hw3_yw3204

wyh

10/6/2018

2.2

No, because the null hypothesis implies that X and Y have no linear association or a negative linear relationship.

2.23

a.

```
gpa <- read.table("CH01PR19.txt", header = FALSE)
names(gpa) <- c("Y", "X")
reg <- lm(Y~X, data = gpa)
coe <- summary(reg)$coefficients
anova(reg)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq F value   Pr(>F)
## X           1  3.588  3.5878  9.2402 0.002917 **
## Residuals 118 45.818  0.3883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the result above, the ANOVA table is like below:

Sum of Squares	df	MSS	F-Stat	P-value	
$SS_{Regression}$	3.588	1	3.5878	9.2402	0.002917
$SS_{Resigual}$	45.818	118	0.3883		
SS_T	49.406	119	0.4151		

b.

MSR estimates $\sigma^2 + \beta_1^2 \Sigma_{i=1}^n (X_i - \bar{X})^2$ and MSE estimates σ^2. Clearly, When $\beta_1 = 0$, they are the same.

c.

The alternative is: $\beta_1 \neq 0$. We build the following statistics: $F = (\frac{\hat{\beta_1}}{se(\hat{\beta_1})})^2$. Since $\frac{\hat{\beta_1}}{se(\hat{\beta_1})} \sim t(n-2)$ and the relation between t-statitic and F-statistic, we have: $F \sim F(1, n-2)$. Thus, the decision rule is: when $F > F(1, n-2, 0.99)$, we reject the null.

```
F_stat <- (coe[2, 1]/(coe[2, 2])) ^ 2
F_ref <- qf(0.99, 1, 118)
F_stat > F_ref
```

```
## [1] TRUE
```

According to the above result, we reject the null.

d.

According to the ANOVA table, the absolute reduction in Y is 3.588, and the relative reduction is $3.588/49.406 = 0.07262276$. The name is coefficient of determination.

e.

```
R2 <- anova(reg)[1, 2] / (anova(reg)[1, 2] + anova(reg)[2, 2])
r <- sqrt(R2)
r
```

```
## [1] 0.2694818
```

Baesd on the above analysis, we assign a positive sign to r and its value is 0.2694818.

f.

R^2 is the fraction of the model-explained variability in Y and r can be interpreted as linear relationship between X and Y. Thus, R^2 has a more clear-cut interpretation.

2.26

a.

```
plastic <- read.table("CH01PR22.txt")
names(plastic) <- c("Y", "X")
reg_1 <- lm(Y~X, data = plastic)
anova(reg_1)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X          1 5297.5  5297.5  506.51 2.159e-12 ***
## Residuals 14  146.4    10.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the result above, the ANOVA table is like below:

Sum of Squares	df	MSS	F-Stat	P-value	
$SS_{Regression}$	5297.5	1	5297.5	506.51	2.159e-12
$SS_{Resigual}$	146.4	14	10.5		
SS_T	5443.9	15	362.9267		

b.

We construct the following F-statistic: $F = (\frac{b1}{se(b1)})^2$, which has a F-distribution with parameter (1, 14). The alternative is: $\beta_1 \neq 0$, and the decision rule is if P-value is smaller than $\alpha = 0.01$, we reject the null.

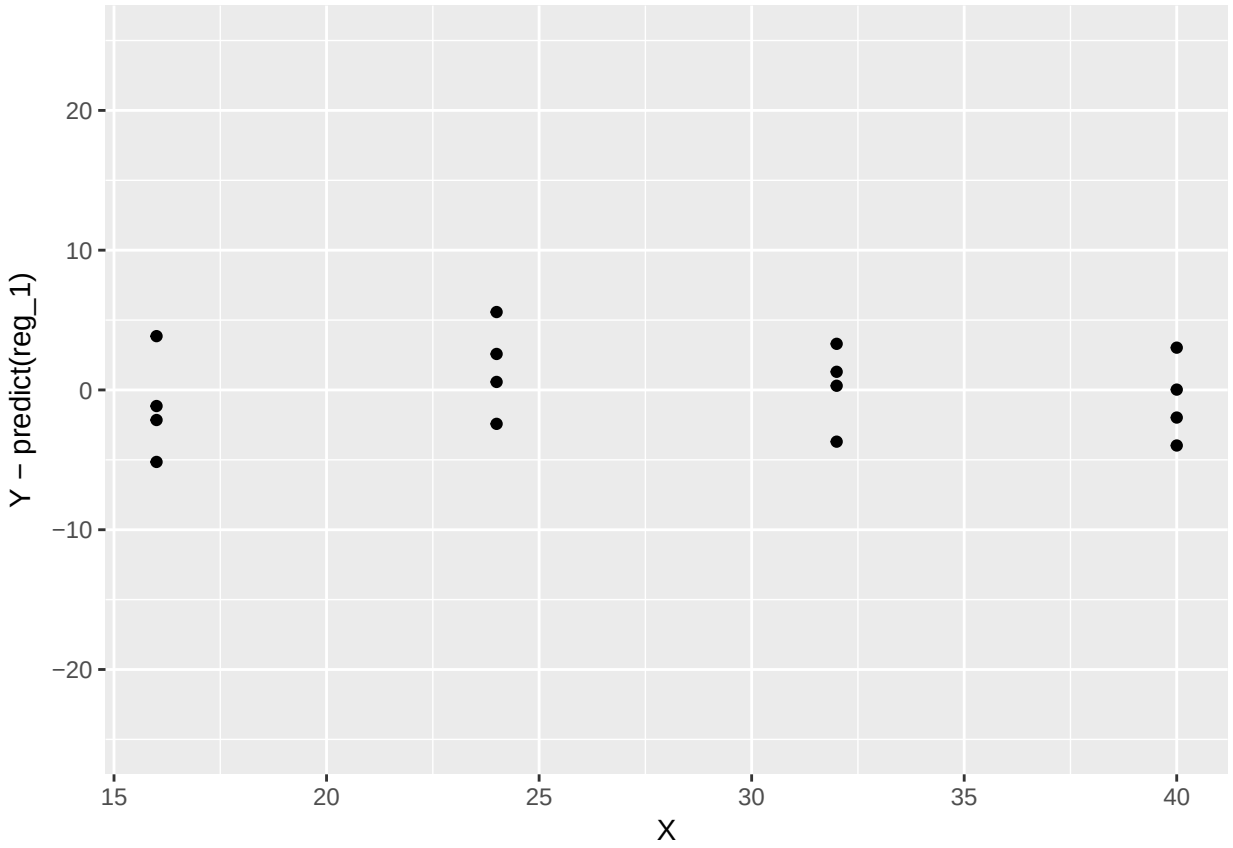
```
coe_1 <- summary(reg_1)$coefficients
F_1 <- (coe_1[2, 1]/coe_1[2, 2])^2
P_1 <- 1-pf(F_1, 1, 14)
P_1 < 0.01
```

```
## [1] TRUE
```

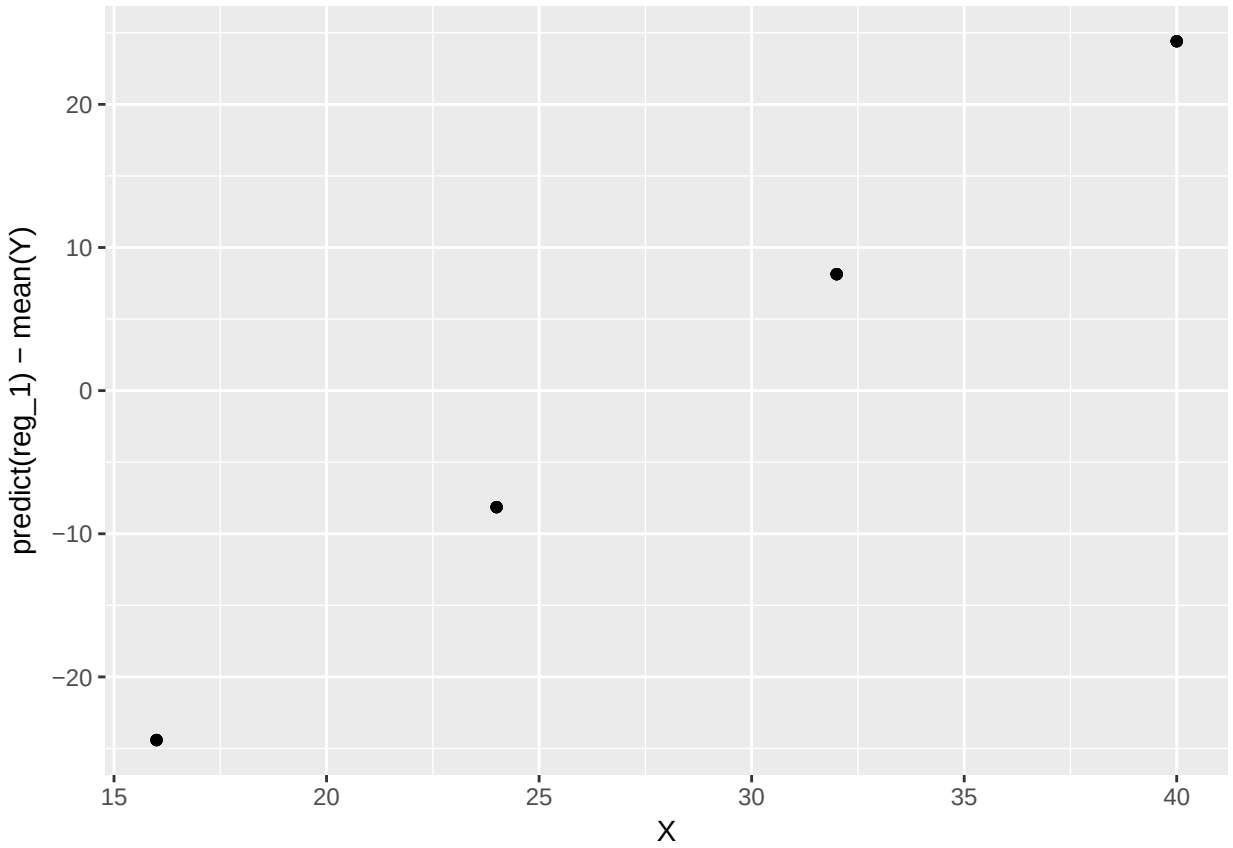
According to the above result, we reject the null.

c.

```
library(ggplot2)
ggplot(plastic, aes(y = Y - predict(reg_1), x = X)) + geom_point() + ylim(c(-25, 25))
```




```
ggplot(plastic, aes(y = predict(reg_1) - mean(Y), x = X)) + geom_point()
```



Clearly, SSR appears to be the larger component of SSTO. This implies the R^2 tends to be close to one.

d.

```
R2_1 <- anova(reg_1)[1, 2] / (anova(reg_1)[1, 2] + anova(reg_1)[2, 2])
r_1 <- sqrt(R2_1)
```

R^2 and r is calculated above and the sign of is positive based on the former analysis.

2.56

a.

For the expected value of MSR, we have:

$\mathbb{E}(MSR)$

$= \mathbb{E}(SSR)$

$= \Sigma_{i=1}^{5}(\hat{Y}_i - \bar{Y})^2$

$= \Sigma_{i=1}^{5}[b_1 X_i + b_0 - (b_1 \bar{X} + b_0)]^2$

$= b_1^2 \Sigma_{i=1}^{5}(X_i - \bar{X})^2$

Since $b_1 \sim N(\beta_1, \frac{\sigma^2}{\Sigma_{i=1}^{5}(X_i - \bar{X})^2})$, we have:

$\mathbb{E}(b_1^2) = \mathbb{E}(b_1)^2 + Var(b_1) = \beta_1^2 + \frac{\sigma^2}{\Sigma_{i=1}^{5}(X_i - \bar{X})^2}$

Thus, the expected value of MSR is:

$\mathbb{E}(MSR) = \mathbb{E}(b_1^2) * \Sigma_{i=1}^5 (X_i - \bar{X})^2$

$= (\beta_1^2 + \frac{\sigma^2}{\Sigma_{i=1}^5 (X_i - \bar{X})^2}) * \Sigma_{i=1}^5 (X_i - \bar{X})^2$

$= \beta_1^2 * \Sigma_{i=1}^5 (X_i - \bar{X})^2 + \sigma^2$

$= 1026.36$

For MSE, since it is an unbiased estimate of σ^2, it is straightforward that: $\mathbb{E}(MSE) = \sigma^2 = 0.36$

b.

We have noticed that the distribution of b_1 is: $b_1 \sim N(\beta_1, \frac{\sigma^2}{\Sigma_{i=1}^5 (X_i - \bar{X})^2})$.

According to this distribution, the larger the denominator of the variance, the more probable that b_1 will be close to the true mean. Thus, it will be worse to make obeservations at the latter points set since their variance is smaller.
While for estimating the mean response, it is the same for both situations. Because the varaince of the true mean are both $\frac{\sigma^2}{n}$.

2.61

We only calculate $\frac{SSR}{SSTO}$ under the situation in which Y_1 is regressed on Y_2, since for another situation, it follows the same procedure.

$(\frac{SSR}{SSTO})_1$

$= \frac{b_1^2 \Sigma (Y_{2i} - \bar{Y_2})^2}{\Sigma (Y_{1i} - \bar{Y_1})^2}$

$= (\frac{\Sigma (Y_{1i} - \bar{Y_1})(Y_{2i} - \bar{Y_2})}{\Sigma (Y_{2i} - \bar{Y_2})^2})^2 * \frac{\Sigma (Y_{2i} - \bar{Y_2})^2}{\Sigma (Y_{1i} - \bar{Y_1})^2}$

$= \frac{(\Sigma (Y_{1i} - \bar{Y_1})(Y_{2i} - \bar{Y_2}))^2}{\Sigma (Y_{1i} - \bar{Y_1})^2 * \Sigma (Y_{2i} - \bar{Y_2})^2}$

For $(\frac{SSR}{SSTO})_2$, it also equals to $\frac{(\Sigma (Y_{1i} - \bar{Y_1})(Y_{2i} - \bar{Y_2}))^2}{\Sigma (Y_{1i} - \bar{Y_1})^2 * \Sigma (Y_{2i} - \bar{Y_2})^2}$.

Thus, it is the same under both occassions.

2.66

a.

```
set.seed(1)
err <- rnorm(5, 0, 5)
X <- seq(4, 20, 4)
Y <- 4*X + 20 + err
reg_2 <- lm(Y~X)
reg_2
```

```
##
## Call:
## lm(formula = Y ~ X)
##
```

```
## Coefficients:
## (Intercept)            X
##      15.661        4.415
```

```r
X_h <- data.frame(X = 10)
predict(reg_2, X_h, se.fit = TRUE, interval = "confidence", level = 0.95)
```

```
## $fit
##        fit      lwr      upr
## 1 59.81546 52.80263 66.82829
##
## $se.fit
## [1] 2.203597
##
## $df
## [1] 3
##
## $residual.scale
## [1] 4.645591
```

Based on the above result, $b_0 = 15.6610$ and $b_1 = 4.4154$. The fitted value is 59.81546 and its confidence interval is $(52.80263, 66.82829)$.

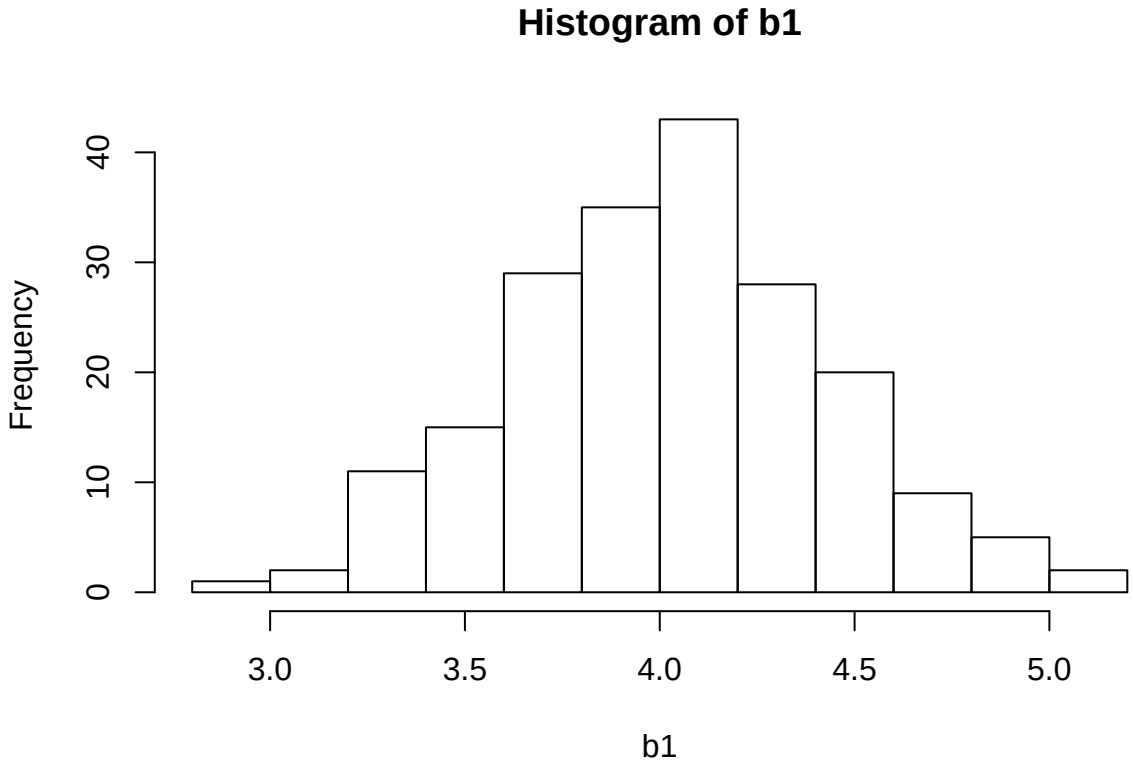
b.

```r
b1 <- c()
CIs <- list()

for (i in 1:200) {
  set.seed(i)
  err <- rnorm(5, 0, 5)
  X <- seq(4, 20, 4)
  Y <- 4*X + 20 + err
  reg <- lm(Y~X)
  ce <- summary(reg)$coefficients
  b1[i] <- ce[2, 1]
  pre <- predict(reg, X_h, se.fit = TRUE, interval = "confidence", level = 0.95)$fit
  CIs[[i]] <- c(pre[2], pre[3])
}
```

c.

```r
hist(b1)
```




```
mean(b1)
```

```
## [1] 4.020114
```

```
sd(b1)
```

```
## [1] 0.4095325
```

Theoretically, the distribution of b_1 is: $b_1 \sim N(4, 0.125)$. And comparing to the result above, they are consistent.

d.

```
p = 0
for(i in 1:200) {
  if(60 >= CIs[[i]][1] && 60 <= CIs[[i]][2]) {
    p = p + 1;
  }
}
p / 200
```

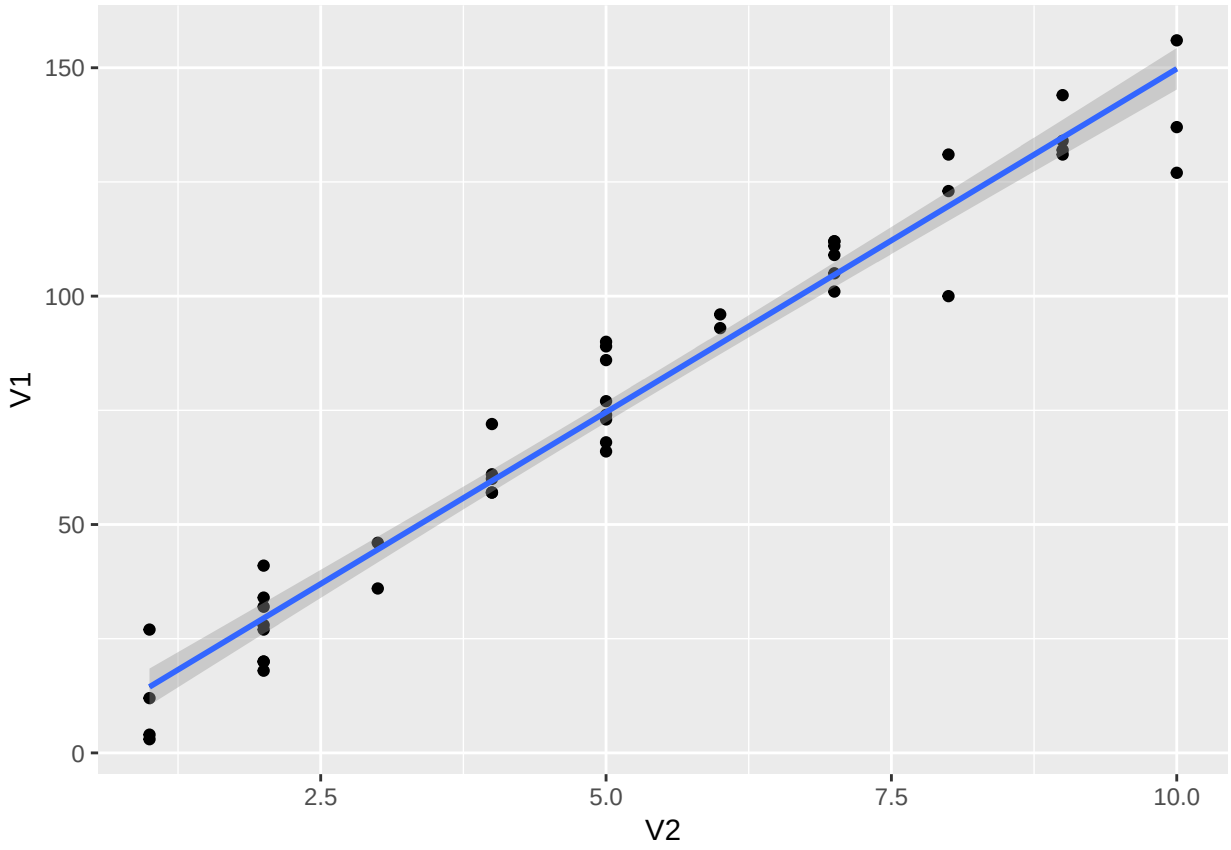
```
## [1] 0.97
```

The proportion is 97% and is consistent with theorectical expectation which is 95%.

2.68

a.

```
copier <- read.table("CH01PR20.txt")
library(ggplot2)
pred <- predict(lm(V1~V2, copier))
ggplot(copier, aes(x = V2, y = V1)) + geom_point() + geom_line(aes(y = pred)) + geom_smooth(method = "lm
```



The plot is provided above.

b.

Observing the superimposed confidence band, it is narrow enough to convince us the regression relationship is well estimated.