Homework Chapter 2

Jinhong Du 15338039

1 $Es\{b_1\}$

$\because \quad \dfrac{(n-2)MSE}{\sigma^2} \sim \chi^2(n-2)$

$\therefore \quad \dfrac{(n-2)SS_{XX}s^2\{b_1\}}{\sigma^2} \sim \chi^2(n-2)$

$\therefore$

$$\begin{aligned}
E\left(\sqrt{\frac{(n-2)SS_{XX}s^2\{b_1\}}{\sigma^2}}\right) &= \int_0^\infty \frac{1}{\Gamma(\frac{n-2}{2})2^{\frac{n-2}{2}}} x^{\frac{n-2}{2}-1}e^{-\frac{x}{2}}x^{\frac{1}{2}}\mathrm{d}x \\
&= \frac{\Gamma(\frac{n-2}{2})2^{\frac{n-2}{2}}}{\Gamma(\frac{n-1}{2})2^{\frac{n-1}{2}}}\int_0^\infty \frac{1}{\Gamma(\frac{n-1}{2})2^{\frac{n-1}{2}}} x^{\frac{n-1}{2}-1}e^{-\frac{x}{2}}\mathrm{d}x \\
&= \frac{\Gamma(\frac{n-2}{2})}{\Gamma(\frac{n-1}{2})2^{\frac{1}{2}}}
\end{aligned}$$

$\therefore$

$$E(s\{b_1\}) = \sqrt{\frac{\sigma^2}{2(n-2)SS_{XX}}}\frac{\Gamma(\frac{n-2}{2})}{\Gamma(\frac{n-1}{2})} \neq \sigma$$

$\therefore \quad s\{b_1\}$ is not a unbiased estimator of σ

2

$$\max_{X_h} \frac{\left(\frac{\hat{Y}_h-\mathbb{E}Y_h}{s\{\hat{Y}_h\}}\right)^2}{2} \sim F(2, n-2),$$

$s\{\hat{Y}_h\} = \sqrt{MSE\left(\frac{1}{n}+\frac{(X_h-\overline{X})^2}{SS_{XX}}\right)}$

$\because$

$$\max_t \frac{(a+bt)^2}{c+dt^2} = \frac{a^2}{c}+\frac{b^2}{d}$$

$\therefore$

$$
\begin{aligned}
\max_i \frac{1}{2}\left(\frac{\hat{Y}_h - \mathbb{E}Y_h}{s\{\hat{Y}_h\}}\right)^2 &= \max_i \frac{1}{2}\frac{[(\hat{\beta}_0 + \hat{\beta}_1 X_h) - (\beta_0 + \beta_1 X_h)]^2}{MSE\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right)} \\
&= \max_i \frac{1}{2}\frac{[(\overline{Y} - \mathbb{E}\overline{Y}) + (\hat{\beta}_1 - \beta_1)(X_h - \overline{X})]^2}{MSE\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right)} \\
&= \frac{1}{2MSE}\frac{(\overline{Y} - \mathbb{E}\overline{Y})^2}{\frac{1}{n}} + \frac{1}{2MSE}\frac{(\hat{\beta}_1 - \beta_1)^2}{\frac{1}{SS_{XX}}} \\
&= \frac{\dfrac{\left(\frac{\overline{Y} - \mathbb{E}\overline{Y}}{\sqrt{\frac{\sigma^2}{n}}}\right)^2 + \left(\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{SS_{XX}}}}\right)^2}{2}}{\dfrac{\frac{SSE}{\sigma^2}}{n-2}}
\end{aligned}
$$

$\because$

$$
\left(\frac{\overline{Y} - \mathbb{E}\overline{Y}}{\sqrt{\frac{\sigma^2}{n}}}\right)^2 \sim \chi^2(1),\ \left(\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{SS_{XX}}}}\right)^2 \sim \chi^2(1),\ \frac{MSE}{\sigma^2} \sim \chi^2(n-2)
$$

$\therefore$

$$
\max_{X_h} \frac{\left(\frac{\hat{Y}_h - \mathbb{E}Y_h}{s\{\hat{Y}_h\}}\right)^2}{2} \sim F(2, n-2)
$$

3

$$
(X, Y) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)
$$

$$
H_0 : \rho = 0.3 \qquad H_1 : \rho \neq 0.3
$$

Do the Fisher z transformation

$$
z' = \frac{1}{2}\ln\left(\frac{1 + r_{12}}{1 - r_{12}}\right)
$$

When n is large, approximately

$$
z' \dot{\sim} N\left(\xi, \frac{1}{n-3}\right)
$$

where

$$
\xi = \frac{1}{2}\ln\left(\frac{1 + \rho_{12}}{1 - \rho_{12}}\right)
$$

$\because$

$$
\rho_{12} = \lim_{\xi \to \infty} \frac{e^{2\xi} - 1}{e^{2\xi} + 1}
$$

$\therefore$

$$
\mathbb{P}\left(\frac{|z' - \xi|}{\sqrt{\frac{1}{n-3}}} < z(1 - \frac{\alpha}{2})\right) = 1 - \alpha
$$

Here, the transform correlation is

$$
z' = \frac{1}{2}\ln\left(\frac{1 + 0.3}{1 - 0.3}\right) = 0.1783375
$$

$\therefore$ the $100(1-\alpha)\%$ confidence interval of ξ is $(z' - z(1-\frac{\alpha}{2})\sqrt{\frac{1}{n-3}}, z' + z(1-\frac{\alpha}{2})\sqrt{\frac{1}{n-3}}) = (c_1, c_2)$

$\therefore$ the $100(1-\alpha)\%$ confidence interval of ρ_{12} is $\left(\frac{e^{2c_1}-1}{e^{2c_1}+1}, \frac{e^{2c_2}-1}{e^{2c_2}+1}\right)$

2.6 Refer to Airfreight breakage Problem 1.21.

(a) Estimate β_1 with a 95 percent confidence interval. Interpret your interval estimate.

```
x <- c(1, 0, 2, 0, 3, 1, 0, 1, 2, 0)
y <- c(16, 9, 17, 12, 22, 13, 8, 15, 19, 11)
data1 <- data.frame(x,y)
fit <- lm('y~x',data1)
confint(fit)
```

```
##                2.5 %    97.5 %
## (Intercept) 8.670370 11.729630
## x           2.918388  5.081612
```

$\because$ $\beta_1 \sim N(\beta_1, \frac{\sigma^2}{SS_{XX}})$

$\therefore$ $\dfrac{b_1 - \beta_1}{\sqrt{\frac{\sigma^2}{SS_{XX}}}} \sim N(0,1)$

$\because$ $\dfrac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}$, b_1 and MSE are independent

$\therefore$ $\dfrac{\frac{b_1-\beta_1}{\sqrt{\frac{\sigma^2}{SS_{XX}}}}}{\frac{(n-2)MSE}{\sigma^2}} = \dfrac{b_1-\beta_1}{\sqrt{\frac{MSE}{SS_{XX}}}} \sim t(n-2)$

$\because$

$$Pr\left\{t(0.025; n-2) < \frac{b_1 - \beta_1}{s\{b_1\}} < t(0.975; n-2)\right\} = 0.95$$

where $s\{b_1\} = \sqrt{\dfrac{MSE}{SS_{XX}}} = 0.469$

$\therefore$ the 95% confident interval of β_1 is $(2.918388, 5.081612)$

(b) Conduct a t test to decide whether or not there is a linear association between number of times a carton is transferred (X) and number of broken ampules (Y). Use a level of significance of .05. State the alternatives, decision rule, and conclusion. What is the P—value of the test?

```
summary(fit)
```

```
##
## Call:
## lm(formula = "y~x", data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.2   -1.2    0.3    0.8    1.8
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  10.2000     0.6633  15.377 3.18e-07 ***
## x             4.0000     0.4690   8.528 2.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.483 on 8 degrees of freedom
## Multiple R-squared:  0.9009, Adjusted R-squared:  0.8885 
## F-statistic: 72.73 on 1 and 8 DF,  p-value: 2.749e-05
```

$$H_0 : \beta_1 = 0 \qquad\qquad H_1 : \beta_1 \neq 0$$
$$t = \frac{b_1 - 0}{s\{b_1\}}$$

If $|t^*| \leqslant t(0.975; 8) = 2.306$ conclude H_0, otherwise H_1.

$$\begin{aligned} t^* &= \frac{b_1 - 0}{s\{b_1\}} \\ &= 8.528 > 2.306 \\ P(|t| < t^*) &= 2.75 \times 10^{-5} \end{aligned}$$

$\therefore$ conclude H_1

(c) β_0 represents here the mean number of ampules broken when no transfers of the shipment are made—i.e., when $X = 0$. Obtain a 95 percent confidence interval for β_0 and interpret it.

$\because$ $\beta_0 = \overline{Y} - \beta_1 \overline{X} \sim N\left(\beta_0, \sigma^2 \frac{\sum_{i=1}^{n} X_i^2}{nSS_{XX}}\right) = N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{SS_{XX}}\right)\right)$, $\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}$, b_0 and MSE are independent

$\therefore$

$$\frac{\frac{b_0 - \beta_0}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{SS_{XX}}\right)}}}{\frac{(n-2)MSE}{\sigma^2}} = \frac{b_0 - \beta_0}{s\{b_0\}} \sim t(n-2)$$

where $s\{b_0\} = \sqrt{MSE\left(\frac{1}{n} + \frac{\overline{X}^2}{SS_{XX}}\right)}$.

$\because$

$$Pr\left\{\left|\frac{b_0 - \beta_0}{s\{b_0\}}\right| < t(0.975; n-2)\right\} = 0.95$$

where $s\{b_0\} = 0.663, b_0 = 10.2$

$\therefore$ the 95% confident interval of β_0 is $(8.670370, 11.729630)$

(d) A consultant has suggested, on the basis of previous experience, that the mean number of broken ampules should not exceed 9.0 when no transfers are made. Conduct an appropriate test, using $\alpha = .025$. State the alternatives, decision rule, and conclusion. What is the P—value of the test?

$$H_0 : \beta_0 \leqslant 9 \qquad H_1 : \beta_1 > 9$$
$$t = \frac{b_0 - 9}{s\{b_0\}}$$

If $t^* \leqslant t(0.975; 8) = 2.306$ conclude H_0, otherwise H_1.

$\because$

$$\begin{aligned} t^* &= \frac{b_0 - 9}{s\{b_1\}} \\ &= 1.810 < 2.306 \\ P(t < t^*) &= 0.053 \end{aligned}$$

$\therefore$ conclude H_0

(e) Obtain the power of your test in part (b) if actually $\beta_1 = 2.0$. Assume $\sigma\{b_1\} = .50$. Also obtain the power of your test in part (d) if actually $\beta_0 : 11$. Assume $\sigma\{b_0\} = .75$.

```
delta1 = abs(2-0)/0.5
print(sprintf('delta1:%f',delta1))
```

```
## [1] "delta1:4.000000"
```

```
print(sprintf('s{b1}:%f',0.4690))
```

```
## [1] "s{b1}:0.469000"
```

```
print(1-pt(qt(0.975,8),8,delta1)+pt(-qt(0.975,8),8,delta1))
```

```
## [1] 0.9367429
```

When $\beta_1 = 2$, $\sigma\{b_1\} = 0.5$, $\frac{b_1}{s\{b_1\}} \sim t(n-2; \delta_1)$,

$$\begin{aligned} Power &= Pr\left\{\left|\frac{b_1}{s\{b_1\}}\right| > t(0.975; n-2)\middle|\delta_1\right\} \\ &= 1 - Pr\left\{-t(0.975; 8) < \frac{b_1}{s\{b_1\}} < t(0.975; 8)\middle|\delta_1\right\} \\ &= 0.9367429 \end{aligned}$$

```
delta0 = abs(11-9)/0.75
print(sprintf('delta0:%f',delta0))
```

```
## [1] "delta0:2.666667"
```

```
print(sprintf('s{b0}:%f',0.6633))
```

```
## [1] "s{b0}:0.663300"
```

```r
print(1-pt(qt(0.95,8),8,delta0))
```

```
## [1] 0.7844117
```

When $\beta_0 = 11$, $\sigma\{b_0\} = 0.75$, $\dfrac{b_0 - 9}{\sigma\{b_0\}} \sim t(n-2;\delta_0)$

$$\begin{aligned} Power &= Pr\left\{\frac{b_0 - 9}{\sigma\{b_0\}} > t(0.975; n-2)\right\} \\ &= 1 - Pr\left\{\frac{b_0 - 9}{\sigma\{b_0\}} \leqslant t(0.975; n-2)\right\} \\ &= 0.7844117 \end{aligned}$$

2.15 Refer to Airfreight breakage Problem 1.21.

(a) Because of changes in airline routes, shipments may have to be transferred more frequently than in the past. Estimate the mean breakage for the following numbers of transfers: $X = 2, 4$. Use separate 99 percent confidence intervals. Interpret your results.

```r
framex = data.frame(x=c(2,4))
pred_a = predict(fit,newdata = framex,se.fit = TRUE,type = "response",
interval = "confidence", level=0.99)
pred_a
```

```
## $fit
##    fit      lwr      upr
## 1 18.2 15.97429 20.42571
## 2 26.2 21.22316 31.17684
##
## $se.fit
##        1        2
## 0.663325 1.483240
##
## $df
## [1] 8
##
## $residual.scale
## [1] 1.48324
```

$\because \quad EY_h = \beta_0 + \beta_1 X_h$

$\therefore \quad \hat{Y}_h = b_0 + b_1 X_h \sim N\left(\beta_0 + \beta_1 X_h, \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right]\sigma^2\right)$

$\because \quad (b_0, b_1, \hat{Y}_h)$ and MSE are independent

$\therefore$

$$\frac{\frac{\hat{Y}_h - EY_h}{\sqrt{\left[\frac{1}{n} + \frac{(X_h-\overline{X})^2}{SS_{XX}}\right]\sigma^2}}}{\sqrt{\frac{\frac{(n-2)MSE}{\sigma^2}}{n-2}}} = \frac{\hat{Y}_h - EY_h}{\sqrt{MSE\left[\frac{1}{n} + \frac{(X_h-\overline{X})^2}{SS_{XX}}\right]}}$$

$$= \frac{\hat{Y} - E\hat{Y}_h}{s\{\hat{Y}_h\}} \sim t(n-2)$$

$\therefore$

$$Pr\left\{\left|\frac{\hat{Y}_h - EY_h}{s\{\hat{Y}_h\}}\right| < t(1-\frac{\alpha}{2}; n-2)\right\} = 1-\alpha$$

Given $X_h = 2$,

$$\hat{Y}_h = 4 \times 2 + 10.2 = 18.2$$

$$s\{\hat{Y}_h\} = 0.663325$$

$$t(0.995; 8) = 3.355$$

Therefore, the predict interval of $\hat{Y}_h$ is $(15.97429, 20.42571)$.

Given $X_h = 4$,

$$s\{\hat{Y}_h\} = 1.483240$$

$$\hat{Y}_h = 4 \times 4 + 10.2 = 26.2$$

$$t(0.995; 8) = 3.355$$

Therefore, the predict interval of $\hat{Y}_h$ is $(21.22316, 31.17684)$.

(b) The next shipment will entail two transfers. Obtain a 99 percent prediction interval for the number of broken ampules for this shipment. Interpret your prediction interval.

```
framex = data.frame(x=c(2))
predict(fit,newdata = framex,se.fit = TRUE,type = "response",
interval = "prediction", level=0.99)
```

```
## $fit
##    fit      lwr      upr
## 1 18.2 12.74814 23.65186
##
## $se.fit
## [1] 0.663325
##
## $df
## [1] 8
##
## $residual.scale
## [1] 1.48324
```

```
pred <- predict(fit,newdata = framex,se.fit = TRUE,type = "response",
interval = "prediction", level=0.99)
s_pred = sqrt(sum(residuals(fit)^2) / df.residual(fit)+(pred$se.fit)^2)
print(sprintf('s{pred}:%f',s_pred))
```

```
## [1] "s{pred}:1.624808"
```

$\because$ $Y_{h(new)} = \beta_0 + \beta_1 X_h + \epsilon_{h(new)}$

$\therefore$ $Y_{h(new)} \sim N(\beta_0 + \beta_1 X_h, \sigma^2)$

$\because$ the prediction of $Y_{h(new)}$ is $\hat{Y}_h = b_0 + b_1 X_h \sim N\left(0, \left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right]\sigma^2\right)$

$\because$ $(Y_{h(new)}, \hat{Y}_h)$ and MSE are independent

$\therefore$

$$\frac{\dfrac{\hat{Y}_{h(new)} - \hat{Y}_h}{\sqrt{\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right]\sigma^2}}}{\sqrt{\dfrac{\dfrac{(n-2)MSE}{\sigma^2}}{n-2}}} = \frac{\hat{Y}_{h(new)} - \hat{Y}_h}{\sqrt{MSE\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right]}}$$

$$= \frac{\hat{Y}_{h(new)} - \hat{Y}_h}{s\{\hat{Y}_h - EY_h\}}$$

$$= \frac{\hat{Y}_{h(new)} - \hat{Y}_h}{s\{pred\}} \sim t(n-2)$$

$\therefore$

$$Pr\left\{\left|\frac{\hat{Y}_h - EY_h}{s\{\hat{Y}_h - EY_h\}}\right| < t(1 - \frac{\alpha}{2}; n-2)\right\} = 1 - \alpha$$

$\because$ $s\{pred\} = 1.624808$

$\therefore$ the prediction interval of $Y_{h(new)}$ is $(12.74814, 23.65186)$.

(c) In the next several days, three independent shipments will be made, each entailing two transfers. obtain a 99 percent prediction interval for the mean number of ampules broken for the three shipments. Convert this interval into a 99 percent prediction interval for the total number of ampules broken in the three shipments.

$\therefore$

$$\frac{\hat{Y}_{h(new)} - \hat{Y}_h}{\sqrt{MSE\left[\frac{1}{3} + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right]}} = \frac{\hat{Y}_{h(new)} - \hat{Y}_h}{s\{predmean\}} \sim t(n-2)$$

```r
framex = data.frame(x=c(2))
predict(fit,newdata = framex,se.fit = TRUE,type = "terms",
interval = "prediction", level=0.99)
```

```
## $fit
##   x
## 1 4
## attr(,"constant")
## [1] 14.2
```

```
##
## $se.fit
##           x
## 1 0.4690416
##
## $lwr
##           x
## 1 -1.219758
## attr(,"constant")
## [1] 14.2
##
## $upr
##          x
## 1 9.219758
## attr(,"constant")
## [1] 14.2
##
## $df
## [1] 8
##
## $residual.scale
## [1] 1.48324
```

```r
pred <- predict(fit,newdata = framex,se.fit = TRUE,type = "response",
interval = "prediction", level=0.99)
s_predmean = sqrt(sum(residuals(fit)^2) / df.residual(fit)/3+(pred$se.fit)^2)
print(sprintf('s{predmean}:%f',s_pred))
```

```
## [1] "s{predmean}:1.624808"
```

```r
print(sprintf('the prediction interval of Yh(new) is (%f,%f)',pred$fit[1]-
qt(.995, df=c(8))*s_predmean,pred$fit[1]+qt(.995, df=c(8))*s_predmean))
```

```
## [1] "the prediction interval of Yh(new) is (14.565427,21.834573)"
```

```r
print(sprintf('the total number of broken ampules is (%f,%f)',(pred$fit[1]-
qt(.995, df=c(8))*s_predmean)*3,(pred$fit[1]+qt(.995, df=c(8))*s_predmean)*3))
```

```
## [1] "the total number of broken ampules is (43.696282,65.503718)"
```

(d) Determine the boundary values of the 99 percent confidence band for the regression line when $X_h = 2$ and when $X_h = 4$. Is your confidence band wider at these two points than the corresponding confidence intervals in part (a)? Should it be?

$\because$

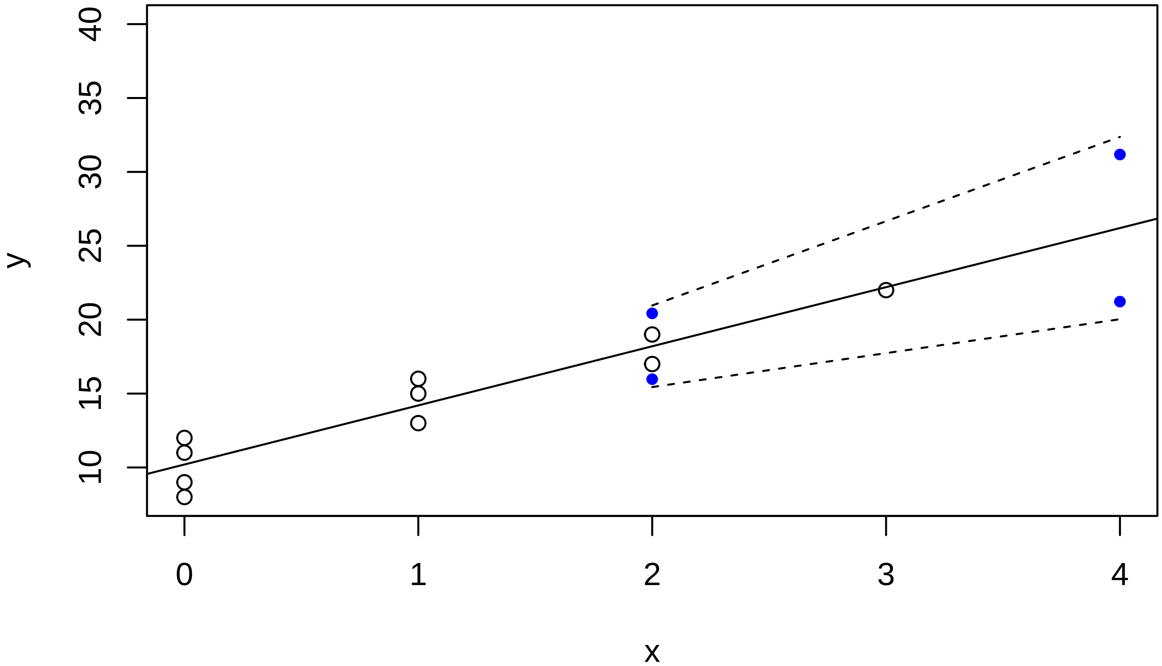
$$W = \sqrt{2F(1-\alpha; 2, n-2)}$$

$\therefore$ the Working-Hotelling Confidence Band is

$$(\hat{Y}_h - Ws\{\hat{Y}_h\}, \hat{Y}_h + Ws\{\hat{Y}_h\})$$

```r
framex = data.frame(x=seq(2,4,2))
preds <- predict(fit, newdata = framex, interval = 'confidence',level=0.99,se.fit = TRUE)
plot(y ~ x, data = data1,xlim=c(0,4),ylim=c(8,40))
abline(fit)
points(seq(2,4,2),pred_a$fit[,2],col='blue',pch=20)
```

```r
points(seq(2,4,2),pred_a$fit[,3],col='blue',pch=20)
#lines(seq(2,4,2), preds$fit[ ,3], lty = 'dashed', col = 'red',type = 'b')
#lines(seq(2,4,2), preds$fit[ ,2], lty = 'dashed', col = 'red',type='b')
w <- sqrt(2*qf(0.99,2,fit$df))
band <- cbind(preds$fit-w*preds$se.fit,preds$fit+w*preds$se.fit)
points(framex$x,band[,1],type = 'l',lty=2)
points(framex$x,band[,4],type = 'l',lty=2)
```




```r
print(sprintf('W=%f',w))
```

```
## [1] "W=4.159113"
```

```r
print(sprintf('Confident band for X=2 is (%f,%f)',band[1,1],band[1,4]))
```

```
## [1] "Confident band for X=2 is (15.441157,20.958843)"
```

```r
print(sprintf('Confident band for X=4 is (%f,%f)',band[2,1],band[2,4]))
```

```
## [1] "Confident band for X=4 is (20.031038,32.368962)"
```

Yes, they are both wider than intervals in (a).

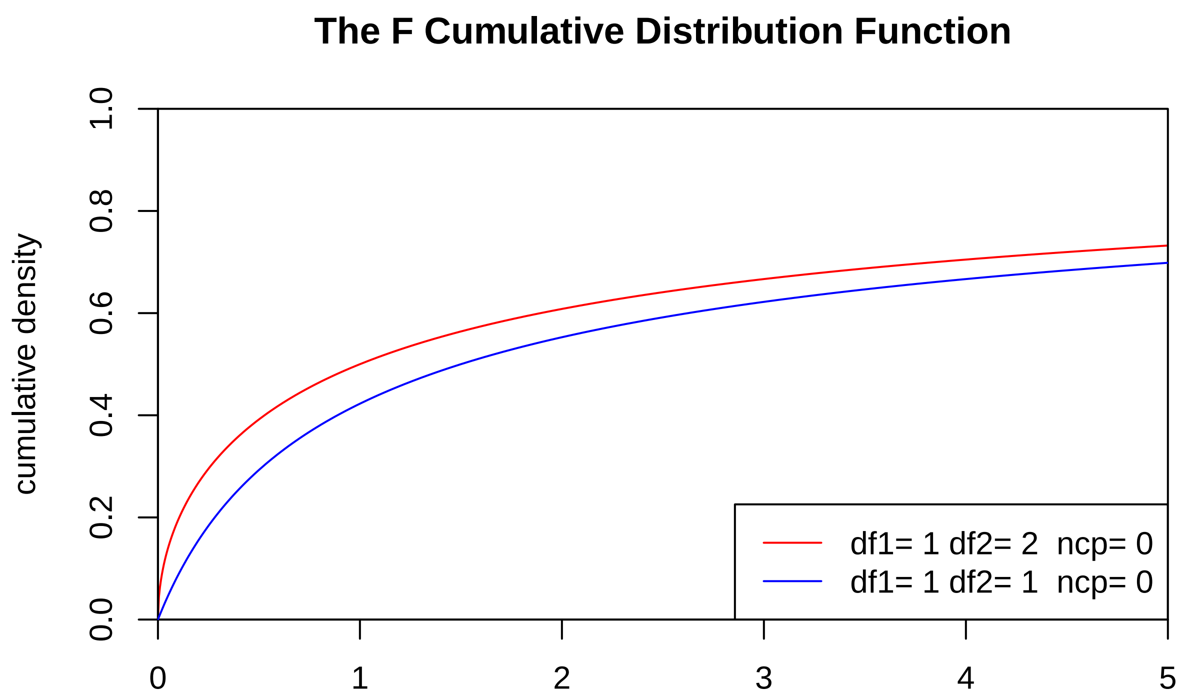
If $F \sim F(1,n-2)$,$T \sim t(n-2)$,then $F = T^2$. Here $F \sim F(2,n-2)$,$T \sim t(n-2)$,,$\sqrt{2F} \geqslant T$ since $F(2,n-2)$ has heavier tail than $F(1,n-1)$.

```r
set.seed(1)
x<-seq(0,5,length.out=1000)
y<-pf(x,1,1,0)

plot(x,y,col="red",xlim=c(0,5),ylim=c(0,1),type='l',
     xaxs="i", yaxs="i",ylab='cumulative density',xlab='',
     main="The F Cumulative Distribution Function")

lines(x,pf(x,2,1,0),col="blue")
```

```r
legend("bottomright",legend=
paste("df1=",c(1,1),"df2=",c(2,1)," ncp=", c(0,0)), lwd=1, col=c("red","blue"))
```



2.25 Refer to Airfreight breakage Problem 1.21.

(a) Set up the ANOVA table. Which elements are additive?

Given

$$H_0 : \beta_1 = 0 \qquad\qquad H_1 : \beta_1 \neq 0$$

$\because$

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

$\therefore$

$$SSTO = SSR + SSE$$

$\because \quad \frac{SSE}{\sigma^2} \sim \chi^2(n-2),\ \frac{SSR}{\sigma^2} = b_1 SS_{XX} \overset{H_0}{\sim} \chi^2(1),\ \frac{SSTO}{\sigma^2} \sim \chi^2(n-1)$ and $\frac{SSE}{\sigma^2} \perp \frac{SSR}{\sigma^2}$

$\therefore$

$$F^* = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{MSR}{MSE} \overset{H_0}{\sim} F(1, n-2)$$

```r
summary(aov(fit))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## x            1  160.0   160.0   72.73 2.75e-05 ***
## Residuals    8   17.6     2.2                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) Conduct a F test to decide whether or not there is a linear association between the number of times a carton is transferred and the number of broken ampules; control the α risk at $.05$. State the alternatives, decision rule, and conclusion.

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

If $F^* \leqslant F(1-\alpha; 1, n-1)$ then conclude H_0. Otherwise conclude H_1.

$\because$

$$\begin{aligned} F^* &= \frac{MSR}{MSE} \\ &= \frac{160}{2.2} \\ &= 72.73 \\ &> F(0.95; 1, 8) = 5.32 \end{aligned}$$

$\therefore$ conclude H_1

(c) Obtain the t^* statistic for the test in part (b) and demonstrate numerically its equivalence to the F^* statistic obtained in part (b).

$\because$

$$b_1 \sim N(\beta_1, \frac{\sigma^2}{SS_{XX}})$$

$\therefore$

$$t^* = \frac{b_1}{\sqrt{\frac{MSE}{SS_{XX}}}} \overset{H_0}{\sim} t(n-2)$$

$$\begin{aligned} t^* &= \frac{MSR}{MSE} \\ &= \frac{160}{2.2} \\ &= 72.73 \\ &> F(0.95; 1, 8) = 5.32 \end{aligned}$$

```
SXX <- sum((data1$x - mean(data1$x))^2)
MSE <- sum(fit$residuals^2)/fit$df
t <- fit$coefficients[2] / sqrt(MSE / SXX)
print(sprintf('t value is %f',t))
```

```
## [1] "t value is 8.528029"
```

```
print(sprintf('square t value is %f=F',t^2))
```

```
## [1] "square t value is 72.727273=F"
```

(d) Calculate R^2 and r. What proportion of the variation in Y is accounted for by introducing X into the regression model?

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}}$$
$$R^2 = r^2$$

```
print(sprintf('r:   %f',cor(data1$x,data1$y)))
```

```
## [1] "r:   0.949158"
```

```
print(sprintf('R^2: %f',cor(data1$x,data1$y)^2))
```

```
## [1] "R^2: 0.900901"
```

```
print(sprintf(
'The proportion of variation in Y accounted for by introducting X into regression model'))
```

```
## [1] "The proportion of variation in Y accounted for by introducting X into regression model"
```

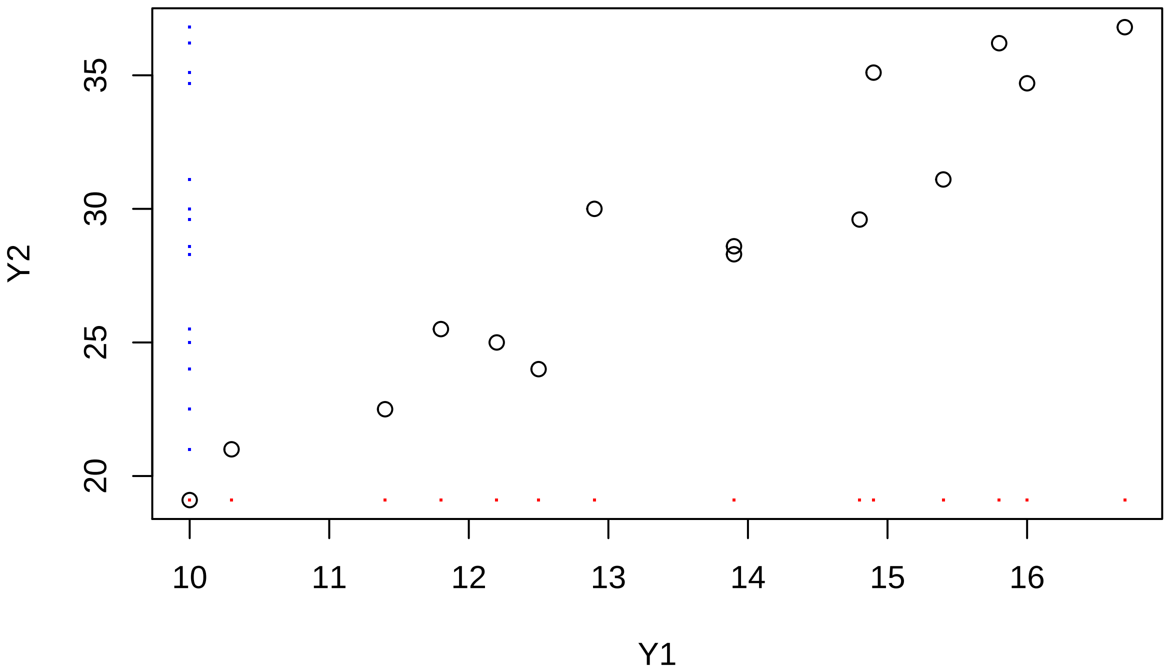
```
print(sprintf(' is %f',cor(data1$x,data1$y)^2))
```

```
## [1] " is 0.900901"
```

2.42 Property assessments. The data that follow show assessed value for property tax purposes ($_1$, in thousand dollars) and sales price (Y_2, in thousand dollars) for a sample of 15 parcels of land for industrial development sold recently in "arm's length" transactions in a tax district. Assume that bivariate normal model (2.74) is appropriate here.

(a) Plot the data in a scatter diagram. Does the bivariate normal model appear to be appropriate here? Discuss.

```
data2 <- read.table("CH02PR42.txt",head=FALSE,col.names = c('Y1','Y2'))
plot(data2)
points(rep(min(data2$Y1),length(data2$Y1)),data2$Y2,col='blue',pch=46)
points(data2$Y1,rep(min(data2$Y2),length(data2$Y2)),col='red',pch=46)
```



The bivariate normal model is apprpriate because it looks lkie Y_1 and Y_2 share a high related coefficient ρ_{12}.

(b) Calculate r_{12}. What parameter is estimated by r_{12}? What is the interpretation of this parameter?

$r_{12} = \dfrac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}}$ is

```
cor(data2)[1,2]
```

```
## [1] 0.9528469
```

ρ_{12} is estimated by r_{12}. It is the related coefficient of Y_1 and Y_2.

(c) Test whether or not Y_1 and Y_2 are statistically independent in the population, using test statistic (2.87) and level of significance .01. State the alternatives, decision rule, and conclusion.

$$H_0 : \rho_{12} = 0 \qquad H_1 : \rho_{12} \neq 0$$

$$\begin{aligned} t^* &= \frac{r_{12}}{\sqrt{\frac{1-r_{12}^2}{n-2}}} \\ &= \frac{b_1}{\sqrt{\frac{MSE}{SS_{XX}}}} \\ &= \frac{b_1}{s\{b_1\}} \overset{H_0}{\sim} t(n-2) \end{aligned}$$

If $|t^*| \leqslant t(1-\frac{\alpha}{2}; n-2)$ then conclude H_0. Otherwise conclude H_1.

Here $t^* = 11.322 > 3.012$, thus conclude H_1.

```r
cor.test(data2$Y1,data2$Y2,conf.level=0.01)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data2$Y1 and data2$Y2
## t = 11.322, df = 13, p-value = 4.187e-08
## alternative hypothesis: true correlation is not equal to 0
## 1 percent confidence interval:
##  0.9525126 0.9531789
## sample estimates:
##       cor
## 0.9528469
```

(d) To test $\rho_{12} = .6$ versus $\rho_{12} \neq .6$, would it be appropriate to use test statistic (2.87)?

No, because the t test is based on $\rho_{12} = 0 \quad \Longleftrightarrow \quad \beta_{12} = \beta_{21} = 0$.

2.46 Refer to Property assessments Problem 2.42. There is some question as to whether or not bivariate model (2.74) is appropriate.

(a) Obtain the Spearman rank correlation coefficient r_s.

Rank $(Y_{11}, Y_{21}, \cdots, Y_{n1})$ from 1 to n and label $(R_{11}, \cdots, R_{n1})$.

Rank $(Y_{12}, Y_{22}, \cdots, Y_{n2})$ from 1 to n and label $(R_{12}, \cdots, R_{n2})$.

$$r_S = \frac{\sum_{i=1}^{n}(R_{i1} - \overline{R}_1)(R_{i2} - \overline{R}_2)}{\sqrt{\sum_{i=1}^{n}(R_{i1} - \overline{R}_1)^2 \sum_{i=1}^{n}(R_{i2} - \overline{R}_2)^2}}$$

```r
cor(data2$Y1,data2$Y2,method = "spearman")
```

```
## [1] 0.9454874
```

(b) Test by means of the Spearman rank correlation coefficient whether an association exists between property assessments and sales prices using test statistic (2.101) with $\alpha = .01$. State the alternatives, decision rule, and conclusion.

To test

$$H_0 : \text{There is no association between } Y_1 \text{ and } Y_2$$
$$H_1 : \text{There is an association between } Y_1 \text{ and } Y_2$$

use

$$t^* = \frac{r_S}{\sqrt{\frac{1-r_S^2}{n-2}}}$$

If $|t^*| \leqslant t(1 - \frac{\alpha}{2}; n-2)$, conclude H_0. Otherwise conclude H_1.

Here, $r_S = 10.46803 > 3.012276$, conclude H_1.

```r
rs <- cor(data2$Y1,data2$Y2,method = "spearman")
t2 = rs/sqrt((1-rs^2)/(length(data2$Y1)-2))
t2
```

```
## [1] 10.46803
```

```r
qt(.995, df=length(data2$Y1)-2)
```

```
## [1] 3.012276
```

(c) How do your estimates and conclusions in parts (a) and (b) compare to those obtained in Problem 2.42?

The Pearson's Correlation Test is more precise than Spearman Rank Correlation Test. Pearson Correlation Coefficient only cares about linear relationship while Spearman Rank Correlation Coefficient cases about association between two groups of data.

2.53 (Calculus needed.)

(a) Obtain the likelihood function for the sample observations $Y_1, \cdots, _n$, given $X_1, \cdots, X_n$ the conditions on page 83 apply.

$\because$

$$Y_i|X_i = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$\therefore$

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2 | X_1, \cdots, X_n, Y_1, \cdots, Y_n) &= f_{Y_1, \cdots, Y_n | X_1, \cdots, X_n}(y_1, \cdots, y_n | x_1, \cdots, x_n) \\ &= \prod_{i=1}^{n} f(y_1|x_1) \\ &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2} \end{aligned}$$

(b) Obtain the maximum likelihood estimators of β_0, β_1 and σ^2. Are the estimators of β_0 and β_1 the same as those in (1.27) when the X_i are fixed?

$$\ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Let

$$\begin{cases} \dfrac{\partial \ln L}{\partial \beta_0} = \dfrac{1}{\sigma^2}\sum\limits_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \dfrac{\partial \ln L}{\partial \beta_1} = \dfrac{1}{\sigma^2}\sum\limits_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \dfrac{\partial \ln L}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\sum\limits_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = 0 \end{cases}$$

we get

$$\begin{cases} \hat{\beta_0} = b_0 \\ \hat{\beta_1} = b_1 \\ \hat{\sigma^2} = \dfrac{\sum\limits_{i=1}^{n}(Y_i - \hat{Y_i})^2}{n} \end{cases}$$

It is the same as (1.27).

2.57 The normal error regression model (2.1) is assumed to be applicable.

(a) When testing $H_0 : \beta_1 = 5$ versus $H_a : \beta_1 \neq 5$ by means of a general linear test, what is the reduced model? What are the degrees of freedom df_R?

The reduced model is

$$Y_i = \beta_0 + 5X_i + \epsilon_i$$

$$df_R = n - 1$$

(b) When testing $H_0 : \beta_0 = 2, \beta_1 = 5$ versus H_a : not both $\beta_0 = 2$ and $\beta_1 = 5$ by means of a general linear test, what is the reduced model? What are the degrees of freedom df_R?

The reduced model is

$$Y_i = 2 + 5X_i + \epsilon_i$$

$$df_R = n$$

2.59 (Calculus needed.)

(a) Obtain the maximum likelihood estimators of the parameters of the bivariate normal distribution in (2.74).

$\because$

$$f(Y_1, Y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} e^{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{Y_1-\mu_1}{\sigma_1}\right)-2\rho_{12}\left(\frac{Y_1-\mu_1}{\sigma_1}\right)\left(\frac{Y_2-\mu_2}{\sigma_2}\right)+\left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2\right]}$$

$\therefore$

$$\begin{aligned} L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho_{12}; y_{11}, \cdots, y_{n1}, y_{12}, \cdots, y_{n2}) &= \prod_{i=1}^{n} f(y_{i1}, y_{i2}) \\ &= \frac{1}{(2\pi\sigma_1\sigma_2)^n\sqrt{1-\rho_{12}^2}} e^{-\frac{1}{2(1-\rho_{12}^2)}\sum_{i=1}^{n}\left[\left(\frac{y_{i1}-\mu_1}{\sigma_1}\right)^2-2\rho_{12}\left(\frac{y_{i1}-\mu_1}{\sigma_1}\right)\left(\frac{y_{i2}-\mu_2}{\sigma_2}\right)+\left(\frac{y_{i2}-\mu_2}{\sigma_2}\right)^2\right]} \end{aligned}$$

(b) Using the results in part (a), obtain the maximum likelihood estimators of parameters of the conditional probability distribution of $_1$ for any value of $_2$ in (2.80).

$$\begin{aligned} &L_{1|2}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho_{12}; y_{11}, \cdots, y_{n1} | y_{12}, \cdots, y_{n2}) \\ =&\frac{L_{12}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho_{12}; y_{11}, \cdots, y_{n1}, y_{12}, \cdots, y_{n2})}{L_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho_{12}; y_{12}, \cdots, y_{n2})} \\ =&\frac{1}{(2\pi\sigma_{1|2}^2)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_{i1}-\alpha_{1|2}-\beta_{12}y_{i2}}{\sigma_{1|2}}\right)^2} \end{aligned}$$

where

$$\begin{cases} \alpha_{1|2} = \mu_1 - \mu_2\rho_{12}\frac{\sigma_1}{\sigma_2} \\ \beta_{12} = \rho_{12}\frac{\sigma_1}{\sigma_2} \\ \sigma_{1|2}^2 = \sigma_1^2(1-\rho_{12}^2) \end{cases}$$

(c) Show that the maximum likelihood estimators of $\alpha_{1|2}$ and β_{12} obtained in part (b) are the same as the least squares estimators (1.10) for the regression coefficients in the simple linear regression model.

$$\ln L_{1|2} = -\frac{n}{2}\ln(2\pi\sigma_{1|2}^2) - \frac{1}{2\sigma_{1|2}^2}\sum_{i=1}^{n}(y_{i1} - \alpha_{1|2} - \beta_{12}y_{i2})^2$$

Let

$$\begin{cases} \dfrac{\partial \ln L_{1|2}}{\partial \alpha_{1|2}} = \frac{1}{\sigma_{1|2}^2}\sum\limits_{i=1}^{n}(y_{i1} - \alpha_{1|2} - \beta_{12}y_{i2}) = 0 \\ \dfrac{\partial \ln L_{1|2}}{\partial \beta_{12}} = \frac{1}{\sigma_{1|2}^2}\sum\limits_{i=1}^{n}(y_{i1} - \alpha_{1|2} - \beta_{12}y_{i2})y_{i2} = 0 \\ \dfrac{\partial \ln L_{1|2}}{\partial \sigma_{1|2}^2} = -\frac{n}{2\sigma_{1|2}^2} + \frac{1}{2\sigma_{1|2}^4}\sum\limits_{i=1}^{n}(y_{i1} - \alpha_{1|2} - \beta_{12}y_{i2})^2 = 0 \end{cases}$$

we get

$$\begin{cases} \hat{\beta}_{12} = \dfrac{\sum\limits_{1}^{n}(Y_{i1} - \overline{Y_1})^2(Y_{i2} - \overline{Y_2})^2}{\sum\limits_{1}^{n}(Y_{i2} - \overline{Y_2})^2} \\ \hat{\alpha}_{1|2} = \overline{Y_1} - \beta_{12}\overline{Y_2} \\ \hat{\sigma}_{1|2}^2 = \dfrac{\sum\limits_{i=1}^{n}(Y_{i1} - \overline{Y_1})^2}{n} \end{cases}$$

It is the same as the least squares estimators for the regression coefficients in the simple linear regression model.

2.60 Show that test statistics (2.17) and (2.87) are equivalent.

$$t^* = \frac{b_1}{s\{b_1\}} \quad \Longleftrightarrow \quad t^* = \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}}$$

$\because$

$$r_{12} = \frac{\sum\limits_{i}^{n}(Y_{i1} - \overline{Y}_1)(Y_{i2} - \overline{Y}_2)}{\sqrt{\left[\sum\limits_{i}^{n}(Y_{i1} - \overline{Y}_1)^2\right]\left[\sum\limits_{i}^{n}(Y_{i2} - \overline{Y}_2)^2\right]}}$$

Let $Y_{i1} = Y_i, Y_{i2} = X_i$, we have

$$r_{12} = \frac{\sum\limits_{i}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sqrt{\left[\sum\limits_{i}^{n}(Y_i - \overline{Y})^2\right]\left[\sum\limits_{i}^{n}(X_i - \overline{X})^2\right]}}$$

$\therefore$

$$
\begin{aligned}
b_1 &= \frac{\sum_i^n (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_i^n (X_i - \overline{X})^2} \\
&= r_{12} \left[\frac{\sum_i^n (Y_i - \overline{Y})^2}{\sum_i^n (X_i - \overline{X})^2} \right]^{\frac{1}{2}} \\
&= r_{12} \sqrt{\frac{SS_{YY}}{SS_{XX}}} \\
MSE &= \frac{SSE}{n-2} \\
&= \frac{SSTO - SSR}{n-2} \\
&= \frac{1}{n-2}(SS_{YY} - b_1^2 SS_{XX}) \\
&= \frac{(1 - r_{12}^2) SS_{YY}}{n-2} \\
s\{b_1\} &= \sqrt{\frac{MSE}{SS_{XX}}} \\
&= \sqrt{\frac{(1 - r_{12}^2) SS_{YY}}{(n-2) SS_{XX}}} \\
t^* &= \frac{b_1}{s\{b_1\}} \\
&= \frac{r_{12}\sqrt{n-2}}{\sqrt{1 - r_{12}^2}}
\end{aligned}
$$