Yonsei University, Department of Applied Statistics

Chapter 3 Linear Regression

YBIGTA Science Team 신보현

January 25, 2019

Contents

3. Liear Regression

선형회귀에서 관심사는 아래와 같다.

- - x와 y사이에 관계가 있을까? 관계가 있다면 linear할까 non-linear할까?
- 관계가 있다면 얼마나 강한 관계가 있을까? 다시 말해서 x가 주어졌을 때, y를 높은 정확도로 예측할 수 있을까?
- 여러개의 x가 있다면 어떤 x가 y에 영향을 미칠까?
- x의 y에 대한 각각의 효과를 어떻게 측정할까?
- 미래의 y를 어떻게 정확히 측정할까?
- synergy effect가 있을까?

지금부터 위의 질문들에 답해보자.

3.1 Simple Linear Regression

단순선형회귀는 우선 x와 y사이에 근사적으로 선형관계가 있다고 추측한다. 수학적으로 이것은 아래 식으로 나타낸다. 때대로 우리는 이것을 regressing of Y onto X이라고 부르기도 한다.

$$Y \approx \beta_0 + \beta_1 X \tag{3.1}$$

여기서 β_0와 β_1은 unknown parameter이므로 training data를 이용해 이 둘의 추정치(estimates)을 만든다. 그에따라서 아래와 같은 식이 성립한다.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3.2}$$

여기서 $\hat{y}$은 x에대한 prediction을 의미한다(prdicted value of the response)

3.1.1 estimating the coefficients

실제에서는 β_0, β_1이 알려져 있지 않기 때문에 data set을 이용해 추정치를 계산해야 한다. 그리고 이 추정치 $\hat{\beta}_0, \hat{\beta}_1$은 주어진 data set를 가장 잘 fit하는 애들로 고른다. 다시 말해서 이 추정치로 이루어진 line은 주어진 data에 가장 가까워야 한다. 이렇게 주어진 data와 가장 가까운 거리를 계산하는(최소화 거리) 방법은 LSE 방법이다. 즉, 다시 말해,

$$(\widehat{\beta_0}, \widehat{\beta_1}) = \underset{\beta_0, \beta_1}{argmax} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\widehat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

위에서 $\hat{y}_i$ 은 x에 대한 y의 prediction(fitted value) 이며 실제 y의 값 (y_i) 와 이 예측치의 차이를 residual(잔차) 라고 한다. 이것은 i번째의 observed response(실제 관측된 데이터. training data) 와 response value(우리의 선형 모델에 의해서 예측된 값) 의 차이다.

$$e_i = y_i - \hat{y}_i$$

그리고 우리는 이 잔차의 제곱합을 RSS(residual sum of squares) 라고 정의한

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

LSE는 이 잔차를 최소로 하는 β_0, β_1 에 대한 추정치 (estimates) 로써 정규방정식에서 β_0, β_1 에 대해 편미분을 한 후 구할 수 있다. . 아래의 추정치는 LSE을 통해 나온 unknown parameter에 대한 estimator이다

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3.4}$$

3.1.2 assessing the accuracy of the coefficient estimates

우리는 2.1에서 x와 y의 진정한 관계 (true relationship) 는$Y = f(X) + \epsilon$ 임을 2장에서 살펴보았는데, 이를 통해서 선형회귀를 표현하면 $Y = \beta_0 + \beta_1 X + \epsilon$ 일 것이다. 이것은 population regression line을 정의하고 x와 y의 true relationship에 대한 가장 best linear 추정치이다. 어떠한 모집단에서 data는 완벽히 선현관계일 수는 없다. 어떤 linear line을 가정했을 때 그것에 대해서 랜덤하게 흩어져 있을 텐데 이것을 보정해주는 것이 바로 ϵ 이다. 다르게 생각하면, x와 y의 관계를 linear하다고 생각을 하고, 그 이외의 요소들은 모두 error인 ϵ 에 넣어버리는 것이다.
예를 들어, 학점에 영향을 미치는 요인이 공부 시간만 있고 이 둘이 선형 관계를 이룬다고 생각해보자. 학점에 영향을 미치는 요인은 공부 시간 뿐만 아니라 다른 여러 요인들이 있을 것이다. 이러한 것들을 모두 ϵ 에 넣음으로써 모델을 학점과 공부 시간의 단순 선형 회귀 모형으로 설계하는 것이다.

이를 정리하면

* X와 Y의 true relationship: $Y = \beta_0 + \beta_1 X + \epsilon$

여기서 true relationship이란 우리가 추정하고자 하는 unknown but fixed function인 f을 의미하고 선형 회귀에서는 이를 linear하게 표현한다.

* ϵ에 대한 가정 : $\epsilon \sim iid\ N[0, \sigma^2]$

* $\epsilon = Y - E[Y]$ (모집단이기 때문에 $\hat{Y}$은 있을 수 없다. 기댓값으으로 대체하여 표현할 수 있음)

* population regression line: $E[Y] = \beta_0 + \beta_1 X$

* unknown parameter: $\beta_0, \beta_1, \sigma^2$

하지만 이는 이론적인 회귀분석이다. 모집단은 관찰할 수 없기 때문이다. 모집단에 위의 관계가 있을 것이라고 생각하고 표본을 통해 population regression line을 estimate한다.

* $e_i = Y_i - \widehat{Y_i}$ (ϵ_i에 대응되는 개념으로 잔차, e_i가 있다. 오차에 대한 estimate라고 생각하면 된다.)

* sample regression line: $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_i$

* estimated parameter through LSE: $\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\sigma^2}$

여기서 ϵ은 위에서 언급했듯이, 실제로 관계에 영향을 주지만 우리가 잡아내지 못하는 요소들이 주는 영향력을 포함한다. 다시 말해 실제 x와 y의 관계는 선형이 아닐 수도 있으며 x가 아닌 다른 요소들이 y의 변화에 영향을 줄 수 있다. ϵ와 x은 독립이라고 가정한다.

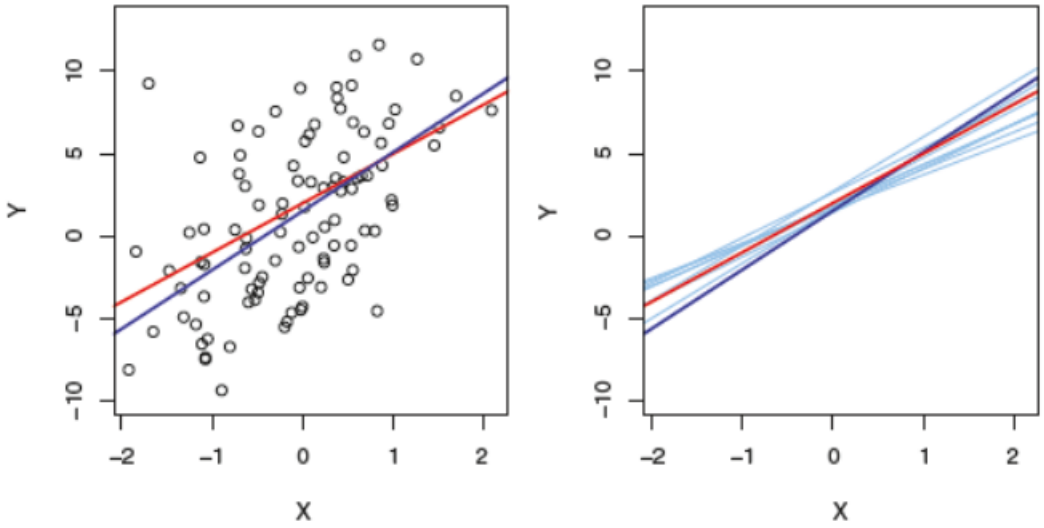

FIGURE 3.3. *A simulated data set.* Left: *The red line represents the true relationship,* $f(X) = 2 + 3X$, *which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for* $f(X)$ *based on the observed data, shown in black.* Right: *The population regression line is again shown in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.*

red line은 true relationship이고 blue line은 LSE로 추정된 선이다. 왼쪽 그림은 $Y = 2 + 3X + \epsilon$라는 population regression line에서 뽑은 100개의 data를 LSE의 기법으로 추정한 선이고 오른쪽은 10여개의 data set을 뽑아서 LSE로 추정시킨 여러 선들이다.

이렇게 true relationship line과 LSE로 추정한 line은 미묘하지만 그래도 차이를 보인다. 그럼에도 우리는 왜 LSE를 통해서 얻은 $\hat{\beta}_0, \hat{\beta}_1$을 쓸까??

mean추정 과정에 대해서 생각해보자. 우리는 random variable Y에 대한 모수 μ을 모르지만 이를 알고 싶어서 n개의 sample을 관측했다. 그렇다면 이것으로부터 우리는 sample mean을 만들 수 있다. 이 sample mean은 하나의 특정한 n개의 관측지 set은 모수 μ을 under/overestimate을 할 수 있다. 하지만 이러한 관측치를 매우 많이 뽑아서 sample mean을 구하고 그 수 많은 sample mean을 평균을 낸다면 그것은 모수 μ와 정확히 일치할 것이다. 이것은 통계학에서 많이 나오는 unbiasedness의 개념으로서, 위의 말로 된 설명을 수식으로 나타내면 $E[\hat{\mu}] = \mu$일 때, $\hat{\mu}$은 μ에 대한 unbiased estimator(불편 추정량: 어느 한쪽에 치우치지 않은 추정량)이고 이 불편 추정량은 true parameter을 그 정의에 의해서 over/under

estimate하지 않을 것이다. 그리고 이것은 $\hat{\beta}_0$과 $\hat{\beta}_1$에서도 적용된다. 따라서 이론적으로는 오른쪽의 10개의 LSE로 추정한 line들의 평균은 red line에 일치해야 한다.

위의 sample mean 얘기를 계속 해보자. 우리는 여러 개의 data set을 뽑아서 계산한 sample mean의 평균은 모수 μ과 가까워진다고 했다. 그렇다면 단 하나의 sample mean은 모수 μ과 얼마나 멀리 떨어져 있을까? 이것은 sample mean이 모수 μ을 얼마나 잘 추정할 것인지에 대한 대답이기도 하다. 일반적으로 우리는 sample mean에 대해 standard error을 계산함으로써 정확도를 계산한다.

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n} \tag{3.7}$$

대략적으로 말하면 standard error는 sample mean이 실제 value μ와 얼마나 다른지를 알려준다. n이 극한으로 갈수록 var은 0이 될 것이고 이 말은 sample mean과 μ가 차이가 0에 수렴한다는 것이고 그것은 $E[sample\,mean] = \mu$와 일맥상통한다. 비슷한 맥락에서 우리는$\hat{\beta}_0$과 $\hat{\beta}_1$이 각각의 true parameter와 얼마나 떨어져 있는지 알려주는 지표는 아래의 Standard Error이다. 여기서 $\sigma^2 = Var(\epsilon)$이다.

$$\begin{aligned} SE(\hat{\beta}_0)^2 &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \\ SE(\hat{\beta}_1)^2 &= \sigma^2 \left[\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \end{aligned} \tag{3.8}$$

일반적으로 σ^2는 알려져 있지 않다. 따라서 이를, $RSS/(n-2) = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)$으로 추정을 하는데, $RSS/(n-2)$ 가 σ^2에 대한 불편 추정량이기 때문이다.

3.1.3 assessing the accuracy of the model

우리는 우리가 세운 모델이 얼마나 데이터에 적합한지 알고 싶다. 선형 회귀에서는 두 가지를 사용한다. RSE와 R^2이다.

**** RSE***

RSE는 우리가 구한 fitted value인 $\hat{y}_i$가 y_i와 얼마나 떨어져 있는지 그 정도를 보여준다. 공식은 다음과 같다.

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.15}$$

RES는 모델이 적합을 잘 못하는 정도를 뜻하기도 한다. 만약 RSE가 작다는 것은 predictions과 true outcome간의 차이는 작다는 것이고 이것은 데이터에 모델이 잘 적합된 것으로 판단할 수 있다.

*** R^2 (결정계수)**

RSE는 범위가 정해져 있지 않은 수치로 표현이 되는데, 이에 따라서 데이터에 얼마나 잘 적합이 되었는지 비교하기가 애매할 때가 있다. 이를 보완한 결정계수는 비율의 형태(설명이 되는 explained var의 비율)를 취한다. 따라서 항상 0과 1사이의 값을 가진다.

$$R^2 = \frac{SSTO - SSE}{SSTO} = 1 - \frac{SSE}{SSTO} = \frac{SSR}{SSTO} \tag{3.17}$$

$$SSTO = \sum(y_i - \bar{y})^2, SSR = \sum(\hat{y}_i - \bar{y})^2, SSE = \sum(y_i - \hat{y}_i)^2$$

$$SSTO: \ Sum\, of\, Total\, Squares$$

$$SSR: \ Sum\, of\, regression\, Squares$$

$$SSE: \ Sum\, of\, error\, Squares$$

SSTO를 자세히 살펴보면 Y에서의 variance를 측정한다. 그리고 resgression이 시행되기 전에 이미 response에 내재된 속성이라고 볼 수 있다. 반면 SSE는 regression을 시행하고 난 뒤 설명이 되지 않은 채로 남겨진 변동성이라고 볼 수 있다. 따라서 SSTO - SSE를 하면 회귀를 함으로써 설명이 되는(또는 제거가 되는) response에서의 변동성이라고 볼 수 있고 R^2은 바로 이 비율을 측정한다.
결정계수가 1에 가깝다면 response에서의 많은 부분이 회귀에 의해서 설명된다는 뜻이고 0에 가깝다면 회귀는 response의 변동성을 설명 못한다는 뜻이다.
하지만 결정계수는 치명적인 단점이 있다. 다중선형회귀로 넘어가면, 변수가 추가 됨에 따라서 그 변수가 의미가 있든, 없든 항상 결정계수는 올라간다. 따라서 추가되는 변수가 noise variable이라도 결정계수는 증가하므로, 결정계수를 절대적인 판단의 기준으로 삼는 것은 좋지 않다. 이후에 모델의 성능을 비교하는 여러 척도를 살펴본다.

3.2 multiple linear regression

다중 선형 회귀 대신 단순 성형 회귀를 여러번 돌리는 것은 어떤 결과를 낼까? 이것은 만족스럽지 못한 결과를 낸다. 첫째로 각각의 선형 회귀는 다른 회귀 방정식과 연관되므로 이것을 어떻게 취합해서 하나의 prediction으로 만들지가 애매하다. 그리고 각각의 회귀 방정식은 회귀 계수를 결정하는데 다른 요소를 무시한다.

아래는 다중 선형 회귀의 형태이다.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{3.19}$$

3.2.1 estimating the regression coefficients

단순 선형 회귀에서와 마찬가지로 x앞의 계수들은 알려져 있지 않고 estimated 되어야 한다. 각 계수를 estimate 하여 바꾼 식은 아래의 식이다.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \tag{3.21}$$

각각의 계수의 추정치는 단순 선형 회귀에서와 같이 LSE를 통해 얻어진다.

단순 선형 회귀와 다중 선형 회귀에서의 결과가 서로 상충할 수 있다. 예를 들어 여기선 신문과 광고 세일에 관해서 단순 선형 회귀를 했을 때는 결과가 유의미 했지만 다중 선형 회귀에서는 p-value가 거의 0에 가까워 무의미한 계수라는 결과가 나왔다. 이는, 단순 선형 회귀에서는 다른 효과들, tv와 radio의 효과를 무시한 채 진행했고 다중 선형 회귀에서는 tv와 radio의 효과를 fixed 한 채로 진행했기 때문이다. response var와 tv, radio, newspaper 간의 상관계수에 주목하자.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

radio와 newspaper간의 상관계수는 0.35로써 만약에 신문에 더 많은 돈을 투자할수록 라디오에서도 더 많은 돈을 투자하는 경향이 있음을 알 수 있다. 이제 다중 선형 회귀가 맞고 신문 광고가 sales에 아무런 영향이 없지만 라디오는 sales를 증가시킨다고 가정을 하자. 그렇다면 시장에서는 라디오에 더 투자할수록 sales는 올라갈 것이고 상관계수 매트릭스에서 보여주듯이 우리는 또한 신문에 더 많은 돈을 투자할 것이다. 따라서 단순 선형 회귀에서 sales와 신문과의 관계만 집중하지만 높은 신문의 value는 높은 sales의 경향이 있겠지만 그것은 사실 신문이 sales에 영향을 미치는 것은 아니다. 신문은 라디오가 sales에 영향을 미치는 것에 무임승차하는 것과 같다.

다른 예시로 상어의 공격와 아이스크림 세일과의 관계를 회귀 분석한다고 하자. 데이터를 통해 둘은 양의 상관관계가 있음이 밝혀졌다. sales와 신문의 관계와 같이. 물론 아무도 아이스크림이 금지되어야 한다고 주장하지는 않는다. 사실은 높은 기온이 사람들을 해변가로

가게 하고 또한 아이스크림을 더 사게 만들었다. 사실 아이스크림은 상어 공격과 아무런 상관이 없고 다만 높은 기온과 상관이 있는 것이다.

3.2.2 some important questions

1. n개의 predictors 중에서 하나라도 response을 예측하는데 유용한 변수가 있을까?

2. 모든 predictors가 y을 설명하는데 도움을 줄까? 아니면 몇 개만 영향을 줄까?

3. 모델이 데이터를 얼마나 잘 적합시킬까?

4. predictor values가 주어졌을 때 어떤 response value를 예측해야할까? 아니면 우리의 예측이 얼마나 정확할까?

***1*번에 대한 설명**

단순 선형 회귀에서 regression이 의미가 있는지 확인하기 위해서 우리는 단순히 $\beta_1 = 0$인지 확인하기만 했다. 즉 다시 말해 β_1이 0과 얼마나 떨어져 있는지 그 정도를 측정했다. 다중 선형 회귀에서는 모든 계수들이 0인지 아닌지가 관심사이다. 이것을 대립가설과 귀무가설로 나타내면 아래와 같다.

$$H_o: \ \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a: \ \text{at least one } \beta_j \text{is non-zero}$$

이 가설은 F 통계량을 계산함으로써 증명할 수 있다.

$$F = \frac{(SSTO - SSE)/p}{SSE/(n-p-1)} \tag{3.23}$$

통계학에서는 귀무가설 하에서의 모델을 Reduced Model이라고 하고 대립가설 하에서의 모델을 Full Model이라고 한다. 이렇게 두 모델을 이용하여 가설검정을 할 수도 있는데 이를 Partial F test라고 한다. Reduced Model과 Full Model은 아래와 같다.

$$Model(R): \ y_i = \beta_0 + \epsilon_i$$

$$Model(F): \ y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i$$

(R) 모델과 (F) 모델 하에서의 SSTO, SSE을 LSE을 통해 구하면 아래와 같다.

$$\widehat{\beta_0}^R = \overline{y} \therefore SSE^R = \sum (y_i - \overline{y})^2 = SSTO^F$$

$$SSE^F = \sum (y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \cdots + \widehat{\beta_p} x_{pi}))^2$$

따라서 위 3.23의 검정 통계량은 $F = \dfrac{(SSE^R - SSE^F)/(df^R - df^F)}{SSE^F/df^F} \sim F(df^R - df^F, df^F)$ 로 바꿔 쓸 수 있다. 이와 같은 Partial F test는 자주 등장하니 꼭 기억해두자.

*** 2번에 대한 설명**

다중 선형 회귀에서는 F통계량을 계산하고 이와 관련된 p-value을 계산 하는 것이 가장 처음의 과제이다. 만약 우리가 최소 하나의 predictors가 response와 연관이 되어 있다고 결론을 내면 우리는 어떤 것이 바로 그것인지 궁금해 할 것이다! 우리는 개개인의 p-value을 볼 수 있지만 p가 매우 크면 잘못된 판단을 할 수도 있다. 이렇게 연관된 변수만 선택하는 과정을 variable selection이라고 부른다. 이것은 챕터 6에서 상세히 다루고 여기서는 간략하게 소개만 한다.

이상적으로는 predictors의 여러 subset을 만들어서 모델에 넣어보는 방법이 있다. 그러면 가장 좋은 성능을 내는 모델을 고르면 된다. 하지만 p가 매우 많으면 경우의 수가 너무 많아진다. 대신 우리는 자동화되고 효율적인 접근식을 선택한다.

-forward selection: 절편만 포함하고 predictors가 없는 모델에서 시작한다. 그리고 p개의 단순 선형 회귀를 진행하고 그 빈 모델에 가장 낮은 RSS를 도출하는 변수를 추가한다. 이 접근 방식은 어떤 stopping line에 도달하면 멈춘다

-backward selection: 모든 변수가 있는 모델에서 시작한다. 그리고 p-value가 가장 높은 것부터 변수를 줄여나간다.

-mixed selection: forward selection과 backward selection과의 혼합물이다. 처음에는 forward selection으로 변수가 없이 시작을 한다. 그리고 어느 한 시점에서 p-value가 어느 선을 넘어가면 변수를 삭제하기 시작한다.

또는 2번에 대한 설명에서 언급된 Partial F test을 사용할 수도 있다. 예를 들어, k 번째 변수의 유의성을 판단하기 위해 아래와 같은 가설검정을 진행한다고 하자.

$$H_0 : \beta_k = 0 \; vs \; H_1 : \beta_k \neq 0$$

$$Model(R) : \; y_i = \beta_0 + \beta_1 x_{1i} + \cdots \beta_{k-1} x_{k-1,i} + \beta_{k+1} x_{k+1,i} + \cdots \beta_p x_{pi} + \epsilon_i$$

$$Model(F) : \; y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i$$

(R) 모델 하에서의 SSE를 R을 통해서, (F) 모델 하에서의 SSE를 R을 통해서 구한 뒤, $F = \frac{(SSE^R - SSE^F)/(df^R - df^F)}{SSE^F/df^F} \sim F(df^R - df^F, df^F)$을 이용하여 해당 변수에 대한 유의성을 판단하면 된다.

**** 3*번에 대한 설명**

단순선형 회귀와 마찬가지로, RSE와 결정계수가 모델의 성능 지표로 활용될 수 있다. 하지만 이들은 그리 좋은 지표가 아니다. 우선 RSE는 앞에서 언급 했듯이, 비율이 아닌 값으로 계산되기 때문에 상대적인 비교가 어려울 수 있다. 예를 들어, kg 단위로 기록된 자료와 g 단위로 기록된 자료의 RSE는 전자의 그것이 더 크겠지만 사실 이 크기의 차이는 적합의 차이에서 비롯된 것은 아니다. 결정계수 또한 변수가 의미 있든 없든, 추가 됨에 따라서 증가하기 때문에 적절하지 못한 지표이다. 그렇다면 어떤 지표를 많이 사용할까?

- Adjusted R^2(수정된 결정계수): $R^2_{a,p} = 1 - \frac{SSE_p(n-p)}{SSTO/(n-1)} = 1 - \frac{MSE_p}{SSTO/(n-1)}$
  수정된 결정 계수가 최대가 되거나, 안정화되기 시작하는 변수들을 선택하자.
- Mallow' C_p: $C_p = \frac{SSE_p}{MSE(X_1, \cdots, X_{r-1})} - (n - 2p)$
  작고 p(변수의 개수)에 가까운 C_p가 좋다.
- AIC_p, BIC_p: $AIC_p = nlogSSE_p - nlogn + 2p$, $BIC_p = nlogSSE_p - nlogn + plogn$
  작은 값이 좋다.
- PRESS: $\sum_{i=1}^n (Y_i - \hat{Y}_{i(i)})^2 = \sum_{i=1}^n \left(\frac{Y_i - \hat{Y}_i}{1 - h_{ii}}\right)^2$ where $\hat{Y}_{i(i)}$ denotes the ith fitted value without ith value is omitted, h_{ii} is diagonal element in Hat matrix
  작은 PRESS 값이 좋다.

**** 4*번에 대한 설명 *(Predictions)***

multiple regression model을 적합했으면 3.21을 바로 사용하면 될 듯 하다. 하지만 여기에는 세 가지 불확실성이 있다.

1. 우리가 추정한 $\hat{\beta}_0, \cdots, \hat{\beta}_p$는 $\beta_0, \cdots, \beta_p$에 대한 추정지이다. 다시 말해서 least squares plane인 $\widehat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$는 true population regression plane인 $f(X) = \beta_0 + \cdots + \beta_p X_p$의 estimate에 불과하다.
2. 물론 $f(X)$에 대해서 linear model을 추정하는 것은 항상 실생활에서 approximation이기 때문에 model bias라고 불리는 reducible error의 위험성이 존재한다. 따라서 우리는 'best'한 linear 'approximation'을 사용한다.

3. 아무리 우리가 $f(X)$와 $\beta_0, \cdots, \beta_p$을 안다고 하더라도 random error인 ϵ때문에 완벽히 예측할 수 없다. 우리는 이것을 챕터2에서 irreducible error라고 불렀다. Y가 $\widehat{Y}$로부터 얼마나 떨어져 있을까? 우리는 이를 위해서 prediction intervals을 사용한다. prediction interval은 confidence보다 항상 넓은데 이는 $f(X)$를 추정의 error(reducible error)와 개개인의 점이 실제 population regression plane과 얼마나 떨어져 있는지의 불확실성(irreducible error)을 포함하기 때문이다.

3.3 Other Considerations in the Regression Model

3.3.1 Qualitative Predictors

우리는 여태까지 수치형 변수만을 생각했지만 항상 그렇지는 않고 범주형 변수도 있을 수 있다.

Predictors with Only Two Levels

예를 들어 우리가 남, 여 사이에 card balance 차이를 알고 싶다고 하자. 여기서는 우선 다른 변수를 무시한다. simple linear regression에서 하나의 변수가 범주형 변수인 경우를 생각해보자. 우리는 그저 두 개의 값을 가지는 indicator 또는 dummy variable을 만든다. 예를 들어서, gender변수에 기반하여 dummy 변수를 생성하면

$$x_i = \begin{cases} 1 & \text{if ith person is female} \\ 0 & \text{if ith person is male} \end{cases} \tag{3.26}$$

이고 이것을 이용하여 regression equation을 만들면 아래와 같다.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if ith person if female} \\ \beta_0 + \epsilon_i & \text{if ith person is male} \end{cases} \tag{3.27}$$

이제 β_0은 남자의 평균 credit card balance으로 해석할 수 있고 $\beta_0 + \beta_1$은 여자의 평균 credit card balance으로 해석할 수 있다.

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

TABLE 3.7. *Least squares coefficient estimates associated with the regression of* `balance` *onto* `gender` *in the* `Credit` *data set. The linear model is given in (3.27). That is, gender is encoded as a dummy variable, as in (3.26).*

table 3.7은 3.27와 관련된 계수 추정치와 다른 정보를 보여준다. 남자에 대한 평균 credit card debt은 509.80(절편)으로 추정되고 여성은 19.73 + 509.80으로 추정된다. 하지만 더비 변수에 대한 p-value가 매우 높음을 주목하자. 이것은 성별 사이에 credit card balance 차이에 대한 통계적인 증거가 없음을 의미한다.

여자를 1로 할지 0으로 할지의 결정은 임의적이고 회귀 적합에 아무런 영향이 없지만 계수에 대한 해석이 달라진다. 만약 우리가 남자를 1, 여자를 0으로 한다면 β_0, β_1은 각각 529.53, -19.73일 것이고 이것은 남자의 credit card balance 추정치가 529.53-19.73=509.80으로 나올 것이다. 대안적으로 0/1 코딩 말고 우리는 다음과 같이 코딩을 할 수도 있다.

$$x_i = \begin{cases} 1 & \text{if ith person is female} \\ -1 & \text{if ith person is male} \end{cases}$$

그리고 이 더비 변수를 활용하여 regression equation을 만들면 다음과 같다.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if ith person is female} \\ \beta_0 - \beta_1 + \epsilon_i & \text{if ith person is male} \end{cases}$$

이제 β_0은 성별 효과를 무시한 전반적인 credit card balance 평균으로 해석되고 β_1은 여자가 위에 위에 있는 평균과 남자가 아래에 있는 평균의 양을 의미한다. 여기서 β_0의 추정치는 519.665인데 이것은 남성과 여성의 평균이다. β_1의 추정치는 9.865인데 이것은 남성과 여성의 평균 차이인 19.73의 반이다. 결국 더미 변수를 어떻게 설정을 하든 마지막 예측은 동일하다. 계수가 어떻게 해석되는지에서 차이가 있을 뿐이다.

Qualitative Predictors with More than Two Levels

만약 범주형 변수가 두 개 이상의 levels을 가진다면 하나의 더미 변수는 모든 값을 대표하지 못할 것이다. 이러한 상황에서 우리는 추가적인 더미 변수를 생성한다. 예를 들어 ethnicity

변수에서 다음과 같이 더미 변수를 생성한다.

$$x_{i1} = \begin{cases} 1 & \text{if ith person is Asian} \\ 0 & \text{if ith person is not Asian} \end{cases} \tag{3.28}$$

$$x_{i2} = \begin{cases} 1 & \text{if ith person is Caucasian} \\ 0 & \text{if ith person is not Caucasian} \end{cases} \tag{3.29}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if ith person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if ith person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if ith person is African American} \end{cases} \tag{3.30}$$

이제 β_0은 African Americans의 평균 credit card balance으로 해석되고 β_1은 Asian과 African American과의 평균 valance차이로 해석되고 β_2은 Caucasian과 African American과의 평균 balacne차이로 해석된다. levels의 갯수보다 항상 한개 더 적은 dummy variable이 있을 것이다. 더미 변수가 없는 level(여기서는 African American)은 baseline이라고 불린다.

결국 범주형 자료에 대해서 회귀분석을 진행할 때, 각 범주를 어떤 식으로 코딩하느냐에 따라서 회귀 계수에 대한 해석이 달라짐에 주의해야 한다.

3.3.2 Extensions of the Linear Model

standard linear regression model인 3.19은 꽤나 좋을 수 있지만 실제 사용하는데 굉장히 제한적인 가정을 한다. 가장 중요한 가정 중 두 개는 predictors와 response간의 관계를 additive하고 linear하다고 보는 것이다. additive 가정은 예측 변수 X_j가 Y에 미치는 변화의 효과는 다른 예측 변수의 값과 독립이라고 본다. linear 가정은 반응 변수 Y에서의 변화는 X_j에서의 한 단위(one-unit)변화에 의한 것인데 이 변화의 정도가 X_j에 상관 없이 항상 상수(constant)라고 본다. 이 책에서 우리는 이러한 두 가정을 relax하는 여러가지 복잡한 방법을 살펴본다. 여기서 우리는 linear model을 확장하기 위한 몇 가지 common lcassical approaches을 살펴본다.

Removing the Additive Assumption

이전 분석에서 우리는 TV와 radio 모두 sales와 연관이 있다고 결론 내렸다. linear model에서는 TV에서의 한 단위 증가에 따른 평균 효과는 항상 β_1인데 이것은 radio에 사용되는 돈과 관련없이 항상 성립한다.

그런데 이러한 simple model은 틀릴 수도 있다. 예를 들어 라디오 광고에 돈을 쓰는 것이 사실은 TV 광고의 효과성을 증가시킨다고 가정해보자. 이러한 상황에서 고정된 $100,000이 있다고 할 때, 라디오에 반을 쓰고 티비에 반을 쓰는 것이 티비나 라디오에 모든 돈을 투자하는 것보다 세일을 더 증가시킬 수 있다. 마케팅에서는 이것을 synergy effect라고 부르고 통계에서는 이것을 interaction effect라고 부른다. 두 개의 변수를 가지고 있는 standard linear regression model을 생각해보자.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

이 모델에 따르면 X_1에서 한 단위 증가는 Y에서 평균적으로 β_1단위 증가를 가져 올 것이다. 여기서 X_2의 존재가 이것을 바꾸지 않음에 주목하자. 다시 말해서 X_2에 상관 없이 β_1 단위 만큼 증가하는 것은 고정된 효과다. 이 모델을 interaction effect을 포함하는 모델로 확장시키면 다음과 같다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \tag{3.31}$$

어떻게 interaction term이 additive 가정을 relax할까? 3.31이 다음과 같이 다시 쓰인 다는 것에 주목하자.

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned} \tag{3.32}$$

여기서 $\tilde{\beta}_1$은 X_2에 따라서 변화하기 때문에 Y에 대한 X_1의 영향이 더 이상 상수가 아니다. X_2를 조정하는 것은 X_1, Y 모두에게 영향을 줄 것이다. 전에는 β_1은 상수였다. 하지만 이제 $\tilde{\beta}_1$은 상수가 아니라 X_2에 따라서 변화한다. 따라서 X_2가 변함에 따라서 X_1, Y에 영향을 줄 것이다.

다시 광고 예제로 돌아오자. 라디오, 티비, 그리고 이 둘의 interaction term을 포함하는 sales에 대한 linear model은 다음과 같은 형태이다.

$$\begin{aligned} sales &= \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times (radio \times TV) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times radio) \times TV + \beta_2 \times radio + \epsilon \end{aligned} \tag{3.33}$$

interaction term을 포함하는 모델이 main effects만을 포함하는 모델보다 더 우수하다. $TV \times radio$ term에 대한 p-value는 매우 낮으므로 이것은 $\beta_3 \neq 0$을 강력하게 지지한다. 다시 말해서 true relationship은 additive가 아니라는 것이 분명하다.3.33에 대한 결정계수는

96.8%인 반면에 TV와 radio만을 변수로 하는 모델의 결정계수는 89.7%이다.

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

TABLE 3.9. *For the* Advertising *data, least squares coefficient estimates associated with the regression of* sales *onto* TV *and* radio*, with an interaction term, as in (3.33).*

table 3.9로부터 TV의 \$1000 증가는 sales의 $(\hat{\beta}_1 + \hat{\beta}_3 \times radio) \times 1000 = 19 + 1.1 \times radio$ 만큼의 증가를 가져오고 radio의 \$1000 증가는 sales의 $(\hat{\beta}_2 + \hat{\beta}_3 \times TV) \times 1000 = 29 + 1.1 \times TV$ 만큼의 증가를 가져온다.

이 예시에서는 티비, 라디오, interaction term과 관련된 p-value는 통계적으로 의미가 있어 이 세개의 변수를 모델에 포함시키는 것이 합당하다. 하지만 interaction term의 p-value는 매우 낮지만 main effects의 p-value는 큰 경우가 있다. hierarchical principle는 우리가 interaction을 모델에 포함시킨다면, 우리는 main effects의 p-value가 유의미하지 않더라고 그것을 포함시켜야 한다고 말한다.

물론 범주형 변수에서도 interactions의 개념은 적용될 수 있다. 사실, 수치형 변수와 범주형 변수의 조합은 좋은 해석을 가져올 수 있다. Section 3.3.1에서 Credit data set에서 income(수치형)과 student(범주형)을 사용해서 balance을 예측하는 상황을 생각해보자. interaction term이 없다면 다음과 같은 형태일 것이다.

$$balance_i \approx \beta_0 + \beta_1 \times income_i + \begin{cases} \beta_2 & \text{if ith person is a student} \\ 0 & \text{if ith person is not a student} \end{cases}$$

$$= \beta_1 \times income_i + \begin{cases} \beta_0 + \beta_2 & \text{if ith person is a student} \\ \beta_0 & \text{if ith person is not a student} \end{cases} \tag{3.34}$$

이것이 두 개의 평행한 선을 적합시킨다는 것에 주목하자. 하나는 학생일때, 하나는 학생이 아닐 때 적합하는 것이다. 기울기는 β_1으로 동일하고 절편만 다른 두 개의 선이다. 이것은 figure 3.7 왼쪽 그림에 나타나 있다.

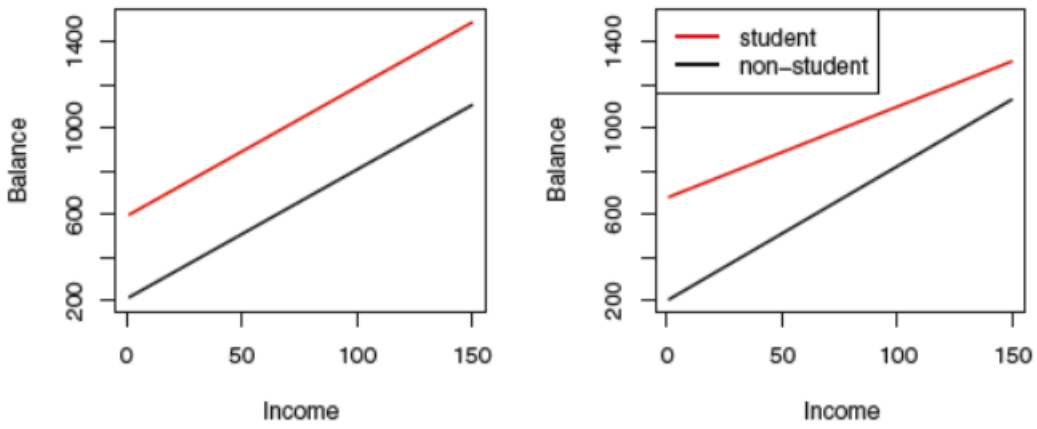

FIGURE 3.7. *For the* `Credit` *data, the least squares lines are shown for prediction of* `balance` *from* `income` *for students and non-students.* Left: *The model (3.34) was fit. There is no interaction between* `income` *and* `student`. Right: *The model (3.35) was fit. There is an interaction term between* `income` *and* `student`.

model now becomes

두 직선이 평행하다는 것은 income에서의 한단위 증가가 개인이 student인지 여부에 의존하지 않음을 의미한다. 이것은 모델의 심각한 한계를 포함하는데 사실 income에서의 변화는 학생일때와 학생이 아닐 때의 credit card balance에서의 차이를 가져오기 때문이다.

이러한 한계는 interaction variable을 추가함으로써 해결될 수 있다.

$$balance \approx \beta_0 + \beta_1 \times income_i + \begin{cases} \beta_2 + \beta_3 \times income_i & \text{if student} \\ 0 & \text{if not student} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times income_i & \text{if student} \\ \beta_0 + \beta_1 \times income_i & \text{if not student} \end{cases} \tag{3.35}$$

다시 우리는 두 개의 regression lines을 가지고 있지만 기울기와 절편이 다르다. 이것은 income에서의 변화가 학생일때, 학생이 아닐 때의 credit card balances에 각각 다르게 영향을 줌을 의미한다. 학생일 때의 기울기가 더 낮음에 주목해보자. 이것은 income의 증가가 학생이 아닐때에 비해서 학생일 때의 credit card balance가 더 적게 증가한다는 것을 말한다.

Non-linear Relationships

이전에도 말했 듯이, 3.19은 반응변수와 예측변수 간에 선형의 관계가 있다고 가정한다. 하지만 어떤 경우는 선형이 아닐 수도 있다. 여기서 우리는 선형이 아닌 관계에 적용할 수 있는 확장된 linear model에 대해서 polynomial regression을 이용하여 얘기해본다.

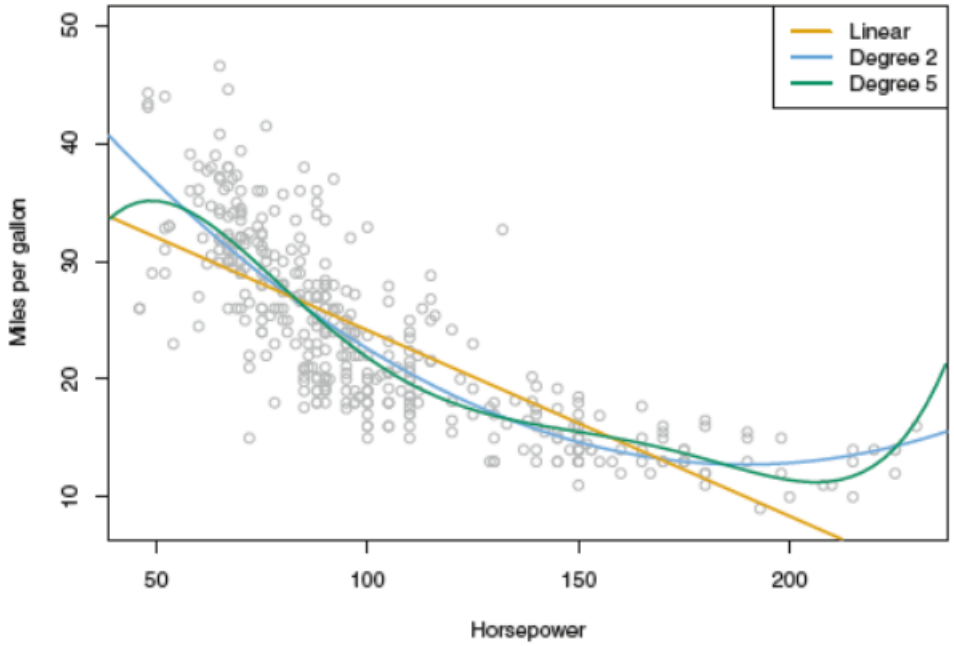

FIGURE 3.8. *The* `Auto` *data set. For a number of cars,* `mpg` *and* `horsepower` *are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes* $\texttt{horsepower}^2$ *is shown as a blue curve. The linear regression fit for a model that includes all polynomials of* `horsepower` *up to fifth-degree is shown in green.*

orange line represents the linear regression fit. There is a pronounced relationship between `mpg` and `horsepower`, but it seems clear that this relationship is in fact non-linear: the data suggest a curved relationship. A simple approach for incorporating non-linear associations in a linear model is to include transformed versions of the predictors in the model. For example, the points in Figure 3.8 seem to have a *quadratic* shape, suggesting that a model of the form

figure 3.8은 horsepower와 mpg간에 관계를 보여준다. 사실 이 관계는 non-linear함이 분명하다. 이러한 non-linear관계를 linear model에 포함시키는 방법은 transformed versions of the predictors를 모델에 포함시키는 것이다. 예를 들어 figure 3.8의 점은 quadratic 모양을

가지는 것 처럼 보이는데 이것은

$$mpg = \beta_0 + \beta_1 \times horsepower + \beta_2 \times horsepower^2 + \epsilon \tag{3.36}$$

의 모델 형태가 더 나은 적합을 가져올 것이다. 3.36은 mpg을 horsepower의 non-linear function을 이용하여 예측하는데 여전히 linear model이다! 다시 말해서 3.36은 그저 $X_1 = horsepower, X_2 = horsepower^2$ 의 multiple linear regression이다. 따라서 우리는 linear regression software을 $\beta_0, \beta_1, \beta_2$ 를 추정하여 non-linear fit의 결과를 내기 위해 사용할 수 있다. 그렇다면 3승, 4승, 5을 이런 애들을 포함시키면 어떻게 될까? figure 3.8에서 초록색 선이 5승까지 포함한 모델인데 불필요하게 wiggly하다. 이러한 방법을 polynomial regression 이라고 부르고 나중에 좀 더 자세히 알아보도록 하자.

3.3.3 Potential Problems

linear regression model을 적합할 때는 다음과 같은 문제점이 생길 수 있다.

1. Non-lineartiy of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity

회귀 분석을 진행한 뒤에는, 초반에 세운 가정이 맞는지 확인하는 절차가 필요하다. 선형성에 대한 가정, 오차의 분산이 상수라는 가정, 오차가 서로 독립이라는 가정, 오차가 정규분포를 따른 다는 과정 (가설검정, 신뢰구간 등의 추론을 할 시에) 을 확인할 필요가 있다. 위의 가정들이 성립하지 않는다면 회귀분석이 가지는 의미가 줄어들기 때문이다.

1. Non-linearity of the Data

<Residual Plot>
linear regression model은 반응변수와 예측변수 간에 straight line 관계가 있다고 가정한다. 하지만 만약 linear 관계와 거리가 멀다면? 우리가 한 모든 것들이 의미가 없어진다.

Residual plots은 이러한 non-linearity을 판단하는 유용한 그래픽 툴이다. residual plot은 $e_i = y_i - \hat{y}_i$와 x_i와의 scatter plot이다. multiple regression model에서는 여러개의 예측변수가 있기 때문에 대신 residuals과 predicted(or fitted) values $\hat{y}_i$와의 plot을 그려본다. 이상적으로, 잔차 그림은 일정한 패턴을 보여서는 안된다. 이는 잔차의 의미를 생각해보면 알 수 있다. 잔차란, 회귀 직선으로 적합을 하고 남은 부분을 의미한다. 이 남은 부분이 일정한 패턴을 보인다면, 적합한 회귀 직선이 이러한 패턴을 잡아내지 못한다는 뜻이다.

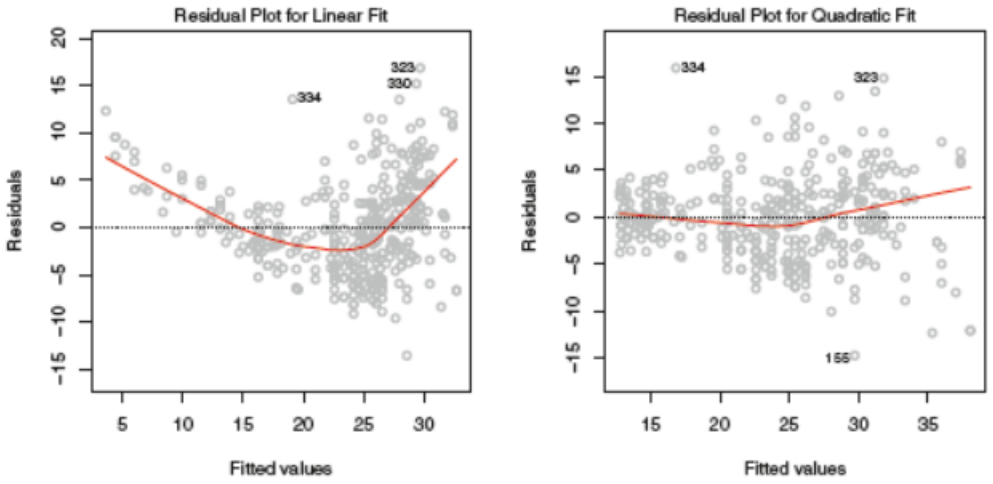

FIGURE 3.9. *Plots of residuals versus predicted (or fitted) values for the* `Auto` *data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend.* Left: *A linear regression of* `mpg` *on* `horsepower`. *A strong pattern in the residuals indicates non-linearity in the data.* Right: *A linear regression of* `mpg` *on* `horsepower` *and* $\texttt{horsepower}^2$. *There is little pattern in the residuals.*

왼쪽 그림은 분명하게 U자 모양이다. 이것은 데이터에서 non-linearity을 강하게 보여준다. 반면에 오른쪽은 quadratic term을 포함하는 linear 모델인데 여기서는 residual의 패턴이 거의 없다. 이것은 quadratic term이 데이터에 대한 적합도를 높여주었다는 것을 보여준다. 종종, 잔차 대신 표준화한 잔차를 사용하기도 한다.

$$e_i^* = \frac{e_i - \overline{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

사실, e_i의 분산의 estimate는 MSE가 아니라 $(1-h_{ii})MSE$ 이다. 따라서 위의 식은 완벽한 표준화가 아니라는 관점에서 Semi Studentized Residuals이라고도 불린다. 만약 residual plot이 데이터에서 non-linear associations을 나타낸다면, $logX, \sqrt{X}, X^2$ 등의 non-linear transformations을 사용할 수 있다. 나중에 더 살펴보자.

<EVP>

2. Correlation of Error Terms

linear regression의 중요 가정 중 하나는 error terms, $\epsilon_1, \cdots, \epsilon_n$이 uncorrelated라는 것이다. 이것이 의미하는 것이 무엇일까? 예를 들어 errors가 uncorrelated라면 ϵ_i가 positive하다는 것은 ϵ_{i+1}에 대해서 정보를 거의 주지 않는다. 추정된 회귀 계수에 대해서 계산되는 standard errors는 error가 독립이라는 가정에 기초한다. 따라서 오차의 독립성을 확인해줘야 한다.

왜 이러한 상관성이 error term에서 일어나는 것일까? 이러한 상관성은 time series data에서 흔히 일어난다. 이것을 판단하기 위해서, 모델의 잔차를 시간에 대한 함수로써 plot을 그려본 후, 어떠한 패턴이 없으면 error는 독립이라는 것이고 그렇지 않다면 error는 correlated라는 것이며 여기서 우리는 잔차에서 tracking을 볼 수 있을 것이다. 다시 말해, 인접한 잔차는 비슷한 값을 가질 것이다. figure 3.10을 살펴보자.

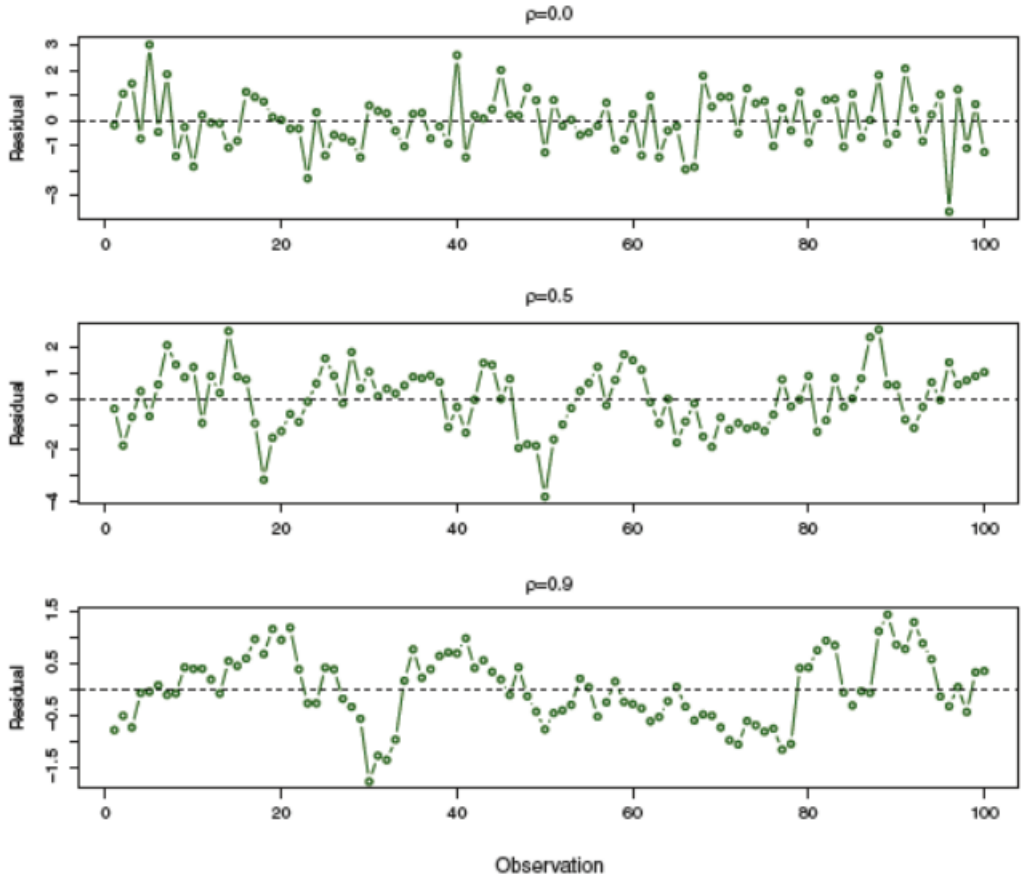

FIGURE 3.10. *Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.*

맨 위는 residual의 패턴이 일정하지 않아 error가 독립인 경우이고 맨 아래는 반대의 경우이다.

3. Non-constant Variance of Error Terms

linear regression model에서의 또 다른 중요한 가정 중 하나는 error terms이 constant variance, $Var(\epsilon_i) = \sigma^2$를 가지고 있다는 것이다. linear model에서의 standard errors, confidence intervals, hypothesis test는 모두 이러한 가정에 의존한다.

불행하게도, 종종 error term의 분산은 상수가 아니다. 예를 들어, error term의 분산은 반응 변수의 값이 증가함에 따라서 증가할 수도 있다. 이러한 non-constant variances, 또는 heteroscedasticity는 잔차 plot에서 funnel shape의 존재로 파악할 수 있다. figure 3.11에 예시가 있다.

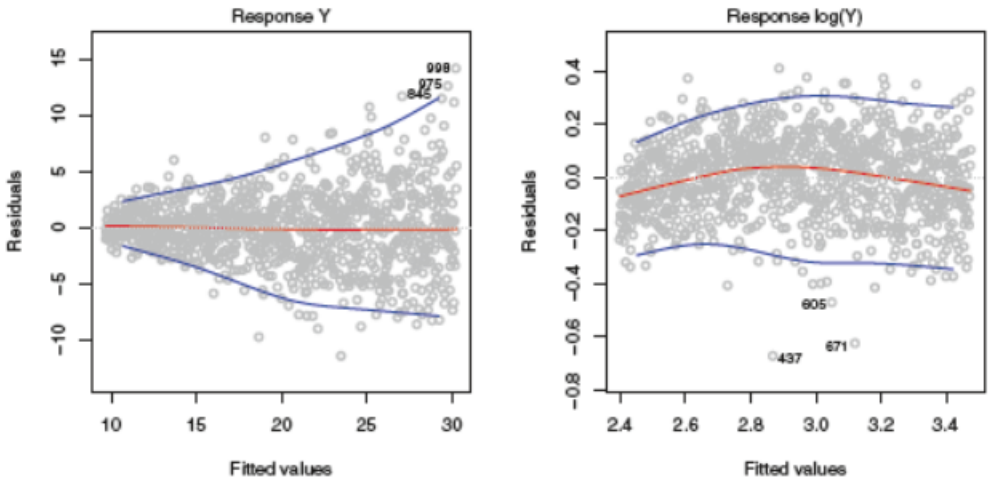

FIGURE 3.11. *Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns.* Left: *The funnel shape indicates heteroscedasticity.* Right: *The response has been log transformed, and there is now no evidence of heteroscedasticity.*

figure 3.11의 왼쪽 그림을 보면 fitted values가 증가할 수록 잔차의 규모도 커지고 있다. 이러한 상황에 마주치면, 가능한 해결책은 반응변수 Y을 $logY, \sqrt{Y}$ 등의 concave function으로 변환하는 것이다. 이러한 변환은 반응 변수를 크게 수축시켜서 heteroscedaticity의 감소를 가져올 것이다. 오른쪽 그림은 $logY$를 이용하여 변환한 결과이다. 이제 잔차는 constant variance을 가지는 것 처럼 보인다.

좀 더 객관적인 방법으로, Brown-Forsythe Test 와 Breusch-Pagan Test(or Cook-Weisberg Test)가 있다.

먼저 Brown-Forshythe 검정을 살펴보자. 이 검정은 잔차의 변동성에 기반한다. 우선 데이터를 X에 따라 두 부분으로 나누는데, 한 그룹이 상대적으로 작은 X을 포함하도록, 다른 그룹은 상대적으로 큰 X을 포함하도록 나눈다. 만약 오차의 분산이 X에 따라서 증가하거나 감소한다면 어느 한 그룹의 잔차가 다른 그룹의 잔차보다 더 변동적일 것이다. 그에 따라서, 어떤 한 그룹의 평균(중앙값) 주위의 편차에 대한 절댓값이 다른 그룹의 그것보다 더 클 것이다. 이러한 아이디어로 two-sample t test을 진행하여 어느 한 그룹의 편차 절대값이 다른 그룹의 그것보다 통계적으로 다른지 검정한다.

이 검정의 장점은, 오차의 정규성에 의존하지 않는 점, 잔차는 일반적으로 정규분포를 따르지 않지만 test statistic인 t^*은 오차의 분산이 상수이고 표본의 개수가 매우 적지 않는

한, 근사적으로 t 분포를 따르는 점이다. (증명은 생략)

$$n = n_1 + n_2$$

$$d_{i1} = |e_{i1} - \tilde{e_1}|, d_{i2} = |e_{i2} - \tilde{e_2}|$$

$$t^*_{BF} = \frac{\overline{d_1} - \overline{d_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \dot{\sim} t(n-2)$$

$$where\, s^2 = \frac{\sum(d_{i1} - \overline{d_1})^2 + \sum(d_{i2} - \overline{d_2})^2}{n-2}$$

t^*_{BF} 값이 충분히 커서 그에 따른 p-value가 유의수준 α보다 작다면, 통계적으로 유의하게, 오차의 분산이 상수가 아니라고 볼 수 있다.

다음으로 Breusch-Pagan 검정을 살펴보자. 이 검정은 Brown-Forshythe 보다는 필요한 가정이 몇 가지 있다. 오차항의 독립, 정규성, 그리고 오차의 분산인 σ_i^2가 X에 따라서 아래와 같은 관계를 가진다고 가정한다.

$$log\sigma_i^2 = \lambda_0 + \lambda_1 X_i$$

위 식은 λ_1의 부호에 따라, X의 level이 변하면 σ_i^2가 증가하거나 감소함을 의미한다. 오차항의 분산이 상수라면 $\lambda_1 = 0$이어야 한다. 이러한 논리에 따라 귀무가설과 대립가설, 그리고 test statistic은 아래와 같다.

$$H_0 : \lambda_1 = 0 \; vs \; H_1 : \lambda_1 \neq 0$$

$$\chi^2_{BP} = \frac{SSR^*}{2} \div \left(\frac{SSE}{n}\right)^2 \dot{\sim} \chi^2(1)$$

$$where\, SSR^*\, is\, SSR\, when\, regression\, of\, e^2\, onto\, X$$

4. Normality(For Inference)

오차의 정규성은 신뢰구간, 가설검정 등 통계적 추론을 위해 필수적으로 확인해야 하는 가정이다. 예를 들어, 어떤 회귀 계수에 대해서 p-value을 말할 때, 이는 귀무가설 하의 '분포', 즉 오차항의 정규분포를 통해 도출된 p-value이므로 이를 위해 정규성의 가정이 충족되는지 확인해야 한다. 회귀분석에서는 이를 Normal Probability Plot으로 확인하고 R output에서 확인할 수 있다. 어떠한 원리인지 간략하게 살펴보자.

$$Suppose\, \epsilon_1, \cdots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$$

$$Scale\ it\ as\ \frac{\epsilon_1}{\sigma}, \cdots, \frac{\epsilon_n}{\sigma} \overset{iid}{\sim} N(0,1)$$

$$Sort\ by\ value\ \frac{\epsilon_{(1)}}{\sigma}, \cdots, \frac{\epsilon_{(n)}}{\sigma} \overset{iid}{\sim} N(0,1)\ [ordered\ statistics]$$

$$think\ of\ cumulative\ probability(as\ empirical\ probability):\ \frac{1}{n}, \cdots, \frac{n}{n}$$

$$compare\ z\ value\ of\ cumulative\ probability\ with\ \frac{\epsilon_{(1)}}{\sigma}, \cdots, \frac{\epsilon_{(n)}}{\sigma}; i.e$$

$$z(\frac{1}{n}), \cdots, z(\frac{n}{n})\ vs\ \frac{\epsilon_{(1)}}{\sigma}, \cdots, \frac{\epsilon_{(n)}}{\sigma}\ but\ z(\frac{n}{n}) = \infty,\ so\ scale\ it\ by$$

$$z(\frac{k-0.375}{n+0.25})\ vs\ \frac{\epsilon_{(k)}}{\sigma}\ where\ we\ estimate\ \sigma\ by\ MSE\ and\ \epsilon_{(k)}\ by\ e_{(k)}$$

$$\therefore z(\frac{k-0.375}{n+0.25}) \approx \frac{e_{(k)}}{MSE}$$

$$MSEz(\frac{k-0.375}{n+0.25}) \approx e_{(k)}$$

Normal probability plot의 x축은 Expected인데 이것이 위 식의 좌변이고 Residual은 위 식의 우변이다. 따라서 좌변과 우변이 비슷할 때, 즉 그래프 상에서 직선의 관계가 형성된다면 두 근사 값이 비슷하다는 뜻이므로 오차의 정규성 가정을 충족한다고 볼 수 있다.

5. Outliers / Influential Data

<Residual Plot>

outlier는 다수의 데이터들이 가지는 성향과 다른 특징을 가지는 데이터를 말한다. 이상치는 회귀 모델에서 영향을 크게 미칠 수도 있으며, 주제에 따라서는 이러한 이상치를 감지하는 것이 중요할 수도 있다.

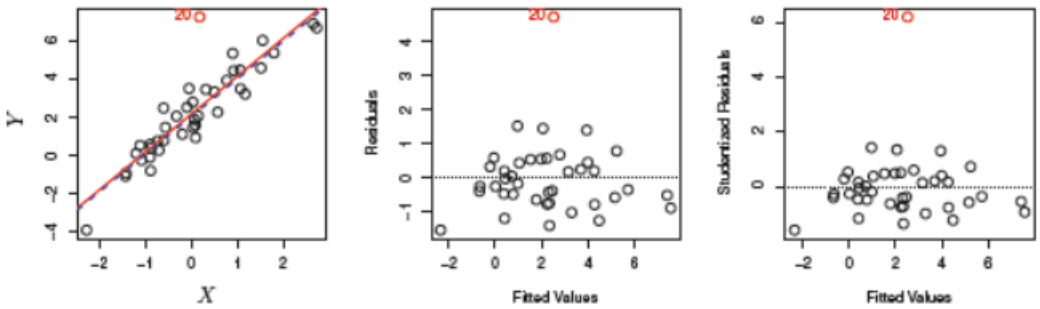

FIGURE 3.12. Left: *The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue.* Center: *The residual plot clearly identifies the outlier.* Right: *The outlier has a studentized residual of 6; typically we expect values between −3 and 3.*

figure 3.12에서 빨간색 포인트는 전형적인 outlier을 나타낸다. 왼쪽 그림에서 빨간색 선은 LSE로 적합한 regression fit이고 파란색 선은 outlier을 제외한 LSE regression fit이다. 이 경우에서는 outlier가 거의 영향을 미치지 않았다. 하지만 outlier가 least squares fit에 큰 영향을 미치지 않더라도 다른 문제가 일어날 수 있다. 예를 들어, 이 예제 에서는 RSE가 outlier를 포함할 때는 1.09지만 제거되었을 때는 0.77이다. RSE는 신뢰구간이나 p-value 계산에 사용되기 때문에 이러한 극적인 변화는 적합의 해석에 대해서 implications을 가진다. 비슷하게 결정계수도 outlier을 포함했을 때 훨씬 내려갔다.

residual plots은 이러한 outliers을 파악하는데 사용된다. figure 3.12 가운데 그림에서 outlier은 분명하게 보인다. 하지만 실제에서는 우리가 어떠한 점을 outlier로 판단하기 전에 잔차가 얼마나 커야 하는지 결정하기 어려울 수 있다. 이러한 문제점을 해결하기 위해 residual plot을 그리는 대신에 studentized residual plot을 그리는데 이는 각각의 잔차 e_i에 그것의 estimated standard error을 나눔으로써 구한다. 이러한 studentized residuals이 3보다 큰 관측치는 outliers일 가능성이 있다.

<Deleted Residuals>
또 다른 방법으로는 Deleted Residuals이 있다. i번째 Deleted Residuals이란, i번째 데이터를 포함시키지 않은 채, 모델을 적합시킨 후, 이 모델을 이용하여 만든 잔차를 의미한다.

$$Let\, d_i = Y_i - \hat{Y}_{i(i)}$$

$$t_i = \frac{d_i}{s(d_i)} = \cdots = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}\right]^2 \sim t(n-p-1)$$

위의 t 분포를 활용하여, $|t_i|$가 큰 데이터를 이상치라고 판단한다.

<DFFITS>
DFFITS는 아래와 같이 정의된다.

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}} = t_i \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

보통, 표본의 개수가 적거나 중간 정도일 때, $DFFITS_i > 1$인 데이터는 눈여겨 봐야하고, 표본의 개수가 많을 때는 $DFFITS_i > 1\sqrt{p/n}$인 데이터를 주목해서 봐야 한다.

<Cook's Distance>
Cook's Distance는 아래와 같이 정의된다.

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE} = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

$D_i > 0.5$인 경우에 조사하는 것이 좋고, $D_i > 1$이라면 항상 조사를 해봐야 한다.

6. High Leverage Points

우리는 방금 outliers는 x_i가 주어졌을 때, 흔하지 않은 y_i를 의미함을 배웠다. 이와는 반대로 high leverage한 관측치는 흔하지 않은 x_i을 가진다.

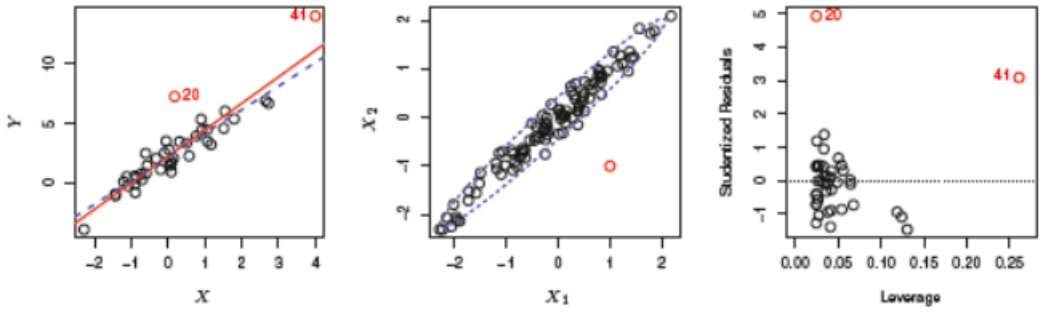

FIGURE 3.13. Left: *Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed.* Center: *The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage.* Right: *Observation* 41 *has a high leverage and a high residual.*

figure 3.13을 보자. 왼쪽 그림에서 관측치 41은 high leverage을 가지고 있다. 빨간 선과 파란 선을 비교해 볼 때, high leverage을 가지는 관측치를 제거하는 것은 least squares line에 큰

영향을 줌을 알 수 있다. 사실, high leverage인 관측치는 추정된 regression line에 상당한 영향을 미친다.

figure 3.13에서 가운데 그림을 보자. 사실, 빨간색 점에서 X_1 값만 보면 이 값은 정상적인 범위에서 떨어져 있지 않다. 마찬가지로 X_2 값도 마찬가지이다. 하지만 이를 동시에 본다면 정상 범위에서 떨어져 있는 값임을 알 수 있다. 여기서는 변수가 2개여서 쉽게 알아낼 수 있었지만 multiple regression에서 고차원이 될 수록 한번에 그림으로 표현하기 어렵기 때문에 이를 알아내기가 어렵다.

다중선형회귀에서 높은 leverage을 가지는 데이터를 찾아내기 위해, h_{ii}를 활용한다. 여기서 h_{ii}는 hat matrix($= X(X^{'}X)^{-1}X^{'}$)의 i번째 diagonal element이다.

$$h_{ii} = \frac{1}{n} + (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{'}(\boldsymbol{X_c^{'}X_c})^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \tag{2}$$

$$where\ \boldsymbol{X}_c\ is\ n \times (p-1)\ design\ matrix\ for\ centered\ variables(without\ intercept)$$

위의 h_{ii}으로부터 x_i가 정확히 centered 되었다면 $h_{ii} = \frac{1}{n}$이 될 것이고 멀어질수록 $\frac{1}{n}$보다 커질 것이며 1에 가까워질수록 $Var(e_i) = (1 - h_{ii})\sigma^2$에서 우변이 0에 가까어지고 $Var(e_i) = 0$이 되는 극단적인 상황을 생각해볼 수 있다. 즉, leverage가 너무 커져서, $y_i = \hat{y}_i$가 되어, 그만큼 해당 데이터의 영향력이 크다는 것을 의미한다.

figure 3.13의 오른쪽 그림은 studentized residuals와 h_i의 plot을 그린 것이다. 관측치 41은 high leverage statistic일 뿐만 아니라 매우 높은 studentized residual을 가진다. 다시 말해서 outlier이기도 하고 또한 high leverage 관측치라는 것이다. 이것은 특히 위험한 조합이다!

7. Collinearity

Collinearity은 독립변수 2개 간의 선형성이 존재하는 것을 의미한다. 예를 들어서, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$의 모델에서 $x_1 = 2x_2$와 같은 경우를 말한다. Multicollineartiy은 이러한 collinearity가 여러 변수들 간에 존재하는 경슬 의미한다. 예를 들어서, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$의 모델에서 $x_3 = -x_1 + 3x_4$와 같은 경우이다.

회귀분석을 진행할 때에 이렇게 다중공선성이 존재한다면 모델에 심각한 영향을 미칠 수 있다. 따라서 이를 감지하기 위해서 VIF라는 지표를 많이 사용한다. VIF는 추정된 회귀계수의 분산이 변수들 간에 다중공선성이 없다고 가정을 했을 때의 분산에 비해서 얼마나 부풀려졌는지를 측정한다. 관용적으로 VIF가 10을 넘는다면 다중공선성이 심각한 문제를 일으키고 있다고 판단한다. VIF는 다음과 같은 공식을 이용하여 계산된다.

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

여기서 $R^2_{X_j|X_{-j}}$ 는 X_j 를 모든 다른 변수에 대한 regression의 R^2 을 의미한다. 만약 $R^2_{X_j|X_{-j}}$ 이 1에 가깝다면 X_j 를 다른 변수들의 선형 결합으로 설명하는 정확도가 높다는 의미이고, 이에 따라 분모는 작아지고, 전체적으로 vif는 커질 것이다.

3.4 The Marketing Plan

이제 우리는 광고 데이터에 대해서 이 챕터 시작에 했던 7개의 질문으로 돌아가자.

1. ***Is there a relationship between advertising sales and budget?***

   이 질문은 multiple regression model을 만든 후 이 모델이 유의미 한지 회귀 계수에 대한 가설검정을 시행하면 된다. section 3.2.2에서 F-statistic을 사용하면 된다고 했다.

2. ***How strong is the relationship?***

   우리는 모델의 정확도를 측정하는 두 가지 척도를 이야기 했다. 첫 번째는 RSE 추정치 인데 이것은 standard deviation of the response from the population regression line이다. 두 번째로 R^2(결정계수) 인데 이것은 predictors에 의해서 설명이 되는 response에서의 variability의 비율을 의미한다.

3. ***Which media contribute to sales?***

   이것에 답하기 위해서 우리는 각 예측 변수의 t-statistic과 관련된 p-value을 살펴볼 수 있다.

4. ***How large is the effect of each medium on sales?***

   우리는 $\hat{\beta}_j$ 의 standard error가 β_j 의 신뢰구간을 설정하는데 사용됨을 살펴보았다. 그리고 우리는 3.3.3에서 다중공선성이 매우 큰 standard error을 초래할 수 있음을 살펴보았다. 얼마나 영향을 끼치는지는 신뢰구간을 통해서 살펴보자

5. ***How accurately can we predict future sales?***

   반응변수는 3.21을 이용하여 예측될 수 있다. 이러한 estimate에 관련된 정확도는 우리가 개개인의 반응변수, 즉 $Y = f(X) + \epsilon$ 을 예측하느냐, 또는 반응변수의 평균, 즉 $f(X)$ 을 예측하느냐에 달려있다. 만약 전자라면 우리는 prediction interval을 사용하고 후자라면 confidence interval을 사용한다. prediction interval은 항상 confidence interval보다 넓을 것인데 그 이유는 걔네는 ϵ(the irreducible error)와 관련된 불확실성까지 포함하기 때문이다.

6. ***Is the relationship linear?***

   우리는 3.3.3에서 residual plots이 non-linearity을 알아내기 위해서 사용됨을 살펴보았다. 만약 관계가 linear하다면 residual plot은 아무런 패턴을 보여주지 않을 것이다. 또한 3.3.2에서 우리는 predictors의 변환을 포함한다면 non-linear 관계를 어느 정도 완화할 수 있을 것이다.

7. ***Is there synergy among the advertising media?***

   standard linear regression 모델은 예측변수와 반응변수간에 additive한 관계가 있다고 가정한다. additive한 모델은 해석하기 쉬운데 이는 각각의 예측 변수의 반응 변수에 대한 효과가 서로 다른 예측변수에 unrelated하다고 보기 때문이다. 하지만 이러한 additive 가정은 실제 데이터에서 비현실적이다. 3.3.2에서 우리는 non-additive 관계를 호환하기 위해 어떻게 interaction term을 포함시키는지 살펴보았다. interaction term의 작은 p-value은 그러한 관계가 있음을 의미한다.

3.5 Comparison of Linear Regression with K-Nearest Neighbors

Chapter 2에서 얘기 했듯이, 그것은 $f(X)$에 대해 linear functional form을 가정하기 때문에, 선형 회귀는 parametric 접근법의 한 예이다. parametric 방법은 여러 장점을 가지고 있다. 적은 수의 계수만을 추정하면 되기 때문에 적합하기가 쉽다. 하지만 이러한 parametric 방법은 분명 단점이 있다. 모델을 설계하는 과정에서 $f(X)$에 대해 강력한 가정을 한다. 만약 가정된 함수의 형태가 사실과 다르고 예측이 우리의 목표라면, parametric 방법은 잘 작동하지 않을 것이다.

반면에 non-parametric 방법은 $f(X)$에 대한 parametric 형태를 가정하지 않고 regression을 하기 위해 대안적이고 더 flexible한 방법을 제시한다. 여기서는 non-parametric의 대표적인 형태인 KNN regression을 살펴본다. KNN regression 방법은 KNN classifier와 깊게 연관되어 있다. 주어진 K와 prediction poin x_0 를 이용하여 KNN regression은 먼저 K개의 x_0와 가장 가까운 K개의 training 관측치 (N_0으로 표기)를 알아낸다. 그리고 $f(x_0)$을 N_0의 training 반응 변수의 평균을 사용하여 추정한다. 다시 말해서

$$\hat{f(x_0)} = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

일반적으로 가장 적합한 K 값은 bias-variance tradeoff에 의존할 것이다. 작은 값의 K는 가장 flexible한 적합을 보여주지만 높은 variance을 가진다. 이러한 변동성은 예측이 온전히 하나의 관측치에 의존하기 때문에 발생한다. 반면에 큰 K의 값은 부드럽고 덜 변동성있는

적합을 보여준다. 예측이 여러 개의 점의 평균에 근거하기 때문에 하나의 관측치를 바꾸는 것은 적은 효과를 가져온다. 하지만 이러한 부드러움은 $f(X)$에 있는 어떠한 구조를 가림으로써 bias를 발생시킬 수 있다.

어떠한 상황에서 parametric 방법이 non-paramtric 방법보다 더 우수할까? parametric에서 가정한 그 모델이 실제 f와 비슷하다면 더 우수할 것이다. 하지만 실생활에서는 true relationship은 알려져 있지 않다. 따라서 이렇게 관계가 알려져 있지 않은 상태에서 만약에 관계가 선형이라면 linear regression보다 살짝 열등하고 non-linear이라면 더 좋은 결과를 내므로 KNN을 시행하는 것이 더 좋다고 결론지을 수 있다. 하지만 p가 커진다면(고차원이 된다면) KNN은 linear regression보다 성능이 더 떨어진다. 사실 차원의 증가는 linear regression에서는 작은 성능 저하를 일으켰지만(test MSE), KNN에서는 ten-fold increase in MSE를 일으켜 차원이 높아질수록 KNN은 linear model에 비해 성능이 급격히 떨어진다. 이러한 이유는 차원이 높아질수록 sample size에서 감소가 일어나기 때문이다. 다시 말해서 $p = 1$일 때 100개의 training 관측치를 가지고 있어서 $f(X)$을 추정하는데 충분한 정보를 가지고 있었지만 $p = 20$이라면 이 100개의 training 관측치를 분산시켜야 한다. 그렇게 되면 어떠한 관측치가 근처의 neighbor가 없는 현상(curse of dimensionality라고 불린다)이 일어날 수 있다. 다시 말해 주어진 test 관측치 x_0에 대해서 이와 가장 가까운 K개이 관측치가 x_0와 매우 멀리 떨어져 있어서 poor KNN fit이 나오는 것이다. 일반적으로 parametric 방법은 predictor당 관측치의 갯수가 적을 때 non-parametric 방법을 압도한다. 그리고 차원이 작다 하더라도, 해석능력 관점에서 KNN 보다는 linear regression을 더 선호할 것이다. KNN의 MSE가 선형 모델보다 살짝 낮다면 우리는 심플한 모델을 얻는 대신에 이런 조금의 정확도 차이를 감수하는 것이다.