.428     47.386  -1.043   0.2969

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


(Dispersion parameter for binomial family taken to be 1)


    Null deviance: 34.372  on 26  degrees of freedom

Residual deviance: 24.817  on 24  degrees of freedom

AIC: 30.817


Number of Fisher Scoring iterations: 5
```

(1) 3.260

(2) 1.338

(3) 2.437

(4) 0.0148

정답: (3)

해설: Wald test statistic은 $\frac{\hat{\beta}_i}{s.e.(\hat{\beta}_i)}$과 같이 정의되므로 $\frac{3.26}{1.338} = 2.437$ 이 됩니다.

3. 확률 $\pi = 0.5$ 일 때 log odds를 계산한 결과로 맞는 것은?

(1) -1

(2) 0

(3) 0.5

(4) 1

정답: (2)

해설: log odds는 $\log\left(\frac{\pi}{1-\pi}\right)$와 같이 정의되므로 $\log\left(\frac{0.5}{1-0.5}\right) = \log 1 = 0$이 됩니다.