**An Introduction to Statistical Learning with Applications in R**

**(R의 응용 프로그램을 사용한 통계 학습 소개)**

Second Edition (두번째 버전)

**Start Date** : 23.01.01

**Student** : Rae Kyeong Jeong

**Professor** : Yujin Chung

**E-mail** : [yujinchung@kyonggi.ac.kr](mailto:yujinchung@kyonggi.ac.kr)

**Member**

~~정재형(17)~~ ~~김가영(19)~~ 김상철(18) 김현주(20) ~~송기현(20)~~ ~~박선현(21)~~ 정래경(19) 배진현(18) 오승태(19)

**Research Assistant Time**

Tuesday 20:00 ~ 21:00 [초반 ISLR]

**Activity Report**

Every week Wednesday and Sunday

**PDF Download Site**

<https://www.statlearning.com/>

**Note**

1. 미리미리 번역이라도 해놓기!

2.

3.

4.

**Study Range**

2.2.1 Measuring the Quality of Fit

2.2.3 K-Nearest Neighbors ~ 2.3 전까지 (Lab 미실시)

3.1.3 Assessing the Accuracy of the Model ~ 3.2

-- 새로 시작 23.03.12 –

2.1.2 How Do We Estimate f?의 Non-Parametric Methods ~ 2.1.3 The Trade-Off Between Prediction Accuracy and Model Interpretability

2.2.2 The Bias-Variance Trade-Off

**2.2.1 맞춤 품질 측정 (Measuring the Quality of Fit) p29 ~ p33**

In order to evaluate the performance of a statistical learning method on a given data set, we need some way to measure how well its predictions actually match the observed data.

주어진 데이터 세트에 대한 통계적 학습 방법의 성능을 평가하려면, 예측이 실제로 관찰된 데이터와 얼마나 잘 일치하는지 측정할 방법이 필요합니다.

That is, we need to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation.

즉, 주어진 관찰에 대한 예측 응답 값이 해당 관찰에 대한 실제 응답 값에 가까운 정도를 정량화해야 합니다.

In the regression setting, the most commonly used measure is the mean squared error (MSE), given by (2.5) where ˆf(xi) is the prediction that ˆf gives for the i th observation.

회귀 설정에서, 가장 일반적으로 사용되는 측도(방법)는 평균 제곱 오차(MSE)로, (2.5)에서 제공됩니다. 여기서 텍스트이(가) 표시된 사진

자동 생성된 설명ˆf(xi)는 ˆf가 i 관측치에 대해 제공하는 예측입니다.

The MSE will be small if the predicted responses are very close to the true responses, and will be large if for some of the observations, the predicted and true responses differ substantially.

MSE는 예측 응답이 실제 응답과 매우 가까우면 작을 것이고, 일부 관측치에 대해 예측 응답과 실제 응답이 크게 다른 경우에는 MSE가 커질 것입니다.

The MSE in (2.5) is computed using the training data that was used to fit the model, and so should more accurately be referred to as the training MSE.

(2.5)의 MSE는 모델을 맞추는 데 사용된 훈련 데이터를 사용하여 계산되므로 그래서 더 정확하게는 훈련 MSE라고 언급해야 합니다.

But in general, we do not really care how well the method works on the training data.

그러나 일반적으로, 우리는 방법이 훈련 데이터에서 얼마나 잘 작동하는지 별로 신경 쓰지 않습니다.

Rather, we are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data.

오히려, 이전에 본 적이 없는 테스트 데이터에 방법을 적용할 때 얻은 예측의 정확도에 관심이 있습니다.

Why is this what we care about?

우리가 이것에 관심을 갖는 이유는 무엇일까요?

Suppose that we are interested in developing an algorithm to predict a stock’s price based on previous stock returns.

이전 주식 수익률을 기반으로 주식 가격을 예측하는 알고리즘을 개발하는 데 관심이 있다고 가정합니다.

We can train the method using stock returns from the past 6 months.

지난 6개월의 주식 수익률을 사용하여 방법을 훈련할 수 있습니다.

But we don’t really care how well our method predicts last week’s stock price.

그러나 우리는 우리의 방법이 지난 주 주가를 얼마나 잘 예측하는지 별로 신경 쓰지 않습니다.

We instead care about how well it will predict tomorrow’s price or next month’s price.

우리는 대신 그것이 내일 가격이나 다음 달 가격을 얼마나 잘 예측할 수 있는지에 관심을 가집니다.

On a similar note, suppose that we have clinical measurements (e.g. weight, blood pressure, height, age, family history of disease) for a number of patients, as well as information about whether each patient has diabetes.

마찬가지로 여러 환자에 대한 임상 측정 (예: 체중, 혈압, 키, 나이, 질병 가족력)과 각 환자의 당뇨병 여부에 대한 정보가 있다고 가정합니다.

We can use these patients to train a statistical learning method to predict risk of diabetes based on clinical measurements.

우리는 이 환자들을 사용하여 임상 측정을 기반으로 당뇨병의 위험을 예측하는 통계적 학습 방법을 훈련할 수 있습니다.

In practice, we want this method to accurately predict diabetes risk for future patients based on their clinical measurements.

실제로, 우리는 이 방법이 임상 측정을 기반으로 미래 환자의 당뇨병 위험을 정확하게 예측하기를 원합니다.

We are not very interested in whether or not the method accurately predicts diabetes risk for patients used to train the model, since we already know which of those patients have diabetes.

우리는 이미 어떤 환자가 당뇨병에 걸렸는지 알고 있기 때문에 이 방법이 모델을 훈련하는 데 사용된 환자의 당뇨병 위험을 정확하게 예측하는지 여부에 별로 관심이 없습니다.

To state it more mathematically, suppose that we fit our statistical learning method on our training observations {(x1, y1),(x2, y2),...,(xn, yn)}, and we obtain the estimate ˆf.

좀 더 수학적으로 표현하자면, 학습 관찰 {(x1, y1),(x2, y2),...,(xn, yn)}에 통계적 학습 방법을 적용하고 추정치 ˆf를 얻는다고 가정합니다.

We can then compute ˆf(x1), ˆf(x2),..., ˆf(xn).

그런 다음 ˆf(x1), ˆf(x2),..., ˆf(xn)을 계산할 수 있습니다.

If these are approximately equal to y1, y2,...,yn, then the training MSE given by (2.5) is small.

이들이 대략 y1, y2,...,yn과 같다면 (2.5)로 주어진 훈련 MSE는 작습니다.

However, we are really not interested in whether ˆf(xi) ≈ yi; \* 근삿값 : ≈

그러나, 우리는 ˆf(xi) ≈ yi인지 여부에 실제로 관심이 없습니다.

instead, we want to know whether ˆf(x0) is approximately equal to y0, where (x0, y0) is a previously unseen test observation not used to train the statistical learning method.

대신, f(x0)가 대략 y0과 같은지 여부를 알고 싶습니다. 여기서 (x0, y0)은 통계적 학습 방법을 훈련하는 데 사용되지 않은 이전에 본 적이 없는 테스트 관찰입니다.

We want to choose the method that gives the lowest test MSE, as opposed to the lowest training MSE.

가장 낮은 훈련 MSE가 아니라(~과 반대로) 가장 낮은 테스트 MSE를 제공하는 방법을 선택하려고 합니다.

In other words, if we had a large number of test observations, we could compute (2.6) the average squared prediction error for these test observations (x0, y0).

즉, 만약 많은 수의 테스트 관찰을 가지고 있다면, 이러한 테스트 관찰(x0, y0)에 대한 평균 제곱 예측 오류 (2.6)를 계산할 수 있습니다.

We’d like to select the model for which this quantity is as small as possible.

이 수량이 가능한 한 적은 모델을 선택하고 싶습니다.

How can we go about trying to select a method that minimizes the test MSE?

테스트 MSE를 최소화하는 방법을 선택하려면 어떻게 해야 합니까?

In some settings, we may have a test data set available that is, we may have access to a set of observations that were not used to train the statistical learning method.

일부 설정에서는, 사용 가능한 테스트 데이터 세트가 있을 수 있습니다. 즉, 통계 학습 방법을 교육하는 데 사용되지 않은 관찰 세트에 액세스(접근, 처리)할 수 있습니다.

We can then simply evaluate (2.6) on the test observations, and select the learning method for which the test MSE is smallest.

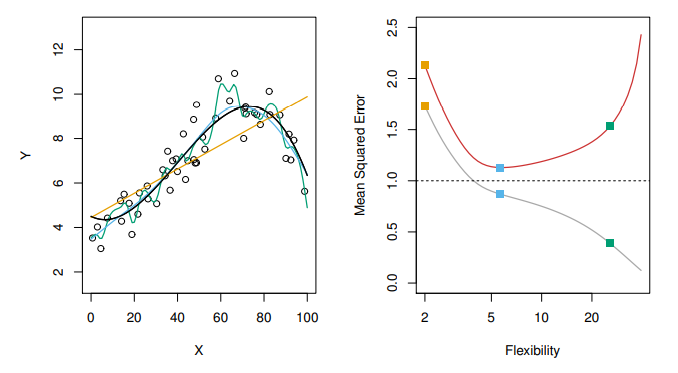
그런 다음 테스트 관찰에서 간단히 평가(2.6)하고, 테스트 MSE가 가장 작은 학습 방법을 선택할 수 있습니다.

But what if no test observations are available?

그러나 테스트 관찰을 사용할 수 없다면 어떻게 될까요?

In that case, one might imagine simply selecting a statistical learning method that minimizes the training MSE (2.5).

이 경우, 훈련 MSE(2.5)를 최소화하는 통계적 학습 방법을 단순히 선택하는 것을 상상할 수 있습니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ[ FIGURE 2.9. 그림 2.9 ]ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

Left: Data simulated from f, shown in black.

왼쪽: 검은색으로 표시된 f에서 시뮬레이션한 데이터

Three estimates off are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).

선형 회귀선(주황색 곡선)과 두 개의 평활화 스플라인 맞춤(파란색 및 녹색 곡선)의 세 가지 추정치가 표시됩니다.

Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line).

오른쪽: 훈련 MSE (회색 곡선), 테스트 MSE (빨간색 곡선) 및 모든 방법에 대한 최소 가능한 테스트 MSE (점선).

Squares represent the training and test MSEs for the three fits shown in the left hand panel.

사각형은 왼쪽 패널에 표시된 세 가지 맞춤에 대한 훈련 및 테스트 MSE들을 나타냅니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

This seems like it might be a sensible approach, since the training MSE and the test MSE appear to be closely related.

훈련 MSE와 테스트 MSE가 밀접하게 관련되어 있는 것처럼 보이기 때문에 이는 합리적인 접근 방식인 것처럼 보입니다.

Unfortunately, there is a fundamental problem with this strategy:

불행하게도, 이 전략에는 근본적인 문제가 있습니다.

there is no guarantee that the method with the lowest training MSE will also have the lowest test MSE.

훈련 MSE가 가장 낮은 방법이 테스트 MSE도 가장 낮다는 보장이 없습니다.

Roughly speaking, the problem is that many statistical methods specifically estimate coefficients so as to minimize the training set MSE.

대략적으로 말하면, 문제는 많은 통계적 방법이 훈련 세트 MSE를 최소화하기 위해 특별히 계수를 추정한다는 것입니다.

For these methods, the training set MSE can be quite small, but the test MSE is often much larger.

이러한 방법의 경우, 훈련 세트 MSE는 매우 작을 수 있지만, 테스트 MSE는 종종 훨씬 큽니다.

Figure 2.9 illustrates this phenomenon on a simple example.

그림 2.9는 이러한 현상을 간단한 예로 보여줍니다.

In the left hand panel of Figure 2.9, we have generated observations from (2.1) with the true f given by the black curve.

그림 2.9의 왼쪽 패널에서, 우리는 검은색 곡선으로 주어진 실제 f를 사용하여 (2.1)에서 관측치을 생성했습니다.

The orange, blue and green curves illustrate three possible estimates for f obtained using methods with increasing levels of flexibility.

주황색, 파란색 및 녹색 곡선은 유연성 수준이 증가하는 방법을 사용하여 얻은 f에 대한 세 가지 가능한 추정치를 보여줍니다.

The orange line is the linear regression fit, which is relatively inflexible.

주황색 선은 상대적으로 융통성이 없는 선형 회귀 적합입니다.

The blue and green curves were produced using smoothing splines, discussed in Chapter 7, with different levels of smoothness.

파란색과 녹색 곡선은 7장에서 논의된 다양한 수준의 평활화 스플라인을 사용하여 생성되었습니다.

It is clear that as the level of flexibility increases, the curves fit the observed data more closely.

유연성 수준이 증가함에 따라, 곡선이 관찰된 데이터에 더 가깝게 맞는다는 것이 분명합니다.

The green curve is the most flexible and matches the data very well;

녹색 곡선은 가장 유연하며 데이터와 매우 잘 일치합니다.

however, we observe that it fits the true f (shown in black) poorly because it is too wiggly.

그러나, 우리는 그것이 너무 흔들리기 때문에 실제 f(검은색으로 표시됨)에 잘 맞지 않는다는 것을 관찰합니다.

By adjusting the level of flexibility of the smoothing spline fit, we can produce many different fits to this data.

평활화 스플라인 적합의 유연성 수준을 조정하여, 이 데이터에 대해 다양한 적합(?)을 생성할 수 있습니다.

We now move on to the right hand panel of Figure 2.9.

이제 그림 2.9의 오른쪽 패널로 이동합니다.

The grey curve displays the average training MSE as a function of flexibility, or more formally the degrees of freedom, for a number of smoothing splines.

회색 곡선은 다수의 평활화 스플라인에 대한 평균 훈련 MSE를 유연성의 함수, 또는 더 공식적으로는 자유도로 표시합니다.

The degrees of freedom is a quantity that summarizes the flexibility of a curve; it is discussed more fully in Chapter 7.

자유도는 곡선의 유연성을 요약하는 양입니다. 7장에서 더 자세히 논의됩니다.

The orange, blue and green squares indicate the MSEs associated with the corresponding curves in the left hand panel.

주황색, 파란색 및 녹색 사각형은 왼쪽 패널의 해당되는 곡선과 관련된 MSE들을 나타냅니다.

A more restricted and hence smoother curve has fewer degrees of freedom than a wiggly curve - note that in Figure 2.9, linear regression is at the most restrictive end, with two degrees of freedom.

더 제한적이고 부드러운 곡선은 흔들거리는 곡선보다 자유도가 적습니다. - 그림 2.9를 주목하면, 선형 회귀는 자유도가 2인 가장 제한적인 끝에 있습니다.

The training MSE declines monotonically as flexibility increases.

훈련 MSE는 유연성이 증가함에 따라 단조롭게 감소합니다.

In this example the true f is non-linear, and so the orange linear fit is not flexible enough to estimate f well.

이 예시에서 실제 f는 비선형이므로, 주황색 선형 맞춤은 f를 잘 추정할 만큼 충분히 유연하지 않습니다.

The green curve has the lowest training MSE of all three methods, since it corresponds to the most flexible of the three curves fit in the left hand panel.

녹색 곡선은 왼쪽 패널에 맞는 세 가지 곡선 중 가장 유연한 곡선에 해당하므로, 세 가지 방법 모두에서 훈련 MSE가 가장 낮습니다.

In this example, we know the true function f, and so we can also compute the test MSE over a very large test set, as a function of flexibility.

이 예시에서, 우리는 실제 함수 f를 알고 있으므로, 매우 큰 테스트 세트에 대한 테스트 MSE를 유연성 함수로 계산할 수도 있습니다.

(Of course, in general f is unknown, so this will not be possible.)

(물론, 일반적으로 f는 알 수 없으므로 불가능합니다.)

The test MSE is displayed using the red curve in the right hand panel of Figure 2.9.

테스트 MSE는 그림 2.9의 오른쪽 패널에 보이는 빨간색 곡선을 사용하여 표시됩니다.

As with the training MSE, the test MSE initially declines as the level of flexibility increases.

교육 MSE와 마찬가지로, 테스트 MSE는 유연성 수준이 증가함에 따라 초기에 감소합니다.

However, at some point the test MSE levels off and then starts to increase again.

그러나, 어느 시점에서 테스트 MSE는 평준화되었다가 다시 증가하기 시작합니다.

Consequently, the orange and green curves both have high test MSE.

결과적으로, 주황색 및 녹색 곡선 모두 테스트 MSE가 높습니다.

The blue curve minimizes the test MSE, which should not be surprising given that visually it appears to estimate f the best in the left hand panel of Figure 2.9.

파란색 곡선은 테스트 MSE를 최소화합니다. 시각적으로 그림 2.9의 왼쪽 패널에서 f를 가장 잘 추정하는 것으로 나타나므로 이는 놀라운 일이 아닙니다.

The horizontal dashed line indicates Var(ϵ), the irreducible error in (2.3), which corresponds to the lowest achievable test MSE among all possible methods.

수평 파선은 (2.3)에서 감소할 수 없는 오차인 Var(ϵ)를 나타내며, 가능한 모든 방법 중에서 달성 가능한 가장 낮은 테스트 MSE에 해당합니다.

Hence, the smoothing spline represented by the blue curve is close to optimal.

따라서, 파란색 곡선으로 표시된 평활화 스플라인이 최적에 가깝습니다.

In the right hand panel of Figure 2.9, as the flexibility of the statistical learning method increases, we observe a monotone decrease in the training MSE and a U-shape in the test MSE.

그림 2.9의 오른쪽 패널에서, 통계적 학습 방법의 유연성이 증가함에 따라, 우리는 훈련 MSE에서 단조 감소와 테스트 MSE에서 U자 모양을 관찰합니다.

This is a fundamental property of statistical learning that holds regardless of the particular data set at hand and regardless of the statistical method being used.

이것은 사용 중인 특정 데이터 세트와 사용 중인 통계 방법에 관계없이 유지되는 통계 학습의 기본 속성입니다.

As model flexibility increases, training MSE will decrease, but the test MSE may not.

모델 유연성이 증가함에 따라, 훈련 MSE는 감소하지만, 테스트 MSE는 그렇지 않을 수 있습니다.

When a given method yields a small training MSE but a large test MSE, we are said to be overfitting the data.

주어진 방법이 작은 학습 MSE를 생성하지만 큰 테스트 MSE를 생성하는 경우, 데이터를 과대적합한다고 합니다.

This happens because our statistical learning procedure is working too hard to find patterns in the training data, and may be picking up some patterns that are just caused by random chance rather than by true properties of the unknown function f.

이것은 우리의 통계적 학습 절차가 훈련 데이터에서 패턴을 찾기 위해 너무 열심히 노력하고 있고, 알려지지 않은 함수 f의 진정한 속성 보다는 무작위 우연(기회)에 의해 발생하는 일부 패턴을 선택하고 있을 수 있기 때문에 발생합니다.

When we overfit the training data, the test MSE will be very large because the supposed patterns that the method found in the training data simply don’t exist in the test data.

훈련 데이터를 과대적합하면, 훈련 데이터에서 발견한 방법이 테스트 데이터에 존재하지 않는 것으로 추정되는 패턴이 있기 때문에 테스트 MSE가 매우 커집니다.

Note that regardless of whether or not overfitting has occurred, we almost always expect the training MSE to be smaller than the test MSE because most statistical learning methods either directly or indirectly seek to minimize the training MSE.

참고로 과적합이 발생했는지 여부에 관계없이, 우리는 대부분의 통계적 학습 방법이 직간접적으로 훈련 MSE를 최소화하려고 하기 때문에 거의 항상 훈련 MSE가 테스트 MSE보다 작을 것으로 예상합니다.

Overfitting refers specifically to the case in which a less flexible model would have yielded a smaller test MSE.

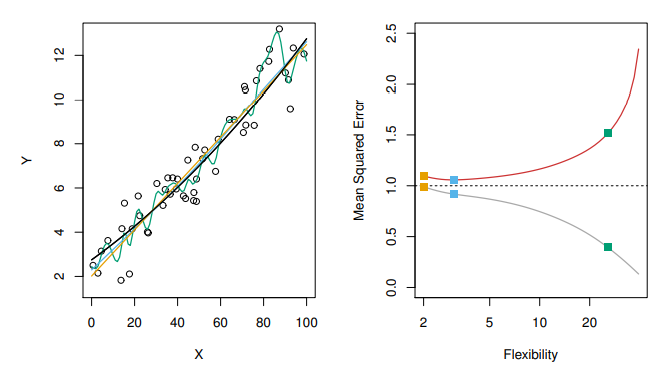
과적합은 특히 덜 유연한 모델이 더 작은 테스트 MSE를 생성했을 경우를 나타냅니다.

Figure 2.10 provides another example in which the true f is approximately linear.

그림 2.10은 실제 f가 대략 선형인 또 다른 예시를 제공합니다.

Again we observe that the training MSE decreases monotonically as the model flexibility increases, and that there is a U-shape the test MSE.

다시 우리는 모델 유연성이 증가함에 따라 훈련 MSE가 단조롭게 감소하고, 테스트 MSE의 U자 모양이 있음을 관찰합니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ[ FIGURE 2.10. 그림 2.10. ]ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

Details are as in Figure 2.9, using a different true f that is much closer to linear.

세부 사항은 그림 2.9와 같으며, 선형에 훨씬 더 가까운 다른 실제 f를 사용합니다.

In this setting, linear regression provides a very good fit to the data.

이 설정에서, 선형 회귀를 사용하면 데이터에 매우 적합합니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

However, because the truth is close to linear, the test MSE only decreases slightly before increasing again, so that the orange least squares fit is substantially better than the highly flexible green curve.

그러나 진실은 선형에 가깝기 때문에, 테스트 MSE는 다시 증가하기 전에 약간만 감소하므로, 주황색 최소 제곱 적합이 실질적으로 유연한 녹색 곡선보다 훨씬 낫습니다.

Finally, Figure 2.11 displays an example in which f is highly non linear.

마지막으로, 그림 2.11은 f가 매우 비선형인 예를 보여줍니다.

The training and test MSE curves still exhibit the same general patterns, but now there is a rapid decrease in both curves before the test MSE starts to increase slowly.

훈련 및 테스트 MSE 곡선은 여전히 동일한 일반 패턴을 나타내지만, 이제 테스트 MSE가 천천히 증가하기 시작하기 전에 두 곡선 모두에서 급격한 감소가 있습니다.

In practice, one can usually compute the training MSE with relative ease, but estimating test MSE is considerably more difficult because usually no test data are available.

실제로, 훈련 MSE는 보통 상대적으로 쉽게 계산할 수 있지만, 일반적으로 테스트 데이터를 사용할 수 없기 때문에 테스트 MSE를 추정하는 것은 훨씬 더 어렵습니다.

As the previous three examples illustrate, the flexibility level corresponding to the model with the minimal test MSE can vary considerably among data sets.

앞의 세 가지 예에서 알 수 있듯이, 최소 테스트 MSE가 있는 모델에 해당하는 유연성 수준은 데이터 세트 간에 상당히 다를 수 있습니다.

Throughout this book, we discuss a variety of approaches that can be used in practice to estimate this minimum point.

이 책 전반에 걸쳐, 우리는 이 최소점을 추정하기 위해 실제로 사용할 수 있는 다양한 접근법에 대해 논의(추론)합니다.

One important method is cross-validation (Chapter 5), which is a method for estimating test MSE using the training data.

중요한 방법 중 하나는 훈련 데이터를 사용하여 테스트 MSE를 추정하는 방법인 교차 검증 (Chapter 5)입니다.

**2.2.3 K-최근접 이웃 K-Nearest Neighbors ~ 2.3 전까지 (Lab 미실시) p39 ~ p42**

In theory, we would always like to predict qualitative responses using the Bayes classifier.

이론적으로, 우리는 항상 베이즈 분류기를 사용하여 질적 반응을 예측하고 싶습니다.

But for real data, we do not know the conditional distribution of Y given X, and so computing the Bayes classifier is impossible.

그러나 실제 데이터의 경우, 우리는 X가 주어진 Y의 조건부 분포를 모르기 때문에 베이즈 분류기를 계산하는 것은 불가능합니다.

Therefore, the Bayes classifier serves as an unattainable gold standard against which to compare other methods.

따라서, 베이즈 분류기는 다른 방법을 비교할 수 있는 달성 불가능한 금본위제 역할을 합니다.

Many approaches attempt to estimate the conditional distribution of Y given X, and then classify a given observation to the class with highest estimated probability.

많은 접근법은 X가 주어진 Y의 조건부 분포를 추정한 다음, 추정 확률이 가장 높은 클래스로 주어진 관측치를 분류하려고 시도합니다.

One such method is the K-nearest neighbors (KNN) classifier.

이러한 방법 중 하나는 KNN (K-nearest Neighbors) 분류기입니다.

Given a positive integer K and a test observation x0, the KNN classifier first identifies the K points in the training data that are closest to x0, represented by N0.

양의 정수 K와 테스트 관측치 x0이 주어지면, KNN 분류기는 N0으로 표시되는 x0에 가장 가까운 훈련 데이터에서 K 점들을 먼저 식별합니다.

It then estimates the conditional probability for class j as the fraction of points in N0 whose response values equal j:

텍스트이(가) 표시된 사진

자동 생성된 설명그런 다음 클래스 j에 대한 조건부 확률을 반응 값이 j와 같은 N0의 점의 부분으로 추정합니다:

Finally, KNN classifies the test observation x0 to the class with the largest probability from (2.12).

마지막으로, KNN은 테스트 관측치 x0을 (2.12)에서 가장 큰 확률을 갖는 클래스로 분류합니다.

Figure 2.14 provides an illustrative example of the KNN approach.

그림 2.14는 KNN 접근법의 예시를 제공합니다.

In the left-hand panel, we have plotted a small training data set consisting of six blue and six orange observations.

왼쪽 패널에서는, 6개의 파란색 및 6개의 주황색 관측치로 구성된 작은 훈련 데이터 세트를 구성했습니다.

Our goal is to make a prediction for the point labeled by the black cross. Suppose that we choose K = 3.

우리의 목표는 검은 십자가로 표시된 점에 대한 예측을 하는 것 입니다. K = 3을 선택한다고 가정합시다.

Then KNN will first identify the three observations that are closest to the cross.

그런 다음 KNN은 먼저 십자가에 가장 가까운 세 개의 관측치를 식별합니다.

This neighborhood is shown as a circle.

이 동네는 원으로 표시되어 있습니다.

It consists of two blue points and one orange point, resulting in estimated probabilities of 2/3 for the blue class and 1/3 for the orange class.

이것은 두 개의 파란색 점과 하나의 주황색 점으로 구성되어 있으며, 결과적으로 파란색 클래스의 경우 2/3, 주황색 클래스의 경우 1/3의 추정 확률이 생성됩니다.

Hence KNN will predict that the black cross belongs to the blue class.

따라서 KNN은 흑십자가 파란색 계열(클래스)에 속한다고 예측할 것입니다.

In the right-hand panel of Figure 2.14 we have applied the KNN approach with K = 3 at all of the possible values for X1 and X2, and have drawn in the corresponding KNN decision boundary.

그림 2.14의 오른쪽 패널에서 우리는 X1과 X2에 대해 가능한 모든 값에 K = 3으로 KNN 접근법을 적용하고, 해당(상응하는) KNN 결정 경계를 그렸습니다.

Despite the fact that it is a very simple approach, KNN can often produce classifiers that are surprisingly close to the optimal Bayes classifier.

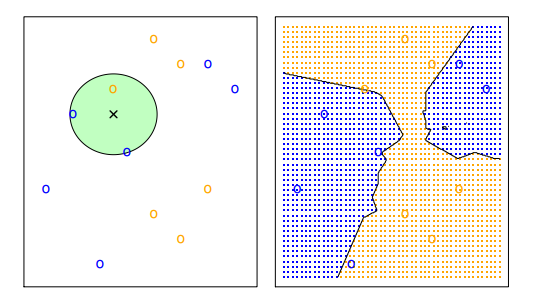
매우 간단한 접근 방식임에도 불구하고, KNN은 종종 최적의 베이즈 분류기에 놀라울 정도로 가까운 분류기를 생성할 수 있습니다.

Figure 2.15 displays the KNN decision boundary, using K = 10, when applied to the larger simulated data set from Figure 2.13.

그림 2.15는 그림 2.13의 더 큰 시뮬레이션 데이터 세트에 적용할 때, K = 10을 사용하여 KNN 결정 경계를 표시합니다.

Notice that even though the true distribution is not known by the KNN classifier, the KNN decision boundary is very close to that of the Bayes classifier.

그래도 실제 분포는 KNN 분류기에 의해 알려져 있지 않지만, KNN 결정 경계는 베이즈 분류기의 그것과 매우 가깝다는 점에 유의해야 합니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ[ FIGURE 2.14. 그림 2.14 ]ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

The KNN approach, using K = 3, is illustrated in a simple situation with six blue observations and six orange observations.

K = 3을 사용하는 KNN 접근법은 6개의 파란색 관측치와 6개의 주황색 관측치가 있는 간단한 상황에서 설명됩니다.

Left: a test observation at which a predicted class label is desired is shown as a black cross.

왼쪽: 예측 클래스 레이블을 원하는 테스트 관측치가 테스트 십자형으로 표시됩니다.

The three closest points to the test observation are identified, and it is predicted that the test observation belongs to the most commonly occurring class, in this case blue.

테스트 관측치에 가장 가까운 세 개의 점이 식별되며, 테스트 관측치가 가장 흔하게 발생하는 클래스에 속할 것으로 예측되며, 이 경우에는 파란색입니다.

Right: The KNN decision boundary for this example is shown in black.

오른쪽: 이 예에 대한 KNN 결정 경계는 검은색으로 표시됩니다.

The blue grid indicates the region in which a test observation will be assigned to the blue class, and the orange grid indicates the region in which it will be assigned to the orange class.

파란색 그리드는 테스트 관찰이 파란색 클래스에 할당될 영역을 나타내고, 주황색 그리드는 테스트 관찰이 주황색 클래스에 할당될 영역을 나타냅니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ[ FIGURE 2.15. 그림 2.15. ]ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

지도이(가) 표시된 사진

자동 생성된 설명

The black curve indicates the KNN decision boundary on the data from Figure 2.13, using K = 10.

검은색 곡선은 K = 10을 사용하여 그림 2.13의 데이터에 대한 KNN 결정 경계를 나타냅니다.

The Bayes decision boundary is shown as a purple dashed line.

베이즈 결정 경계는 보라색 점선으로 표시됩니다.

The KNN and Bayes decision boundaries are very similar.

KNN과 베이즈의 결정 경계는 매우 유사합니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ[ FIGURE 2.16. 그림 2.16. ] ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

지도이(가) 표시된 사진

자동 생성된 설명

A comparison of the KNN decision boundaries (solid black curves) obtained using K = 1 and K = 100 on the data from Figure 2.13.

그림 2.13의 데이터에서 K = 1 및 K = 100을 사용하여 얻은 KNN 결정 경계(단단한(?) 검정색 곡선)를 비교합시다.

With K = 1, the decision boundary is overly flexible, while with K = 100 it is not sufficiently flexible.

K = 1인 경우에는, 결정 경계가 지나치게 유연한 반면, K = 100인 경우에는 충분히 유연하지 않습니다.

The Bayes decision boundary is shown as a purple dashed line.

베이즈 결정 경계는 보라색 점선으로 표시됩니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

The test error rate using KNN is 0.1363, which is close to the Bayes error rate of 0.1304.

KNN을 사용한 테스트 오류율은 0.1363으로, 0.1304의 Bayes 오류율에 가깝습니다.

The choice of K has a drastic effect on the KNN classifier obtained.

K개의 선택은 획득한 KNN 분류기에 급격한 영향을 미칩니다.

Figure 2.16 displays two KNN fits to the simulated data from Figure 2.13, using K = 1 and K = 100.

그림 2.16은 K=1 및 K=100을 사용하여, 그림 2.13의 시뮬레이션 데이터에 대한 두 개의 KNN 적합치를 표시합니다.

When K = 1, the decision boundary is overly flexible and finds patterns in the data that don’t correspond to the Bayes decision boundary.

K = 1인 경우, 결정 경계는 지나치게 유연하며 데이터에서 베이즈 결정 경계와 일치하지 않는 패턴을 찾습니다.

This corresponds to a classifier that has low bias but very high variance.

이는 편향은 낮지만 분산이 매우 높은 분류기에 해당합니다.

As K grows, the method becomes less flexible and produces a decision boundary that is close to linear.

K가 성장함에 따라, 방법은 유연성이 떨어지고 선형에 가까운 결정 경계를 생성합니다.

This corresponds to a low-variance but high-bias classifier.

이는 작은 분산이지만 높은 편향 분류기에 해당합니다.

On this simulated data set, neither K = 1 nor K = 100 give good predictions: they have test error rates of 0.1695 and 0.1925, respectively.

이 시뮬레이션된 데이터 집합에서, K = 1 또는 K = 100은 각각 0.1695 및 0.189의 검정 오류율을 가지고 있으므로 좋은 예측을 제공하지 않습니다.

Just as in the regression setting, there is not a strong relationship between the training error rate and the test error rate.

회귀 설정에서와 마찬가지로, 훈련 오류율과 검사 오류율 사이에는 강력한 관계가 없습니다.

With K = 1, the KNN training error rate is 0, but the test error rate may be quite high.

K = 1인 경우, KNN 훈련 오류율은 0이지만, 테스트 오류율이 상당히 높을 수 있습니다.

In general, as we use more flexible classification methods, the training error rate will decline but the test error rate may not.

일반적으로, 우리는 보다 유연한 분류 방법을 사용함에 따라, 훈련 오류율은 감소하지만 테스트 오류율은 감소하지 않을 수 있습니다.

In Figure 2.17, we have plotted the KNN test and training errors as a function of 1/K.

그림 2.17에서, 우리는 1/K의 함수로 KNN 테스트 및 훈련 오류를 표시했습니다.

As 1/K increases, the method becomes more flexible.

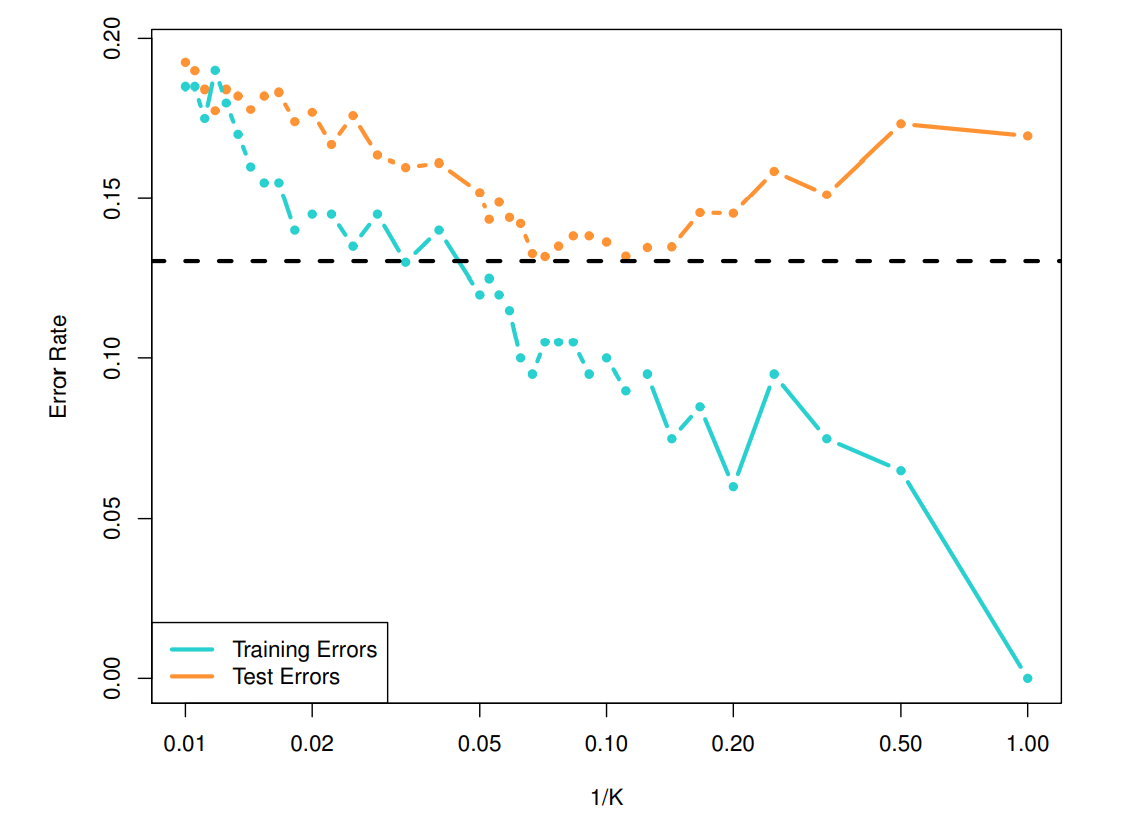
1/K가 증가할수록, 방법이 더 유연해집니다.

As in the regression setting, the training error rate consistently declines as the flexibility increases.

회귀 설정에서와 마찬가지로, 훈련 오류율은 유연성이 증가함에 따라 지속적으로 감소합니다.

However, the test error exhibits a characteristic U-shape, declining at first (with a minimum at approximately K = 10) before increasing again when the method becomes excessively flexible and overfits.

그러나, 테스트 오류는 U자형의 특성(특성)을 나타내며, 방법이 지나치게 유연해지고 과적합될 때 다시 증가하기 전에 처음에는 감소합니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ[ FIGURE 2.17. 그림 2.17. ] ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

The KNN training error rate (blue, 200 observations) and test error rate (orange, 5,000 observations) on the data from Figure 2.13, as the level of flexibility (assessed using 1/K on the log scale) increases, or equivalently as the number of neighbors K decreases.

그림 2.13의 데이터에 대한 KNN 훈련 오류율 (파란색, 관측치 200개)과 테스트 오류율 (주황색, 관측치 5,000개)은 유연성 수준 (로그 척도의 1/K를 사용하여 평가)이 증가하거나 동등하게 이웃 K의 수가 감소함에 따라 나타납니다.

The black dashed line indicates the Bayes error rate.

검은색 점선은 베이즈 오류율을 나타냅니다.

The jumpiness of the curves is due to the small size of the training data set.

곡선의 점프력은 훈련 데이터 세트의 작은 크기 때문입니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

In both the regression and classification settings, choosing the correct level of flexibility is critical to the success of any statistical learning method.

회귀 및 분류 설정 모두에서, 정확한 유연성 수준을 선택하는 것은 모든 통계 학습 방법의 성공에 중요합니다.

The bias-variance tradeoff, and the resulting U-shape in the test error, can make this a difficult task.

치우침-분산 거래와 테스트 오류의 결과 U자형은 이것을 어려운 작업으로 만들 수 있습니다.

In Chapter 5, we return to this topic and discuss various methods for estimating test error rates and thereby choosing the optimal level of flexibility for a given statistical learning method.

5장에서는, 우리는 이 주제로 돌아가 테스트 오류율을 추정하고 주어진 통계 학습 방법에 대한 최적의 유연성 수준을 선택하기 위한 다양한 방법에 대해 논의합니다.

**3.1.3 모델의 정확성 평가 Assessing the Accuracy of the Model ~ 3.2 p68 ~ p72**

Once we have rejected the null hypothesis (3.12) in favor of the alternative hypothesis (3.13), it is natural to want to quantify the extent to which the model fits the data.

귀무 가설(3.12)을 대립 가설(3.13)에 유리하도록 기각했으면, 모형이 데이터를 적합시키는 정도를 정량화하려는 것은 당연합니다.

The quality of a linear regression fit is typically assessed using two related quantities:

선형 회귀 적합선의 품질은 일반적으로 두 가지 관련 양을 사용하여 평가됩니다:

the residual standard error (RSE) and the R2 statistic. \* R2 = R제곱

잔차 표준 오차 (RSE) 및 R2 통계량입니다.

Table 3.2 displays the RSE, the R2 statistic, and the F-statistic (to be described in Section 3.2.2) for the linear regression of number of units sold on TV advertising budget.

표 3.2는 TV 광고 예산에서 판매된 유닛 수의 선형 회귀에 대한 RSE, R2 통계량 및 F-통계 (섹션 3.2.2에서 설명)를 보여줍니다.

**Residual Standard Error 잔차 표준 오류**

Recall from the model (3.5) that associated with each observation is an error term ϵ.

모형(3.5)에서 각 관측치와 관련된 오차항 ϵ을 떠올립니다.

Due to the presence of these error terms, even if we knew the true regression line (i.e. even if β0 and β1 were known), we would not be able to perfectly predict Y from X.

이러한 오차항의 존재로 인해 실제 회귀선을 알고 있더라도 (즉, β0과 β1이 알려져 있더라도), X에서 Y를 완벽하게 예측할 수 없을 것입니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ[ TABLE 3.2. 표 3.2. ] ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

테이블이(가) 표시된 사진

자동 생성된 설명

For the Advertising data, more information about the least squares model for the regression of number of units sold on TV advertising budget.

광고 데이터의 경우, TV 광고 예산에서 판매된 단위 수의 회귀 분석에 대한 최소 제곱 모형에 대한 자세한 내용입니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

The RSE is an estimate of the standard deviation of ϵ.

RSE는 ϵ의 표준 편차에 대한 추정치입니다.

Roughly speaking, it is the average amount that the response will deviate from the true regression line.

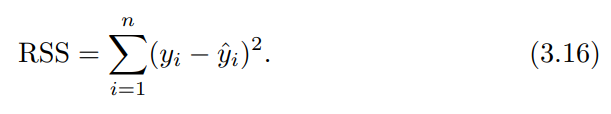
대략적으로 말하면, 반응이 실제 회귀선에서 벗어나는 평균 양입니다.

It is computed using the formula (3.15).

텍스트, 시계이(가) 표시된 사진

자동 생성된 설명그것은 공식 (3.15)를 사용하여 계산됩니다.

Note that RSS was defined in Section 3.1.1, and is given by the formula (3.16).

RSS는 섹션 3.1.1에서 정의되었으며, 공식(3.16)으로 주어집니다.

In the case of the advertising data, we see from the linear regression output in Table 3.2 that the RSE is 3.26.

광고 데이터의 경우 표 3.2의 선형 회귀 출력에서 RSE가 3.26임을 알 수 있습니다.

In other words, actual sales in each market deviate from the true regression line by approximately 3,260 units, on average.

다시 말해, 각 시장의 실제 판매량이 실제 회귀선에서 평균 약 3260대 이탈한 셈입니다.

Another way to think about this is that even if the model were correct and the true values of the unknown coefficients β0 and β1 were known exactly, any prediction of sales on the basis of TV advertising would still be off by about 3,260 units on average.

이에 대해 생각할 수 있는 또 다른 방법은 모델이 정확하고 알려지지 않은 계수 β0과 β1의 실제 값이 정확하게 알려져 있더라도 TV 광고를 기반으로 한 판매 예측은 평균적으로 약 3,260대가 여전히 빗나간다는 것이다.

Of course, whether or not 3,260 units is an acceptable prediction error depends on the problem context.

물론, 3,260 단위가 허용 가능한 예측 오차인지 아닌지 여부는 문제의 맥락에 따라 달라집니다.

In the advertising data set, the mean value of sales over all markets is approximately 14,000 units, and so the percentage error is 3,260/14,000 = 23 %.

광고 데이터 집합에서, 모든 시장의 판매 평균 값은 약 14,000개이므로, 백분율 오차는 3,260/14,000 = 23%입니다.

The RSE is considered a measure of the lack of fit of the model (3.5) to the data.

RSE는 데이터에 대한 모형(3.5)의 적합성 결여에 대한 측도로 간주(고려)됩니다.

If the predictions obtained using the model are very close to the true outcome values — that is, if ˆyi ≈ yi for i = 1,...,n — then (3.15) will be small, and we can conclude that the model fits the data very well.

모형을 사용하여 얻은 예측값이 실제 결과값에 매우 가깝다면 — 즉, i = 1, ..., n에 대해 ˆyi ≈ yi이면 — (3.15)은 작으며, 모형이 데이터를 매우 잘 적합한다는 결론을 내릴 수 있습니다.

On the other hand, if ˆyi is very far from yi for one or more observations, then the RSE may be quite large, indicating that the model doesn’t fit the data well.

반면에, 하나 이상의 관측치에서 ˆyi가 yi에서 매우 멀리 떨어져 있으면, RSE가 상당히 클 수 있으므로 모형이 데이터를 잘 적합하지 않음을 나타냅니다.

**R2 Statistic R2 통계 (R의 제곱)**

The RSE provides an absolute measure of lack of fit of the model (3.5) to the data.

RSE는 데이터에 대한 모형(3.5)의 적합성 결여에 대한 절대 측도를 제공합니다.

But since it is measured in the units of Y , it is not always clear what constitutes a good RSE.

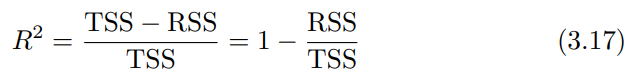
그러나 Y 단위로 측정되기 때문에, 좋은 RSE를 구성하는 것이 항상 명확하지는 않습니다.

The R2 statistic provides an alternative measure of fit.

R2 통계량은 적합성의 대안적인 측도를 제공합니다.

It takes the form of a proportion — the proportion of variance explained—and so it always takes on a value between 0 and 1, and is independent of the scale of Y .

텍스트이(가) 표시된 사진

자동 생성된 설명이것은 비율 (설명된 분산의 비율) 의 형태를 취하므로 항상 0과 1 사이의 값을 가지며, Y의 척도와는 독립적입니다.

To calculate R2, we use the formula (3.17) where [TSS 공식] is the total sum of squares, and RSS is defined in (3.16).

우리는 R2를 계산하기 위해, 공식 (3.17)을 사용하며 여기서 [TSS 공식]은 총 제곱합이고 RSS는 (3.16)에 정의됩니다.

TSS measures the total variance in the response Y , and can be thought of as the amount of variability inherent in the response before the regression is performed.

TSS는 반응 Y의 총 분산을 측정하며, 회귀 분석을 수행하기 전에 반응에 내재된 변동성의 양으로 간주할 수 있습니다.

In contrast, RSS measures the amount of variability that is left unexplained after performing the regression.

반대로, RSS는 회귀 분석을 수행한 후 설명되지 않은 변동성의 양을 측정합니다.

Hence, TSS - RSS measures the amount of variability in the response that is explained (or removed) by performing the regression, and R2 measures the proportion of variability in Y that can be explained using X.

따라서, TSS - RSS는 회귀 분석을 수행하여 설명된 (또는 제거된) 반응의 변동성 양을 측정하고, R2는 X를 사용하여 설명할 수 있는 Y의 변동성 비율을 측정합니다.

An R2 statistic that is close to 1 indicates that a large proportion of the variability in the response is explained by the regression.

R2 통계량이 1에 가까우면 반응 변수의 상당 부분이 회귀 분석에 의해 설명된다는 것을 나타냅니다.

A number near 0 indicates that the regression does not explain much of the variability in the response;

숫자가 0에 가까우면 회귀 분석에서 반응의 변동성을 많이 설명하지 못한다는 것을 나타냅니다;

this might occur because the linear model is wrong, or the error variance σ2 is high, or both.

이 문제는 선형 모형이 잘못되었거나 오차 분산 σ2가 높거나 같기 때문에 발생할 수 있습니다.

In Table 3.2, the R2 was 0.61, and so just under two-thirds of the variability in sales is explained by a linear regression on TV.

표 3.2에서, R2는 0.61이었고, 따라서 매출 변동의 2/3 미만은 TV의 선형 회귀로 설명됩니다.

The R2 statistic (3.17) has an interpretational advantage over the RSE (3.15), since unlike the RSE, it always lies between 0 and 1.

R2 통계량(3.17)은 RSE와 달리 항상 0과 1 사이에 있기 때문에 RSE(3.15)에 비해 해석상 이점이 있습니다.

However, it can still be challenging to determine what is a good R2 value, and in general, this will depend on the application.

그러나, 좋은 R2 값이 무엇인지 결정하는 것은 여전히 어려울 수 있으며, 일반적으로, 이것은 애플리케이션에 따라 달라질 것입니다.

For instance, in certain problems in physics, we may know that the data truly comes from a linear model with a small residual error.

예를 들어, 물리학의 특정 문제에서, 우리는 데이터가 작은 잔차 오차가 있는 선형 모델에서 진정으로 나온다는 것을 알 수 있습니다.

In this case, we would expect to see an R2 value that is extremely close to 1, and a substantially smaller R2 value might indicate a serious problem with the experiment in which the data were generated.

이 경우, 1에 매우 가까운 R2 값이 나타날 것으로 예상되며, R2 값이 상당히 작으면 데이터가 생성된 실험에 심각한 문제가 나타날 수 있습니다.

On the other hand, in typical applications in biology, psychology, marketing, and other domains, the linear model (3.5) is at best an extremely rough approximation to the data, and residual errors due to other unmeasured factors are often very large.

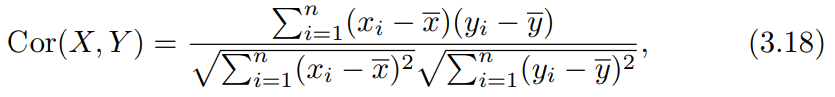
반면, 생물학, 심리학, 마케팅 및 기타 영역의 일반적인 응용에서 선형 모델(3.5)은 기껏해야 데이터에 대한 극히 대략적인 근사치이며, 다른 측정되지 않은 요인으로 인한 잔류 오차가 매우 큰 경우가 많습니다.

In this setting, we would expect only a very small proportion of the variance in the response to be explained by the predictor, and an R2 value well below 0.1 might be more realistic!

이 설정에서는, 우리는 예측 변수에 의해 설명되는 반응의 분산 중 매우 작은 부분만 예상하며, 0.1보다 훨씬 작은 R2 값이 더 현실적일 수 있습니다!

The R2 statistic is a measure of the linear relationship between X and Y .

R2 통계량은 X와 Y 사이의 선형 관계에 대한 측도입니다.



Recall that correlation, defined as (3.18) is also a measure of the linear relationship between X and Y .

(3.18)로 정의된 상관관계는 또한 X와 Y 사이의 선형 관계에 대한 측도라고 불러집니다.

This suggests that we might be able to use r = Cor(X, Y ) instead of R2 in order to assess the fit of the linear model.

이것은 우리가 선형 모델의 적합성을 평가하기 위해 R2 대신 r = Cor(X, Y )를 사용할 수 있음을 제안합니다.

In fact, it can be shown that in the simple linear regression setting, R2 = r2.

실제로, 단순 선형 회귀 설정에서 R2 = r2임을 알 수 있습니다.

In other words, the squared correlation and the R2 statistic are identical.

즉(다시 말해), 상관 제곱과 R2 통계량이 동일합니다.

However, in the next section we will discuss the multiple linear regression problem, in which we use several predictors simultaneously to predict the response.

그러나, 다음 섹션에서 우리는 여러 예측 변수를 동시에 사용하여 반응을 예측하는 다중 선형 회귀 문제에 대해 논의(토론, 설명)합니다.

The concept of correlation between the predictors and the response does not extend automatically to this setting, since correlation quantifies the association between a single pair of variables rather than between a larger number of variables.

예측 변수와 반응 변수 간의 상관 관계는 상관 관계가 더 많은 변수 간의 연관성이 아니라 단일 변수 쌍 간의 연관성을 정량화하기 때문에 이 설정으로 자동 확장되지 않습니다.

We will see that R2 fills this role.

우리는 R2가 이 역할을 수행하는 것을 볼 것입니다.

**3.2 Multiple Linear Regression 3.2 다중 선형 회귀 분석**

Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable.

단순 선형 회귀 분석은 단일 예측 변수를 기반으로 반응을 예측하는 데 유용한 방법입니다.

However, in practice we often have more than one predictor.

그러나, 실제로는 예측 변수가 두 개 이상인 경우가 많습니다.

For example, in the Advertising data, we have examined the relationship between sales and TV advertising.

예를 들어, 광고 데이터에서, 우리는 매출과 TV 광고의 관계를 조사했습니다.

We also have data for the amount of money spent advertising on the radio and in newspapers, and we may want to know whether either of these two media is associated with sales.

우리는 또한 라디오와 신문 광고에 사용된 금액에 대한 데이터가 있으며, 이 두 매체 중 하나가 판매와 관련이 있는지 여부를 알고 싶을 수도 있습니다.

How can we extend our analysis of the advertising data in order to accommodate these two additional predictors?

이 두 가지 추가 예측 변수를 수용하기 위해 광고 데이터 분석을 어떻게 확장할 수 있을까요?

One option is to run three separate simple linear regressions, each of which uses a different advertising medium as a predictor.

한 가지 옵션은 각각 다른 광고 매체를 예측 변수로 사용하는 세 개의 개별 단순 선형 회귀 분석을 실행하는 것입니다.

For instance, we can fit a simple linear regression to predict sales on the basis of the amount spent on radio advertisements.

예를 들어, 우리는 라디오 광고에 지출된 양(금액)을 기준으로 매출을 예측하기 위해 간단한 선형 회귀 분석을 적합 시킬 수 있습니다.

Results are shown in Table 3.3 (top table).

결과는 표 3.3 (상단 표)과 같습니다.

We find that a $1,000 increase in spending on radio advertising is associated with an increase in sales of around 203 units.

우리는 라디오 광고에 대한 1,000달러의 지출 증가가 약 203대의 판매 증가와 관련이 있다는 것을 발견했습니다.

Table 3.3 (bottom table) contains the least squares coefficients for a simple linear regression of sales onto newspaper advertising budget.

표 3.3 (아래 표)에는 신문 광고 예산에 대한 판매의 단순 선형 회귀에 대한 최소 제곱 계수가 포함되어 있습니다.

A $1,000 increase in newspaper advertising budget is associated with an increase in sales of approximately 55 units.

신문 광고 예산의 1,000달러 증가는 약 55개의 판매 증가와 관련이 있습니다.

However, the approach of fitting a separate simple linear regression model for each predictor is not entirely satisfactory.

그러나, 각 예측 변수에 대해 별도의 단순 선형 회귀 모형을 적합시키는 접근 방식은 완전히 만족스럽지는 않습니다.

First of all, it is unclear how to make a single prediction of sales given the three advertising media budgets, since each of the budgets is associated with a separate regression equation.

우선, 세 개의 광고매체 예산을 고려할 때 각각의 예산은 별도의 회귀 방정식과 연관되어 있기 때문에 어떻게 하나의 매출 예측을 할 수 있을지가 불분명합니다.

Second, each of the three regression equations ignores the other two media in forming estimates for the regression coefficients.

둘째, 세 개의 회귀 방정식 각각은 회귀 계수에 대한 추정치를 형성할 때 다른 두 매체를 무시합니다.

We will see shortly that if the media budgets are correlated with each other in the 200 markets in our data set, then this can lead to very misleading estimates of the association between each media budget and sales.

우리는 만약 우리의 데이터 세트의 200개 시장에서 미디어 예산이 서로 상관되어 있다면, 이는 각 미디어 예산과 매출 간의 연관성에 대한 매우 오해의 소지가 있는 추정치로 이어질 수 있음을 곧 알게 될 것입니다.

Instead of fitting a separate simple linear regression model for each predictor, a better approach is to extend the simple linear regression model (3.5) so that it can directly accommodate multiple predictors.

각 예측 변수에 대해 별도의 단순 선형 회귀 모형을 적합시키는 대신, 단순 선형 회귀 모형(3.5)을 확장하여 여러 예측 변수를 직접 수용할 수 있도록 하는 것이 더 좋은 방법입니다.

We can do this by giving each predictor a separate slope coefficient in a single model.

우리는 단일 모형에서 각 예측 변수에 별도의 기울기 계수를 제공하여 이 작업을 수행할 수 있습니다.

In general, suppose that we have p distinct predictors.

일반적으로, 뚜렷한 예측 변수 P가 있다고 가정합니다.

테이블이(가) 표시된 사진

자동 생성된 설명ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ[ TABLE 3.3. 표 3.3. ] ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

More simple linear regression models for the Advertising data.

광고 데이터에 대한 보다 단순한 선형 회귀 모형입니다.

Coefficients of the simple linear regression model for number of units sold on Top: radio advertising budget and Bottom: newspaper advertising budget.

상단: 라디오 광고 예산 and 하단: 신문 광고 예산에서 판매된 유닛 수에 대한 단순 선형 회귀 모델의 계수.

A $1,000 increase in spending on radio advertising is associated with an average increase in sales by around 203 units, while the same increase in spending on newspaper advertising is associated with an average increase in sales by around 55 units.

라디오 광고에 대한 지출이 1,000달러 증가하는 것은 약 203대의 평균적인 판매 증가와 관련이 있는 반면, 신문 광고에 대한 지출의 동일한 증가는 약 55대의 평균적인 판매 증가와 관련이 있습니다.

(Note that the sales variable is in thousands of units, and the radio and newspaper variables are in thousands of dollars.)

(참고로 판매 변수는 수천 단위, 라디오 및 신문 변수는 수천 달러 단위입니다.)

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

Then the multiple linear regression model takes the form (3.19) where Xj represents the jth predictor and βj quantifies the association between that variable and the response.

그런 다음 다중 선형 회귀 모형은 (3.19) 형식을 취하며, 여기서 Xj는 j번째 예측 변수를 나타내고 βj는 해당 변수와 반응 사이의 연관성을 정량화합니다.

We interpret βj as the average effect on Y of a one unit increase in Xj , holding all other predictors fixed.

우리는 βj를 다른 모든 예측 변수를 고정한 채, Xj의 한 단위 증가의 Y에 대한 평균 효과로 해석합니다.

In the advertising example, (3.19) becomes (3.20).

광고 예제에서, (3.19)는 (3.20)이 됩니다.

**2.1.2 How Do We Estimate f?의 Non-Parametric Methods**

**~ 2.1.3 The Trade-Off Between Prediction Accuracy and Model Interpretability p23 ~ p26**

**Non-Parametric Methods 비모수적 방법**

Non-parametric methods do not make explicit assumptions about the functional form of f.

비모수적 방법은 f의 함수 형식에 대해 명시적인 가정을 하지 않습니다.

Instead they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly.

대신 그들은 너무 거칠거나 흔들리지 않고 가능한 한 데이터 포인트에 가까워지는 f의 추정치를 찾습니다.

Such approaches can have a major advantage over parametric approaches:

이러한 접근 방식은 모수적 접근 방식에 비해 큰 이점을 가질 수 있습니다:

by avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f.

f에 대한 특정 함수 형식의 가정을 피함으로써, 그들은 f에 대해 더 넓은 범위의 가능한 모양을 정확하게 맞출 수 있는 잠재력이 있습니다.

Any parametric approach brings with it the possibility that the functional form used to estimate f is very different from the true f, in which case the resulting model will not fit the data well.

모든 모수적 접근 방식은 f를 추정하는 데 사용되는 함수 형식이 실제 f와 매우 다를 가능성이 있으며, 이 경우 결과 모델이 데이터에 잘 맞지 않을 것입니다.

In contrast, non-parametric approaches completely avoid this danger, since essentially no assumption about the form of f is made.

대조적으로, 비모수적 접근법은 본질적으로 f의 형식에 대한 가정이 이루어지지 않기 때문에 이러한 위험을 완전히 피합니다.

But non-parametric approaches do suffer from a major disadvantage:

그러나 비모수적 접근 방식에는 큰 단점이 있습니다:

since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

그들은 f를 추정하는 문제를 적은 수의 매개변수로 줄이지 않기 때문에, f에 대한 정확한 추정치를 얻기 위해서는 매우 많은 수의 관찰 (모수적 접근 방식에 일반적으로 필요한 것보다 훨씬 더 많은 수)이 필요합니다.

An example of a non-parametric approach to fitting the Income data is shown in Figure 2.5.

소득 데이터를 맞추는 비모수적 접근 방식의 예가 그림 2.5에 나와 있습니다.

A thin-plate spline is used to estimate f.

박판 스플라인(TPS)은 f를 추정하는 데 사용됩니다.

This approach does not impose any prespecified model on f.

이 접근법은 사전 지정된 모델을 f에 적용하지 않습니다.

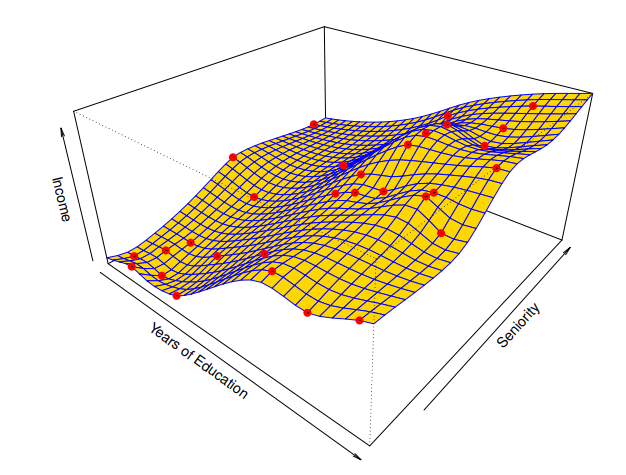
It instead attempts to produce an estimate for f that is as close as possible to the observed data, subject to the fit

대신 맞춤(fit)에 따라 관찰된 데이터에 가능한 한 가까운 f에 대한 추정치를 생성하려고 시도합니다.

that is, the yellow surface in Figure 2.5—being smooth.

즉, 그림 2.5의 매끄러운(부드러운) 노란색 표면입니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ[ Figure 그림 2.6. ] ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ



A rough thin-plate spline fit to the Income data from Figure 2.3.

그림 2.3의 Income 데이터에 알맞는(적합한) 대략적인 박판 스플라인(TPS).

This fit makes zero errors on the training data.

이 맞춤(fit)은 훈련 데이터에서 오류를 전혀 발생시키지 않습니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

In this case, the non-parametric fit has produced a remarkably accurate estimate of the true f shown in Figure 2.3.

이러한 경우, 비모수적 맞춤(fit)은 그림 2.3에 표시된 실제 f의 매우 정확한 추정치를 생성했습니다.

In order to fit a thin-plate spline, the data analyst must select a level of smoothness.

박판 스플라인(TPS)을 맞추려면, 데이터 분석가가 평활도 수준을 선택해야 합니다.

Figure 2.6 shows the same thin-plate spline fit using a lower level of smoothness, allowing for a rougher fit.

그림 2.6은 더 낮은 수준의 평활도를 사용하여 더 거친 맞춤(fit)을 허용하는 동일한 박판스플라인(TPS) 맞춤(fit)을 보여줍니다.

The resulting estimate fits the observed data perfectly!

결과 추정치는 관찰된 데이터와 완벽하게 일치합니다!

However, the spline fit shown in Figure 2.6 is far more variable than the true function f, from Figure 2.3.

그러나, 그림 2.6에 표시된 스플라인 맞춤(fit)은 그림 2.3의 실제 함수 f보다 훨씬 가변적입니다.

This is an example of overfitting the data, which we discussed previously.

이것은 이전에 논의한 데이터 과적합의 예입니다.

It is an undesirable situation because the fit obtained will not yield accurate estimates of the response on new observations that were not part of the original training data set.

얻어낸 맞춤(fit)이 원래 훈련 데이터 세트의 일부가 아닌 새로운 관찰에 대한 응답의 정확한 추정치를 산출하지 못하기 때문에 바람직하지 않은 상황입니다.

We discuss methods for choosing the correct amount of smoothness in Chapter 5.

우리는 5장에서 올바른 평활도를 선택하는 방법에 대해 설명합니다.

Splines are discussed in Chapter 7.

스플라인은 7장에서 설명합니다.

As we have seen, there are advantages and disadvantages to parametric and non-parametric methods for statistical learning.

우리가 본 바와 같이, 통계 학습을 위한 모수적 및 비모수적 방법에는 장점과 단점이 있습니다.

We explore both types of methods throughout this book.

우리는 이 책 전체에서 두 가지 유형의 방법을 탐구합니다.

**2.1.3 The Trade-Off Between Prediction Accuracy and Model Interpretability**

**2.1.3 예측 정확도와 모델 해석 가능성 간의 트레이드오프(균형)**

Of the many methods that we examine in this book, some are less flexible, or more restrictive, in the sense that they can produce just a relatively small range of shapes to estimate f.

우리는 이 책에서 검토하는 많은 방법 중 일부는 f를 추정하기 위해 상대적으로 작은 범위의 모양만 생성할 수 있다는 점에서 덜 유연하거나 더 제한적입니다.

For example, linear regression is a relatively inflexible approach, because it can only generate linear functions such as the lines shown in Figure 2.1 or the plane shown in Figure 2.4.

예를 들어 선형 회귀는 그림 2.1에 표시된 선이나 그림 2.4에 표시된 평면과 같은 선형 함수만 생성할 수 있기 때문에 상대적으로 융통성이 없는 접근 방식입니다.

Other methods, such as the thin-plate splines shown in Figures 2.5 and 2.6, are considerably more flexible because they can generate a much wider range of possible shapes to estimate f.

그림 2.5 및 2.6에 표시된 박판 스플라인(TPS)과 같은 다른 방법은, f를 추정하기 위해 훨씬 더 넓은 범위의 가능한 모양을 생성할 수 있기 때문에 훨씬 더 유연합니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ[ Figure 그림 2.7. ] ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

텍스트이(가) 표시된 사진

자동 생성된 설명

A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods.

다양한 통계적 학습 방법을 사용하여 유연성과 해석 가능성 사이의 트레이드오프(균형)를 나타냅니다.

In general, as the flexibility of a method increases, its interpretability decreases.

일반적으로, 방법의 유연성이 증가하면 해석 가능성이 감소합니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

One might reasonably ask the following question:

합리적으로 다음과 같은 질문을 할 수 있습니다:

why would we ever choose to use a more restrictive method instead of a very flexible approach?

왜 우리는 매우 유연한 접근 방식 대신 더 제한적인 방법을 사용할까요?

There are several reasons that we might prefer a more restrictive model.

보다 제한적인 모델을 선호하는 몇 가지 이유가 있습니다.

If we are mainly interested in inference, then restrictive models are much more interpretable.

만약 우리가 주로 추론에 관심이 있다면, 제한적 모델이 훨씬 더 해석하기 쉽습니다.

For instance, when inference is the goal, the linear model may be a good choice since it will be quite easy to understand the relationship between **Y** and **X1, X2,...,Xp**.

예를 들어, 추론이 목표인 경우, **Y**와 **X1, X2,...,Xp** 사이의 관계를 이해하기가 매우 쉽기 때문에 선형 모델이 좋은 선택일 수 있습니다.

In contrast, very flexible approaches, such as the splines discussed in Chapter 7 and displayed in Figures 2.5 and 2.6, and the boosting methods discussed in Chapter 8, can lead to such complicated estimates of f that it is difficult to understand how any individual predictor is associated with the response.

대조적으로, 7장에서 논의되고 그림 2.5와 2.6에 표시된 스플라인과 8장에서 논의된 부스팅 방법과 같이, 매우 유연한 접근 방식은 f의 복잡한 추정치를 초래할 수 있으므로, 개별 예측 변수가 반응과 어떻게 연관되어 있는지 이해하기 어렵습니다.

Figure 2.7 provides an illustration of the trade-off between flexibility and interpretability for some of the methods that we cover in this book.

그림 2.7은 우리가 이 책에서 다루는 일부 방법에 대한 유연성과 해석 가능성 사이의 트레이드오프(균형)을 보여줍니다.

Least squares linear regression, discussed in Chapter 3, is relatively inflexible but is quite interpretable.

3장에서 논의된 최소 제곱 선형 회귀는 상대적으로 융통성이 없지만 꽤 해석 가능합니다.

The lasso, discussed in Chapter 6, relies upon the linear model (2.4) but uses an alternative fitting procedure for estimating the coefficients **β0, β1,..., βp**.

6장에서 논의된 lasso는 선형 모델(2.4)에 의존하지만 계수 **β0, β1,..., βp**를 추정하기 위해 대체 맞춤(fit) 절차를 사용합니다.

The new procedure is more restrictive in estimating the coefficients, and sets a number of them to exactly zero.

새로운 절차는 계수를 추정하는 데 더 제한적이며, 그리고 계수의 수를 정확히 0으로 설정합니다.

Hence in this sense the lasso is a less flexible approach than linear regression.

따라서 이러한 의미에서 lasso는 선형 회귀보다 덜 유연한 접근 방식입니다.

It is also more interpretable than linear regression, because in the final model the response variable will only be related to a small subset of the predictors ㅡ namely, those with nonzero coefficient estimates.

또한 최종 모형에서 반응 변수는 예측 변수의 작은 부분 집합 (즉, 계수 추정치가 0이 아닌 부분 집합)에만 관련되기 때문에 선형 회귀 분석보다 더 해석하기 쉽습니다.

Generalized additive models (GAMs), discussed in Chapter 7, instead extend the linear model (2.4) to allow for certain non-linear relationships.

7장에서 논의된 GAMs(일반화된 가법 모형)은 대신 특정 비선형 관계를 허용하도록 선형 모델(2.4)을 확장합니다.

Consequently, GAMs are more flexible than linear regression.

결과적으로, GAMs은 선형 회귀보다 더 유연합니다.

They are also somewhat less interpretable than linear regression, because the relationship between each predictor and the response is now modeled using a curve.

그들은 또한 각 예측 변수와 응답 사이의 관계가 이제 곡선을 사용하여 모델링되기 때문에 선형 회귀보다 해석하기가 다소 어렵습니다.

Finally, fully non-linear methods such as **bagging, boosting, support vector machines with non-linear kernels, and neural networks (deep learning)**, discussed in Chapters 8, 9, and 10, are highly flexible approaches that are harder to interpret.

마지막으로, 8장, 9장, 10장에서 논의한 **배깅, 부스팅, 비선형 커널이 있는 서포트 벡터 머신, 신경망(딥 러닝)**과 같은 완전 비선형 방법은 해석하기 어려운 매우 유연한 접근 방식입니다.

We have established that when inference is the goal, there are clear advantages to using simple and relatively inflexible statistical learning methods.

우리는 추론이 목표일 때, 간단하고 상대적으로 융통성이 없는 통계 학습 방법을 사용하는 것이 분명한 이점이 있음을 확인했습니다.

In some settings, however, we are only interested in prediction, and the interpretability of the predictive model is simply not of interest.

그러나, 일부 설정에서는, 우리는 오직 예측에만 관심이 있고, 예측 모델의 해석 가능성은 전혀 관심이 없습니다.

For instance, if we seek to develop an algorithm to predict the price of a stock, our sole requirement for the algorithm is that it predict accurately ㅡ interpretability is not a concern.

예를 들어, 만약 우리가 주식 가격을 예측하는 알고리즘을 개발하려는 경우, 알고리즘에 대한 우리의 유일한 요구 사항은 정확하게 예측하는 것입니다. ㅡ 해석 가능성은 문제가 되지 않습니다.

In this setting, we might expect that it will be best to use the most flexible model available.

이러한 설정에서는, 우리는 사용 가능한 가장 유연한 모델을 사용하는 것이 가장 좋을 것이라고 기대할 수 있습니다.

Surprisingly, this is not always the case!

놀랍게도, 항상 그런 것은 아닙니다!

We will often obtain more accurate predictions using a less flexible method.

우리는 종종 덜 유연한 방법을 사용하여 보다 정확한 예측을 얻는 경우가 있습니다.

This phenomenon, which may seem **counterintuitive** at first glance, has to do with the potential for overfitting in highly flexible methods.

언뜻 보기에 **직관에 반하는** 것처럼 보일 수 있는 이 현상은, 매우 유연한 방법에서 과대적합의 가능성과 관련이 있습니다.

We saw an example of overfitting in Figure 2.6.

우리는 그림 2.6에서 과대적합의 예를 보았습니다.

We will discuss this very important concept **further** in Section 2.2 and throughout this book.

우리는 이 매우 중요한 개념을 2.2절과 이 책 전반에 걸쳐 **더 자세히** 논의할 것입니다.

**2.2.2 The Bias-Variance Trade-Off 편향-분산 트레이드 오프(절충) p33 ~ p36**

The U-shape observed in the test MSE curves (Figures 2.9 - 2.11) turns out to be the result of two competing properties of statistical learning methods.

테스트 MSE 곡선이 U자 모양(그림 2.9~2.11)을 보이는 것은 통계학습방법의 두 가지 상충되는 성질 때문입니다.

\* MSE : 평균제곱오차

Though the mathematical proof is beyond the scope of this book, it is possible to show that the expected test MSE, for a given value x0, can always be decomposed into the sum of three fundamental quantities: the variance of ˆf(x0), the squared bias of ˆf(x0) and the variance of the error terms ϵ.

이것을 수학적으로 증명하는 것은 이 책의 범위를 벗어나지만, 주어진 값 x0에 대한 기대 테스트 MSE는 항상 세 가지의 기본적 수량인 ^f(x0)의 분산, ^f(x0)의 제곱 편향 및 오차항 ϵ의 분산의 합으로 분해될 수 있음을 보여줄 수 있습니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ[ Figure 그림 2.11. ] ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

차트이(가) 표시된 사진

자동 생성된 설명

Details are as in Figure 2.9, using a different f that is far from linear.

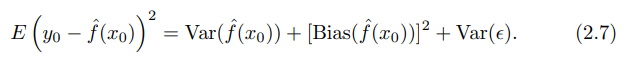
직선과는 상당히 다른 f를 사용하여 그림 2.9에서와 동일한 3가지 적합 결과를 보여줍니다.

In this setting, linear regression provides a very poor fit to the data.

이 설정에서, 선형 회귀는 데이터에 매우 적합되지 않습니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

That is, 즉, 다음 식이 성립됩니다.



Here the notation E(‘’) defines the expected test MSE at x0, and refers to the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets, and tested each at x0.

여기서 표기법 E(‘’)는 기대 테스트 MSE에 대한 정의로, 많은 훈련 세트(자료)들을 사용하여 f를 반복적으로 추정하고 각각을 x0에서 테스트한 경우 얻을 수 있는 테스트 MSE의 평균을 나타냅니다.

The overall expected test MSE can be computed by averaging E(‘’) over all possible values of x0 in the test set.

전체 기대 테스트 MSE는 테스트 세트(자료)들의 모든 가능한 x0 값에 대해 E(‘’)를 평균하여 계산할 수 있습니다.

Equation 2.7 tells us that in order to minimize the expected test error, we need to select a statistical learning method that simultaneously achieves low variance and low bias.

수학식 2.7은 기대 테스트 오차를 최소화하기 위해서는, 우리는 낮은 분산과 낮은 편향을 동시에 달성하는 통계 학습 방법을 선택할 필요가 있음을 알려줍니다.

Note that variance is inherently a nonnegative quantity, and squared bias is also nonnegative.

분산은 본질적으로 음이 아니고, 편향 제곱도 음이 아닙니다.

Hence, we see that the expected test MSE can never lie below Var(ϵ), the irreducible error from (2.3).

따라서, 우리는 기대 테스트 MSE가 (2.3)의 축소 불가능한 오류인 Var(ϵ) 보다 작을 수 없다는 것을 알 수 있습니다.

What do we mean by the variance and bias of a statistical learning method?

통계 학습 방법의 분산과 편향은 무엇을 의미할까요?

Variance refers to the amount by which ˆf would change if we estimated it using a different training data set.

분산은 만약 우리가 다른 훈련 데이터 세트(자료)들을 사용하여 추정한 경우, ˆf 가 변경되는 정도를 나타냅니다.

Since the training data are used to fit the statistical learning method, different training data sets will result in a different ˆf.

훈련 데이터는 통계 학습 방법에 적합하도록 사용되기 때문에, 훈련 데이터 세트마다 ˆf 가 다릅니다.

But ideally the estimate for f should not vary too much between training sets.

그러나 이상적으로는 f에 대한 추정치가 훈련 세트(자료) 간에 너무 많이 달라지지 않아야 합니다.

However, if a method has high variance then small changes in the training data can result in large changes in ˆf.

그러나, 만약 분산이 높은 방법의 경우, 훈련 데이터의 작은 변경(변화)으로 인해 ˆf 가 크게 변경될 수 있습니다.

In general, more flexible statistical methods have higher variance.

일반적으로, 통계 방법의 유연성이 높을수록 분산이 더 높습니다.

Consider the green and orange curves in Figure 2.9.

그림 2.9의 녹색 및 주황색 곡선을 살펴봅시다.

The flexible green curve is following the observations very closely.

유연한 녹색 곡선은 관측치를 매우 가깝게 따르고 있습니다 (아주 잘 따라가고 있습니다).

It has high variance because changing any one of these data points may cause the estimate ˆf to change considerably.

이러한 데이터 점 중 하나를 변경하면 추정치 ˆf 가 크게 변경될 수 있기 때문에 분산이 높습니다.

In contrast, the orange least squares line is relatively inflexible and has low variance, because moving any single observation will likely cause only a small shift in the position of the line.

반대로, 주황색 최소 제곱선은 하나의 관측치를 이동하면 선의 위치가 약간만 변경될 수 있기 때문에 비교적 유연하지 않고 분산이 낮습니다.

On the other hand, bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.

반면, 편향은 매우 복잡할 수 있는 실생활 문제를 훨씬 단순한 모델에 의해 근사시킴으로써 도입(발생)되는 오류를 말합니다.

For example, linear regression assumes that there is a linear relationship between Y and X1, X2,...,Xp.

예를 들어, 선형 회귀는 Y와 X1, X2, ..., Xp 사이에 선형 관계가 있다고 가정합니다.

It is unlikely that any real-life problem truly has such a simple linear relationship, and so performing linear regression will undoubtedly result in some bias in the estimate of f.

어떤 실생활 문제도 그렇게 간단한 선형 관계를 가질 가능성은 거의 없으므로, 선형 회귀를 수행하면 의심할 여지 없이 f의 추정치에 약간의 편향이 생깁니다.

In Figure 2.11, the true f is substantially non-linear, so no matter how many training observations we are given, it will not be possible to produce an accurate estimate using linear regression.

그림 2.11에서, 실제 f는 실질적으러 (상당히) 비선형이므로, 아무리 많은 훈련 관측치가 주어지더라도, 선형 회귀를 사용하여 정확한 추정치를 생성할 수 없습니다.

In other words, linear regression results in high bias in this example.

즉, 이 예제에서는 선형 회귀를 통해 높은 편향이 발생합니다.

However, in Figure 2.10 the true f is very close to linear, and so given enough data, it should be possible for linear regression to produce an accurate estimate.

그러나, 그림 2.10에서 실제 f는 선형에 매우 가깝기 때문에, 충분한 데이터가 주어지면, 선형 회귀는 정확한 추정치를 생성할 수 있어야 합니다.

Generally, more flexible methods result in less bias.

일반적으로, 더 높은 유연한 방법은 더 적은 편향을 초래합니다.

As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease.

일반적으로, 보다 유연한 방법을 사용할수록 분산은 증가하고 편향은 감소합니다.

The relative rate of change of these two quantities determines whether the test MSE increases or decreases.

이 두 양의 상대적 변화율에 따라 테스트 MSE의 증가 또는 감소 여부가 결정됩니다.

As we increase the flexibility of a class of methods, the bias tends to initially decrease faster than the variance increases.

통계방법의 유연성을 증가시키면, 처음에는 분산이 증가하는 것보다 편향이 더 빨리 감소하는 경향이 있습니다.

Consequently, the expected test MSE declines.

결과적으로, 예상되는 테스트 MSE는 감소합니다.

However, at some point increasing flexibility has little impact on the bias but starts to significantly increase the variance.

그러나, 어느 시점에서 유연성을 높이는 것은 편향에 거의 영향을 미치지 않지만 분산이 상당히 증가하기 시작합니다.

When this happens the test MSE increases. Note that we observed this pattern of decreasing test MSE followed by increasing test MSE in the right-hand panels of Figures 2.9–2.11.

이러한 경우 테스트 MSE가 증가합니다. 우리는 그림 2.9–2.11의 오른쪽 패널에서 테스트 MSE가 감소한 후 테스트 MSE가 증가하는 이러한 패턴을 관찰했습니다.

The three plots in Figure 2.12 illustrate Equation 2.7 for the examples in Figures 2.9–2.11.

그림 2.12의 세 그림은 그림 2.9–2.11의 예에 대한 방정식 2.7을 보여줍니다.

In each case the blue solid curve represents the squared bias, for different levels of flexibility, while the orange curve corresponds to the variance.

각각의 경우 파란색 실선 곡선은 다양한 유연성 수준에 대한 제곱 편향을 나타내며, 주황색 곡선은 분산에 해당합니다.

The horizontal dashed line represents Var(ϵ), the irreducible error.

수평 점선은 축소 불가능한 오차인 Var(ϵ)를 나타냅니다.

Finally, the red curve, corresponding to the test set MSE, is the sum of these three quantities.

마지막으로, 빨간색 곡선은 테스트 MSE에 해당하는며, 이 세 가지 오차의 합입니다.

In all three cases, the variance increases and the bias decreases as the method’s flexibility increases.

세 가지 경우 모두, 방법의 유연성이 증가함에 따라 분산이 증가하고 편향이 감소합니다.

However, the flexibility level corresponding to the optimal test MSE differs considerably among the three data sets, because the squared bias and variance change at different rates in each of the data sets.

그러나, 최적의 테스트 MSE를 제공하는 유연성 수준은 세 데이터 세트 간에 상당히 다릅니다. 왜냐하면 이는 각 데이터 세트(자료)에서 제곱 편향과 분산이 서로 다른 속도로 변하기 때문입니다.

In the left-hand panel of Figure 2.12, the bias initially decreases rapidly, resulting in an initial sharp decrease in the expected test MSE.

그림 2.12의 왼쪽 패널에서, 편향은 초기에 급격히 감소하여, 기대 테스트 MSE의 초기 급격한 감소를 초래합니다.

On the other hand, in the center panel of Figure 2.12 the true f is close to linear, so there is only a small decrease in bias as flexibility increases, and the test MSE only declines slightly before increasing rapidly as the variance increases.

반면, 그림 2.12의 중앙 패널에서 실제 f는 선형에 가깝기 때문에, 유연성이 증가함에 따라 편향이 약간 감소할 뿐이며, 분산이 증가함에 따라 테스트 MSE는 빠르게 증가하기 전에 약간 감소할 뿐입니다. (약감 감소하다가 증가한다)

Finally, in the right-hand panel of Figure 2.12, as flexibility increases, there is a dramatic decline in bias because the true f is very non-linear.

마지막으로, 그림 2.12의 오른쪽 패널에서, 유연성이 증가함에 따라, 실제 f는 매우 비선형적이기 때문에 편향이 극적으로 감소합니다.

There is also very little increase in variance as flexibility increases.

또한 유연성이 증가함에 따라 분산이 거의 증가하지 않습니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ[ Figure 그림 2.12. ] ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

차트, 도표이(가) 표시된 사진

자동 생성된 설명

Squared bias (blue curve), variance (orange curve), Var(ϵ) (dashed line), and test MSE (red curve) for the three data sets in Figures 2.9–2.11.

그림 2.9–2.11의 세 데이터 집합에 대한 제곱 편향(파란색 곡선), 분산(주황색 곡선), Var(ϵ)(점선) 및 테스트 MSE(빨간 곡선).

The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

수직 점선은 가장 작은 테스트 MSE에 해당하는 유연성 수준을 나타냅니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

Consequently, the test MSE declines substantially before experiencing a small increase as model flexibility increases.

결과적으로, 테스트 MSE는 모델 유연성이 증가함에 따라 약간의 증가를 경험하기 전에 상당히 감소합니다.

The relationship between bias, variance, and test set MSE given in Equation 2.7 and displayed in Figure 2.12 is referred to as the bias-variance trade-off.

방정식 2.7에 제시되고 그림 2.12에 표시된 편향, 분산 및 테스트 MSE 사이의 관계를 편향-분산 트레이드오프(절충)이라고 합니다.

Good test set performance of a statistical learning method requires low variance as well as low squared bias.

통계 학습 방법이 테스트 세트(자료)에 대해 좋은 성능을 내려며 낮은 제곱 편향 뿐만 아니라 낮은 분산도 필요합니다.

This is referred to as a trade-off because it is easy to obtain a method with extremely low bias but high variance (for instance, by drawing a curve that passes through every single training observation) or a method with very low variance but high bias (by fitting a horizontal line to the data).

이를 트레이드오프(절충)라고 하는데, 이는 편차가 매우 낮지만 분산이 높은 방법 (예: 모든 단일 훈련 관측치를 통과하는 곡선을 그리는 방법)이나 분산이 매우 낮지만 편향이 높은 방법 (데이터에 수평선을 적합시키는 방법)을 쉽게 얻을 수 있기 때문입니다.

The challenge lies in finding a method for which both the variance and the squared bias are low.

문제는 분산과 제곱 편향이 모두 낮은 방법을 찾는 것이 어렵다는 것 입니다.

This trade-off is one of the most important recurring themes in this book.

이 트레이드오프(절충)는 이 책에서 가장 중요한 반복 주제 중 하나입니다.

In a real-life situation in which f is unobserved, it is generally not possible to explicitly compute the test MSE, bias, or variance for a statistical learning method.

f가 관찰되지 않는 실생활 상황에서는, 일반적으로 통계 학습 방법에 대한 테스트 MSE, 편향 또는 분산을 명시적으로 계산할 수 없습니다.

Nevertheless, one should always keep the bias-variance trade-off in mind.

그럼에도 불구하고, 편향-분산 트레이드오프(절충)를 항상 염두에 두어야 합니다.

In this book we explore methods that are extremely flexible and hence can essentially eliminate bias.

이 책에서 우리는 매우 유연성이 높아, 따라서 본질적으로 편견을 제거할 수 있는 방법을 탐구합니다.

However, this does not guarantee that they will outperform a much simpler method such as linear regression.

그러나, 이것이 선형 회귀와 같은 훨씬 간단한 방법을 능가한다(나을 것이다)는 것을 보장하지는 않습니다.

To take an extreme example, suppose that the true f is linear.

극단적인 예를 들어, 실제 f가 선형이라고 가정합니다.

In this situation linear regression will have no bias, making it very hard for a more flexible method to compete.

이러한 상황에서 선형 회귀는 편향이 없으므로, 보다 유연한 방법으로 좋은 성능을 내기는 매우 어렵습니다.

In contrast, if the true f is highly non-linear and we have an ample number of training observations, then we may do better using a highly flexible approach, as in Figure 2.11.

대조적으로, 실제 f가 매우 비선형적이고 우리가 충분한 수의 훈련 관측치를 가지고 있다면, 우리는 그림 2.11과 같이 매우 유연한 접근법을 사용하여 더 잘 할 수 있습니다.

In Chapter 5 we discuss cross-validation, which is a way to estimate the test MSE using the training data.

5장에서는 우리는 훈련 데이터를 사용하여 테스트 MSE를 추정하는 방법인 교차 검증에 대해 논의합니다.

**3.1.2 Assessing the Accuracy of the Coefficient Estimates**

**계수 추정값의 정확도 평가 p63 ~ p68**

Recall from (2.1) that we assume that the true relationship between X and Y takes the form **Y = f(X) + ϵ** for some unknown function f, where **ϵ** is a mean-zero random error term.

(2.1)에서 X와 Y의 실제 상관관계는 어떤 알려지지 않은 함수 f에 대해 **Y = f(X) + ϵ**의 형태를 가지며, **ϵ**은 평균이 0인 랜덤 오차항입니다.



If f is to be appproximated by a linear function, then we can write this relationship as

만약 f가 선형함수로 근사된다면 이 관계를 다음과 같이 나타낼 수 있습니다.



Here **β0** is the intercept term — that is, the expected value of Y when X = 0, and **β1** is the slope — the average increase in Y associated with a one-unit increase in X.

여기서, **β0**는 절편 — 즉, X = 0일 때 Y의 기대값이고 **β1**은 기울기 — X의 한 유닛 증가에 연관된 Y 의 평균 증가입니다.

The error term is a catch-all for what we miss with this simple model:

오차항은 이러한 단순한 모델로 나타낼 때 수반되는 여러 가지 한계(miss)를 위한 것입니다:

the true relationship is probably not linear, there may be other variables that cause variation in Y , and there may be measurement error.

실제 관계는 아마도 선형적이지 않을 수 있고, Y 값의 변화를 초래하는 다른 변수들이 있을 수 있으며, 측정 오차가 있을 수 있습니다.

We typically assume that the error term is independent of X.

우리는 오차항을 보통 X와 독립이라고 가정합니다.

The model given by (3.5) defines the population regression line, which is the best linear approximation to the true relationship between X and Y .

(3.5)의 모델은 모회귀선을 정의하며, X와 Y의 실제 상관관계에 가장 잘 맞는 선형 근사입니다.

The least squares regression coefficient estimates (3.4) characterize the least squares line (3.2).

최소제곱 회귀계수의 추정치 (3.4)는 최소제곱직선 (3.2)를 결정합니다.

텍스트이(가) 표시된 사진

자동 생성된 설명중요!



The left-hand panel of Figure 3.3 displays these two lines in a simple simulated example.

그림 3.3의 왼쪽 패널은 간단한 모의 데이터를 이용해 이러한 두 직선을 나타냅니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

차트이(가) 표시된 사진

자동 생성된 설명

FIGURE 3.3. 그림 3.3 A simulated data set. 모의 자료.

Left: The red line represents the true relationship, **f(X)=2+3X**, which is known as the population regression line.

왼쪽: 붉은색 선은 모회귀선으로 알려진 실제 상관관계 **f(X)=2+3X**를 나타냅니다.

The blue line is the least squares line; it is the least squares estimate for **f(X)** based on the observed data, shown in black.

파란선은 최소 제곱선으로, 검은색의 관측 데이터를 기반으로 계산한 **f(X)**에 대한 최소제곱 추정치입니다.

Right: The population regression line is again shown in red, and the least squares line in dark blue.

오른쪽: 모회귀선은 붉은색, 최소 제곱선은 진한 파란색으로 표시됩니다.

In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations.

연한 파란색의 10개의 최소 제곱선은 각각 다른 랜덤 관측치셋을 기반으로 계산됩니다.

Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.

각 최소 제곱선은 다르지만 이들의 평균은 모회귀선과 상당히 가깝습니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

We created 100 random Xs, and generated 100 corresponding Y s from the model

우리는 100개의 X 값을 랜덤으로 생성하고 아래 모델로부터 100개의 대응하는 Y 값을 생성하였습니다.



where **ϵ** was generated from a normal distribution with mean zero.

여기서, **ϵ**은 평균이 0인 정규분포로부터 생성되었습니다.

The red line in the left-hand panel of Figure 3.3 displays the true relationship, **f(X) = 2+3X**, while the blue line is the least squares estimate based on the observed data.

그림 3.3에서 왼쪽 패널의 붉은색 직선은 실제 상관관계 **f(X) = 2+3X**를 나타낸 것이고, 푸른색 직선은 관측된 데이터에 근거한 최소제곱 추정값입니다.

The true relationship is generally not known for real data, but the least squares line can always be computed using the coefficient estimates given in (3.4).

실재하는 데이터의 경우, 실제 상관관계는 일반적으로 알려져 있지 않지만, 최소 제곱선은 (3.4)의 계수 추정값을 사용하여 항상 계산할 수 있습니다.

In other words, in real applications, we have access to a set of observations from which we can compute the least squares line; however, the population regression line is unobserved.

다시 말하면, 실제 응용에서는 관측 자료를 사용하여 최소 제곱선을 계산할 수 있습니다; 하지만, 모회귀선은 관측되지 않습니다.

In the right-hand panel of Figure 3.3 we have generated ten different data sets from the model given by (3.6) and plotted the corresponding ten least squares lines.

그림 3.3의 오른쪽 패널은 (3.6)의 모델을 사용하여 생성한 10개의 서로 다른 데이터셋에 대응하는 10개의 최소 제곱선을 나타낸 것입니다.

Notice that different data sets generated from the same true model result in slightly different least squares lines, but the unobserved population regression line does not change.

동일한 실제 모델을 사용하여 생성된 다른 데이터셋은 약간 다른 최소 제곱선을 가지지만 모회귀선은 동일합니다.

At first glance, the difference between the population regression line and the least squares line may seem subtle and confusing.

언뜻 보기에, 모회귀선과 최소 제곱선 사이의 차이는 매우 작고 구별하기 어려울 수 있습니다.

We only have one data set, and so what does it mean that two different lines describe the relationship between the predictor and the response?

우리는 자료가 하나밖에 없는데, 두 개의 다른 직선이 설명변수와 반응변수의 상관관계를 기술하는 것은 무엇을 의미할까요?

Fundamentally, the concept of these two lines is a natural extension of the standard statistical approach of using information from a sample to estimate characteristics of a large population.

근본적으로, 이 두 직선의 개념은 표본의 정보를 사용하여 큰 모집단의 특징을 추정하는 표준 통계적 방법의 자연스런 확장입니다.

For example, suppose that we are interested in knowing the population mean **µ** of some random variable Y .

예를 들어, 어떤 확률변수 Y의 모평균 **µ**를 알고자 한다고 해봅시다.

Unfortunately, **µ** is unknown, but we do have access to n observations from Y , y1,...,yn, which we can use to estimate **µ**.

불행하게도, **µ**는 알려져 있지 않지만, 우리는 Y의 n개 관측치, y1,..., yn을 알 수 있고 이것을 사용하여 **µ**를 추정할 수 있습니다.

A reasonable estimate is **ˆµ = ¯y**, where **[식 mean]** is the sample mean.

합리적인 추정값은 **ˆµ = ¯y** 이고, 여기서 **[식 mean]**는 표본평균입니다.

The sample mean and the population mean are different, but in general the sample mean will provide a good estimate of the population mean.

표본평균과 모평균은 다르지만, 일반적으로 표본평균은 모평균의 좋은 추정값을 제공합니다.

In the same way, the unknown coefficients **β0** and **β1** in linear regression define the population regression line.

마찬가지로, 선형회귀의 알려지지 않은 계수 **β0**와 **β1**은 모회귀선을 정의합니다.

We seek to estimate these unknown coefficients using **βˆ0** and **βˆ1** given in (3.4).

우리는 이러한 알려지지 않은 계수를 (3.4)의 **βˆ0** 와 **βˆ1**을 사용하여 추정하고자 합니다.

These coefficient estimates define the least squares line.

이 계수 추정값들은 최소 제곱선을 정의합니다.

The analogy between linear regression and estimation of the mean of a random variable is an apt one based on the concept of bias.

선형회귀와 확률변수의 평균값 추정 비유(유추, 추정)는 편향의 개념에서 보면 적절한 것입니다.

If we use the sample mean **ˆµ** to estimate **µ**, this estimate is unbiased, in the sense that on average, we expect **ˆµ** to equal **µ**.

만약 우리가 표본 평균 **ˆµ**을 사용하여 **µ**를 추정한다면, **ˆµ**은 평균적으로 **µ**와 동일하다고 기대된다는 점에서 이 추정값은 편향되지 않은 것입니다.

What exactly does this mean?

이것은 정확히 무엇을 의미할까요?

It means that on the basis of one particular set of observations y1,...,yn, **ˆµ** might overestimate **µ**, and on the basis of another set of observations, **ˆµ** might underestimate **µ**.

이것은 어떤 하나의 특정 관측치셋 y1,..., yn에 대해서 **ˆµ**은 **µ**를 과대추정할 수 있고, 또 다른 관측치셋에 대해서는 **ˆµ**이 **µ**를 과소추정할 수 있다는 것을 의미합니다.

But if we could average a huge number of estimates of **µ** obtained from a huge number of sets of observations, then this average would exactly equal **µ**.

그러나, 만약 우리가 아주 많은 수의 관측치셋으로부터 얻은 **µ**의 추정값들을 평균할 수 있으면 이 평균 값은 **µ**와 정확하게 동일한 값이 될 것입니다.

Hence, an unbiased estimator does not systematically over- or under-estimate the true parameter.

그러므로, 비편향 추정량은 실제 파라미터(모수)를 조직적으로 과대추정 또는 과소추정하는 것이 아닙니다.

The property of unbiasedness holds for the least squares coefficient estimates given by (3.4) as well:

비편향 성질은 (3.4)의 최소제곱 계수추정에 대해서도 성립합니다:

if we estimate **β0** and **β1** on the basis of a particular data set, then our estimates won’t be exactly equal to **β0** and **β1**.

만약 우리가 특정 데이터셋에 대해 **β0**와 **β1**을 추정하면, 그 추정값은 **β0** 및 **β1**과 정확하게 일치하지는 않을 것입니다.

But if we could average the estimates obtained over a huge number of data sets, then the average of these estimates would be spot on!

하지만, 만약 우리가 아주 많은 수의 데이터셋에 대해 얻은 추정값들을 평균할 수 있으면, 이 추정값들의 평균값은 정확하게 일치할 것입니다!

In fact, we can see from the right-hand panel of Figure 3.3 that the average of many least squares lines, each estimated from a separate data set, is pretty close to the true population regression line.

사실, 우리는 그림 3.3의 오른쪽 패널을 보면 다른 데이터 셋으로부터 추정된 최소 제곱선들의 평균은 실제 모회귀선에 매우 근접한다는 것을 볼 수 있습니다.

We continue the analogy with the estimation of the population mean **µ** of a random variable Y .

확률변수 Y의 모평균 **µ**의 추정에 대한 비유(유추, 추정)를 계속해봅시다.

A natural question is as follows: how accurate is the sample mean **ˆµ** as an estimate of **µ**?

자연스러운 질문은 다음과 같습니다: 표본평균 **ˆµ**이 **µ**의 추정값으로 얼마나 정확할까요?

We have established that the average of **ˆµ**’s over many data sets will be very close to **µ**, but that a single estimate **ˆµ** may be a substantial underestimate or overestimate of **µ**.

우리는 많은 수의 데이터셋에 대한 **ˆµ**의 평균은 **µ**에 아주 근접하지만, 하나의 추정값 **ˆµ** 은 **µ**를 상당히 과소추정 또는 과대추정할 수 있습니다.

How far off will that single estimate of **ˆµ** be? 하나의 추정값 **ˆµ**은 **µ**와 얼마나 다를까요?

In general, we answer this question by computing the standard error of **ˆµ**, written as **SE(ˆµ)**.

일반적으로, 이 질문에 대한 답은 **SE(ˆµ)**으로 표현하는 **ˆµ**의 표준오차를 계산하는 것입니다.

We have 표준오차에 대한 잘 알려진 식은 아래와 같습니다.



where **σ** is the standard deviation of each of the realizations **yi** of Y .

여기서 **σ**는 Y의 값 **yi**의 표준편차입니다.

Roughly speaking, the standard error tells us the average amount that this estimate **ˆµ** differs from the actual value of **µ**.

대체로 표준오차는 추정값 **ˆµ**이 **µ**의 실제값과 평균 어느 정도 다른 지를 말합니다.

Equation 3.7 also tells us how this deviation shrinks with n — the more observations we have, the smaller the standard error of **ˆµ**.

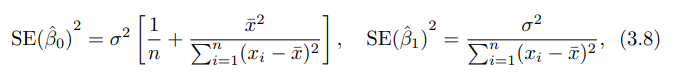
식 (3.7)은 또한 **n이 증가함에 따라 ˆµ이** 편차가 얼마나 줄어드는지를 말해줍니다. — 관측치 수가 많을수록 **ˆµ**의 표준오차가 작아집니다.

In a similar vein, we can wonder how close **βˆ0** and **βˆ1** are to the true values **β0** and **β1**.

유사한 맥락으로, **βˆ0**와 **βˆ1**이 얼마나 **β0**와 **β1**에 근접할 수 있는지 궁금할 수 있습니다.

To compute the standard errors associated with **βˆ0** and **βˆ1**, we use the following formulas:

**βˆ0**와 **βˆ1**의 표준오차를 계산하기 위해서는 다음 식을 사용합니다.



where **σ2 = Var(ϵ)**  여기서 **σ2 = Var(ϵ)**입니다.

For these formulas to be strictly valid, we need to assume that the errors **ϵi** for each observation have common variance **σ2** and are uncorrelated.

이 식들이 유효하려면, 우리는 각 관측치에 대한 오차 **ϵi**가 공통의 분산 **σ2**과 상관관계가 없다는 가정이 필요합니다.

This is clearly not true in Figure 3.1, but the formula still turns out to be a good approximation.

그림 3.1의 경우 이것은 명백히 사실이 아니지만, 이 식들은 여전히 좋은 근사치입니다.

Notice in the formula that **SE(βˆ1)** is smaller when the **xi** are more spread out; intuitively we have more leverage to estimate a slope when this is the case.

위 식에서 **SE(βˆ1)**은 **xi**가 넓게 퍼질수록 더 작아집니다; 직관적으로 이 경우에는 기울기를 추정할 영향력이 더 큽니다.

We also see that **SE(βˆ0)** would be the same as **SE(ˆµ)** if **¯x** were zero (in which case **βˆ0** would be equal to **¯y**).

또한, 만약 **¯x**가 0이면 (이 경우 **βˆ0** 은 **¯y** 와 동일할 것임) **SE(βˆ0)**은 **SE(ˆµ)**와 동일하게 될 것이라는 것을 알 수 있습니다.

In general, **σ2** is not known, but can be estimated from the data.

일반적으로 **σ2**은 알려져 있지 않지만, 데이터로부터 추정할 수 있습니다.

This estimate of **σ** is known as the residual standard error, and is given by the formula **[식 RSE]**.

**σ**의 추정치는 잔차 표준오차로 알려져 있으며, **[식 RSE]**으로 구해집니다.

Strictly speaking, when **σ2** is estimated from the data we should write **^SE(βˆ1)** to indicate that an estimate has been made, but for simplicity of notation we will drop this extra “hat”.

엄밀히 말해, **σ2** 이 추정될 때 우리는 추정값이라는 것을 나타내기 위해 **^SE(βˆ1)**으로 표현해야 합니다. 하지만 표기의 단순함을 위해 우리는 추가적인 “해트(hat)" 기호를 생략할 것입니다.

Standard errors can be used to compute confidence intervals.

표준오차는 신뢰구간을 계산하는 데 사용될 수 있습니다.

A 95 % confidence interval is defined as a range of values such that with 95 % probability, the range will contain the true unknown value of the parameter.

95% 신뢰 구간은 이 값의 범위가 95%의 확률로 파라미터(모수)의 알려지지 않은 실제값을 포함하게 될 것으로 정의됩니다.

The range is defined in terms of lower and upper limits computed from the sample of data.

이러한 범위는 데이터 표본으로부터 계산된 하한값과 상한값으로 정의됩니다.

A 95% confidence interval has the following property: if we take repeated samples and construct the confidence interval for each sample, 95% of the intervals will contain the true unknown value of the parameter.

95% 신뢰 구간에는 다음과 같은 특성이 있습니다: 만약 우리가 반복되는 표본을 추출하여 각 표본에 대한 신뢰 구간을 구성하면 구간의 95%에 모수의 실제 알 수 없는 값이 포함됩니다.

For linear regression, the 95 % confidence interval for **β1** approximately takes the form

선형회귀의 경우, **β1**에 대한 95% 신뢰구간은 대략 아래과 같은 형태를 가집니다.



That is, there is approximately a 95 % chance that the interval will contain the true value of **β1**.

즉, 아래의 구간은 대략 95%의 확률로 **β1**의 실제값을 포함할 것입니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

Similarly, a confidence interval for **β0** approximately takes the form  
마찬가지로, **β0**에 대한 신뢰구간은 대략 다음의 형태를 가집니다.



In the case of the advertising data, the 95 % confidence interval for **β0** is [6.130, 7.935] and the 95 % confidence interval for **β1** is [0.042, 0.053].

앞의 광고 데이터에서, **β0**에 대한 95% 신뢰구간은 [6.130, 7.935]이고 **β1**에 대한 95% 신뢰구간은 [0.042,0.053]입니다.

Therefore, we can conclude that in the absence of any advertising, sales will, on average, fall somewhere between 6,130 and 7,935 units.

그래서, 광고를 전혀 하지 않으면 평균판매량은 6,130과 7,940대 사이의 어떤 값으로 떨어진다고 결론을 내릴 수 있습니다.

Furthermore, for each $1,000 increase in television advertising, there will be an average increase in sales of between 42 and 53 units.

더욱이, TV 광고 투자가 매 1천 달러 증가할 경우, 판매량은 평균 42와 53대 사이로 판매가 증가할 것입니다.

Standard errors can also be used to perform hypothesis tests on the coefficients.

표준오차는 또한 계수들에 대한 가설검정을 하는 데 사용될 수 있습니다.

The most common hypothesis test involves testing the null hypothesis of

가장 흔히 사용되는 가설검정은 귀무가설(null hypothesis)과 대립가설(alternative hypothesis)을 검정합니다.

H0 : There is no relationship between X and Y (3.12) Ho: X와 Y 사이에 상관관계가 없습니다

versus the alternative hypothesis 대결: 그러면 대립가설은 다음과 같습니다.

Ha : There is some relationship between X and Y . (3.13) Ha: X와 Y 사이에 어떤 상관관계가 있습니다.

Mathematically, this corresponds to testing H0 : **β1 = 0** Versu Ha **: β1 ≠ 0**,

수학적으로, 이것은Ho: **β1 = 0**인지 Ha : **β1 ≠ 0**인지를 검정하는 것과 같습니다.

since if **β1 = 0** then the model (3.5) reduces to **Y = β0 + ϵ**, and X is not associated with Y .

만약 **β1 = 0**이면 모델 (3.5)는 **Y = β0 + ϵ**이 되므로 X는 Y와 관련이 없습니다.

To test the null hypothesis, we need to determine whether **βˆ1**, our estimate for **βˆ1**, is sufficiently far from zero that we can be confident that β1 is non-zero.

귀무가설을 검정하려면 **βˆ1**이 0이 아니라고 확신 할 수 있을 만큼 **βˆ1**에 대한 추정값이 0과 충분히 다른지를 결정해야 합니다.

How far is far enough? 0과 얼마나 달라야 충분할까요?

This of course depends on the accuracy of **βˆ1** — that is, it depends on **SE(βˆ1).**

물론 이것은 **βˆ1**의 정확도에 따라 다릅니다. — 즉, 이것은 **SE(βˆ1)**에 따라 다릅니다.

If **SE(βˆ1)** is small, then even relatively small values of **βˆ1** may provide strong evidence that **β1 ̸= 0**, and hence that there is a relationship between X and Y .

만약 **SE(βˆ1)**이 작으면, **βˆ1**이 비교적(상대적) 작아도 **β1 ̸= 0**이고 따라서 X와 Y는 서로 상관되어 있다는 강한 증거가 될 수 있습니다.

In contrast, if **SE(βˆ1)** is large, then **βˆ1** must be large in absolute value in order for us to reject the null hypothesis.

반대로, 만약 **SE(βˆ1)**이 크면, 귀무가설을 기각하기 위해 **βˆ1**의 절대값이 커야 합니다.

In practice, we compute a t-statistic, given by

실제로는, 우리는 아래와 같이 주어지는 t-통계량을 계산합니다.



which measures the number of standard deviations that **βˆ1** is away from 0.

위 식은 **βˆ1**이 0이 아닌 표준편차의 수를 측정합니다.

If there really is no relationship between X and Y , then we expect that (3.14) will have a t-distribution with n-2 degrees of freedom.

만약 X와 Y 사이에 아무 상관관계가 없으면, (3.14)는 자유도가 n- 2인 t분포를 가질 것입니다.

\* 자유도를 n-2로 두는 이유는 회귀1 수업에서 배움. [계수 2개를 제외한 값임]

The t-distribution has a bell shape and for values of n greater than approximately 30 it is quite similar to the standard normal distribution.

t분포는 종모양을 가지며 n이 대략 30보다 크면 정규분포와 아주 유사합니다.

Consequently, it is a simple matter to compute the probability of observing any number equal to **|t|** or larger in absolute value, assuming **β1 = 0**.

따라서, **β1 = 0**이라고 가정하면, 어떤 값이 **|t|**와 같거나 큰 경우를 관측할 확률을 계산하는 것은 간단합니다.

We call this probability the p-value. 우리는 이 확률을 p-값이라고 부릅니다.

Roughly speaking, we interpret the p-value as follows: a small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance, in the absence of any real association between the predictor and the response.

간단히 말해, 우리는 p-값을 다음과 같이 해석합니다: p-값이 작다는 것은 설명변수와 반응변수 사이에 어떠한 실질적인 상관성이 없는데도 우연에 의해 의미 있는 상관성이 관측될 가능성이 거의 없음을 나타냅니다.

[즉, 귀무가설 기각 -> 상관관계가 있다!]

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

테이블이(가) 표시된 사진

자동 생성된 설명

TABLE 3.1. 표 3.1.

For the Advertising data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget.

광고 자료의 TV 광고예산에 대한 판매량의 회귀에서 최소제곱 모델의 계수를 보여주고 있습니다.

An increase of $1,000 in the TV advertising budget is associated with an increase in sales by around 50 units.

TV 광고예산이 1천 달러 증가하면 관련된 판매량 증가는 약 50 유닛입니다. (1000 \* 0.0475 ~> 50)

(Recall that the sales variable is in thousands of units, and the TV variable is in thousands of dollars.)

(sales 변수의 단위는 1천 유닛이고 TV 변수는 1천 달러입니다.)

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

Hence, if we see a small p-value, then we can infer that there is an association between the predictor and the response.

그러므로, 만약 우리는 p-값이 작으면 설명변수와 반응변수 사이에 상관성이 있다고 유추할 수 있습니다.

We reject the null hypothesis — that is, we declare a relationship to exist between X and Y — if the p-value is small enough.

우리는 귀무가설을 기각합니다 — 즉, 우리는X와 Y 사이에 상관관계가 있다고 합니다. — 만약 p-값이 충분이 작으면

Typical p-value cutoffs for rejecting the null hypothesis are 5% or 1%, although this topic will be explored in much greater detail in Chapter 13.

귀무가설을 기각하기 위한 일반적인 p-값은 5% 또는 1%이지만, 이 주제는 13장에서 훨씬 더 자세히 설명될 것입니다.

When n = 30, these correspond to t-statistics (3.14) of around 2 and 2.75, respectively.

n = 30인 경우, (3.14)의 t-통계량으로 약 2와 2.75에 각각 해당합니다.

Table 3.1 provides details of the least squares model for the regression of number of units sold on TV advertising budget for the Advertising data.

표 3.1은 광고 자료에서 TV 광고예산에 따른 판매량의 최소제곱 회귀모델에 대한 상세사항을 나타낸 것입니다.

Notice that the coefficients for **βˆ0** and **βˆ1** are very large relative to their standard errors, so the t-statistics are also large;

표를 살펴보면, **βˆ0**와 **βˆ1**에 대한 계수들은 그들의 표준오차에 비해 상당히 큰 값이며, 그래서 t-통계량도 큽니다;

the probabilities of seeing such values if H0 is true are virtually zero.

만약 H0가 참이면 이러한 값을 관측할 확률은 거의 0입니다.

Hence we can conclude that **β0 ̸= 0** and **β1 ̸= 0**.

그러므로 우리는 **β0 ̸= 0** 이고 **β1 ̸= 0**라고 결론을 내릴 수 있습니다.

**3.3 ~ 3.3.2 Removing the Additive Assumption 전까지 p83 ~ p87**

**3.3 Other Considerations in the Regression Model 3.3 회귀모델에서 다른 고려할 사항**

**3.3.1 Qualitative Predictors 3.3.1 질적 설명변수(예측변수)**

In our discussion so far, we have assumed that all variables in our linear regression model are quantitative.

지금까지 논의한 바로는, 우리는 선형회귀모델의 모든 변수는 양적(quantitative)이라고 가정하였습니다.

But in practice, this is not necessarily the case; often some predictors are qualitative.

하지만 실제로는, 반드시 그런 것은 아닙니다; 보통 설명 변수들이 질적(qualitative)인 경우도 있습니다.

For example, the Credit data set displayed in Figure 3.6 records variables for a number of credit card holders.

예를 들어, 그림 3.6의 표시된 Credit 자료(데이터 세트)는 신용 카드 보유자의 수에 대한 변수를 기록합니다.

The response is balance (average credit card debt for each individual) and there are several quantitative predictors: age, cards (number of credit cards), education (years of education), income (in thousands of dollars), limit (credit limit), and rating (credit rating).

반응변수는 balance(개인의 평균 신용카드 대금)와 몇 가지 양적 설명변수: age, cards(신용카드 수), education(교육 연수), income (수천 달러), limit(신용 한도), rating(신용 등급)을 기록한 것입니다.

Each panel of Figure 3.6 is a scatterplot for a pair of variables whose identities are given by the corresponding row and column labels.

그림 3.6의 각 패널은 한 쌍의 변수에 대한 산점도(scatter plot)를 나타낸 것으로, 각 쌍의 변수는 대응하는 행과 열 라벨에 의해 주어집니다.

For example, the scatterplot directly to the right of the word “Balance” depicts balance versus age, while the plot directly to the right of “Age” corresponds to age versus cards.

예를 들어, 단어 “Balance"의 바로 오른쪽 산점도는 balance와 age를 나타낸 것이고, “Age"의 바로 오른쪽 산점도는 age와 cards를 나타낸 것입니다.

In addition to these quantitative variables, we also have four qualitative variables: own (house ownership), student (student status), status (marital status), and region (East, West or South).

이러한 양적 변수 뿐만 아니라, 우리는 또한 4개의 질적 변수: own(주택 소유), student(학생 상태), status(혼인 상태), region(동, 서, 남)이 있습니다.

**Predictors with Only Two Levels 레벨(수준) 수가 2인 설명변수**

Suppose that we wish to investigate differences in credit card balance between those who own a house and those who don’t, ignoring the other variables for the moment.

우리는 다른 변수들은 고려하지 않고 집을 소유한 사람(소유자들)과 소유하지 않은 사람(미소유자들)의 신용카드 대금 차이를 조사한다고 가정하겠습니다.

If a qualitative predictor (also known as a factor) only has two levels, or possible values, then incorporating it into a regression model is very simple.

만약 질적인 설명변수(또한 factor라고 알려진)가 단지 두 개의 레벨 또는 가능한 값을 가지면, 이것을 회귀모델에 포함하는 것은 아주 간단합니다.

We simply create an indicator or dummy variable that takes on two possible numerical values.

우리는 단순히 두 개의 가능한 값을 가지는 지시변수(indicator variable) 또는 가(더미)변수(dummy variable)를 생성합니다.

For example, based on the own variable, we can create a new variable that takes the form

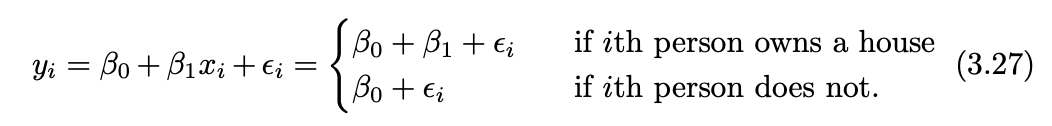
예를 들어, 우리는 own 변수를 기반으로 다음과 같은 형태의 새로운 변수를 만들 수 있습니다.

텍스트, 폰트, 화이트, 라인이(가) 표시된 사진

자동 생성된 설명

and use this variable as a predictor in the regression equation. This results in the model

그리고 이 변수를 회귀식의 한 설명변수로 사용하면, 다음 모델이 얻어집니다.



Now β0 can be interpreted as the average credit card balance among those who do not own, β0 + β1 as the average credit card balance among those who do own their house, and β1 as the average difference in credit card balance between owners and non-owners.

여기서 β0 는 미소유자들의 평균 신용카드 대금이고, β0 + β1 은 소유자들의 평균 신용카드 대금이고, β1은 비소유자들과 소유자들의 신용카드 대금에 있어서의 평균 차이로 해석될 수 있습니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

텍스트, 도표이(가) 표시된 사진

자동 생성된 설명

FIGURE 3.6. 그림 3.6.

The Credit data set contains information about balance, age, cards, education, income, limit, and rating for a number of potential customers.

Credit 자료(데이터 셋)는 다수의 잠재적 고객에 대한 balance, age, cards, education, income, limit, rating 정보를 포함합니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

Table 3.7 displays the coefficient estimates and other information associated with the model (3.27).

표 3.7은 모델 (3.27)와 관련된 계수 추정치와 다른 연관된 정보를 보여줍니다.

The average credit card debt for non-owners is estimated to be $509.80, whereas owners are estimated to carry $19.73 in additional debt for a total of $509.80 + $19.73 = $529.53.

미소유자들의 평균 신용카드 대금은 509.8 달러로 추정되고, 소유자들은 19.73 달러 더 높은 509.8 + 19.73 = 529.53 달러로 추정됩니다.

However, we notice that the p-value for the dummy variable is very high.

하지만, 우리는 가변수에 대한 p-값이 매우 높다는 사실을 눈여겨 보겠습니다.

This indicates that there is no statistical evidence of a difference in average credit card balance based on house ownership.

이것은 평균 신용카드 대금에 있어서 집을 가지고 있는지에 따른 차이가 존재한다는 통계적 증거가 없음을 나타냅니다.

The decision to code owners as 1 and non-owners as 0 in (3.27) is arbitrary, and has no effect on the regression fit, but does alter the interpretation of the coefficients.

(3.27)에서 소유자를 1로 미소유자를 0으로 정한 것은 임의로 결정한 것으로 회귀적합에 전혀 영향을 주지 않으며, 단지 계수들에 대한 해석만 달라집니다.

If we had coded non-owners as 1 and owners as 0, then the estimates for β0 and β1 would have been 529.53 and −19.73, respectively, leading once again to a prediction of credit card debt of $529.53 − $19.73 = $509.80 for nonowners and a prediction of $529.53 for owners.

만약 미소유자들을 1로 소유자들을 0으로 정하면 β0 와 β1 에 대한 추정값은 각각 529.53과 -19.73이 되어, 앞에서와 마찬가지로 신용카드 대금에 대한 예측값은 미소유자가 529.53-19.73 = 509.8 달러이고 소유자가 529.53 달러가 됩니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

TABLE 3.7. 표 3.7.

Least squares coefficient estimates associated with the regression of balance onto own in the Credit data set.

Credit 자료(데이터 세트)에서 own에 대한 balance의 회귀와 관련된 최소제곱 계수 추정치를 나타내고 있습니다.

The linear model is given in (3.27).

이 선형모델은 (3.27)에 주어집니다.

That is, ownership is encoded as a dummy variable, as in (3.26).

즉, 소유권은 (3.26)에서와 같이 가변수로 코딩됩니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

Alternatively, instead of a 0/1 coding scheme, we could create a dummy variable

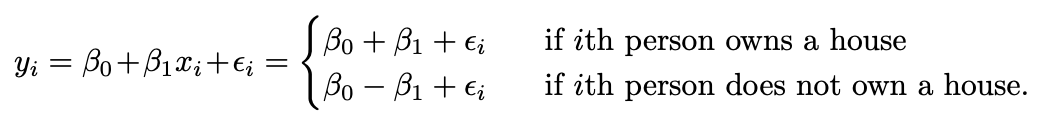
또는, 0/1로 코딩하는 방식 대신에, 우리는 가변수를 아래와 같이 만들 수 있습니다.

텍스트, 폰트, 화이트, 라인이(가) 표시된 사진

자동 생성된 설명

and use this variable in the regression equation. This results in the model

그리고 이 변수를 회귀식에 사용하면 모델은 다음과 같이 표현됩니다.



Now β0 can be interpreted as the overall average credit card balance (ignoring the house ownership effect), and β1 is the amount by which house owners and non-owners have credit card balances that are above and below the average, respectively.

이 경우, β0은 (소유권 효과를 고려하지 않은) 전체 평균 신용카드 대금으로 해석될 수 있고, β1은 각각 소유자들은 평균보다 높고 미소유자들은 평균보다 낮은 그 양을 나타냅니다.

In this example, the estimate for β0 is $519.665, halfway between the non-owner and owner averages of $509.80 and $529.53.

이 예에서, β0 에 대한 추정값은 519.665 달러가 될 것인데, 이 값은 미소유자들과 소유자들 평균인 509.8과 529.53 사이의 중간값입니다.

The estimate for β1 is $9.865, which is half of $19.73, the average difference between owners and non-owners.

β1에 대한 추정값은 9.865 달러가 될 것이며, 이 값은 소유자들과 미소유자들 평균값의 차인 19.73 달러의 절반입니다.

It is important to note that the final predictions for the credit balances of owners and non-owners will be identical regardless of the coding scheme used.

중요한 점은 소유자들과 미소유자들의 신용카드 대금에 대한 최종 예측값은 사용한 코딩 방식에 관계없이 동일하다는 것입니다.

The only difference is in the way that the coefficients are interpreted.

유일하게 다른 점은 계수들을 해석하는 방식입니다.

**Qualitative Predictors with More than Two Levels 레벨 수가 3 이상인 질적 설명변수**

When a qualitative predictor has more than two levels, a single dummy variable cannot represent all possible values.

질적 설명변수의 레벨 수가 2보다 클 때, 하나의 가변수로는 가능한 모든 값을 나타낼 수 없습니다.

In this situation, we can create additional dummy variables.

이러한 경우, 우리는 가변수를 하나 더 만들 수 있습니다.

For example, for the region variable we create two dummy variables.

예를 들어, region 변수에 대해 우리는 2개의 가변수를 생성합니다.

텍스트, 폰트, 화이트, 라인이(가) 표시된 사진

자동 생성된 설명

The first could be (3.28) and the second could be (3.29).

첫 번째는 (3.28**남**)로 나타낼 수 있고 두 번째는 (3.29**서**) 같이 표현할 수 있습니다.

텍스트, 폰트, 화이트, 서예이(가) 표시된 사진

자동 생성된 설명

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진

자동 생성된 설명

TABLE 3.8. 표 3.8.

Least squares coefficient estimates associated with the regression of balance onto region in the Credit data set.

Credit 자료(데이터 셋)에서 region에 대한 balance의 회귀와 관련된 최소제곱 계수 추정치를 나타내고 있습니다.

The linear model is given in (3.30).

이 선형모델은 (3.30)에 주어집니다.

That is, region is encoded via two dummy variables (3.28) and (3.29).

즉, region은 2개의 가변수 (3.28)과 (3.29)를 통해 코딩됩니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

Then both of these variables can be used in the regression equation, in order to obtain the model (3.30).

그러면 이 가변수 둘 다 회귀식에 사용하여, (3.30) 모델을 얻을 수 있습니다.

텍스트, 폰트, 화이트, 대수학이(가) 표시된 사진

자동 생성된 설명

Now β0 can be interpreted as the average credit card balance for individuals from the East, β1 can be interpreted as the difference in the average balance between people from the South versus the East, and β2 can be interpreted as the difference in the average balance between those from the West versus the East.

이 때, β0은 East의 평균 신용카드 대금이고, β1은 South와 East 간 평균 카드 대금의 차이이고, β2는 West와 East 간 평균 카드 대금의 차이로 해석될 수 있습니다.

There will always be one fewer dummy variable than the number of levels.

가변수의 개수는 항상 레벨 수보다 하나 작을 것입니다.

The level with no dummy variable—East in this example is known as the baseline.

가변수가 없는 레벨이 예제의 East는 기준(baseline)으로 알려져 있습니다.

From Table 3.8, we see that the estimated balance for the baseline, East, is $531.00.

표 3.8을 살펴보면, 우리는 기준인 East에 대한 추정 balance는 531 달러입니다.

It is estimated that those in the South will have $18.69 less debt than those in the East, and that those in the West will have $12.50 less debt than those in the East.

South는 East보다 카드 대금이 18.69 달러 작을 것으로 추정되고, West는 East보다 12.5 달러 적은 카드 대금을 가질 것으로 추정됩니다.

However, the p-values associated with the coefficient estimates for the two dummy variables are very large, suggesting no statistical evidence of a real difference in average credit card balance between South and East or between West and East.

하지만, 두 가변수에 대한 계수 추정치와 연관된 p-값은 아주 커서, South와 East 또는 West와 East 간에 신용카드 대금에 실질적 차이가 있다는 통계적 증거가 없음을 말합니다.

Once again, the level selected as the baseline category is arbitrary, and the final predictions for each group will be the same regardless of this choice.

다시 말해, 기준으로 선택된 레벨은 임의로 정했으며, 각 그룹에 대한 최종 예측은 선택된 기준에 관계없이 동일할 것입니다.

However, the coefficients and their p-values do depend on the choice of dummy variable coding.

하지만, 계수들과 그 p값들은 가변수의 코딩 선택에 의존합니다.

Rather than rely on the individual coefficients, we can use an F -test to test H0 : β1 = β2 = 0; this does not depend on the coding.

개별 계수에 의존하지 않고, 우리는 F-검정을 사용하여 H0: β1 = β2=0 을 검정할 수 있습니다; 이 F-검정은 가변수 코딩에 의존적이지 않습니다.

This F-test has a p-value of 0.96, indicating that we cannot reject the null hypothesis that there is no relationship between balance and region.

여기서 F-검정 p-값은 0.96으로, balance 와 region 사이에 상관관계가 존재하지 않는다는 귀무가설을 기각할 수 없음을 나타냅니다.

Using this dummy variable approach presents no difficulties when incorporating both quantitative and qualitative predictors.

이러한 가변수 방식은 양적 설명변수와 질적 설명변수를 둘 다 포함하는 경우에 어려움 없이 사용 할 수 있습니다.

For example, to regress balance on both a quantitative variable such as income and a qualitative variable such as student, we must simply create a dummy variable for student and then fit a multiple regression model using income and the dummy variable as predictors for credit card balance.

예를 들어, income과 같은 양적 변수와 student와 같은 질적 변수 둘 다에 대한 balance의 회귀를 구하려면, 우리는 단순히 student에 대한 가변수를 만들어 income과 이 가변수를 신용카드 대금에 대한 설명변수로 사용하여 다중회귀모델을 적합해야 합니다.

There are many different ways of coding qualitative variables besides the dummy variable approach taken here.

여기서 사용하는 가변수 방식 외에도 질적 변수를 코딩하는 방법은 많습니다.

All of these approaches lead to equivalent model fits, but the coefficients are different and have different interpretations, and are designed to measure particular contrasts.

이러한 방법들은 모두 동일한 모델적합을 얻지만, 계수 및 해석이 같지 않으며 특정한 차이(contrasts)를 측정하도록 고안됩니다.

This topic is beyond the scope of the book.

이 주제는 책의 범위를 벗어나 여기서 더 이상 다루지 않습니다.

**3.4~3.5 p103 ~ p110**

**3.4 The Marketing Plan 3.3 마케팅 플랜**

We now briefly return to the seven questions about the Advertising data that we set out to answer at the beginning of this chapter.

우리는 이장의 처음에 언급했던 Advertising 자료에 대한 일곱 가지 질문으로 돌아가 보려 합니다.

**1. Is there a relationship between sales and advertising budget?**

**1. 광고예산과 판매 사이에 상관관계가 있는가?**

This question can be answered by fitting a multiple regression model of sales onto TV, radio, and newspaper, as in (3.20), and testing the hypothesis H0 : βTV = βradio = βnewspaper = 0.

이 질문에 대한 답은 (3.20)에서처럼 TV, radio, newspaper에 따른 sales의 다중회귀모델을 적합하고 가설 H0 : βTV = βradio = βnewspaper = 0을 검정함으로써 얻을 수 있습니다.

In Section 3.2.2, we showed that the F-statistic can be used to determine whether or not we should reject this null hypothesis.

3.2.2절에서 살펴보았듯이, F-통계량은 귀무가설을 기각해야 하는지 결정하는 데 사용될 수 있습니다.

In this case the p-value corresponding to the F-statistic in Table 3.6 is very low, indicating clear evidence of a relationship between advertising and sales.

표 3.6의 F-통계량에 대응하는 p-값은 매우 낮은데, 이것은 광고와 판매량 사이에 상관관계가 존재한다는 명백한 증거가 됩니다.

**2. How strong is the relationship?**

**2. 광고예산과 판매 사이에 얼마나 강한 상관관계가 있는가?**

We discussed two measures of model accuracy in Section 3.1.3.

3.1.3절에서 모델 정확도를 나타내는 두 가지 측도를 살펴보았습니다.

First, the RSE estimates the standard deviation of the response from the population regression line.

첫째, RSE는 모회귀선으로부터 반응변수의 표준편차를 추정합니다.

For the Advertising data, the RSE is 1.69 units while the mean value for the response is 14.022, indicating a percentage error of roughly 12%.

Advertising 자료에서, RSE는 1,69 유닛이고 반응변수에 대한 평균값은 14,022입니다. 이것은 대략 12%의 오차에 해당합니다.

Second, the R2 statistic records the percentage of variability in the response that is explained by the predictors.

둘째, R2 통계량은 설명변수들에 의해 설명되는 반응변수의 변동을 백분율로 기록합니다.

The predictors explain almost 90 % of the variance in sales.

설명변수들은 sales의 분산 중 거의 90%를 설명합니다.

The RSE and R2 statistics are displayed in Table 3.6.

RSE와 R2 통계량은 표 3.6에 표시되어 있습니다.

**3. Which media are associated with sales?**

**3. 어느 매체가 판매에 기여하는가?**

To answer this question, we can examine the p-values associated with each predictor’s t-statistic (Section 3.1.2).

이 질문에 답하기 위해 각 설명변수의 t-통계량과 연관된 p-값을 조사합니다(3.1.2절).

In the multiple linear regression displayed in Table 3.4, the p-values for TV and radio are low, but the p-value for newspaper is not.

표 3.4에 나타낸 다중선형회귀에서 TV와 radio에 대한 p값들은 낮지만, 신문에 대한 p-값은 그렇지 않습니다.

This suggests that only TV and radio are related to sales.

이것은 TV와 radio만이 sales와 상관관계가 있다는 것을 의미합니다.

In Chapter 6 we explore this question in greater detail.

우리는 6장에서 이 문제에 대해 더욱 더 상세히 살펴볼 것입니다.

**4. How large is the association between each medium and sales?**

**4. 판매에 대한 각 매체의 효과는 얼마나 되는가?**

We saw in Section 3.1.2 that the standard error of βˆj can be used to construct confidence intervals for βj.

3.1.2절에서 보았듯이 βˆj 의 표준오차는 βj에 대한 신뢰구간을 구하는 데 사용될 수 있습니다.

For the Advertising data, we can use the results in Table 3.4 to compute the 95 % confidence intervals for the coefficients in a multiple regression model using all three media budgets as predictors.

광고 데이터의 경우, 우리는 표 3.4의 결과를 사용하여 세 가지 미디어 예산을 모두 예측자로 사용하여 다중 회귀 모델의 계수에 대한 95% 신뢰 구간을 계산할 수 있습니다.

The confidence intervals are as follows: (0.043, 0.049) for TV, (0.172, 0.206) for radio, and (−0.013, 0.011) for newspaper.

신뢰구간은 TV의 경우 (0.043, 0.049), radio의 경우 (0.172, 0.206), 그리고 newspaper의 경우 (-0.013, 0.011)입니다.

The confidence intervals for TV and radio are narrow and far from zero, providing evidence that these media are related to sales.

TV와 radio에 대한 신뢰구간은 좁고 영과 멀리 떨어져 있기 때문에, 이러한 매체들이 sales와 관련되어 있다는 증거를 제공합니다.

But the interval for newspaper includes zero, indicating that the variable is not statistically significant given the values of TV and radio.

그러나 newspaper에 대한 신뢰 구간은 영을 포함하는데, 이것은 주어진 TV와 radio 값에 대해 newpaper 변수는 통계적으로

유의하지 않다는 것을 나타냅니다.

We saw in Section 3.3.3 that collinearity can result in very wide standard errors.

우리는 3.3.3절에서 보았듯이, 공선성은 매우 넓은 표준오차를 초래할 수 있습니다.

Could collinearity be the reason that the confidence interval associated with newspaper is so wide?

newspaper와 관련된 신뢰구간이 그렇게 넓은 것은 공선성 때문일 수 있을까요?

The VIF scores are 1.005, 1.145, and 1.145 for TV, radio, and newspaper, suggesting no evidence of collinearity.

TV, radio, newspaper에 대한 VIF 값은 각각 1.005, 1.145, 1.145로, 공선성의 증거가 없음을 시사합니다.

In order to assess the association of each medium individually on sales, we can perform three separate simple linear regressions.

판매량에 대한 각 매체의 개별 상관성을 평가하기 위해, 우리는 세 개의 다른 단순선형회귀를 수행할 수 있습니다.

Results are shown in Tables 3.1 and 3.3.

결과는 표 3.1과 3.3에 보여줍니다.

There is evidence of an extremely strong association between TV and sales and between radio and sales.

TV와 sales, 그리고 radio와 sales 사이에 아주 강한 상관관계가 있다는 증거가 있습니다.

There is evidence of a mild association between newspaper and sales, when the values of TV and radio are ignored.

TV와 radio 값을 고려하지 않을 경우 newspaper와 sales 사이에 약간의 상관성이 있다는 증거가 있습니다.

**5. How accurately can we predict future sales?**

**5. 미래의 판매량에 대해 얼마나 정확하게 예측할 수 있는가?**

The response can be predicted using (3.21).

반응변수 값은 (3.21)을 사용하여 예측할 수 있습니다.

The accuracy associated with this estimate depends on whether we wish to predict an individual response, Y = f(X) + ε, or the average response, f(X) (Section 3.2.2).

이 추정치와 연관된 정확도는 예측하고자 하는 것이 개별 반응변수 값 Y = f(X) + ε인지 또는 평균 반응변수 값 f(X) 인지에 따라 다릅니다(3.2.2 절).

If the former, we use a prediction interval, and if the latter, we use a confidence interval.

개별 반응변수 값을 예측하고자 한다면 예측구간을 사용하고, 평균 반응변수 값을 예측하고자 한다면 신뢰구간을 사용합니다.

Prediction intervals will always be wider than confidence intervals because they account for the uncertainty associated with ε, the irreducible error.

예측구간은 축소불가능 오차(irreducible error) ε와 관련된 불확실성을 포함하기 때문에 항상 신뢰구간보다 더 넓습니다.

**6. Is the relationship linear?**

**6. 상관관계가 선형적인가?**

In Section 3.3.3, we saw that residual plots can be used in order to identify non-linearity.

3.3.3절에서 보았듯이, 잔차 그래프는 비선형성을 식별하는 데 사용될 수 있습니다.

If the relationships are linear, then the residual plots should display no pattern.

만약 상관관계가 선형적이면, 잔차 그래프에는 패턴이 없어야 합니다.

In the case of the Advertising data, we observe a non-linear effect in Figure 3.5, though this effect could also be observed in a residual plot.

Advertising 자료의 비선형적 효과가 그림 3.5에서 관찰되는데, 이 효과는 잔차 그래프에서도 관찰할 수 있습니다.

In Section 3.3.2, we discussed the inclusion of transformations of the predictors in the linear regression model in order to accommodate non-linear relationships.

3.3.2절에서는 비선형 상관관계를 수용하기 위해 선형회귀모델의 설명변수들을 변환하는 것에 대해 살펴보았습니다.

**7. Is there synergy among the advertising media?**

**7. 광고 매체 사이에 시너지가 있는가?**

The standard linear regression model assumes an additive relationship between the predictors and the response.

표준선형회귀모델은 설명변수들과 반응변수 사이에 가산적 상관관계를 가정합니다.

An additive model is easy to interpret because the association between each predictor and the response is unrelated to the values of the other predictors.

가산적 모델은 해석하기 쉬운데, 그 이유는 반응변수에 대한 각 설명변수의 효과가 다른 설명변수들의 값과 상관되어 있지 않기 때문입니다.

However, the additive assumption may be unrealistic for certain data sets.

하지만, 가산적 가정은 자료에 따라 현실적이지 않을 수도 있습니다.

In Section 3.3.2, we showed how to include an interaction term in the regression model in order to accommodate non-additive relationships.

3.3.2 절에서 우리는 비가산적 상관관계를 수용하기 위해 회귀모델에 상호작용 항을 포함하는 방법에 대해 살펴보았습니다.

A small p-value associated with the interaction term indicates the presence of such relationships.

상호작용 항과 연관된 p-값이 작으면 이러한 상관관계가 존재한다는 것을 나타냅니다.

Figure 3.5 suggested that the Advertising data may not be additive.

그림 3.5는 Advertising 자료가 가산적이지 않을 수 있음을 시사합니다.

Including an interaction term in the model results in a substantial increase in R2, from around 90 % to almost 97 %.

모델에 상호작용 항을 포함하는 것은 R2 값이 약 90%에서 거의 97%까지 상당히 증가하게 합니다.

**3.5 Comparison of Linear Regression with K-Nearest Neighbors**

**3.5 선형회귀와 K-최근접이웃의 비교**

As discussed in Chapter 2, linear regression is an example of a parametric approach because it assumes a linear functional form for f(X).

2장에서 다루었듯이, 선형회귀는 f(X)를 선형함수 형태라고 가정하기 때문에 모수적(parametric) 기법의 한 예입니다.

Parametric methods have several advantages.

모수적 방법은 몇 가지 장점이 있습니다.

They are often easy to fit, because one need estimate only a small number of coefficients.

추정해야 할 계수의 수가 적기 때문에 적합하기도 쉽습니다.

In the case of linear regression, the coefficients have simple interpretations, and tests of statistical significance can be easily performed.

선형회귀의 경우, 계수들에 대한 해석이 간단하고, 통계적 유의성을 쉽게 검정할 수 있습니다.

But parametric methods do have a disadvantage: by construction, they make strong assumptions about the form of f(X).

그러나 모수적 방법은 f(X)의 형태에 대한 강한 가정을 근거로 구성된다는 단점이 있습니다.

If the specified functional form is far from the truth, and prediction accuracy is our goal, then the parametric method will perform poorly.

만약 명시된 함수형태가 실제와 많이 다르고, 목적이 예측 정확도이면, 모수적 방법은 좋은 결과를 얻지 못할 것입니다.

For instance, if we assume a linear relationship between X and Y but the true relationship is far from linear, then the resulting model will provide a poor fit to the data, and any conclusions drawn from it will be suspect.

예를 들어, X와 Y의 상관관계를 선형이라고 가정하지만 실제 상관관계가 선형적이지 않으면, 결과 모델은 데이터에 잘 적합되지 않을 것이고, 이 모델로부터 얻은 결론은 모두 의심스러울 것입니다.

In contrast, non-parametric methods do not explicitly assume a parametric form for f(X), and thereby provide an alternative and more flexible approach for performing regression.

반대로, 비모수적 방법은 f(X)에 대한 모수적 형태를 명시적으로 가정하지 않고, 대안적이고 더욱 유연한 또 다른 회귀 수행 기법을 제공합니다.

We discuss various non-parametric methods in this book.

우리는 이 책에서는 다양한 비모수적 방법을 다룹니다.

Here we consider one of the simplest and best-known non-parametric methods, K-nearest neighbors regression(KNN regression).

여기서 고려하는 것은 가장 단순하고 잘 알려진 비모수적 방법 중 하나인 K-최근접이웃회귀(KNN 회귀)입니다.

The KNN regression method is closely related to the KNN classifier discussed in Chapter 2.

KNN 회귀방법은 2장에서 다룬 KNN 분류기와 밀접하게 관련되어 있습니다.

Given a value for K and a prediction point x0, KNN regression first identifies the K training observations that are closest to x0, represented by N0.

주어진 K 값과 예측 포인트 x0에 대해, KNN 회귀는 먼저 x0에 가장 가까운 K개의 훈련 관측치 N0을 식별합니다.

It then estimates f(x0) using the average of all the training responses in N0.

그 다음에 N0 내의 모든 훈련 관측치들에 대한 반응변수 값들의 평균을 사용하여 f(x0)을 추정합니다.

In other words,

즉, 다음을 계산합니다.

폰트, 화이트, 친필, 상징이(가) 표시된 사진

자동 생성된 설명

Figure 3.16 illustrates two KNN fits on a data set with p = 2 predictors.

그림 3.16은 설명변수의 수가 p = 2인 데이터셋에 대한 두 개의 KNN 적합을 보여줍니다.

The fit with K = 1 is shown in the left-hand panel, while the right-hand panel corresponds to K = 9.

왼쪽 패널의 적합은 K = 1인 경우이고, 오른쪽 패널은 K = 9인 경우입니다.

We see that when K = 1, the KNN fit perfectly interpolates the training observations, and consequently takes the form of a step function.

우리는 K = 1일 때 KNN 적합은 훈련 관측치들을 완벽하게 보간하고, 그 결과 계단함수의 형태가 됩니다.

When K = 9, the KNN fit still is a step function, but averaging over nine observations results in much smaller regions of constant prediction, and consequently a smoother fit.

K= 9일 때도 KNN 적합은 여전히 계단함수이지만, 9개 관측치의 평균을 이용하면 상수 예측 영역이 훨씬 작아지고 결과적으로 적합은 더 평활해집니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

다채로움, 아동 미술이(가) 표시된 사진

자동 생성된 설명

FIGURE 3.16. 그림 3.16

Plots of f^(X) using KNN regression on a two-dimensional data set with 64 observations (orange dots).

64개 관측치(오렌지색 점)의 2차원 데이터셋에 대해 KNN 회귀를 이용한 f^(X)의 그래프입니다.

Left: K = 1 results in a rough step function fit. 왼쪽: K = 1인 경우로 계단함수 적합이 됩니다.

Right: K = 9 produces a much smoother fit. 오른쪽: K= 9인 경우로 훨씬 평활한 적합이 됩니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

In general, the optimal value for K will depend on the bias-variance tradeoff, which we introduced in Chapter 2.

일반적으로, 최적의 K 값은 2 장에서 소개된 편향-분산 절충(bias-variance tradeoff)에 따라 다를 것입니다.

A small value for K provides the most flexible fit, which will have low bias but high variance.

K 값이 작으면 적합이 유연해져 편향은 낮지만 분산이 클 것입니다.

This variance is due to the fact that the prediction in a given region is entirely dependent on just one observation.

이 차이는 주어진 영역의 예측치가 단 하나의 관찰에만 전적으로 의존한다는 사실 때문입니다.

In contrast, larger values of K provide a smoother and less variable fit; the prediction in a region is an average of several points, and so changing one observation has a smaller effect.

반대로, K 값이 클수록 적합은 더 평활해지고 변동이 줄어듭니다; 즉, 주어진 영역의 예측값은 여러 점들의 평균값이어서 하나의 관측치 변경에 의한 영향이 작아집니다.

However, the smoothing may cause bias by masking some of the structure in f(X).

하지만, 평활화는 f(X) 구조의 일부를 감춤(masking)으로써 편향을 초래할 수 있습니다.

In Chapter 5, we introduce several approaches for estimating test error rates.

5장에서, 우리는 검정오차율을 추정하는 몇 가지 기법을 소개합니다.

These methods can be used to identify the optimal value of K in KNN regression.

이 기법들은 KNN 회귀에서 최적의 K 값을 찾는 데 사용될 수 있습니다.

In what setting will a parametric approach such as least squares linear regression outperform a non-parametric approach such as KNN regression?

어떤 경우에 최소제곱선형회귀와 같은 모수적 방식이 KNN 회귀와 같은 비모수적 방식보다 더 나을까요?

The answer is simple: the parametric approach will outperform the non-parametric approach if the parametric form that has been selected is close to the true form of f.

답은 간단합니다: 모수적 방식은 선택된 모수 형태가 f의 실제 형태에 가까운 경우 비모수적 방식보다 더 나은 결과를 낼 것입니다.

Figure 3.17 provides an example with data generated from a one-dimensional linear regression model.

그림 3.17은 1차원 선형회귀모델을 사용하여 생성한 데이터의 예입니다.

The black solid lines represent f(X), while the blue curves correspond to the KNN fits using K = 1 and K = 9.

검은색 실선은 f(X)을 나타내고, 파란색 곡선은 K= 1과 K= 9를 사용한 KNN 적합에 각각 해당합니다.

In this case, the K = 1 predictions are far too variable, while the smoother K = 9 fit is much closer to f(X).

여기서, K= 1인 경우 예측치들의 변동이 너무 크고, K= 9인 더 평활한 적합이 f(X)와 훨씬 더 가깝습니다.

However, since the true relationship is linear, it is hard for a non-parametric approach to compete with linear regression: a non-parametric approach incurs a cost in variance that is not offset by a reduction in bias.

하지만, 실제 상관관계가 선형적이므로, 비모수적 기법이 선형회귀와 비슷한 결과를 내기는 어렵습니다: 즉, 비모수적 기법이 초래한 분산 증가는 편향 감소로 상쇄되지 않습니다.

The blue dashed line in the left- hand panel of Figure 3.18 represents the linear regression fit to the same data.

그림 3.18의 왼쪽 패널에 도시한 파란색 파선은 동일한 데이터에 대한 선형회귀적합을 나타냅니다.

It is almost perfect.

그것은 거의 완벽하게 일치합니다.

The right-hand panel of Figure 3.18 reveals that linear regression outperforms KNN for this data.

그림 3.18의 오른쪽 패널은 이 데이터의 경우 선형회귀가 KNN보다 낫다는 것을 보여줍니다.

The green solid line, plotted as a function of 1/K, represents the test set mean squared error (MSE) for KNN.

1/K의 함수로 나타낸 녹색 실선은 KNN에 대한 평균제곱오차(MSE)를 나타냅니다.

The KNN errors are well above the black dashed line, which is the test MSE for linear regression.

KNN 오차는 선형회귀에 대한 MSE인 검은색 파선보다 훨씬 위에 있습니다.

When the value of K is large, then KNN performs only a little worse than least squares regression in terms of MSE.

K 값이 크면, MSE 측면에서 KNN은 최소제곱회귀에 비해 단지 약간 더 나쁩니다.

It performs far worse when K is small.

K 값이 작은 경우에는 검정(성능)이 훨씬 떨어집니다.

In practice, the true relationship between X and Y is rarely exactly linear.

현실적으로, X와 Y의 실제 상관관계가 정확하게 선형적인 경우는 거의 없습니다.

Figure 3.19 examines the relative performances of least squares regression and KNN under increasing levels of non-linearity in the relationship between X and Y.

그림 3.19는 X와 Y 사이의 비선형성 수준을 높이면서 최소제곱회귀와 KNN의 상대적 성능을 조사한 것입니다.

In the top row, the true relationship is nearly linear.

맨 위 행 (위쪽 그림)의 경우, 실제 상관관계가 거의 선형입니다.

In this case we see that the test MSE for linear regression is still superior to that of KNN for low values of K.

이 경우 우리는 선형회귀에 대한 검정 MSE는 K 값이 작으면 여전히 KNN보다 더 낫습니다.

However, for K ≥ 4, KNN outperforms linear regression.

하지만, K ≥ 4 경우, KNN이 선형회귀보다 더 낫습니다.

The second row illustrates a more substantial deviation from linearity.

두번째 행 (아랫쪽 그림)은 선형성과의 더 큰 편차(비선형적인 경우를 말하는 거임)를 보여줍니다.

In this situation, KNN substantially outperforms linear regression for all values of K.

이 경우, KNN은 모든 K 값에 대해 대체로 선형회귀다 훨씬 낫습니다.

Note that as the extent of non-linearity increases, there is little change in the test set MSE for the non-parametric KNN method, but there is a large increase in the test set MSE of linear regression.

비모수적 KNN 방법에 대한 검정 MSE는 비선형성의 정도가 증가해도 거의 변화가 없지만, 선형회귀의 검정 MSE는 크게 증가합니다.

Figures 3.18 and 3.19 display situations in which KNN performs slightly worse than linear regression when the relationship is linear, but much better than linear regression for non-linear situations.

그림 3.18과 3.19에 의하면, KNN은 상관관계가 선형적일 때는 선형회귀보다 약간 나쁘고 상관관계가 비선형적인 경우에는 선형회귀보다 훨씬 더 낫습니다.

In a real life situation in which the true relationship is unknown, one might suspect that KNN should be favored over linear regression because it will at worst be slightly inferior to linear regression if the true relationship is linear, and may give substantially better results if the true relationship is non-linear.

실제 상관관계를 모르는 현실적인 환경에서는, 실제 관계가 선형이면, 최악의 경우 선형 회귀보다 약간 열등하고, 실제 관계가 비선형이면 훨씬 더 나은 결과를 제공할 수 있기 때문에, 선형회귀 대신 KNN을 사용해야 한다고 생각할지도 모릅니다.

But in reality, even when the true relationship is highly non-linear, KNN may still provide inferior results to linear regression.

하지만 실제로는, 실제 상관관계가 심하게 비선형적인 경우에도, 여전히 KNN이 선형회귀보다 못한 결과를 줄 수도 있습니다.

In particular, both Figures 3.18 and 3.19 illustrate settings with p = 1 predictor.

특히, 그림 3.18과 3.19는 둘 다 설명변수의 수가 p = 1인 경우를 보여줍니다.

But in higher dimensions, KNN often performs worse than linear regression.

그러나 차원이 높은 경우, KNN은 보통 선형회귀보다 나쁜 성능을 제공합니다.

Figure 3.20 considers the same strongly non-linear situation as in the second row of Figure 3.19, except that we have added additional noise predictors that are not associated with the response.

그림 3.20은 그림 3.19의 두번째 행(아래쪽)과 동일한 정도의 심한 비선형적인 경우를 고려하며 추가로 반응변수와 상관되지 않은 잡음(noise) 설명변수를 갖습니다.

When p = 1 or p = 2, KNN outperforms linear regression.

p = 1 또는 p = 2인 경우, KNN은 선형회귀보다 더 나은 결과를 줍니다.

But for p = 3 the results are mixed, and for p ≥ 4 linear regression is superior to KNN.

그러나 p = 3이면 결과가 엇갈리고, p ≥ 4 이면 선형회귀가 KNN보다 우수합니다.

In fact, the increase in dimension has only caused a small deterioration in the linear regression test set MSE, but it has caused more than a ten-fold increase in the MSE for KNN.

사실, 차원증가는 선형회귀의 검정 MSE는 약간 밖에 증가시키지 않지만, KNN의 MSE는 10배 이상 증가시킵니다.

This decrease in performance as the dimension increases is a common problem for KNN, and results from the fact that in higher dimensions there is effectively a reduction in sample size.

차원이 증가함에 따라 성능이 나빠지는 것은 KNN 방식의 공통적 문제이고, 원인은 고차 원으로 갈수록 표본크기가 실질적으로 줄어드는 효과가 있기 때문입니다.

In this data set there are 50 training observations; when p = 1, this provides enough information to accurately estimate f(X).

p = 1일 때 이 데이터셋에는 50개의 훈련 관측치가 있어 f(X)를 정확하게 추정할 만큼 충분한 정보가 제공됩니다.

However, spreading 50 observations over p = 20 dimensions results in a phenomenon in which a given observation has no nearby neighbors—this is the so-called curse of dimensionality.

하지만, p=20인 차원에 걸쳐 50개의 관측값을 분산하면 주어진 관측치에 가까운 이웃이 없는 현상이 발생한다.ㅡ이것을 차원의 저주라고 합니다.

That is, the K observations that are nearest to a given test observation x0 may be very far away from x0 in p-dimensional space when p is large, leading to a very poor prediction of f(x0) and hence a poor KNN fit.

즉, 주어진 검정 관측치 x0에 가장 가까운 K개의 관측치들은 p가 클 때 p차원 공간의 x0으로부터 아주 멀리 떨어져 있으므로, f(x0)의 예측값이 아주 나쁘게하고 따라서 좋지 않은 KNN 적합을 얻게 됩니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

라인, 도표, 그래프, 경사이(가) 표시된 사진

자동 생성된 설명

FIGURE 3.17. 그림 3.17

Plots of f^(X) using KNN regression on a one-dimensional data set with 50 observations.

50개 관측치의 1차원 데이터셋에 대해 KNN 회귀를 이용한 f^(x)의 그래프.

The true relationship is given by the black solid line.

실제 상관관계는 검은색 실선으로 주어집니다.

Left: The blue curve corresponds to K = 1 and interpolates (i.e. passes directly through) the training data.

왼쪽: 파란색 곡선은 K=1인 경우로 훈련 데이터를 보간합니다(즉, 직접 지나간다).

Right: The blue curve corresponds to K = 9, and represents a smoother fit.

오른쪽: 파란색 곡선은 K=9인 경우로 훨씬 평활한 적합을 나타냅니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

라인, 도표, 그래프, 평행이(가) 표시된 사진

자동 생성된 설명

FIGURE 3.18. 그림 3.18

The same data set shown in Figure 3.17 is investigated further.

그림 3.17에 표시된 동일한 데이터셋에 대해 추가로 살펴봅니다.

Left: The blue dashed line is the least squares fit to the data.

왼쪽: 파란색 파선은 이 데이터에 대한 최소제곱적합입니다.

Since f(X) is in fact linear (displayed as the black line), the least squares regression line provides a very good estimate of f(X).

f(X)가 실제로 선형이므로(검은색 선으로 표시), 최소제곱회귀선은 f(X)에 대해 상당히 좋은 추정치를 제공합니다.

Right: The dashed horizontal line represents the least squares test set MSE, while the green solid line corresponds to the MSE for KNN as a function of 1/K (on the log scale).

오른쪽: 수평 파선은 최소제곱 검정 MSE를 나타내고, 녹색 선은 1/K(로그 스케일에서)의 함수로 KNN의 MSE에 해당합니다.

Linear regression achieves a lower test MSE than does KNN regression, since f(X) is in fact linear.

f(X)가 사실상 선형이기 때문에 선형 회귀는 KNN 회귀보다 더 낮은 검정 MSE를 달성합니다.

For KNN regression, the best results occur with a very large value of K, corresponding to a small value of 1/K.

KNN회귀의 경우, 작은 값 1/K에 해당하는 매우 큰 K값에서 최상의 결과가 발생합니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

도표, 라인, 그래프, 텍스트이(가) 표시된 사진

자동 생성된 설명

FIGURE 3.19. 그림 3.19

Top Left: In a setting with a slightly non-linear relationship between X and Y (solid black line), the KNN fits with K = 1 (blue) and K = 9 (red) are displayed.

왼쪽 위: X와 Y 사이에 약간의 비선형적 상관관계(검은색 선)가 있는 설정에서 K=1(파란색)과 K=9(붉은색)인 KNN 적합.

Top Right: For the slightly non-linear data, the test set MSE for least squares regression (horizontal black) and KNN with various values of 1/K (green) are displayed.

오른쪽 위: 약간 비선형적인 데이터에서 최소제곱회귀(수평의 검은 파선)와 여러가지 1/K 값의 KNN(녹색)에 대한 검정 MSE.

Bottom Left and Bottom Right: As in the top panel, but with a strongly non-linear relationship between X and Y .

왼쪽 및 오른쪽 아래: 위쪽 패널과 마찬가지이며, 다만 이번에는 X와 Y 사이에 강한 비선형적 상관관계가 있습니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

도표, 라인, 그래프, 텍스트이(가) 표시된 사진

자동 생성된 설명

FIGURE 3.20. 그림 3.20

Test MSE for linear regression (black dashed lines) and KNN (green curves) as the number of variables p increases.

변수 p의 수가 증가함에 따른 선형회귀 (검은색 파선)와 KNN (녹색 곡선)의 검정 MSE.

The true function is non–linear in the first variable, as in the lower panel in Figure 3.19, and does not depend on the additional variables.

실제 함수는 그림 3.19의 아래 패널과 같이 첫 번째 변수에 비선형적이고, 추가적인 변수에는 의존적이지 않습니다.

The performance of linear regression deteriorates slowly in the presence of these additional noise variables, whereas KNN’s performance degrades much more quickly as p increases.

선형회귀의 성능은 이러한 추가적인 잡음변수의 존재로 인해 천천히 나빠지고, 이에 반해, KNN의 성능은 p가 증가함에 따라 급격히 나빠집니다.

ㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡㅡ

As a general rule, parametric methods will tend to outperform non-parametric approaches when there is a small number of observations per predictor.

일반적으로, 설명변수당 관측치의 수가 작으면 모수적 방법이 비모수적 방법보다 더 나은 결과를 제공합니다.

Even when the dimension is small, we might prefer linear regression to KNN from an interpretability standpoint.

차원이 낮은 경우에도, 우리는 해석력 관점에서 선형회귀를 KNN보다 선호할 수 있습니다.

If the test MSE of KNN is only slightly lower than that of linear regression, we might be willing to forego a little bit of prediction accuracy for the sake of a simple model that can be described in terms of just a few coefficients, and for which p-values are available.

만약 KNN의 검정 MSE가 선형회귀보다 조금밖에 작지 않으면, 우리는 단지 몇 개의 계수로 설명이 가능하고 p값을 사용할 수 있는 단순한 모델을 위해 약간의 예측 정확도를 포기할 수도 있을 것입니다.