REPORT

수강과목	:	회귀분석(I)
담당교수	:	김충락
학 과	:	통계학과
학 번	:	201611531
이 름	:	정호재
제출일자	:	2019.05.20

Regression Analysis (I)
Project 1.

Due May 20, 2019

You may use any statistical packages like R, minitab, spsss, sas, etc.

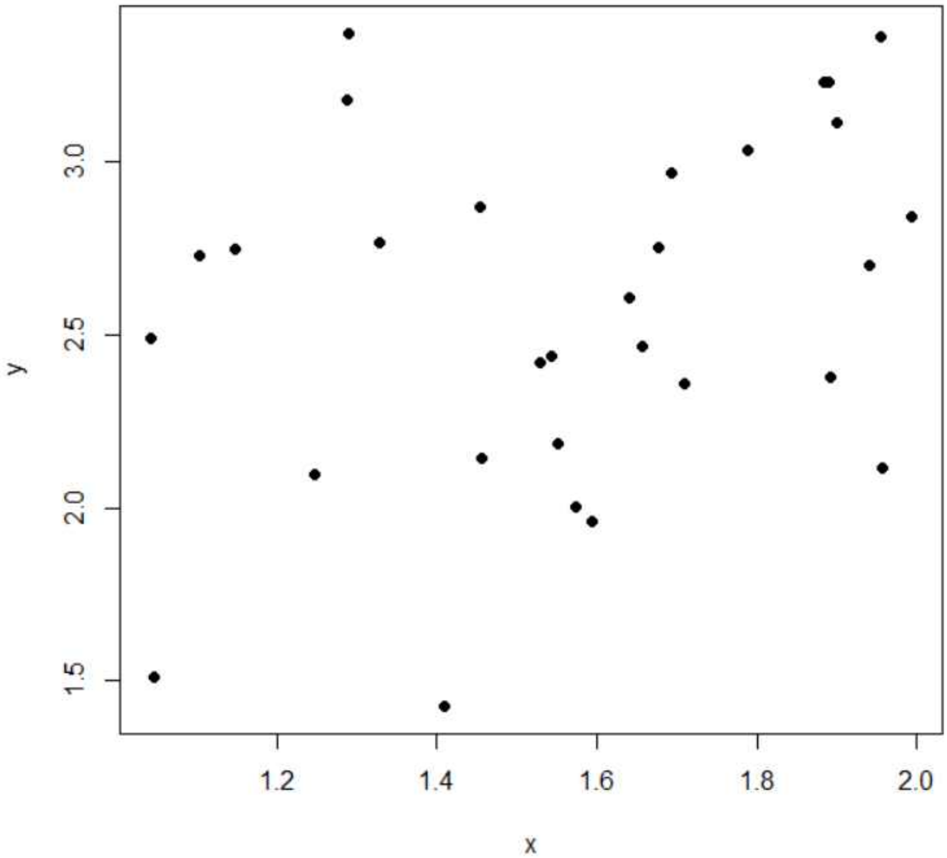
Generation of random numbers

Let $\beta_0 = \beta_1 = 1$. For $i = 1, \cdots, 30$,
$X_i \sim U(1,2)$
$\epsilon_i \sim N(0, 0.5^2)$
$Y_i \leftarrow \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, \cdots, n$

Problems

1.(10) Obtain a scatter plot for the generated random numbers.

```
> x<-runif(30,1,2)
> e<-rnorm(30,0,0.5)
> y<-1+1*x+e
> plot(x,y,type="p",pch=19)
```



2.(30) Compute 95% C.I for β_0 and β_1, and interpret your results.
way1)

Sxx와 Sxy의 식을 이용하여 $\hat{\beta}_0$과 $\hat{\beta}_1$을 구한다.

```
> Sxx<-sum((x-mean(x))^2)
> Sxy<-sum((x-mean(x))*(y-mean(y)))
> b1_hat<-Sxy/Sxx
> b0_hat<-mean(y)-b1_hat*mean(x)
> b0_hat
[1] 0.3127671
> b1_hat
[1] 1.384534
```

σ가 알려져 있지 않기 때문에 s를 사용하여 $\hat{\beta}_0$과 $\hat{\beta}_1$의 표준오차를 추정한다.

```
> s<-sqrt(sum(e^2)/(30-2))
> se_b0<-s*sqrt((1/30)+(mean(x)^2)/Sxx)
> se_b1 <- s/sqrt(Sxx)
```

유의수준이 95%이므로 이때의 t-값을 구한다.

```
> t <- qt(0.975,30-2)
```

$\hat{\beta}_1 - t_{a/2}(n-2)SE(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + t_{a/2}(n-2)SE(\hat{\beta}_1)$
$\hat{\beta}_0 - t_{a/2}(n-2)SE(\hat{\beta}_0) < \beta_0 < \hat{\beta}_0 + t_{a/2}(n-2)SE(\hat{\beta}_0)$ 이므로

```
> 0.3127671+t*se_b0
[1] 1.513299
> 0.3127671-t*se_b0
[1] -0.8877646
> 1.384534+t*se_b1
[1] 2.134254
> 1.384534-t*se_b1
[1] 0.6348138
```

따라서 유의수준 95%에서의 β_0과 β_1의 범위는 아래와 같다.

0.3127671-t*se_b0 < β_0 < 0.3127671+t*se_b0
-0.8877646< β_0 <1.513299

1.384534-t*se_b1 < β_1 < 1.384534+t*se_b1
0.6348138< β_1 <2.134254

way2)
R에 내장되어있는 회귀함수를 사용하여 β_0과 β_1의 유의수준을 구한다.

```
> fit<-lm(y~x)
> confint(fit,level=0.95)

                 2.5 %     97.5 %
(Intercept)  -0.8496633  1.475197
x             0.6586078  2.110460
```

y절편(Intercept, β_0)에 대한 95%의 신뢰구간은 (-0.8496633, 1.475197)
기울기(β_1)에 대한 95%의 신뢰구간은 (0.6586078, 2.110460) 이다.

R에서 식을 이용하여 값을 구할 때 반올림을 하기 때문에 약간의 차이가 있는 것을 알 수 있다.

```
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.41167 -0.18694  0.00569  0.27667  0.84745

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3128     0.5675   0.551 0.585902
x             1.3845     0.3544   3.907 0.000539 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5319 on 28 degrees of freedom
Multiple R-squared:  0.3528,    Adjusted R-squared:  0.3297
F-statistic: 15.26 on 1 and 28 DF,  p-value: 0.0005391

> coef(fit)
 (Intercept)          x
  0.3127671   1.3845341
```

회귀분석 결과 Coefficients: 부분을 살펴보면 y절편(Intercept)은 0.3128, 기울기는 1.3845이다.(더 정확한 값은 coef(fit)을 참고) p값은 0.000539으로 0.05보다 낮다 (95%의 신뢰구간) 즉 y와 x는 $\hat{y} = 0.3128 + 1.3845 \times x$ 과 같은 관계식이 성립한다.

결과의 마지막 세 줄에서
Residual standard error(0.5319)라는 것은 이 모형을 사용하여 x로부터 y를 예측했을 때 평균 0.5319의 오차가 생긴다는 뜻이다.

Multiple R squared가 0.3528이라는 것은 이 모형은 y 분산의 35.28%를 설명해준다는 뜻이다.

F-statistic의 p-value 값은 0.0005391으로 0.05보다 작기 때문에 이 회귀식은 회귀분석 모델 전체에 대해 통계적으로 의미가 있다고 볼 수 있다.

3.(30) Compute 95% C.I for $E(Y)=\beta_0+\beta_1 x$ at $x=1.5$, and interpret your results.

way1)

$SE(\hat{Y}(x))=s\sqrt{\frac{1}{n}+\frac{(x-\overline{X})^2}{Sxx}}$ 이므로

```
> se_y <- s*sqrt((1/30)+(1.5-mean(x))^2/Sxx)
```

$(\hat{\beta}_0+\hat{\beta}_1 x)-t_{a/2}(n-2)SE(\hat{Y}(x)) < E(Y) < (\hat{\beta}_0+\hat{\beta}_1 x)+t_{a/2}(n-2)SE(\hat{Y}(x))$ 이므로

```
> (0.3127671+1.384534*1.5)-t*se_y
[1] 2.17604
> (0.3127671+1.384534*1.5)+t*se_y
[1] 2.603096
```

따라서 $x=1.5$일 때 $E(Y)=\beta_0+\beta_1 x$의 95%의 유의수준은 아래와 같다.

2.17604<E(Y)<2.603096

way2)

R에 내장되어있는 회귀함수를 사용하여 범위를 구한다.

```
> predict(fit, newdata=data.frame(x=1.5),interval="confidence",level=0.95)

       fit      lwr      upr
1 2.389568 2.182817 2.596319
```

X가 특정한 값 x를 가질 때 반응변수의 평균 $E(Y)=\beta_0+\beta_1 x$에 대한 추정

주어진 X의 값 X=x에서 반응 변수 Y의 평균 $E(Y)=\beta_0+\beta_1 x$에 대한 추정을 하기 위해 이에 대한 점 추정치와 이 추정치의 분산을 알아야한다.
우선 Y의 평균값 $E(Y)=\beta_0+\beta_1 x$에 대한 점 추정치로 X=x에서 Y의 예측치 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$을 고려하는 것이 바람직하다.
왜냐하면 이 통계량의 기댓값은 불편성을 만족하기 때문이다.
따라서 E(Y)의 신뢰구간은 Coefficients estimate를 적용하여 구할 수 있다.

95%의 신뢰구간으로 E(Y)를 추정한 결과
E(Y)의 값은 2.389568이고 (2.182817, 2.596319)의 신뢰구간을 가진다.

R에서 식을 이용하여 값을 구할 때 반올림을 하기 때문에 약간의 차이가 있는 것을 알 수 있다.

4.(30) Based on (X_i, Y_i), $i=1,\cdots,30$, compute p-value for testing $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$, where $\rho = Corr(X, Y)$.

```
> cor.test(x,y,level=0.95)

        Pearson's product-moment correlation

data:  x and y
t = 4.2556, df = 28, p-value = 0.000211
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3441363 0.8051792
sample estimates:
      cor
0.6267055
```

유의확률(p-value) 값이 0.000211으로 0.05미만임으로(95%의 신뢰구간) y와 x의 상관이 통계적으로 유의하다.

cor이 양수 0.6267055이므로, y와 x는 한 변수가 증가하면 다른 변수가 증가하는 정비례 관계임을 알 수 있다.