Data Mining HW 1

202055364 황 성 윤

Exercises for Linear Regression

1. 설명변수가 1개(X)이고, 반응변수가 1개(Y)인 데이터를 가지고 있다고 하자. ($n=100$) 그리고 다음의 두 모형(linear regression, cubic regression)을 적합하고자 한다.

$Y=\beta_0+\beta_1X+\epsilon$ ... (1)

$Y=\beta_0+\beta_1X+\beta_2X^2+\beta_3X^3+\epsilon$ ... (2)

(a) 실제 X, Y가 선형(linear)관계가 있다고 가정하자. 모델 (1), (2)의 SSE(잔차제곱합)의 크기를 비교할 수 있는지 설명하여라.

solve) 잔차제곱합 SSE는 다음과 같이 정의된다.

$$SSE=\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2$$

즉, 데이터에 있는 각 관측치에 대해 실제 Y_i의 값에서 모형을 통해서 얻어지는 예측값 $\widehat{Y}_i$의 값을 뺀 뒤 제곱해서 합한 값이다. 만약 모델이 반응변수의 값을 실제 반응변수의 값과 가깝게 예측한다면 SSE의 값은 줄어들게 된다. 하지만 실제 X, Y가 선형(linear)관계가 있다고 가정한다면 X, Y 사이에 선형관계가 있다는 가정 하에 만들어진 모델 (1)이 모델 (2)보다 더 합리적인 모델이라고 말할 수 있다. 하지만, 실제 SSE의 값은 모델 (2)가 모델 (1)보다 더 작게 산출될 것이다. 그러므로 이러한 상황에서는 모델 (1), (2)의 SSE의 크기를 비교하는 것이 큰 의미가 없다고 판단된다.

(b) 실제 X, Y가 비선형(non-linear)관계가 있다고 가정하자. 대신 실제 모형에 대한 정보는 없다. 모델 (1), (2)의 SSE(잔차제곱합)의 크기를 비교할 수 있는지 설명하여라.

solve) 만약 실제 X, Y가 비선형(non-linear)관계가 있다고 가정한다면 좀 더 복잡한 모델 (2)가 모델 (1)보다 반응변수의 값을 실제 반응변수의 값에 더 가깝게 예측하게 되고 이에 따라 SSE의 값도 큰 폭으로 줄어들게 될 것이다. 그렇기 때문에 이러한 상황에서는 (a)의 상황과는 다르게 SSE의 값을 통해 두 모델을 평가한다는 것이 합리적이라고 할 수 있다.

2. 'Auto.csv' 데이터를 이용하여 단순선형 회귀 모형을 적합한다.

(a) 반응변수는 mpg, 설명변수는 horsepower로 하는 단순선형회귀모형(simple linear regression model)을 적합한 후 summary() 함수의 결과를 확인하고 다음의 물음에 답하여라.

solve) 함수를 실행하면 다음과 같은 output을 보여준다.

model : $Y = \beta_0 + \beta_1 X + \epsilon$ (Y : mpg, X : horsepower)

```
Call:
lm(formula = mpg ~ horsepower, data = auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,     Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

i. 두 변수 사이에 관계가 있는가?

solve) 모형을 적합한 결과 회귀계수의 추정값은 $\widehat{\beta_1} = -0.1578$ 이고 이에 대한 유의확률은 $p-value < 2\times 10^{-16}$ 으로서 일반적인 유의수준 $\alpha = 0.05$보다 굉장히 작은 값이다. 따라서 이 회귀계수는 두 변수 사이에 매우 유의미한 관계가 있음을 보여준다.

ii. 두 변수 사이의 관계는 얼마나 강한가?

solve) i.에서 설명한 바와 같이 회귀계수의 추정값은 $\widehat{\beta_1} = -0.1578$ 로서 음의 값이다. 하지만 이에 대한 절대값의 크기가 크다고 할 수는 없으므로 약한 음의 관계가 있다고 할 수 있다.

iii. 두 변수는 음의 관계가 있는가? 양의 관계가 있는가?

solve) i.과 ii.를 통해 설명한 바와 같이 두 변수 사이에는 약한 음의 관계가 있으며 매우 유의하다고 할 수 있다.

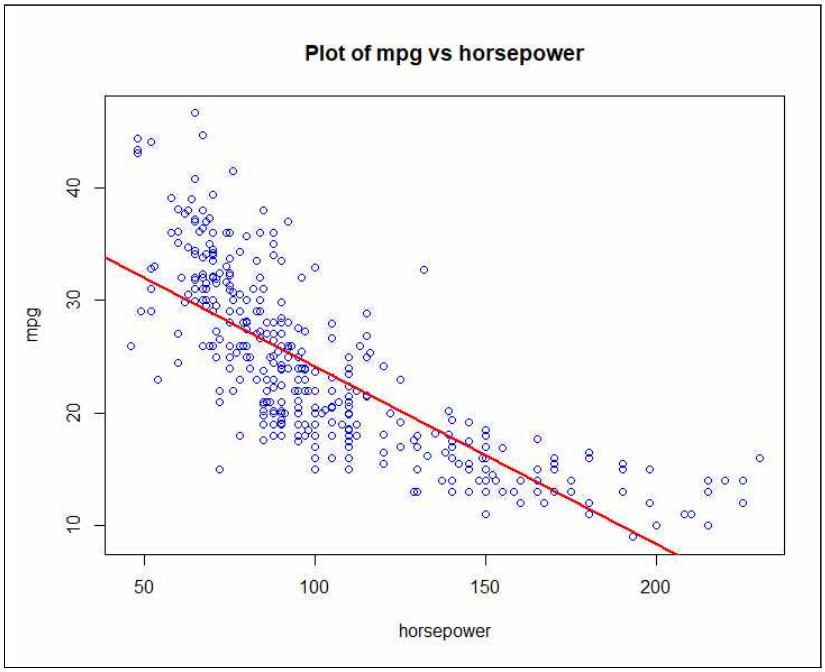
iv. horsepower의 값이 98일 때, mpg의 예측값은 무엇인가, 그리고 95% 신뢰구간은 무엇인가?

solve) output을 통해 얻어지는 예측모델은 다음과 같다.

$\widehat{Y} = 39.9359 - 0.1578X$

이 모델을 통해 horsepower의 값이 98인 경우 R의 함수 predict()를 통해 mpg의 예측값을 구해보면 24.4671이 나오고 95% 신뢰구간은 (23.9731, 24.9611) 이다.

(b) 설명변수와 반응변수의 산점도를 그리고, 회귀직선을 추가하여라. (abline() 사용)
solve)



3. 이 문제는 다중공선성(collinearity)에 관련한 것이다.

(a) R에 다음의 명령문을 실행하여라.

```
set.seed(1)
x1 <- runif(100)
x2 <- 0.5*x1+rnorm(100)/10
y <- 2+2*x1+0.3*x2+rnorm(100)
```

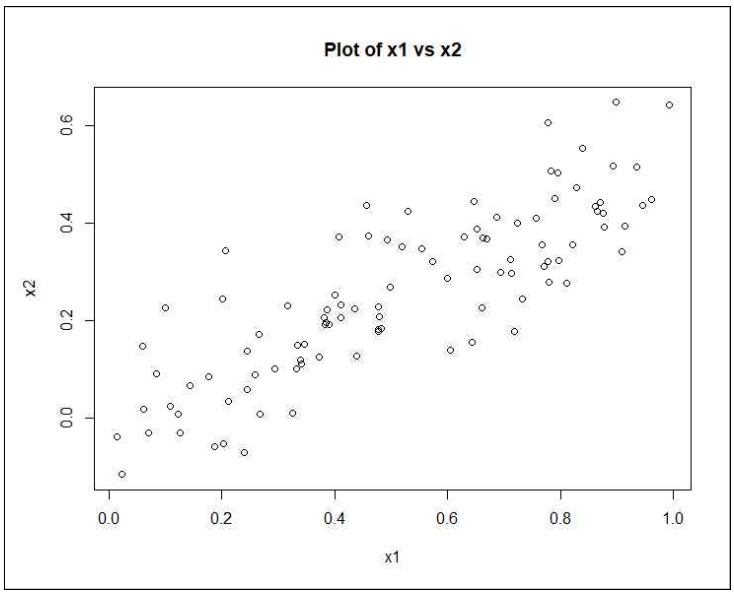
마지막 줄이 두개의 설명변수를 이용한 중회귀모형이다. 회귀모형을 쓰시오. (β 등을 이용하여)

solve) $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i,\ \epsilon_i \sim i.i.d\, N(0, \sigma^2)$

여기에서 각 모수의 참값은 $\beta_0 = 2, \beta_1 = 2, \beta_2 = 0.3, \sigma^2 = 1$이고, $n = 100$이다.

(b) 두 설명변수 x_1과 x_2 사이에 상관관계(correlation)가 있는가? 산점도를 그려서 확인하여라.

solve) 두 설명변수 x_1과 x_2에 대한 산점도는 다음과 같다.



위의 산점도를 통하여 두 설명변수 x_1과 x_2 사이에는 강한 양의 상관관계가 있다고 해석할 수 있다.

(c) 생성된 데이터를 이용하여 (a) 모형의 회귀계수를 추정하여라. 실제 회귀계수와 추정된 회귀계수와 비교하여라. H_0 : $\beta_1 = 0$을 기각할 수 있는가? H_0 : $\beta_2 = 0$을 기각할 수 있는가?
solve) (a) 모형에 대한 적합결과는 다음과 같다.
(각 모수의 참값은 $\beta_0 = 2, \beta_1 = 2, \beta_2 = 0.3, \sigma^2 = 1$임.)

```
Call:
lm(formula = y ~ x1 + x2, data = dt)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8311 -0.7273 -0.0537  0.6338  2.3359

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
x1            1.4396     0.7212   1.996   0.0487 *
x2            1.0097     1.1337   0.891   0.3754
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.056 on 97 degrees of freedom
Multiple R-squared:  0.2088,    Adjusted R-squared:  0.1925
F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

모형적합을 통해 추정된 회귀계수의 값은 $\widehat{\beta_1} = 1.4396$, $\widehat{\beta_2} = 1.0097$ 이다. 이는 실제 회귀계수와 비교했을 때 적지 않은 차이가 있음을 알 수 있다. 그리고 각 회귀계수에 대한 유의확률의 값이 모두 유의수준 $\alpha = 0.05$ 보다 작은 값이므로 두 귀무가설 H_0 : $\beta_1 = 0$과 H_0 : $\beta_2 = 0$을 모두 기각할 수 있음을 알 수 있다.

(d) 이번에는 x_1만을 이용한 단순선형회귀 모형을 적합하여라. 결과를 분석하여라. H_0 : $\beta_1 = 0$을 기각할 수 있는가?
solve) 모형적합 결과는 다음과 같다.

```
Call:
lm(formula = y ~ x1, data = dt)

Residuals:
     Min       1Q   Median       3Q      Max
-2.89495 -0.66874 -0.07785  0.59221  2.45560

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
x1            1.9759     0.3963   4.986 2.66e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.055 on 98 degrees of freedom
Multiple R-squared:  0.2024,    Adjusted R-squared:  0.1942
F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

모형적합 결과 x_1에 대한 유의확률의 값이 유의수준 $\alpha = 0.05$ 보다 매우 작은 값이므로 귀무가설 H_0 : $\beta_1 = 0$을 기각할 수 있다.

(e) 이번에는 x_2만을 이용한 단순선형회귀 모형을 적합하여라. 결과를 분석하여라. H_0 : $\beta_2 = 0$을 기각할 수 있는가?

solve) 모형적합 결과는 다음과 같다.

```
Call:
lm(formula = y ~ x2, data = dt)

Residuals:
     Min       1Q   Median       3Q      Max
-2.62687 -0.75156 -0.03598  0.72383  2.44890

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
x2            2.8996     0.6330    4.58 1.37e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.072 on 98 degrees of freedom
Multiple R-squared:  0.1763,    Adjusted R-squared:  0.1679
F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

모형적합 결과 x_2에 대한 유의확률의 값이 유의수준 $\alpha = 0.05$ 보다 매우 작은 값이므로 귀무가설 H_0 : $\beta_2 = 0$을 기각할 수 있다.

(f) (c)-(e)의 결과가 서로 모순되는가? 설명하여라.

solve) (c), (d), (e)의 결과를 살펴보면 각 설명변수를 하나씩만 사용하여 회귀모형을 적합하면 두 설명변수 x_1과 x_2 모두 반응변수에 대해 유의한 영향을 미친다는 결과를 얻게 된다. 하지만, 두 설명변수 모두를 고려한 다중회귀모형을 적합하면 반응변수 y에 대해 설명변수 x_1은 유의한 영향력이 있지만 설명변수 x_2는 그렇지 않음을 알 수 있다. 이는 설명변수 사이에 연관성으로 인해 발생하는 다중공선성에 의하여 서로 모순되는 결과가 나왔기 때문이다.

(g) 새로운 데이터가 관측되었다고 하자.(이 데이터는 잘못 측정된 것이다.)

```
x1 <- c(x1, 0.1)
x2 <- c(x2, 0.8)
y <- c(y, 6)
```

추가된 데이터를 이용하여 (c)-(e)를 다시 적합하여라. 결과가 어떻게 달라졌는가? 각 모형에서 새로운 데이터는 이상점인가? (잔차가 기존에 있는 데이터에 비해 많이 큰가?) 아니면 영향점인가? (추가된 데이터로 인해 회귀계수의 값이 많이 바뀌었는가?) 설명하여라.

solve) (1) 영향점 여부 판단

먼저 (c)를 다시 적합하면 다음과 같다.

```
Call:
lm(formula = y ~ x1 + x2, data = dtw)

Residuals:
     Min       1Q   Median       3Q      Max
-2.73348 -0.69318 -0.05263  0.66385  2.30619

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
x1            0.5394     0.5922   0.911  0.36458
x2            2.5146     0.8977   2.801  0.00614 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.075 on 98 degrees of freedom
Multiple R-squared:  0.2188,    Adjusted R-squared:  0.2029
F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06
```

이를 통해 새로운 데이터가 적용되기 전의 결과와 비교했을 때 추정된 회귀계수의 값의 변화가 큼을 알 수 있고 유의성 여부도 바뀌었음을 알 수 있다. 이에 따라 새로운 데이터는 영향점이라고 봐도 충분하다고 판단된다.

그리고 (d)를 다시 적합한 결과는 다음과 같다.

```
Call:
lm(formula = y ~ x1, data = dtw)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8897 -0.6556 -0.0909  0.5682  3.5665

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
x1            1.7657     0.4124   4.282 4.29e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.111 on 99 degrees of freedom
Multiple R-squared:  0.1562,    Adjusted R-squared:  0.1477
F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05
```

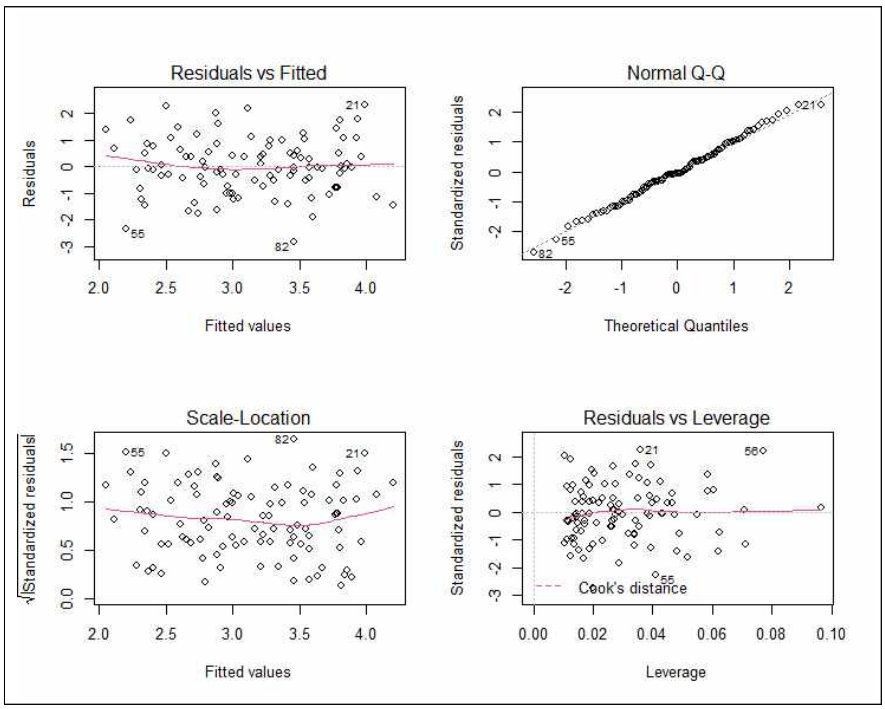
또 (e)를 다시 적합한 결과는 다음과 같다.

```
Call:
lm(formula = y ~ x2, data = dtw)

Residuals:
     Min       1Q   Median       3Q      Max
-2.64729 -0.71021 -0.06899  0.72699  2.38074

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
x2            3.1190     0.6040   5.164 1.25e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.074 on 99 degrees of freedom
Multiple R-squared:  0.2122,    Adjusted R-squared:  0.2042
F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

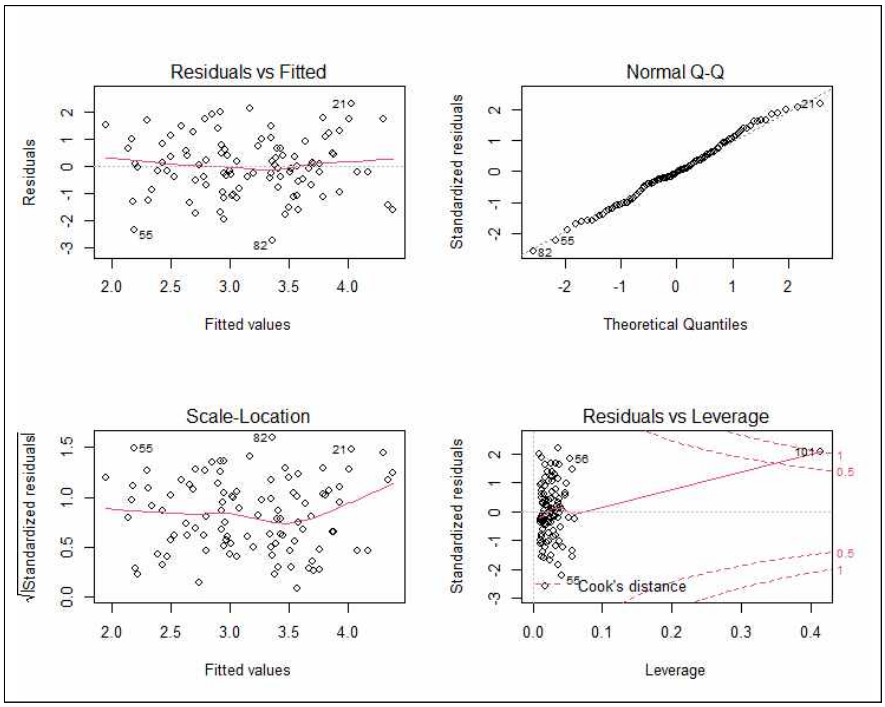
위의 두가지 결과에서는 새로운 데이터가 적용되기 전의 결과와 비교했을 때 추정된 회귀계수의 값이 그렇게 큰 변화를 일으키지는 않았다. 하지만, 중회귀모형을 적합한 경우에는 차이가 심하게 발생하였으므로 새로운 데이터가 영향점이라고 주장할 만한 근거는 충분하다고 보여진다.

(2) 이상점 여부 판단

새로운 데이터가 적용되기 전에 적합한 중회귀모형에 대한 잔차도는 다음과 같다.



그리고 새로운 데이터가 적용된 이후에 적합한 중회귀모형에 대한 잔차도는 다음과 같다.



위의 두가지 결과를 비교해보면 잔차의 값은 크게 변하지 않았음을 알 수 있다. 다만 새로운 데이터의 영향으로 인해 Residual vs Leverage 그림에서 새로운 데이터의 영향력이 크게 부각되어 나타난다는 차이가 있다. 그러므로 새로운 데이터를 이상점이라고 주장하기에는 근거가 부족하다고 판단된다.

Exercises for Logistic Regression

1. 두개의 설명변수 (x_1 = 공부시간, x_2 = 학부평점)를 이용하여 A학점을 받을 확률을 예측하기 위해 로지스틱 회귀모형을 적합하였다. 추정된 회귀계수는 $\widehat{\beta_0} = -6, \widehat{\beta_1} = 0.05, \widehat{\beta_2} = 1$이다.

Logistic regression model : $\log\left(\frac{p(x_1,x_2)}{1-p(x_1,x_2)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

(a) 40시간 공부하고, 평점이 3.5인 학생이 A를 받았을 확률을 예측하여라.
solve) 위에 있는 로지스틱 회귀모형과 추정된 회귀계수의 값을 적용하여 A학점을 받을 확률은 다음과 같은 식으로 추정할 수 있다.

$$\widehat{p(x_1,x_2)} = \frac{\exp(-6+0.05x_1+x_2)}{1+\exp(-6+0.05x_1+x_2)}$$

위의 식에 $x_1 = 40$, $x_2 = 3.5$를 대입하여 구한 확률의 예측값은 약 0.3775이다. (약 37.8%)

(b) 평점이 3.5인 학생은 얼마나 공부를 해야 A를 받을 확률이 50%를 넘을 것인가?
solve) 이 문제는 다음과 같은 부등식을 해결하는 문제와 동일하다.

$$\widehat{p(x_1, x_2 = 3.5)} = \frac{\exp(-6+0.05x_1+3.5)}{1+\exp(-6+0.05x_1+3.5)} > 0.5$$

이 부등식을 풀면 $x_1 > 50$이다. 즉, 이 학생은 공부시간이 50시간을 넘어야 A를 받을 확률이 50%를 넘게 된다.

2. 다음은 odds에 관한 문제이다.

(a) 신용카드결재 문제에서 결재를 하지 못하는 경우(default)에 대한 odds가 0.37인 사람들이 실제로 default할 확률은 평균적으로 얼마인가?
solve) 이 문제는 다음과 같은 식에서 $p(X)$의 값을 구하는 것과 같다.

$$\frac{p(X)}{1-p(X)} = \exp(\beta_0 + \beta_1 X) = 0.37$$

결과적으로 default할 확률은 평균적으로 $p(X) = \frac{37}{63} \approx 0.5873$, 즉 58.7%이다.

(b) 어떤 개인이 default할 확률이 16% 라고 하자. 그 사람이 default할 odds는 얼마인가?
solve) 이 문제에서 주어진 default할 확률은 $p(X) = 0.16$이므로 그 사람이 default할 odds는 다음과 같이 계산된다.

$$\frac{p(X)}{1-p(X)} = \frac{0.16}{1-0.16} \approx 0.1905$$