3. 선형 회귀(Linear Regression)

supervised learning의 아주 간단한 방법인 선형회귀. 다른 통계방법에 비해 간단하나 해석력이 뛰어나 여전히 널리쓰이고 있고 다른 방법들의 기초가 되는 지식이다. 고로 많이 알아두면 알아둘 수록 좋다. 참고한 자료는 'Applied Linear Regression Models 4th edition'.

개인적으로 몰랏던 부분은 (!)로 표시.

개인적인 참고를 위한 좀더 세부적인 설명은 다음의 형식을 맞춰 적었다.

> 세부적인 설명을 적는 부분

독립변수와 종속변수의 관계, 예를 들어 TV,라디오, 뉴스 광고 지출과 총 판매량간의 관계를 밝히고자 할때 다음의 질문이 중요하다.

1. 실제로 광고 지출과 총 판매량간에 관계가 있는가?
2. 그 관계가 얼마나 뚜렷한가?
3. TV, 라디어, 뉴스 모두 관계가 있는가, 아니라면 그 중 누가 관계가 있는가?
4. 이를 통한 예측은 어느정도 정확하다 할수 있는가?
5. 실제 관계가 선형관계인가?
6. 각 매체간에 상호작용 효과는 없는가?(TV에 지출한것이 신문에 지출한 것에도 영향을 미치는경우)

이와 같은 질문에 대하여 선형회귀가 어떠한 답을 내줄 수 있는지를 알아볼 것이다.

3.1 단순 선형회귀 (simple linear regression)

독립변수와 종속변수의 관계를 설명할때 가장 많이 쓰이는 가장 단순한 그 모델. (모두의 머릿속에 떠오르는 그게 단순 선형회귀 맞다.) 식으로 써보자면

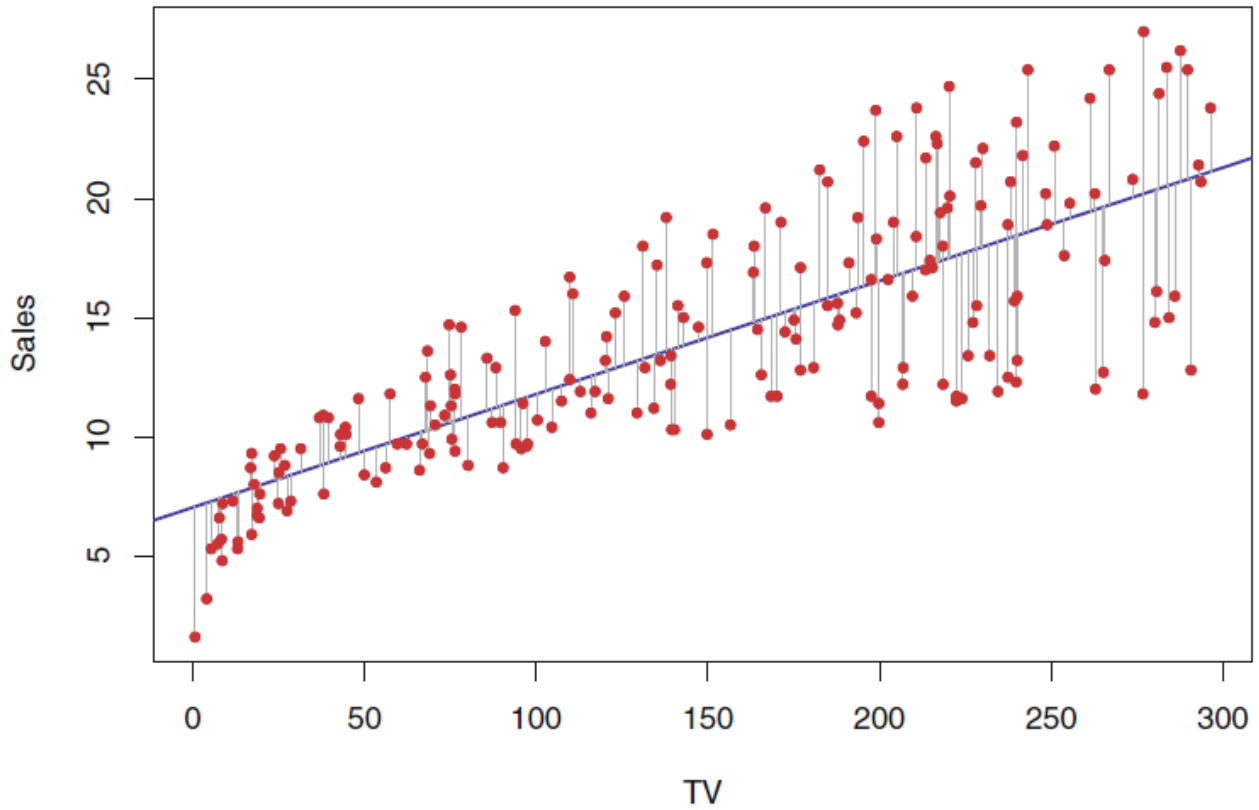
$$Y = \beta_0 + \beta_1 X + \epsilon$$

이를 영어로는 **'regressing Y on(to) X'** 라고 표현한다.

이때 β_0과 β_1이 우리가 추정하고자 하는 모델의 계수, 바꿔말하면 **parameter**고, 우리가 추정한 계수는 역시나 $\hat{\beta}_0$과 $\hat{\beta}_1$으로 쓴다. (추정한 값을 간단하게 b_0의 형태로 쓰기도 한다)

그럼 어떻게 계수들을 구하나?

간단하게 우리가 가진 데이터에 우리의 선형 식이 최대한 잘 맞도록 계수를 구한다. '잘 맞도록'하는 방법으로는 가장 대표적으로 **least squares**방법으 쓰인다.(다른 방법은 6장에서 다룬다!) 이는 결국 앞장에서 나왔던 MSE를 최소화하는 계수를 구하는 것과 같다.



앞장에도 나왔던 그림. 저기서 빨간색이 우리가 가진 데이터, 파란색이 우리가 추정한 함수. 추정된 함수와 가지고 있는 데이터의 차를 잔차라고 하는데, i번째 데이터에 대한 잔차를 e_i라고 표현한다. 데이터와 추정된 함수가 얼마나 잘 맞는지는 잔차들을 제곱(square)해서 구한다. 이를 RSS(sum of squares residual) 혹은 SSE(sum of square error)이라고 한다.

$$RSS = e_1^2 + \ldots + e_n^2$$

그럼 이걸 최소화하는 계수를 어떻게 구하냐? 간단! 해당 식이 제곱형태이니 미분해서 0이 되는 지점을 찾으면 된다.

각 parameter에 대해 미분하고 간단하게 정리하여 나온 least square를 통해 추정된 계수는 다음과 같다.

$$b_0 = \bar{y} - b_1 \bar{x}$$

$b_1 = \frac{\sum x_i(y_i - \bar{y})}{\sum x_i(x_i - \bar{x})}$

> b_0의 경우
>
> $\frac{\partial RSS}{\partial b_0} = \frac{\partial \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2}{\partial b_0} = 0$이 되는 값을 찾으면 된다.
>
> $-2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0$
>
> $\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0$
>
> $\sum(y_i) - nb_0 - \sum b_1 x_i = 0$
>
> $\sum(y_i) - \sum b_1 x_i = nb_0$
>
> $b_0 = \frac{\sum(y_i)}{n} - \frac{\sum b_1 x_i}{n}$
>
> $\therefore b_0 = \bar{y} - b_1\bar{x}$
>
> b_1의 경우
>
> $\frac{\partial RSS}{\partial b_1} = \frac{\partial \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2}{\partial b_1} = 0$이 되는 값을 찾으면 된다.
>
> $-2\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0$
>
> $\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0$
>
> $\sum(x_i y_i) - b_0\sum x_i - \sum b_1 x_i^2 = 0$
>
> $\sum x_i(y_i - \bar{y}) - b_1\sum x_i(x_i - \bar{x}) = 0, \because b_0 = \bar{y} - b1\bar{x}$
>
> $\therefore b_1 = \frac{\sum x_i(y_i - \bar{y})}{\sum x_i(x_i - \bar{x})}$, 이는 $\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})(x_i - \bar{x})}$로도 나타낼 수 있어(전개하면 똑같다) $b_1 = \frac{Sxy}{Sxx}$라고 쓰기도 한다 (sum of x&y, sum of x&x)

이 중 기울기, 즉 b_1은 독립변수 x가 한단위 증가 했을때 종속변수 $\hat{y}(= \hat{f}(x))$가 얼마나 변하는지를 의미한다. TV와 매출의 예시에서 $b_1 = 0.475$라면 TV광고에 1달라 더 쓸때마다 평균 매출은 0.475개 늘어난다는 것을 의미한다. (추정된 선형 모델이 이렇게 예측한다는 것이지 실제 세상에선 당근 오차가 있다.)

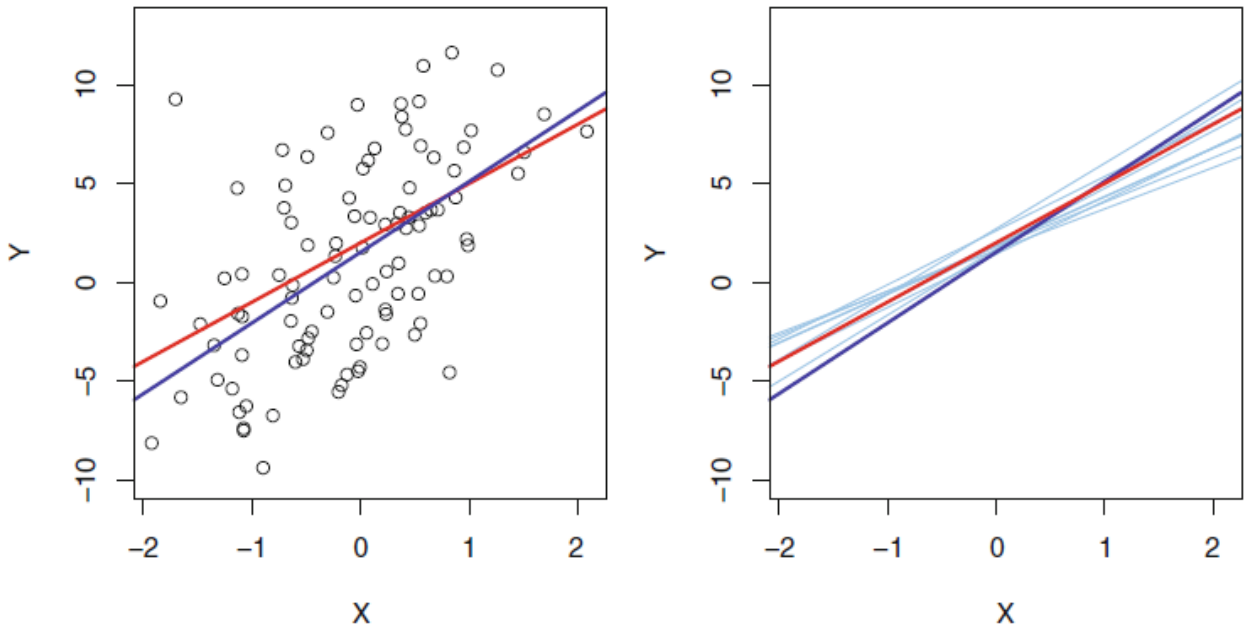
그러나 우리가 가진 데이터는 세상 모든 데이터가 아니라 한계가 있기 때문에, 이를 토대로 추정한 '추정된 회귀선'도 실재의 데이터를 기준으로 그은 선형 '모회귀선'과는 어느정도 차이가 있을 수 밖에 없다. 그러나 어쩔 수 없다. 우리의 주어진 데이터로는 이것이 최선이기에.

다만, Least Square의 아주 중요한 성질중 하나가 least square 방법을 통해 추정된 회귀선은 $unbiased$하다는 것이다. 즉 만약 수많은 데이터 셋들에 대하여 수많은 least square적합을 하면, 이 선들의 **평균선**은 알지 못하는 **진짜 관계**와 같아질 것이라는 것이다.

예를 들어 X와 Y간의 실제 관계가 다음과 같다고 하자

$Y = 2 + 3X + \epsilon$ 여기서 ϵ은 평균이 0이고 많은 경우에 분산이 같은(등분산인) 정규분포를 가정한다. (앞에서도 말했듯이 이게 parametric방법. 이 가정이 물론 현실에서 틀릴 수도 있다.분석 후 이에 대한 검토가 이뤄져야한다.)

위의 식에서 무작위로 100개의 data를 10번 뽑으면, random noise ϵ으로 인해 아주 조금씩 다른 10개의 dataset(크기는 100)이 뽑힐것.



위 그림의 왼쪽에서 빨간선은 진짜 관계선, 즉 true f이다. ($Y = 2 + 3X + \epsilon$) 이를 토대로 10번 랜덤하게 뽑은 데이터에 적합한 10개의 '추정선'들이 오른쪽 그림에 있다. 선들 하나 하나는 빨간선과 조금씩 오차가 있지만, 이들을 모아 **평균**을 낸다면 진짜 빨간선과 점점 더 가까워 지는 것을 볼 수 있다. 즉 우리의 추정선의 **기대값**은 진짜 선형 모회귀선의 값이다.

이는 β_0, β_1에 대해서도 적용된다. 추정된 회귀계수의 기대값 true 값과 일치한다.

그럼 기대되는 값이 참값과 같다면 다음 질문은 '얼마나 그 기대값이 믿을만한가?'이다. 이는 추정된 값($\hat{y}_0, \hat{\beta}_0, \hat{\beta}_1$)의 **분산**을 구하는 문제가 된다. (기대값이 같아도 분산은 당근 천차만별일 수 있다.) 이 '분산'역시, 진짜값은 알 수 없고 우리가 가진 표본의 '표본분산'을 사용 해야한다. least square를 통해 구한 계수의 표본분산(standard error)은 다음과 같다

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(이는 사실 앞에서 말했던 **'오차의 분산이 같다'**라는 가정이 있어야만 성립한다.)

식을 보면 알겠지만, $\sum(x_i - \bar{x})^2$이 클수록 즉 x의 변동이 클수록, 즉 x들이 멀리 퍼져있을 수록 $\hat{\beta}_1$의 SE도 확실하게 낮아지는데, 직관적으로 데이터가 $\bar{x}$에서 멀리 퍼져 있으면 기울기를 예측하는데 더 도움이 될것이라고 이해가 가능하다.(b0 역시 도움이 되지만, 회귀분석에서 주된 관심은 b1에 있다.)

여기서 $\sigma^2 = Var(\epsilon)$이다. 그러나 σ^2는 안 알려진 값이기에 SSE로 이를 추정을 하는데, 이 때 2개의 회귀계수를 추정하고자 하였기에 자유도를 2 잃어 자유도$df = n - 2$이다. (그리고 이렇게 바꿔서 추정을 해서 분포는 Normal에서 t분포로 바뀌게 된다.)

> (글고 $E[MSE] = \sigma$는 분포가정이 없이 가능하지만, t분포를 통한 추정은 $\epsilon \sim N$이라는 가정이 있기에 가능하다. 책 48참고)
>
> $\frac{(y_i - \beta_0 - \beta_1 x_i)}{\sigma} \sim N(0,1)$
>
> $\frac{\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sigma^2} \sim \chi^2[n-2]$
>
> $\therefore \frac{SSE}{\sigma^2} \sim \chi^2[n-2]$

> $\therefore E[\frac{SSE}{\sigma^2}] = n-2$

고로 앞장에서 언급하였던 MSE가 σ^2 추정에 사용되는 신비한 결과. ($MSE = \frac{SSE}{(n-2)} = \hat{\sigma}^2$) 그러나 직관적으로 생각해보면 분산을 '추정'한다는 개념에서는 우리의 추정 선에서의 변동을 보는 MSE를 쓰는것이 이해될 수 있다.

또한 추정된 σ는 거기에 제곱근을 씌워, RSE, 혹은 RMSE(root MSE)라고 쓴다. $RMSE = \sqrt{\frac{SSE}{(n-2)}} = \hat{\sigma}$, 책에선 RSE. 헷갈리면 MSE만 기억해도 된다.

이 추정된 분산을 토대로 회귀 계수에 대한 신뢰구간이나 가설검정을 할 수 있는데, 95% 신뢰구간은 다음과 같다.

$[\hat{\beta}_1 - t_{0.25,n-2} * SE(\hat{\beta}_1), \hat{\beta}_1 + t_{0.25,n-2} * SE(\hat{\beta}_1)]$

이는 b0의 경우에도 마찬가지 이다. 위의 예에서 b1의 95%신뢰구간이 [0.042,0.053]이 나왔을때 이는 **95%의 신뢰수준으로 TV광고의 1달러 증가가 [0.042,0.053]개의 평균 판매액의 증가를 낸다**라고 말할 수 있다.

가설검정은 대표적으로 X와 Y간에 관계가 있는지를 보는데 이는 결국

$H_0 : \beta_1 = 0$을 검정하는것으로 귀결된다.

위에서 언급했듯이 $\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$이 $t(n-2)$분포를 따르므로,

$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$을 구하여 t분포상의 p-value(귀무가설 하에서 이런 사건이 일어날 확률이 얼마나 되는지를 의미하는 확률)를 구하면 된다. b0에 대한 검정도 위의 식에서 1을 0으로 바꿔서 똑같이 진행하면 된다. 지나치게 작은(유의수준보다 작은) p-value는 귀무가설을 기각할 만한 근거를 의미한다. 위의 가설에서는 'X와 Y가 관계 없다'(우리가 세운 X와 Y의 모델은 쓸모없다)가 귀무가설이었으니 p-value가 작을수록 좋다.

모델의 평가

우리가 세운 모델이 쓸모가 있다고 결론이 나면, 두번째로는 얼마나 잘 적합을 하는지를 봐야한다. 이는 앞에서 언급되었던 MSE를 통해 이루어 진다

$MSE = \frac{SSE}{(n-2)} = \hat{\sigma}^2$, (책에선 RMSE(root MSE)를 썻다. RMSE를 사용하는 이유는 아래에서 나오지만, 바로 Y의 단위에 맞추어 생각할 수 있는 scaling의 의미가 있기 때문이다.)(???F통계량이 모델의 정확도 평가로 쓰일 수 있는가)

RSE는 우리의 데이터를 우리 모델이 얼마나 잘 설명할 수 있는가를 의미하는데, 예시에서 RSE가 3.26이라면 실제 판매량 데이터가 우리 모델에서 3.26개 정도 차이날 것이라는 의미이다. 이 차이가 얼마나 큰것인지는 역시 문제에 따라 다르다.(예측변수가 1,2,3..이라면 3.26은 큰것이지만 예측변수가 10000,2000등이라면 3.26은 굉장히 작은것. 단위에 따라 다르다, 즉 scaling을 포함하지 않았다) MSE가 추정된 '변동'이라는 의미에서, RSE는 true regression line에서 벗어난 정도의 '추정'이 라고 말할 수 있다.

이를 비율로써 계산한 지표가 R^2인데, **'전체 변동중 설명된 변동의 비율'**을 의미한다. 비율이므로 당연히 0~1사이 값이다. 이를 좀더 설명하자면

$SST = \sum(y_i - \bar{y})^2$, 즉 regression을 하기도 전에 있던, 전체의 평균에서 각 값들이 얼마나 왓다갔다 하는가를 의미하는 지표이다. (왜 전체 변동을 $\bar{y}$에서의 거리로 구하나? - 표본분산을 구할때 sample mean을 빼주는걸 생각하면 됨. 전체 분산 중 우리가 설명한 분산) 이 중 우리가 예측한 '왓다갓다 하는 정도', 즉 데이터의 변동은 우리의 추정된 함수선일 것이고(SSR로 자주 표현), 추정된 함수선이 예측하지 못한 변동은 앞에서 다룬 SSE($SSE = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$)이다. 고로 $SST = SSR + SSE$ 이다.

> notation에 대해 정리하자면
>
> $RSS = SSE$, $RSE = \sqrt{MSE}$, $TSS = SST$이다.

'전체 변동중 설명된 변동의 비율'인 $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$이다. R^2가 높을 수록, SSE가 낮을 수록 데이터의 변동을 잘 설명하고 있다고 보면 된다. 그러나 안타깝게도 **얼마나 높은 R^2가 만족스러운 적합이냐에는 답이 없다**. 그도 그럴것이, 물리학 등 분야에서는 높은 수치가 기준이 될것이나, 마케팅, 사회학 등 분야에서는 비교작 낮은 수치의 R^2만으로도 의미가 있다고 판단 될 것이다.

또한 R^2는 단순회귀 문제에서는 '얼마나 선형관계가 있느냐'를 보는 문제와 같아지기에, 우리의 데이터의 X와 Y의 $corelation$을 보는것과 같아진다! 즉, 단순선형회귀에 국한해서, $R^2 = r^2$이다. 추가로 단순선형회귀에선 cor(X,Y)나 cor($\hat{Y}, Y$)를 구하는거나 같아진다. Y.hat이 어차피 X의 선형결합이기 때문.

3.2 다중 선형 회귀

독립변수가 많을때는 어쩔까? 단순히 각각의 단순선형회귀를 여러개 하면? 안된다. 왜 안될까?(문제)

1. 우선 애초에, 3개의 독립변수가 다 주어졌을때, 3개의 모델로는 최종 어떤 Y를 결정해야 할지 정할 수가 없다.
2. 각각에 대한 적합은 **독립변수 끼리의 상호작용**을 배제해 버린다. 즉, 서로 correlated된 경우 문제가 심각해진다. (correlated란 서로 관계가 높아서, 한쪽을 알면 다른 한쪽도 어느정도 설명할 수 있다는것. ex-TV광고에 많이 투자한 회사는 대체적으로 신문광고에도 많이 투자했을것)

고로, 여러 독립변수를 한번에 고려할 수 있는 다중 선형회귀를 사용한다. 식은 아주 간단하다

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

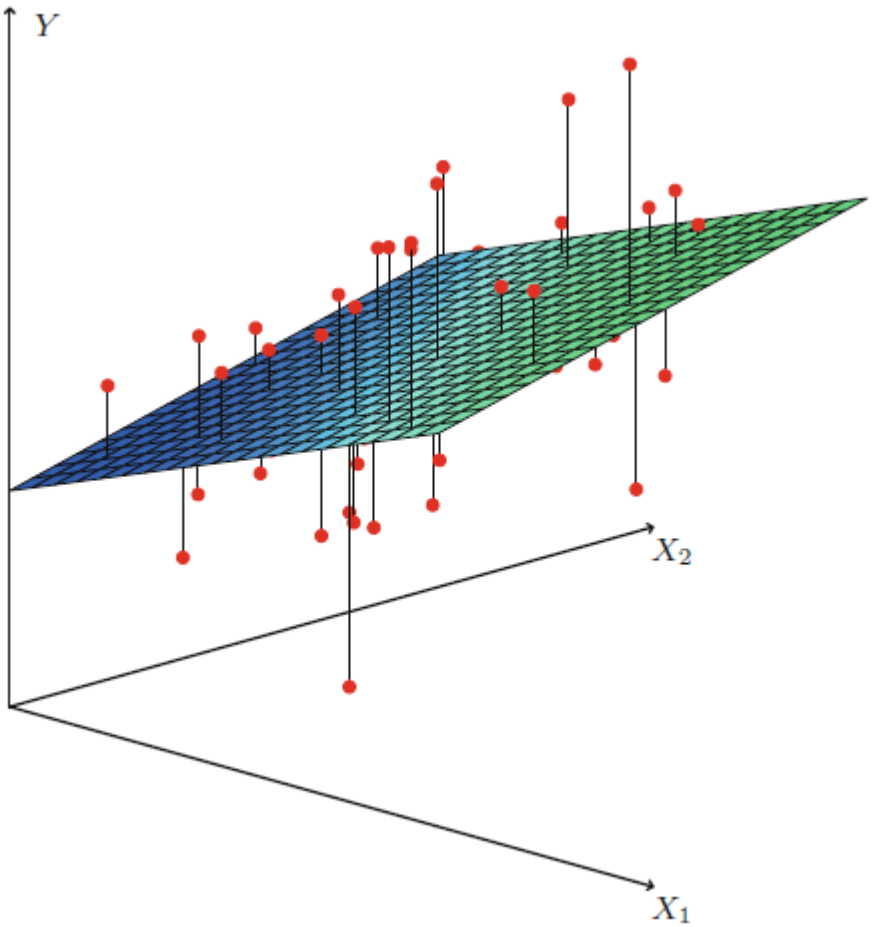
여기서 β_i는 **다른 독립변수들이 고정되어 있을때(!)** X_i의 한단위 증가가 Y의 '평균'에 미치는 영향이다. (문제 : 실제 상황에서는 키가 크면 몸무게가 늘어난다. 이렇게 실제에선 다른변수가 고정되어 있을 수가 없는데 그럼 β_i는 무슨 의미가 있을까?- 답: 회귀계수에 대한 해석을 단순회귀와 완벽하게 같은 이유로 이해하려면 변수들이 완전한 선형독립, 혹은 eigen vector여야만 가능하다. 그러나 현실에서는 변수간에 완전한 독립인 경우가 거의 없기에, 기본적으로 어쩔수 없다고 판단하고 분석을 진행한다. 이 정도가 크다 판단되면 상호작용항을 넣어주는데 이 경우 $(\beta_i + \beta_{ij} X_j) X_i$와 같이 다른 변수가 증가함을 반영한다)

모델의 평가

모델의 평가는 역시 똑같은 방식으로 한다. 예측한 값y_i와 $\hat{y}_i$이 얼마나 차이가 날지. 식으로 쓰면 다음과 같다.

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned}$$

변수가 2개인 경우 그림으로 나타낼 경우 좀더 직관적으로 와닿을 수 있다. 이제 2차원 평면에서의 1차원 선이 아니라 3차원 공간에서의 2차원 평면과 데이터의 차이를 의미하게 되었다. (사람의 인식의 한계가 3차원이기에, X1, X2 두개의 독립변수에 대한 multi linear regression을 본다.)



모델의 계수

이 경우 해당 RSS를 최소화하는 $\hat{\beta}_0, \ldots, \hat{\beta}_p$은 역시나 least square로 구하는데, 이 경우 matrix 연산이 들어간다. (matrix 연산은 별게 아니라 계산을 한꺼번에 하기 편한것. 사실 b0,b1에 대해 미분했듯이 전부 다 미분하고 풀면 답은 같게 나온다)

참고로 matrix에서의 least square는 다음과 같다(일일이 해보면 여러개의 편미분분을 연립방적식 하는거랑 같다)

$$X'Xb = X'Y$$

> $Y = Xb$에서
>
> $$\begin{pmatrix} y_1 \\ y_2 \\ .. \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ .. & .. & .. \\ 1 & x_{1n} & x_{2n} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}$$
>
> (n * 1) = (n * 3) (3* 1)이다.
>
> $X'Xb = X'Y$ 이 식은 이렇게 풀린다. 일일이 해봄
>
> $$X'X = (3*n)(n*3) = \begin{pmatrix} n & \sum x_1 & \sum x_2 \\ \sum x_1 & \sum x_1^2 & \sum x_1 \sum x_2 \\ \sum x_2 & \sum x_1 \sum x_2 & \sum x_2^2 \end{pmatrix}$$

$$X'Y = (3*n)(n*1) = \begin{pmatrix} \sum y \\ \sum x_1 y \\ \sum x_2 y \end{pmatrix}$$

$$\therefore X'Xb = X'Y$$

$$= \begin{pmatrix} n & \sum x_1 & \sum x_2 \\ \sum x_1 & \sum x_1^2 & \sum x_1 \sum x_2 \\ \sum x_2 & \sum x_1 \sum x_2 & \sum x_2^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} \sum y \\ \sum x_1 y \\ \sum x_2 y \end{pmatrix}$$

이를 전개해보면

$nb_0 + b_1 \sum x_1 + b_2 \sum x_2 = \sum y$등의 식이 나오는데, 이는 결국 단순선형에서 최소제곱의 정규방정식을 푸는것과 같게 나온다.

실제로 다중회귀를 해보면, 각각 하나씩 넣었을때는 유의하다(p-value가 작게)나온 변수일지라도 함께 들어가면 별로 안 중요한 경우가 생긴다. 아래 예시.

	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.*

단순회귀에서는 newspaper역시 p-value가 유의하게 나왔다.

그러나 correlation matrix에서 보았을때 news는 radio랑 correlation계수가 높았고, 다중 회귀에서의 p-value가 엄청나게 높다.(즉 newspaper의 회귀계수가 0이라할 충분한 근거가 있다.) 다른 중요한 변수들(TV, 라디오 광고비용)이 이미 안중요한 변수(신문 광고비용)이 할 수 있던 설명을 다 해버렸기에, 그 변수는 더 이상 정보를 주는 변수가 아닌 것이다. 이는 신문 광고는 sale과 indirect한 영향을 주는 변수였다고 설명을 할 수도 있는데, indirect하지만 관계가 있는 대표적인 예로 아이스크림 판매량과 물놀이 사고가 있다. (더위라는 잠복변수)

다중회귀에서의 질문들

다중회귀에서는 다음과 같은 질문들이 주로 던져지는데,

1. p개의 독립변수 중 하나라도 유의한가? (==모든 회귀계수가 전부0은 아닌가?)
2. 그럼 모든p가 유의한가, 그중 몇개만 유의한가?
3. 우리 데이터에 얼마나 잘 적합됐는가
4. 우리의 예측은 얼마나 정확할 것인가?

1. p개의 독립변수 중 하나라도 유의한가? (==모든 회귀계수가 전부0은 아닌가?)

이는 다음을 검정하는 문제로 귀결되고,

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

이는 다음의 통계량을 통해 결정된다.

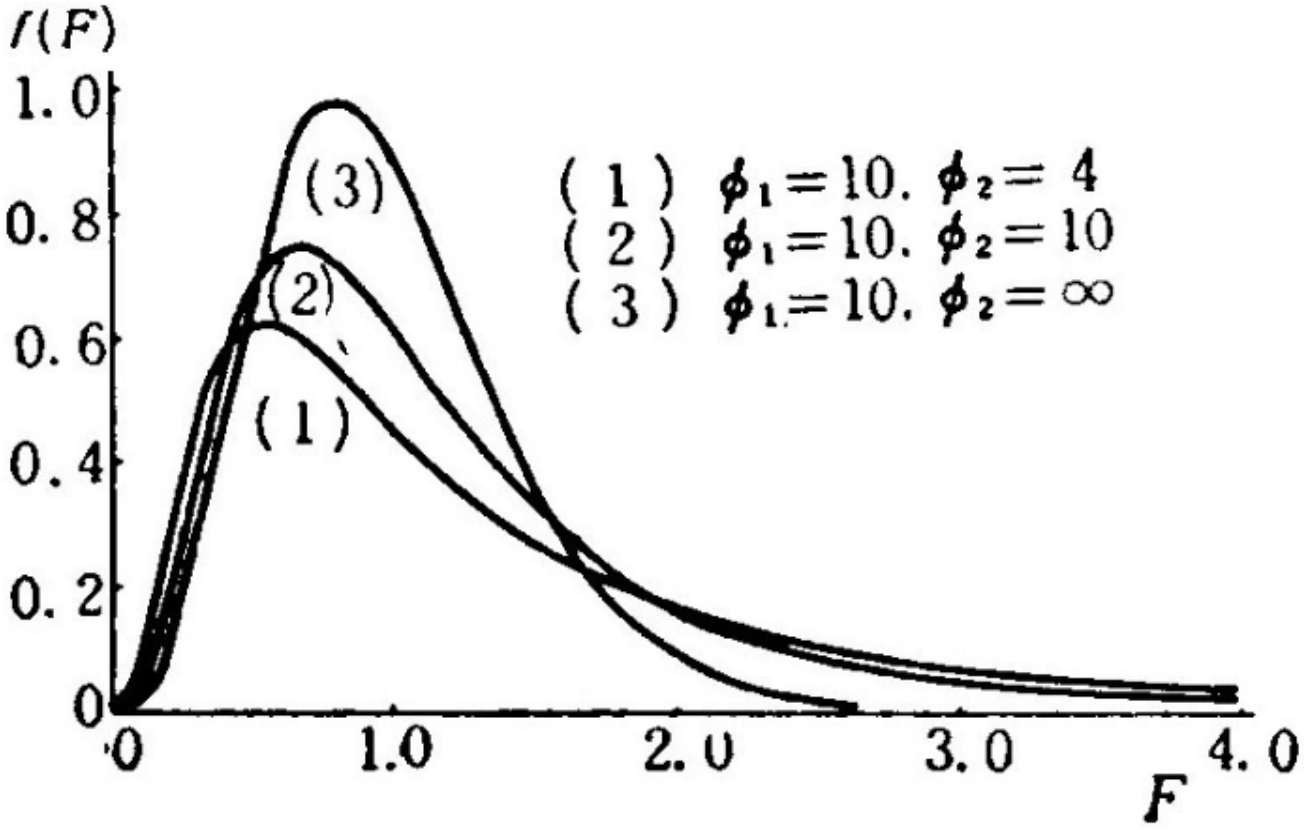
$F = \frac{(SST-SSE)/p}{SSE/(n-(p+1))} = \frac{MSR}{MSE}$ (MSR은 SSR을 df로 나눈것. 여기서 df는 독립변수의 갯수)

> 그러나 뒤에 나올 q개의 계수만을 검정하는 데에도 일반화되서 쓰기에는 다음 식이 짱이다.
>
> $F = \frac{SSE(R)-SSE(F)}{df_R-df_F} \div \frac{SSE(F)}{df_F} \sim F(df_R - df_F, df_F)$
>
> 저기서 R과 F는 각각 Reduced model, Full model이고, 쉽게 말해 R은 귀무가설 하의 모델, F는 모든 변수가 들어간 모델이라 보면 된다. (책의 278쪽)
>
> SSE의 df는 $(n - [b_0 \sim b_{p\text{갯수}}])$이므로 $df_R - df_F = q$(검정하고자 하는 변수 갯수) 라고 볼 수 있다.

앞에서도 나왔듯이 $E[SSE/(n-(p+1))] = \sigma^2$이고, (기대값은 바로 나오지만, 이에 대한 분포가정을 하려면 error term의 normal 가정이 있어야 한다.)

귀무가설 하에서 $E[\frac{SST-SSE}{p}] = \sigma^2$이기에, 귀무가설이 참이라면 이는 1에 가까운 값(단순회귀일때 SSR의 기대값을 구해보자!!!)을 가질 것이다.

구체적으로 위의 통계량 F는 자유도가 $(p, n-(p+1))$인 F분포를 따른다. F분포는 이렇게 생김.



q개의 변수에 대한 검정은 다음을 통한다.

$F = \frac{RSS_0-RSS/q}{RSS/(n-p-1)}$ 여기서 RSS_0은 q개 없애고 적합시킨 모델(Reduced model)

하나의 변수에 대한 검정에서는 $F = t^2$이다.

Q. 왜 전체에 대한 F검정을 하느냐, 그냥 각 계수에 대한 t-test p-value를 보면 안돼냐?

A.위험하다. 특히 변수가 많을때. 변수p가 100개 이고 모두가 유의미하지 않은 변수($\beta_i = 0$)일지라도 **5개 변수(0.05) 정도가 확률적으로 '유의한' p-value를 가질수도.(!)** 그러나 F검정은 변수의 갯수도 고려를 하기에 전체에 대한 p-value를 구할 수 있음.-- ANOVA에서 3C2하지 않는 이유와 일맥상통!!!!봐봐봐

2. 그럼 모든p가 유의한가, 그중 몇개만 유의한가?

전체에 대한 F검정이 유의하다 나오면(즉 전부다 쓸모 없지는 않다.), 그중 어떤 변수를 써야할까? 이때 쓰이는게 변수선택법. 6장에서 더 자세히 다룬다. 추가적으로 모델을 평가하기 위한 기준으론 AIC, BIC, Mallow's C_P등이 있다.

변수선택법에 대해 간단하게 설명하자면

- 전진 선택법 :
  - 1) 아무 변수도 포함되지 않은 모델에서
  - 2) t통계량이 제일 유의한 변수 한개를 넣는다(기준은? 한 변수에 관해서는 F통계량, t통계량, 오차 제곱합 감소 다 같은 결과를 낸다, 암꺼나로 해도 됨)
  - 3)해당 변수를 넣은 상태에서 2번을 계산하여 또 하나를 넣는다. (변수 하나가 기본으로 들어가 있으니 p-value가 달라짐)
  - 4) 더이상 중요한 변수가 없으면(t 통계량이 유의한게 없으면) 멈춘다
- 후진 제거법 :
  - 1) 모든 변수가 있는 모델에서
  - 2) 제일 큰 p-value를 가진 변수를 지움
  - 3) 남은 p-1개 변수로 또 p-value계산해서 뺀다
  - 4) 더이상 뺄 변수가 없으면 멈춘다
- 전진 단계적 회귀(mixed selection)
  - 1) 아무 변수도 포함되지 않은 모델에서
  - 2) t통계량이 제일 유의한 변수 한개를 넣는다
  - 3) 해당 변수를 **넣은 상태에서 p-value를 계산해서** 유의미 하지 않은 변수를 **지운다** (이때 들어오기 위한 p-value임계점과 나가는 임계점을 다르게 한다. 보통 어렵게 들어오고($\alpha = 0.1$) 쉽게 뺀다($\alpha = 0.15$))
  - 4) 다시 중요한 순대로 새로운 변수를 넣는다.
  - 5) 더이상 넣을 변수도, 뺄 변수도 없으면 멈춤

당근 mixed인 전진 단계적 회귀가 젤 좋다.

3. 우리 데이터에 얼마나 잘 적합됐는가

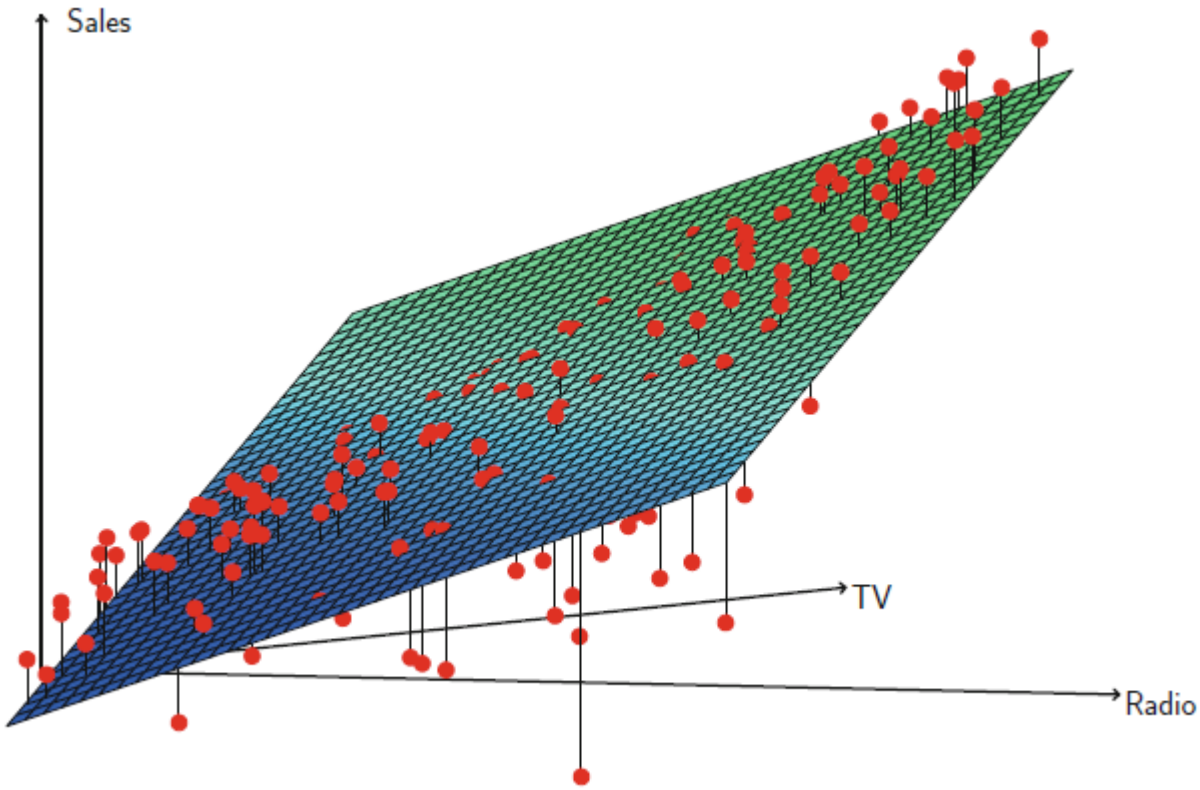
당근 MSE와 R^2로 구한다. 근데 다중 회귀에선 R^2이 cor(X,Y)를 구하는건 더이상 될 수 없고(변수가 1개가 아니니까) cor($\hat{Y}, Y$)의 제곱과 같다. 앞에서 나왔듯이 least square는 주어진 자료 하에서 이 cor을 최대로 만드는 linear 모델을 찾아준다.

그러나 R^2는 변수를 추가하면 할수록 항상 증가한다. 아주 쓸모없는 변수이더라도 least square에 따라 주어진 데이터에서 조금이라도 더 적합을 하게 된다. (몸무게 예측에 안드로메다 태양 흑점수를 추가한다해도!) 이는 주어진 데이터에 지나치게 과적합(overfitting)을 하게 되는것으로, 이러한 성질때문에 모델을 비교할때는 R^2외에 변수의 숫자도 고려한 지표를 사용한다. **변수를 추가함으로써 생기는 위험**이 (모델의 복잡도 혹은 overfit) **추가함으로써 생기는 설명력**보다 큰지를 고려하는것이다. AIC, BIC등등

또 MSE역시 변수의 갯수를 고려한다.

$MSE = \frac{SSE}{n-(p+1)}$ 다시 상기. 사용된 변수갯수(p)에 b_0까지 해서 p+1개.

단순한 수치 외에, 사실 그림이 이를 판단하는데에 좋다. (실제로는 잔차그림이 많이 쓰인다.)



그림을 보면 TV와 Radio에 따른 sales에서 애초에 빨간점의 분포가 약간 곡선(convex)를 띄고 있어 '선형'이 아님을 볼 수 있다. 이는 TV와 Radio에 둘다 투자할 경우 '시너지'가 발생해서 sale이 더 올라간다는 직관적인 생각과도 일치한다. '시너지' 혹은 '상호작용'에 대해서는 뒷 부분에서 설명한다.

4. 우리의 예측은 얼마나 정확할 것인가?

실제 함수 관계가 위처럼 선형이 아닐 수도 있지만, 실제 관계(f)가 선형임에도 우리가 예측한 선형모델($\hat{f}$)과는 차이가 있을 수 있다.

모델의 정확도를 판단할때는 우선적으로 후자, **즉 실제 관계가 선형인 경우 우리 모델과 얼마나 다를지** 에 집중한다. (위의 그림처럼 실제 관계가 선형인지 아닌지는 모델이 다 세워진 후에 진단을 한다.) f와 $\hat{f}$가 얼마나 다를지는 **'신뢰구간'**(confidence interval)으로 구한다

또 다른 주요한 예측은 특정한 실제 값y_0을 예측하고자 할때 y_0와 $\hat{f}$가 얼마나 차이날지 이다. y_0은 실제 함수f에서도 random error를 띄며 발생할 것이기에, 이를 예측하는 **예측 신뢰구간(Prediction interval)은 Confidence interval보다 클 수 밖에 없다**(!)

이 둘이 헷갈릴 수 있는데, f는 fixed된 unknown함수 선, 즉 특정 x에 대해서 'fixed'된 $f(x)$가 있는 것이고, y_0는 $f(x)+\epsilon$, 즉 그 함수선 위의 점을 중심으로 random error(=irreducible error)를 띄며 여러가지로 나타날 수 있다. 우리는 $\hat{f}$을 근거로 예측을 할 수밖에 없기에, **'개별 값에 대한 예측' 에 대한 변동은 [f와 차이 나는 $\hat{f}$의 변동 + 개별 값 자체의 random error]**이다.

> 구체적으로 말하자면
>
> **신뢰구간의 경우**
>
> $\sigma^2[\hat{f}(x_0))] = \sigma^2[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}]$이고 이때 σ^2는 MSE로 추정된다.(이 과정에서 자유도를 2잃고 t분포가 된다)
>
> 이때 신뢰구간은 $\hat{f}(x_0) \pm t(1-\alpha/2; n-2) * SE[\hat{f}(x_0)]$이 된다.

> **개별 예측값에 대한 예측 신뢰구간의 경우**
>
> 개별값 y_0이 추정된 평균$\hat{f}(x_0)$와 얼마나 떨어져 있는지를 $pred$라고 하면
>
> $\sigma^2[pred] = \sigma^2[y_0 - \hat{f}(x_0)] = \sigma^2 + \sigma^2[\hat{f}(x_0)]$
>
> $= \sigma^2 + \sigma^2[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}] = \sigma^2[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}]$
>
> 이 된다. 이때 σ는 MSE로 추정된다.(이 과정에서 자유도를 2잃고 t분포가 된다)
>
> 이때 신뢰구간은 $\hat{f}(x_0) \pm t(1-\alpha/2; n-2) * SE[pred]$이 된다.
>
> (책 57, 62쪽, 책에선 $\hat{Y}_h$를 추정된 함수, $\hat{Y}_{h(new)}$를 개별 값으로 notation씀)

책의 예시에서 TV에 10000만, radio에 2000만을 투자 할 경우 true $f(x)$에 대한 95% confidence interval은 [10,985,11,528] 이었다. 이는 **해당 interval을 100개의 dataset에 대해 구할경우 그중 95개가 true $f(x)$를 포함할것**이라는 의미.

한편 특정 city, 즉 개별 예측에 대한 prediction interval은 [7,930, 14,580]으로, 더 넓어졌다. 이는 **95%신뢰 수준으로 true y_0(!)을 포함할 것**이라는 의미.

3.3 그 외 회귀모형에서 고려할 것들

질적변수

쉽게 말해 범주를 말한다. ({남자, 여자}, {소득 상, 중, 하} 등등) 이 경우 범주 class의 갯수(남여의 경우 2, 소득의 경우 3)에서 **1을 뺀** 각각의 변수(더미변수)를 만들어서 적합한다.

변수는 해당 class에 속하면 1, 아니면 0인 값을 넣는식으로 즉 true냐 false냐를 나타내는 변수를 만든다.

책의 예를 들어 credit card balance를 예측하기 위한 독립변수로 {인종 아시아인, 백인, 아프리카인}이 있다면, 이런 식의 더미 변수를 만들어준다.

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

즉 [아시아인 인지 아닌지], [백인 인지 아닌지]를 나타내는 변수를 만든것. 3개가 아닌 2개의 변수를 만든 이유는 두 개의 변수가 모두 0이면 필연적으로 3번째 class를 의미하는것이기도 하고, 사실 필연적으로 그렇기에 3번째 변수를 넣으면 변수간에 **완벽한 함수관계**가 생겨버려 문제가 생긴다. (matrix연산에서 inverse matrix가 아예 안만들어진다.)

> matrix 연산에서 정규방정식은 X'X의 inverse를 구해야 한다. 그러나 rank(X'X)=rank(X)인데 X가 오나벽한 함수관계를 포함하고 있으면
>
> (X는 애초에 정방이 아니니 inverse matrix에 대해선 논할 수 없다.)

이 경우 위의 더미 변수들은 {0,1}의 값만을 가질 수 있기에, 결과적으로 각 범주에 따른 3개의 함수선이 나오게 된다.

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

이때 맨 오른쪽 식을 보면 알 수 있듯이 β_2는 백인과 아프리카인의 평균 credit balance의 차이를 의미하게 된다.

> Q. 왜 $x_{i1} = (1, 2, 3)$의 형태로 안 만들까?
>
> A. 그 경우 true, false를 나타내는 0,1아니라 '숫자'의 의미가 들어가 버려, 각 범주의 차, 예를 들어 E(아시아인)-E(백인)=E(백인)-E(아프리카인) 등의 명확한 관계를 이미 가정해버리게 된다. 이는 더미변수가 0,1 값을 가지는 이유. 같은 맥락에서 범주가 2개, 즉 더미변수가 한개뿐이라면 0,1이 아니어도 된다. 어차피 선2개일 것이기에, β_{i1}가 우리의 coding에 맞게 잘 조절되어 나온다.

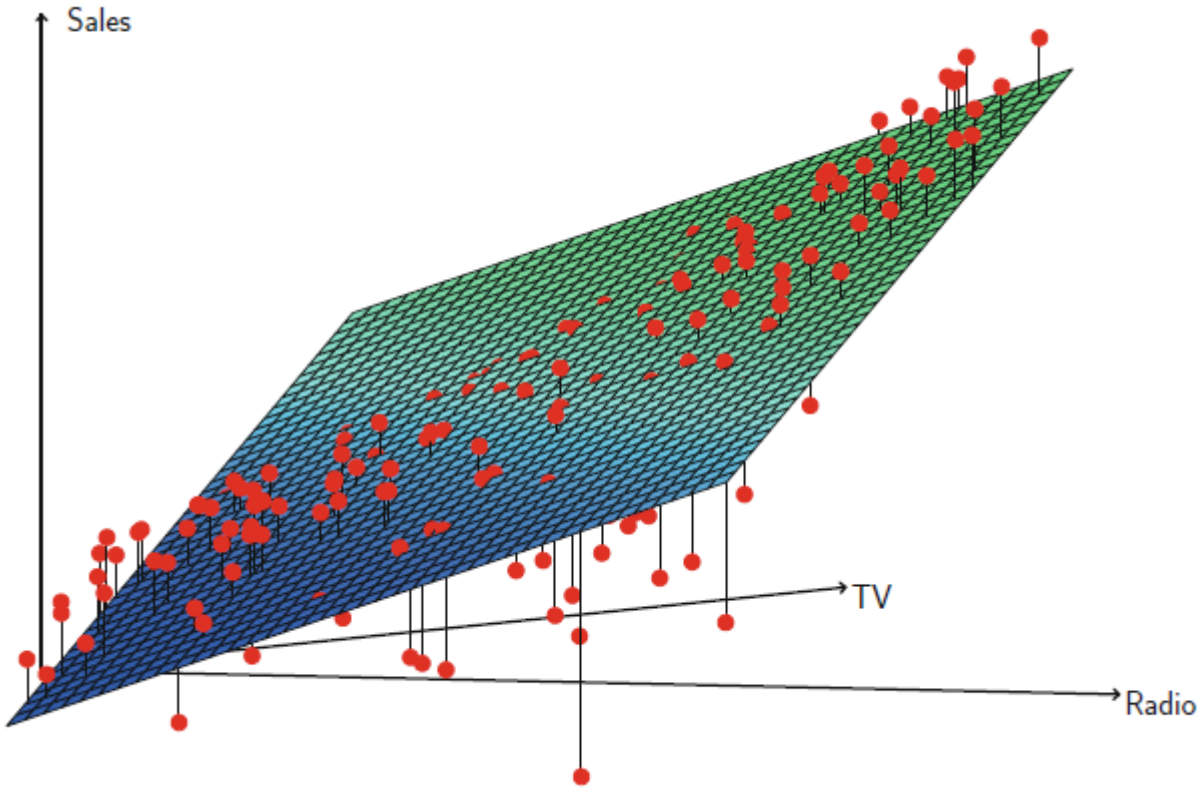
이 경우도 각 회귀계수가 0인지를 p-value를 통해 보며 실제 저 분류가 분석에 유의미한지를 볼 수 있는데, 범주가 2개 이상이어서 **더미변수가 여러개인 경우 개별 p-value가 아닌 $H_0 : \beta_1 = \beta 2 = 0$ 식의 F검정**을 해야 한다(!)

선형 모델의 확장

선형 모델의 가장 중요한 가정 2개. 1)additive : 각각의 예측변수(X)들은 서로 독립이다 2) linear : 선형

additive

선형 모델은 하나의 X_i증가로 인한 Y의 변화는 다른 X_j에 관계없이 상수라고 가정하지만, 그러지 않을때도 있다. TV와 radio에 동시에 투자 했을때 시너지로 인해 sale이 각각 그만큼 투자한 것의 합보다 더 많이 증가하는 경우. 이를 통계에선 **상호작용 효과**라고 한다. 대표적으로 이런 그림.



이 경우 상호작용을 인정하고, **상호 작용항**을 넣어준다.

기존 모델 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$에서 상화작용항을 넣으면

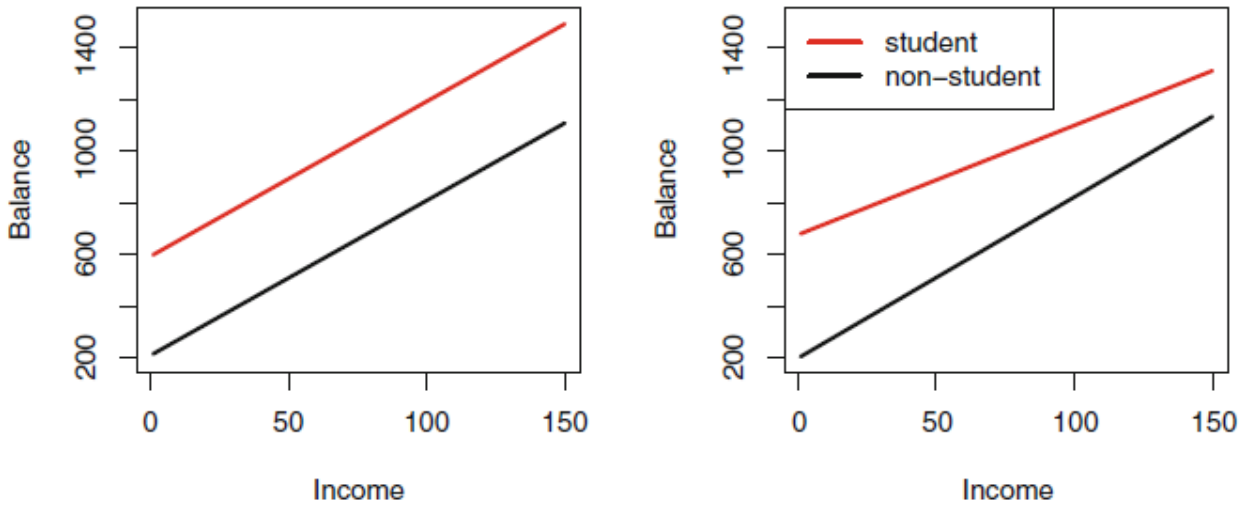
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$ 이렇게 바뀐다.

아주 간단하지만, 이 경우 X_1의 한단위 증가로 인한 Y의 변동은 더이상 상수β_1이 아니다. 이는 다음과 같이 나타내면 더 이해하기 편하다

$Y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$

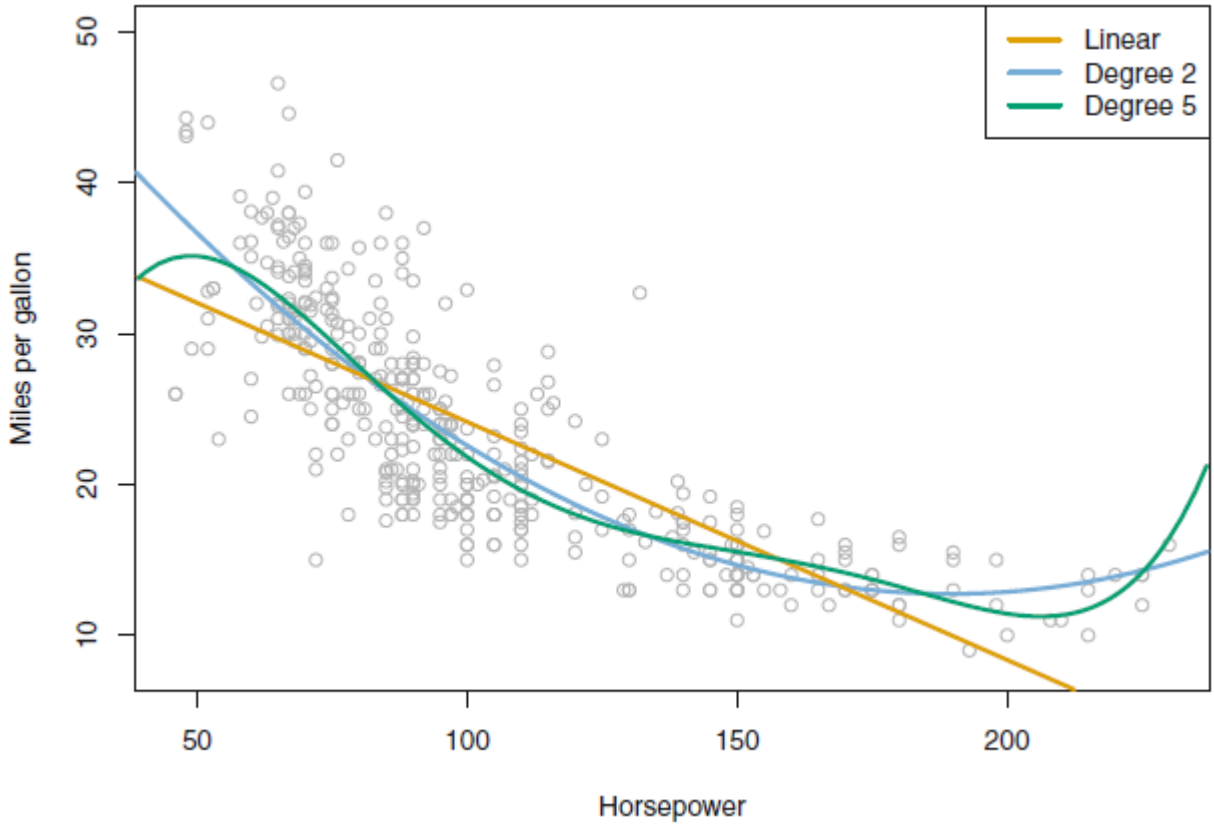
이 경우 X_1의 한단위 증가로 인한 Y의 변동은 $(\beta_1 + \beta_3 X_2)$로, X_2에 따라서 변동하게 되었다. X_2에 대한 경우도 반대로 똑같이 이해하면 된다. β_3역시 p-value를 계산하여 상호작용항이 유의한지 안한지를 판단한다. 이 경우 main항(예시에선 TV, Radio)중 하나는 유의하지 않은데 상호작용항(TV * Radio)이 유의하다 나오는 경우가 있을 수 있는데, **계층적 구조 원칙(archical principle)**에 따라 **p-value가 유의하지 않더라도** 관련된 모든항을 모델에 포함한다. $X_1 X_2$자체가 X_1과 연관되어 있기에, 이들을 빼는건 의도한 의미를 왜곡하게 된다(!).

상호작용항은 질적변수*양적변수의 경우도 그대로 작용한다. 질적변수만 있을때는 기울기는 동일하고 절편이 다른 여러개의 직선이었지만(왼쪽 그림), 상호작용항이 있으면 '기울기도 다른'여러개의 직선이 나오게 된다. (오른쪽 그림)



linear

선형성에 대한 가정을 덜어내는것은 다항회귀(polynomial regression)를 통해 이루어 진다. 쉽게 말해 x^2, x^3등도 고려하는 것이다. 다음 장에서 이에 대해 좀더 다룰 것이다.



이를 보면 우선, 1차 선형은 잘 안맞는 것을 알 수 있다.

2차 다항 회귀의 식은 다음과 같다.

$$mpg = \beta_0 + \beta_1 * horsepower + \beta_2 * horsepower^2 + \epsilon$$

그러나 이 경우도 $horsepower^2$를 X_2처럼 생각하고, 기존의 선형회귀 처럼 적합하면 된다.(만약 완벽한 비선형관계면??? !!!)

진단(!)

진단의 주요과제는 다음과 같다.

1. 반응변수가 선형이 아닐 수도 있을것.
2. error term이 독립이 아니라 correlated되어 있을 수 있을것.
3. error term이 등분산이 아닐수도 있을것.
4. outlier가 있을수도 있을것
5. High-leverage point가 잇을수도 있을것
6. collinearity가 잇을수도 잇을것

여기에선 선형회귀가 최종목적이 아니기에 간단하게만 다루고 넘어간다.

1. 반응변수가 선형이 아닐 수도 있을것.

선형이라 가정했는데 선형이 아니면..도루묵이 된다. 이를 판단하기 위해선 '잔차그림'이 쓰인다. 잔차 ($e_i = y_i - \hat{y}_i$) 와 y_i에 대한 그림을 그리는 식이다.

잔차 그림은 우리의 가정 하에서 특별한 패턴 없이 random하게 분포되어 있어야 하는데, U자형 같이 특정한 패턴이 뚜렷하게 보이면 우리의 가정을 수정할 만한 근거가 될 수 있다. non-linear가 의심될 경우 대표적으로 $\sqrt{X}, X^2$ 등의 변형을 취해준다. (뒷장에서 더 다룸)

2. error term이 독립이 아니라 correlated되어 있을 수 있을것.

우리의 가정은 모든 X수준에서의 error term $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$이 서로 uncorrelate되어 있다는것. 즉 ϵ_{i-1}이 양수가 나오던 음수가 나오던 얼마던지 관계 없이 ϵ_i는 자유롭게 등장한다는 것이다. 만약 실제 error term이 correlated되어 있다면 우리의 추정에 사용되는 표준편차(standard error)를 **더 작게 추정**하게 된다. 이는 신뢰구간 등에 큰 영향을 미치게 된다.

극단적인 예로 같은 데이터를 실수로 2번 집어넣게 되면, (각점에 대해 완벽하게 correlated된 점이 1쌍씩 있게됨) 추정된 회귀계수는 동일하나 standard error는 $\sqrt{(n-2)}$에서 $\sqrt{(2n-2)}$으로 대폭 작게 추정 되게 된다.(!)

그러나 time series data등에서는 빈번하게 error term이 correlated되어 있다. 이를 보기 위해선 역시나 잔차그림이 활용된다. time에 따른 잔차를 그려보는 것

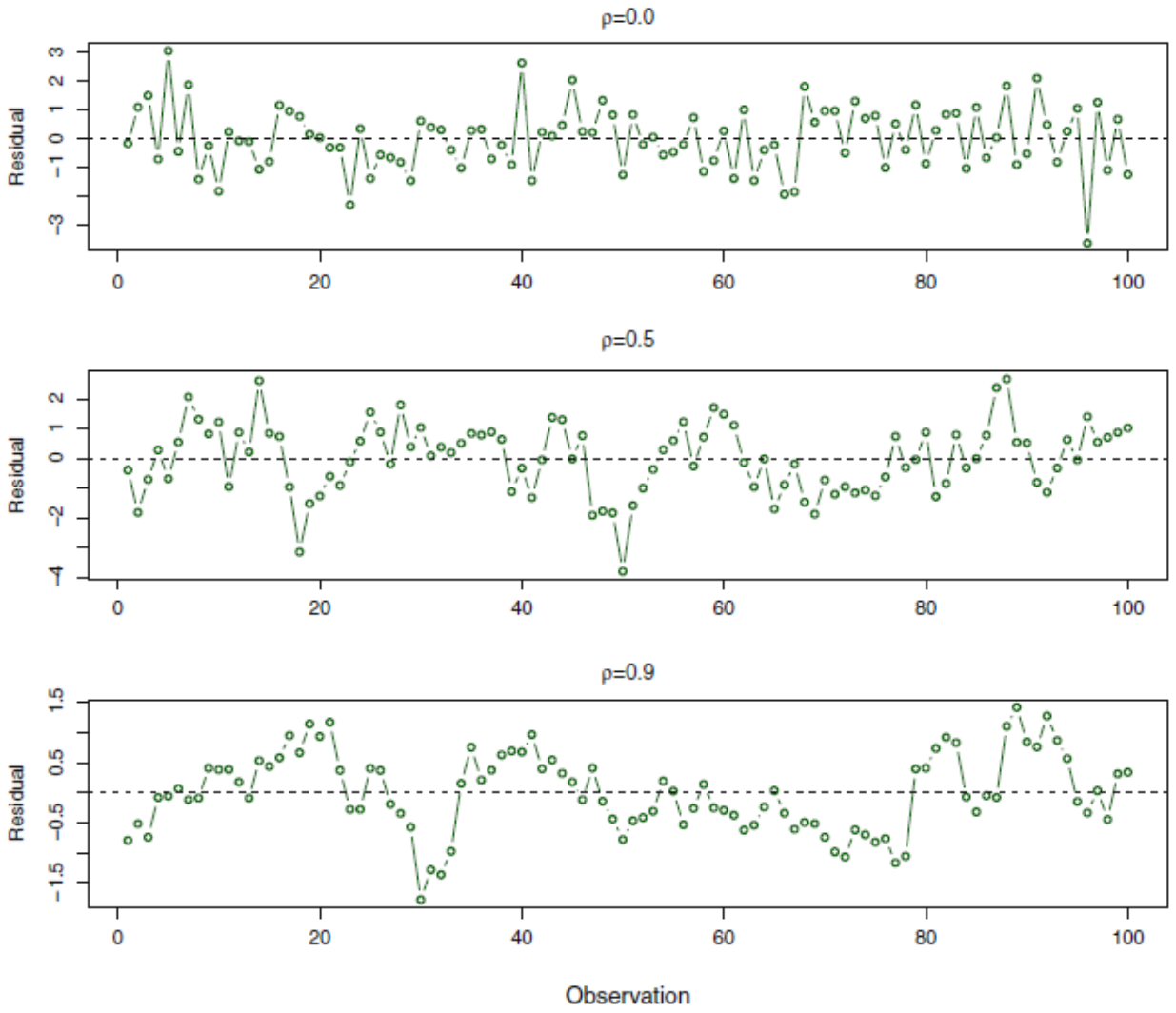
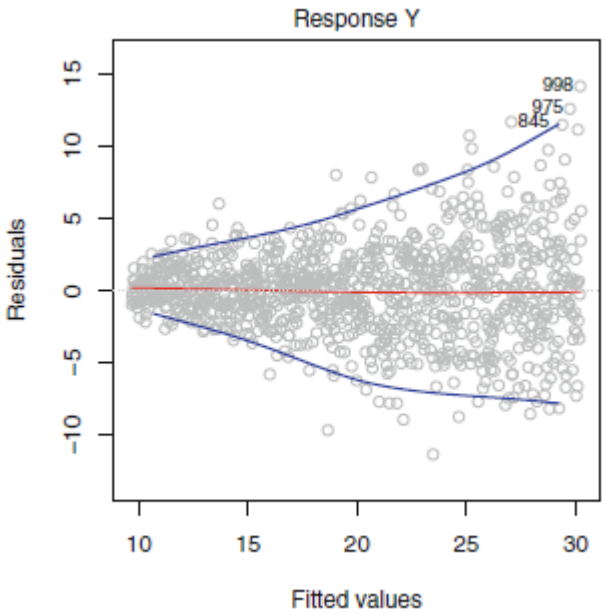


FIGURE 3.10. *Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.*

위 그림을 보면 random하게 지그재그로 분포되어 있는 경우와 시간에 따라 일정한 선이 있는 아래그림이 있다. time series가 아니더라도, **같은 가족에게서 자료가 나온 경우, 같은 생활환경의 사람에서 자료가 수집된 경우** 등 correlated된 경우가 있을 수 있다. error의 uncorrelation은 선형회귀의 중요한 가정이기에, timeseries를 다루기 위한 여러 방법들이 고안 되었다. (???이러고 끝??. timeseries에는 회귀 노노)

3. error term이 등분산이 아닐수도 있을것

모든 X수준에서의 error term $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$은 $Var(\epsilon_i) = \sigma^2$라고 가정하였다. 그러나 실제에서는 잔차그림을 그려보면 일정한 범위 내에서 random하게 분포하는게 아니라(이상적인 잔차) 나팔모양으로 퍼져나가는 등 '이분산성'을 띄는 경우가 있다. 이 경우 다른 통계방법을 쓰기도 하고, 선형회귀를 유지하는 방안으로는 Y에 대해 변환을 해주는 것으로 $logY, \sqrt{Y}$등의 방법이 있다.



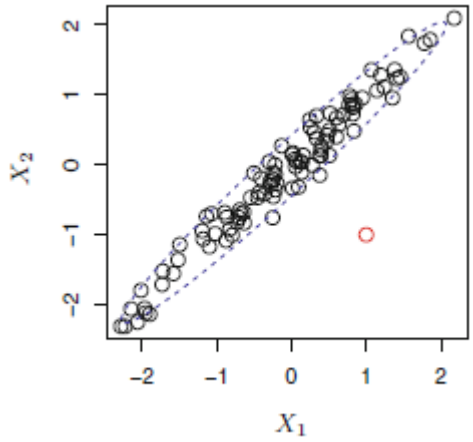
혹은, 자료에 대한 추가적인 정보가 있으면 **weighted least squares**(머지???추가적으로 찾아보자)를 할 수도 있다. 예를 들어 i수준에서 자료가 n_i개 있다면 $\sigma_i^2 = \sigma/n_i$로 가중치를 줘서 계산하는 형식이다.

4. outlier가 있을수도 잇을것. + 5. High-leverage point가 잇을수도 있을것

이 둘의 개념이 헷갈릴 수 있는데, outlier는 **종속변수(Y)가 멀리 떨어져 있는 경우이고**, high-leverage point는 **독립변수(X)가 동떨어져 있는 경우**이다.(!) (더 깊이 알고싶다면 여기) 둘다 좋은건 아닌데, high -leverage point이 더욱 우리의 예측한 선에 상당한 영향을 주는 점이다. 동떨어진 y보다 동떨어진 x가 오히려 예측 선에 큰 영향을 끼치는데, x에 대해 점이 '단 하나' 밖에 없기에 우리의 예측 선이 그 점을 지나가려 노력하기 때문이다. (그래서 지렛대 점이라고 한다) 물론 outlier면서 high leverage point 경우도 많다. 걔네는 우리 예측선에 엄청 영향을 미침.

비록 예측한 선의 회귀계수 자체에 영향을 못주었다 하더라도, outlier는 MSE에 영향을 미쳐 신뢰구간 등 분석에 영향을 미치게 된다. 분석자는 따라서 outlier가 왜 생겼는지(잘못된 데이터인지, 미처 파악못한 흐름인지) 등을 생각하고 모델에 포함할지 제외할지 결정해야 한다. outlier의 파악에도 잔차그림이 쓰이는데, 얼마나 벗어난것인지를 판단하기 위해 단순 잔차그림이 아닌 e_i를 $SE[e_i]$로 나눈 studentized residual이 사용되기도 한다.

그러나 젤 중요한건 high-leverage point인데 우리의 예측에 엄청나게 영향을 미치기 때문, 이는 변수가1개, 2개일 때는 찾기 쉬우나 다중회귀로 가면 눈으로는 볼 수 없다. (그림은 변수가 2개일 경우. 각각 한변수의 관점에서 보았을때는 벗어나는 점을 잡을 수 없다는 것에 주의)



따라서 각 관측값의 leverage를 판단하는 통계량이 따로 있다.

(단순선형회귀의 경우)

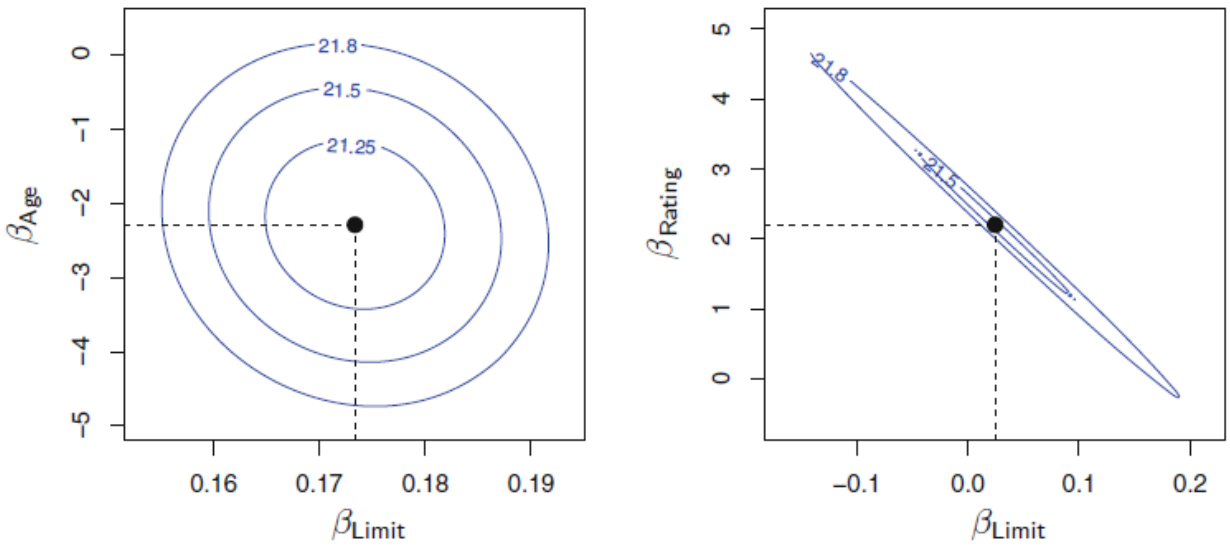
$$h_j = \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

식에서도 x값들의 평균에서 멀리 떨어질 수록 high leverage가 된다는 것을 볼 수 있다.

> (다중선형회귀의 경우)
>
> $h_{ii} = (H)_{ii}$
>
> $H = X(X^TX)^{-1}X^T$ 역시나 계산해보면 single경우와 같아진다

6. collinearity가 잇을수도 잇을것

collinearity는 두 변수, 혹은 '두개 이상'의 변수가 서로 밀접한 **선형관계**가 있는것. 만약 X1과 X2가 거의 1:1로 연결이 되어있다면, 이 경우 Y변수에 대한 **해당 변화가 X1에서 온것인지 X2에서 온것인지** 제대로 알 수 없게되고, 이는 b1, b2에 대한 신뢰계수의 문제로 확장된다.



예시 설명은 않고, Limit 과 Rating이 거의 완벽한 선형결합이다는 것만. 위 그림은 Limit과 Age에 대해 적합한것과 Limit과 rating에 대해 적합한 것인데, 여러 회귀계수들에 대해 같은 SSE를 기준으로 등고선을 그린것 이다. 왼쪽 그림의 경우 최저점(least square로 구한 회귀계수)을 중심으로 원만하게 둥글어 자료의 se를 고려해도 어느정도 자신이 있다 할 수 있지만, 오른쪽 그림은 least square에서 조금만 벗어나도 SSE가 천차 만별이어서, **데이터가 조금**

만 변동(변화)해도 least square점이 천차 만별로 바뀌게 될 것임을 예상할 수 있다. 이는 회귀계수 β_j에 대한 표준편차가 커짐을 의미하고, 이는 $H_0 : \beta_j = 0$을 제대로 기각할 수 없다, 즉 검정의 power가 떨어짐을 의미한다.

이를 방지하기 위해 collinearity를 측정하고자 하는데, 1) 상관행렬을 본다. 그러나 이 방법은 여러개의 변수가 동시에 correlated된 경우를 잡아내지 못해서 2)VIF(다중 공산성)을 체크한다. VIF는 **선형관계가 완벽하게 없을때 최소 1**이다. 실제 분석에서는 **5~10을 넘으면 문제가 있다고 본다**. 심각한 다중공산성을 감지한 경우 1) 단순히 한 변수를 버리거나, 2) 두 변수의 표준화된 값의 평균을 넣는 등 collinear 변수를 결합하는 방식을 택한다.

3. 4 KNN(K-Nearest Neighbors)regression 과의 비교

linear regression은 모수적 방법이다. 앞 장에서도 설명했지만, 모수적(parametric) 방법의 경우 가정된 틀 안의 parameter를 추정하는것으로 문제가 단순화되어 더욱 수월하다는 장점이 있다. 그러나 역시나 가정된 함수가 틀릴경우 도루묵이라는 위험이 있다.

반대로 비모수적 방법은 특정한 가정을 하지 않고 유연한 적합을 하는 것이다. **대표적인 비모수적 방법이 KNN regression**이다. 앞장에서 KNN classifier를 설명했었는데, 이와 근본이 같다. 정해진 상수 K에서, 예측변수 x_0이 주어지면 가지고 있는 데이터에서 가장 가까운 K개의 데이터를 살펴본다. 그리곤 단순히 그들의 Y값을 평균내어 예측을 하는 것이다. 식으로 나타내면 다음과 같다. (N_0이 K개의 점이다.)

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i.$$

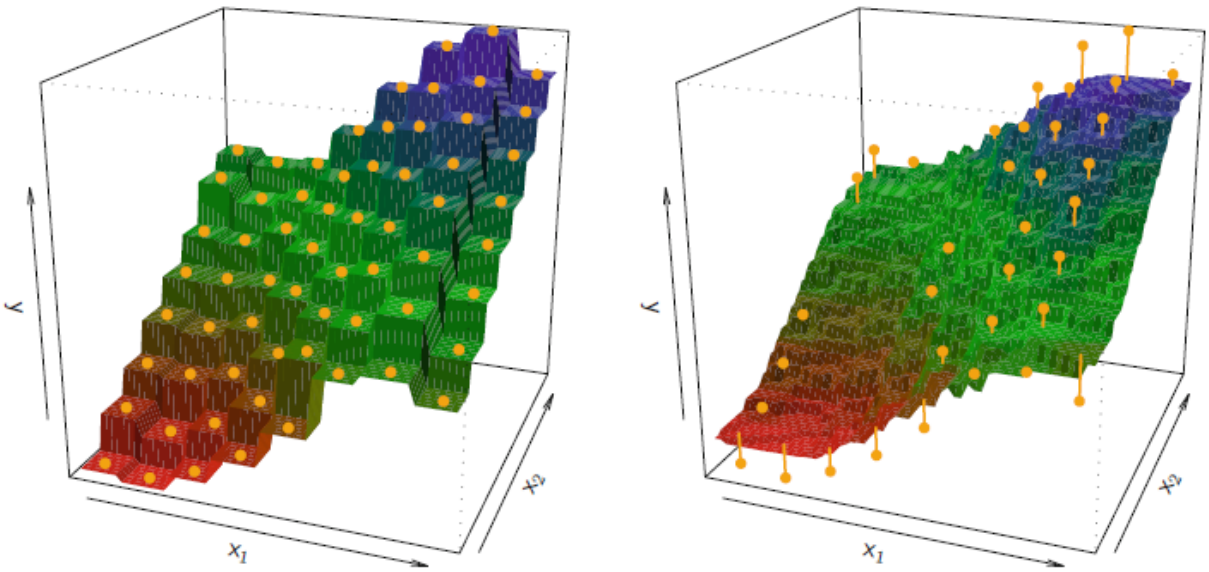
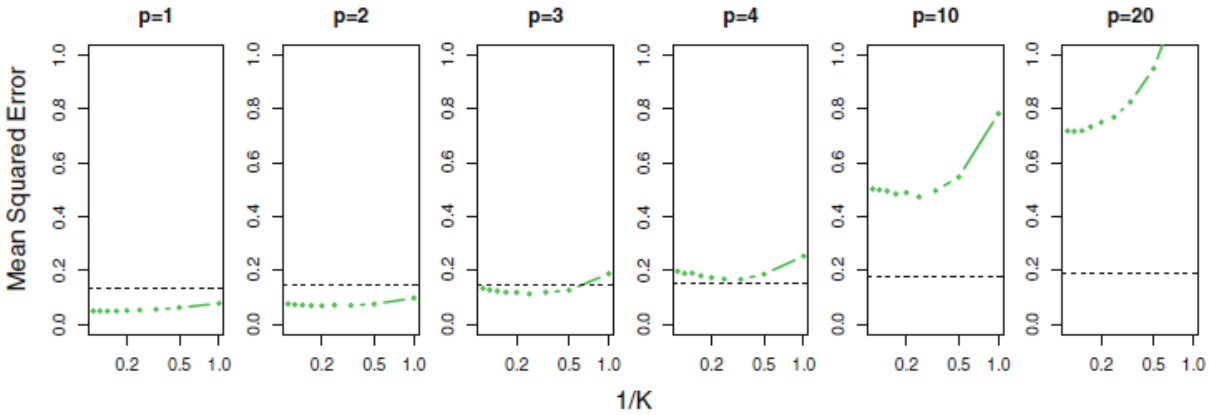


FIGURE 3.16. *Plots of $\hat{f}(X)$ using KNN regression on a two-dimensional data set with 64 observations (orange dots).* Left: $K = 1$ *results in a rough step function fit.* Right: $K = 9$ *produces a much smoother fit.*

위 그림은 각각 K=1개, K=9개로 설정한 KNN regression이다. K=1일때는 각 공간에서 주어진 데이터를 완전히 반영하는 계단형식이고(high Var, low bias), K가 커지면 좀더 많은 데이터를 보며 곡선형태(low Var, high bias)가 되어간다. 앞장에서도 말했듯이, 최적의 K개 설정은 bias-variance trade-off에 따라 **그때그때 다르다**.

당연한 얘기지만, 실제 함수가 linear에 가까울땐 linear regression이 더 좋다. 비모수방식의 유연성은 linear를 정확하게 예측하기 힘들다. 비모수 방식은 그자체로 variance가 큰 방식이기에 K를 늘린다 해도 linear regression만큼의 low Var를 만족할 수 없기 때문.

그러나 놀랍게도 실제 함수가 linear가 아닐때에도, **다중회귀**에 들어가면 **linear regression이 KNN보다 더 좋은 결과를 낸다.(!)** 다음의 그림은 변수갯수 p가 1부터 점점 증가함에 따른 linear regression의 test MSE(회색 점선)과 KNN regression의 여러 flexibility에 따른 test MSE(초록색 곡선)이다.



차원이 2,3차원일때는 KNN이 더 뛰어나지만 차원이 올라갈수록 linear regression은 거의 변화가 없고 KNN은 점점 성능이 눈에 띄게 떨어지게 된다.

이는 유명한 **'차원의 저주(!)'**라는 문제때문이다. 우리개 100개의 training data를 가지고 있을때, p=1, 즉 Y변수까지 해서 2차원일때는 주변 neighbor를 보고 판단할 충분한 자료가 있다. 그러나 차원이 p=20, 즉 21차원이 되버리면, 정확한 판단을 할만한 충분한 비슷한 neighbor 자체가 없어지는 것이다. 즉, 고차원에서는 input x_0에서 가장 가까운 K개들이 멀리 떨어져 있는 점들 밖에 없는 것이다. **쉽게 말하자면** (키: 178, 성별 :남자)인 데이터는 비슷한 near neighbor 데이터가 많겠지만 (키: 178, 성별 :남자, 나이:26, 취미:기타, 수입:300만, 학력:고졸....)인 데이터는 비슷한 near neighbor자체가 없다는 것이다. 이는 결국 KNN의 정확도를 눈에 띄게 떨어뜨리는 결과를 가져온다. 그밖에 linear regression은 해석력이 좋다는 장점 역시 강점으로 작용한다.