STAT 420 Fall 2015

Examples for 09/22/2015

1. Consider the following data set:

x_1	x_2	y
0	1	11
11	5	15
11	4	13
7	3	14
4	1	0
10	4	19
5	4	16
8	2	8

Consider the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i., \qquad i = 1, \ldots, 8.$$

where e_i's are i.i.d. $N(0, \sigma_e^2)$.

Then $$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 8 & 56 & 24 \\ 56 & 496 & 200 \\ 24 & 200 & 88 \end{bmatrix}, \qquad \mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 96 \\ 740 \\ 336 \end{bmatrix},$$

$$C = (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.7125 & -0.025 & -0.1375 \\ -0.025 & 0.025 & -0.05 \\ -0.1375 & -0.05 & 0.1625 \end{bmatrix},$$

a) Obtain the least-squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 0.7125 & -0.025 & -0.1375 \\ -0.025 & 0.025 & -0.05 \\ -0.1375 & -0.05 & 0.1625 \end{bmatrix} \begin{bmatrix} 96 \\ 740 \\ 336 \end{bmatrix} = \begin{bmatrix} \mathbf{3.7} \\ \mathbf{-0.7} \\ \mathbf{4.4} \end{bmatrix}.$$

$$\text{SYY} = \sum (y - \bar{y})^2 = 240, \qquad \text{RSS} = \sum (y - \hat{y})^2 = 76.4,$$

b) Perform the significance of the regression test at a 5% level of significance.

$H_0: \boldsymbol{\beta_1 = \beta_2 = 0}$. $\qquad$ H_a : at least one of β_1 and β_2 is significantly different from 0.

Source	SS	DF	MS	F
Regression	163.6	$p-1=2$	81.8	**5.3534**
Error (Residual)	76.4	$n-p=5$	15.28	
Total	240	$n-1=7$		

Critical Value: $F_{0.05}(2,5) = \mathbf{5.79}$. Reject H_0 if $F > 5.79$.

Decision: **Do NOT Reject H_0.**

c) Test $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$ at $\alpha = 0.10$. Find the p-value.

$\hat{\text{Var}}(\hat{\beta}_1) = C_{11} \times s^2 = 0.025 \times 15.28 = 0.382$.

Test Statistic: $t = \dfrac{-0.7 - 0}{\sqrt{0.382}} \approx \mathbf{-1.1326}$.

$n - p = \mathbf{5}$ d.f.

$$\begin{array}{ccccc} -1.476 & < & -1.1326 & < & -0.727 \\ -t_{0.10}(5) & < & t & < & -t_{0.25}(5) \end{array}$$

$0.10 <$ left tail < 0.25

p-value = 2 tails. $\mathbf{0.20} <$ p-value $< \mathbf{0.50}$.

(p-value $\approx$ 0.308765.)

Do NOT Reject H_0.

d) Test $H_0: \beta_2 = 0$ vs. $H_a: \beta_2 \neq 0$ at $\alpha = 0.05$. Find the p-value.

$\hat{\text{Var}}(\hat{\beta}_2) = C_{22} \times s^2 = 0.1625 \times 15.28 = 2.483$.

Test Statistic: $t = \dfrac{4.4 - 0}{\sqrt{2.483}} \approx \mathbf{2.792}$.

$n - p = \mathbf{5}$ d.f.

2.517	<	2.792	<	3.365
$t_{0.025}(5)$	<	t	<	$t_{0.01}(5)$

$0.01 <$ right tail < 0.025

p-value = 2 tails. **0.02** < p-value < **0.05**.

(p-value ≈ 0.03834.)

Reject $\mathbf{H_0}$.

a), b), c), d)

```
> x1 = c( 0,11,11, 7, 4,10, 5, 8)
> x2 = c( 1, 5, 4, 3, 1, 4, 4, 2)
> y =  c(11,15,13,14, 0,19,16, 8)
>
> fit = lm(y ~ x1 + x2)
>
> summary(fit)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
   1    2    3    4    5    6    7    8
 2.9 -3.0 -0.6  2.0 -5.3  4.7 -1.8  1.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.7000     3.2995   1.121   0.3131
x1           -0.7000     0.6181  -1.133   0.3088
x2            4.4000     1.5758   2.792   0.0383 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.909 on 5 degrees of freedom
Multiple R-Squared: 0.6817,     Adjusted R-squared: 0.5543
F-statistic: 5.353 on 2 and 5 DF,  p-value: 0.05717
```

b) $\mathrm{H}_0: \beta_1 = \beta_2 = 0$ vs. $\mathrm{H}_a:$ not H_0.

Full: $\mathrm{Y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$,

Null: $\mathrm{Y}_i = \beta_0 \qquad\qquad\qquad\qquad + e_i$.

```
> anova(lm(y~1),fit)
Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x1 + x2
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1      7 240.0
2      5  76.4  2     163.6 5.3534 0.05717 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c) $\quad H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0.$

Full: $\quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$,

Null: $\quad Y_i = \beta_0 \qquad\qquad + \beta_2 x_{i2} + e_i$.

```
> fitc = lm(y ~ x2)
> anova(fitc,fit)
Analysis of Variance Table

Model 1: y ~ x2
Model 2: y ~ x1 + x2
  Res.Df  RSS Df Sum of Sq      F Pr(>F)
1      6 96.0
2      5 76.4  1      19.6 1.2827 0.3088
```

d) $\quad H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0.$

Full: $\quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$,

Null: $\quad Y_i = \beta_0 + \beta_1 x_{i1} \qquad\qquad + e_i$.

```
> fitd = lm(y ~ x1)
> anova(fitd,fit)
Analysis of Variance Table

Model 1: y ~ x1
Model 2: y ~ x1 + x2
  Res.Df    RSS Df Sum of Sq     F  Pr(>F)
1      6 195.54
2      5  76.40  1    119.14 7.797 0.03834 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\mathrm{Y}_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i\,p-1} + e_i, \quad i = 1, 2, \ldots, n,$$

$$\mathrm{E}(e_i) = 0, \quad \mathrm{Var}(e_i) = \sigma^2, \quad \mathrm{Cov}(e_i, e_j) = 0, \;\; i \neq j.$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}, \qquad \mathrm{E}(\boldsymbol{e}) = \mathbf{0}, \quad \mathrm{Var}(\boldsymbol{e}) = (\!(\mathrm{Cov}(e_i, e_j))\!)_{ij} = \sigma^2 \mathbf{I}_n.$$

$$\mathbf{Y} = \begin{bmatrix} \mathrm{Y}_1 \\ \mathrm{Y}_2 \\ \cdots \\ \mathrm{Y}_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1\,p-1} \\ 1 & x_{21} & x_{22} & \ldots & x_{2\,p-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{n\,p-1} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_{p-1} \end{bmatrix}, \quad \boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{bmatrix}.$$

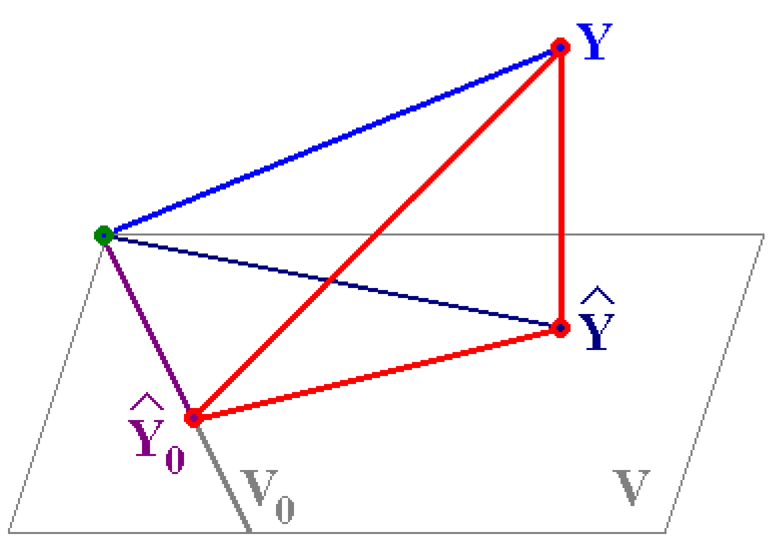
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \qquad \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}.$$

Let $\mathrm{V}_0 = \{ a\,\mathbf{1}, \;\; a \in \mathbf{R} \}$, where $\mathbf{1} = [1, 1, \ldots, 1]^{\mathrm{T}} \in \mathrm{R}^n$,

$$\mathrm{V} = \{ a_0\,\mathbf{1} + a_1 \boldsymbol{x}_{\mathbf{1}} + \ldots + a_{p-1} \boldsymbol{x}_{\boldsymbol{p-1}}, \quad a_0, a_1, \ldots, a_{p-1} \in \mathbf{R} \}.$$

$$\dim(\mathrm{V}_0) = 1, \qquad \dim(\mathrm{V}) = p.$$



$$\sum_{i=1}^{n}\left(\mathrm{Y}_i-\hat{\mathrm{Y}}_{0i}\right)^2=\sum_{i=1}^{n}\left(\mathrm{Y}_i-\hat{\mathrm{Y}}_i\right)^2+\sum_{i=1}^{n}\left(\hat{\mathrm{Y}}_i-\hat{\mathrm{Y}}_{0i}\right)^2 \qquad \hat{\mathbf{Y}}_{\mathbf{0}}=\left[\overline{\mathrm{Y}},\overline{\mathrm{Y}},\ldots,\overline{\mathrm{Y}}\right]^{\mathrm{T}}$$

The significance of the regression test: $\quad \mathrm{H}_0\colon\ \beta_1=\beta_2=\ldots=\beta_{p-1}=0$

Source	SS	df	MS	F
Regr	$\sum_{i=1}^{n}\left(\hat{\mathrm{Y}}_i-\hat{\mathrm{Y}}_{0i}\right)^2$	$\dim(\mathrm{V})-\dim(\mathrm{V}_0)$ $p-1$	$\dfrac{\text{SS Regr}}{\text{df Regr}}$	$\dfrac{\text{MS Regr}}{\text{MS Resid}}$
Resid	$\sum_{i=1}^{n}\left(\mathrm{Y}_i-\hat{\mathrm{Y}}_i\right)^2$	$n-\dim(\mathrm{V})$ $n-p$	$\dfrac{\text{SS Resid}}{\text{df Resid}}$	– estimator for σ^2
Total	$\sum_{i=1}^{n}\left(\mathrm{Y}_i-\hat{\mathrm{Y}}_{0i}\right)^2$	$n-\dim(\mathrm{V}_0)$ $n-1$		

More general: $\quad q<p \qquad \mathrm{H}_0\colon\ \beta_q=\beta_{q+1}=\ldots=\beta_{p-1}=0$

Let $\quad \mathrm{V}_0=\{\,a_0\,\mathbf{1}+a_1\,\boldsymbol{x}_{\mathbf{1}}+\ldots+a_p\,\boldsymbol{x}_{\boldsymbol{q-1}},\quad a_0,a_1,\ldots,a_{q-1}\in\mathbf{R}\,\},$

$\mathrm{V}=\{\,a_0\,\mathbf{1}+a_1\,\boldsymbol{x}_{\mathbf{1}}+\ldots+a_p\,\boldsymbol{x}_{\boldsymbol{p-1}},\quad a_0,a_1,\ldots,a_{p-1}\in\mathbf{R}\,\}.$

	SS	df	MS	F
Diff.	$\sum_{i=1}^{n}\left(\hat{\mathrm{Y}}_i-\hat{\mathrm{Y}}_{0i}\right)^2$	$\dim(\mathrm{V})-\dim(\mathrm{V}_0)$ $p-q$	…	…
Full	$\sum_{i=1}^{n}\left(\mathrm{Y}_i-\hat{\mathrm{Y}}_i\right)^2$	$n-\dim(\mathrm{V})$ $n-p$	…	
Null	$\sum_{i=1}^{n}\left(\mathrm{Y}_i-\hat{\mathrm{Y}}_{0i}\right)^2$	$n-\dim(\mathrm{V}_0)$ $n-q$		

e) Construct a 90% prediction interval for the value of Y at $x_{01} = 2$ and $x_{02} = 3$.

$$X_0^T = [\ 1 \quad 2 \quad 3\ ] \qquad \hat{Y}_0 = 1 \times 3.7 + 2 \times (-0.7) + 3 \times 4.4 = 15.5.$$

$$X_0^T \mathbf{C} X_0 = [1 \quad 2 \quad 3] \cdot \begin{bmatrix} 0.7125 & -0.025 & -0.1375 \\ -0.025 & 0.025 & -0.05 \\ -0.1375 & -0.05 & 0.1625 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = 0.75.$$

$$[\,1 + X_0^T \mathbf{C} X_0\,]\, s^2 = (1 + 0.75) \times 15.28 = 26.74.$$

$$t_{0.05}(5) = 2.015. \qquad 15.5 \pm 2.015 \times \sqrt{26.74} \qquad \mathbf{15.5 \pm 10.42}$$

$$\mathbf{(5.08, 25.92)}$$

```
> predict.lm(fit,data.frame(x1=2,x2=3),interval=c("prediction"),level=0.90)
      fit      lwr      upr
[1,] 15.5 5.080037 25.91996
```

f) Construct a 90% confidence interval for the mean response at $x_{01} = 8$ and $x_{02} = 5$.

$$X_0^T = [\ 1 \quad 8 \quad 5\ ] \qquad \hat{Y}_0 = 1 \times 3.7 + 8 \times (-0.7) + 5 \times 4.4 = 20.1.$$

$$X_0^T \mathbf{C} X_0 = [1 \quad 8 \quad 5] \cdot \begin{bmatrix} 0.7125 & -0.025 & -0.1375 \\ -0.025 & 0.025 & -0.05 \\ -0.1375 & -0.05 & 0.1625 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 8 \\ 5 \end{bmatrix} = 0.60.$$

$$[\,X_0^T \mathbf{C} X_0\,]\, s^2 = (0.60) \times 15.28 = 9.168.$$

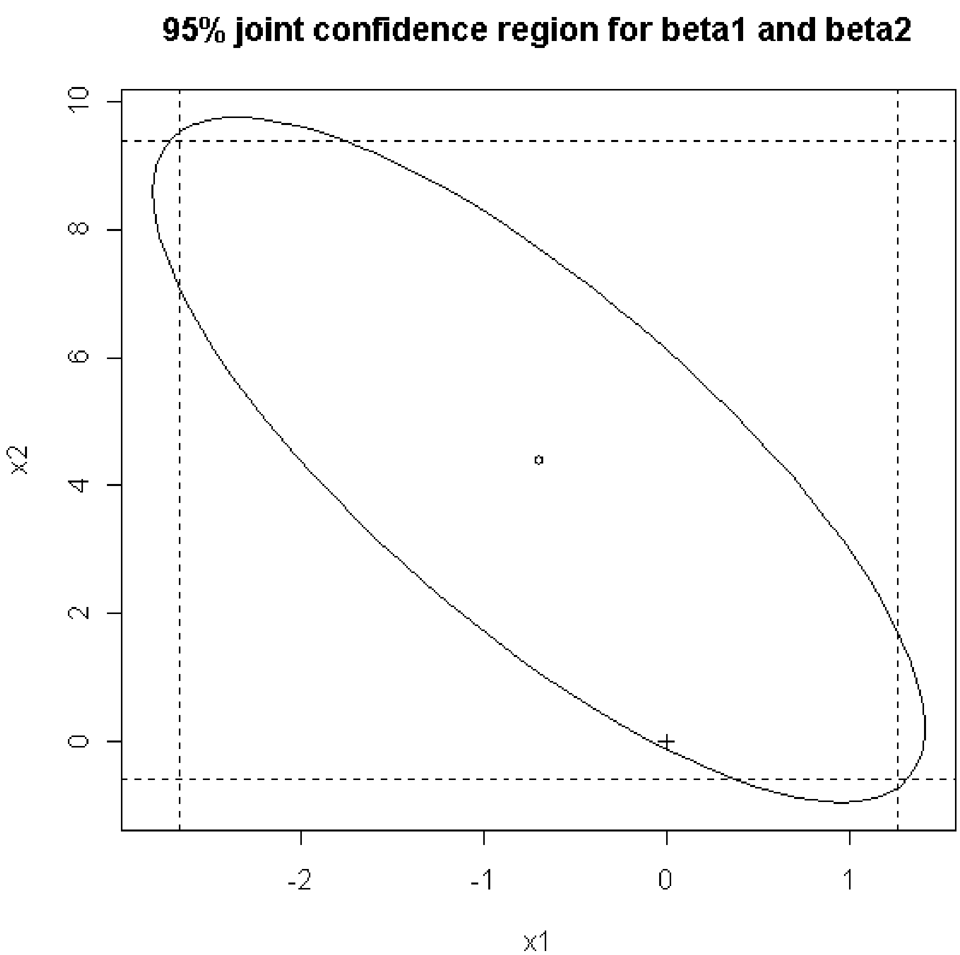
$$t_{0.05}(5) = 2.015. \qquad 20.1 \pm 2.015 \times \sqrt{9.168} \qquad \mathbf{20.1 \pm 6.1}$$

$$\mathbf{(14.0, 26.2)}$$

```
> predict.lm(fit,data.frame(x1=8,x2=5),interval=c("confidence"),level=0.90)
      fit      lwr      upr
[1,] 20.1 13.99869 26.20131
```

```
> x1 = c( 0,11,11, 7, 4,10, 5, 8)
> x2 = c( 1, 5, 4, 3, 1, 4, 4, 2)
> y =  c(11,15,13,14, 0,19,16, 8)
>
> X = cbind( rep(1,8), x1, x2 )
> X
       x1 x2
[1,] 1  0  1
[2,] 1 11  5
[3,] 1 11  4
[4,] 1  7  3
[5,] 1  4  1
[6,] 1 10  4
[7,] 1  5  4
[8,] 1  8  2
>
> t(X) %*% X
       x1  x2
    8  56  24
x1 56 496 200
x2 24 200  88
>
> solve( t(X) %*% X )
               x1      x2
    0.7125 -0.025 -0.1375
x1 -0.0250  0.025 -0.0500
x2 -0.1375 -0.050  0.1625
>
> t(X) %*% y
   [,1]
     96
x1  740
x2  336
>
> solve( t(X) %*% X ) %*% t(X) %*% y
   [,1]
    3.7
x1 -0.7
x2  4.4
```

```
> plot(ellipse(fit,c(2,3),level=0.95),type="l")
> title("95% joint confidence region for beta1 and beta2")
> points(fit$coeff[2],fit$coeff[3])
> points(0,0,pch=3)
>
> confint(fit,"x1",level=0.975)
      1.25 %   98.75 %
x1 -2.655164 1.255164
> confint(fit,"x2",level=0.975)
       1.25 % 98.75 %
x2 -0.5847103 9.38471
> abline(v=confint(fit,"x1",level=0.975),lty=2)
> abline(h=confint(fit,"x2",level=0.975),lty=2)
```



Consider $\mathbf{H_0 : \beta_1 = \beta_2 = 0}$ vs. H_1 : at least one of β_1 and β_2 is not zero.

Recall part (b): **Do NOT Reject $\mathbf{H_0}$** at $\alpha = 0.05$ (p-value = 0.05717).

Note that the origin (0, 0) lies **inside** the 95% joint confidence region for β_1 and β_2.

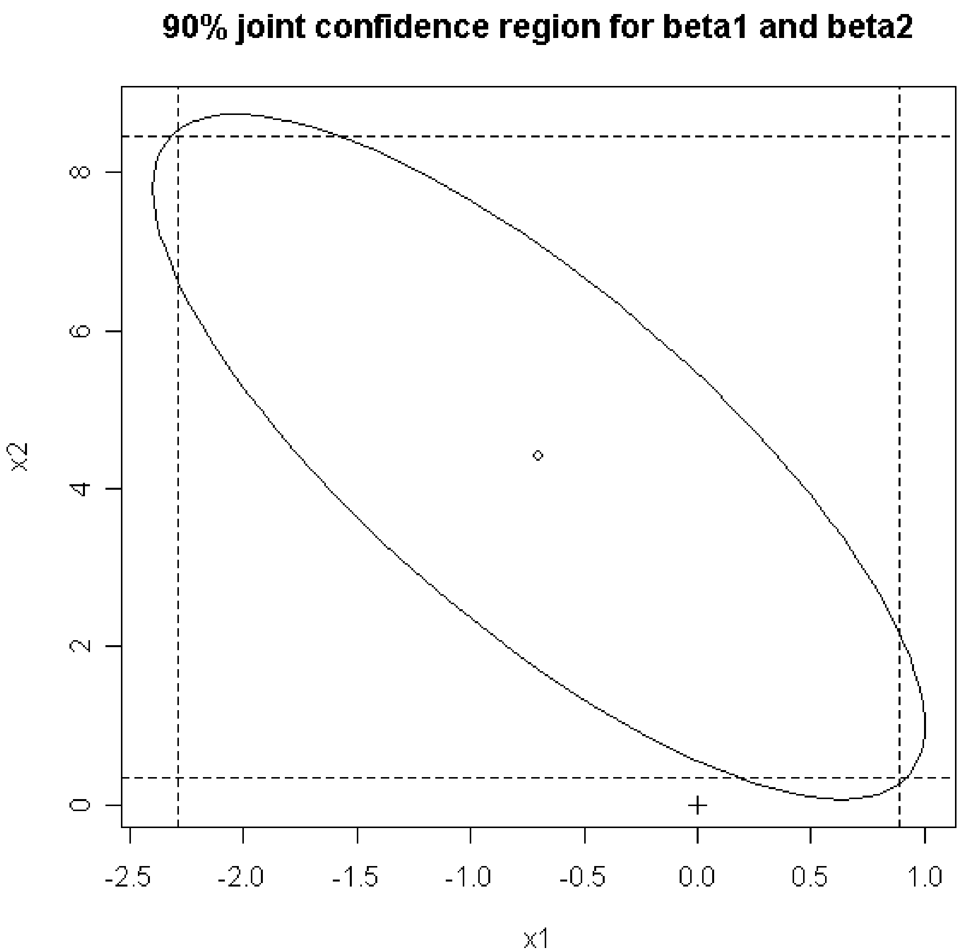
```
> plot(ellipse(fit,c(2,3),level=0.9),type="l")
> title("90% joint confidence region for beta1 and beta2")
> points(fit$coeff[2],fit$coeff[3])
> points(0,0,pch=3)
>
> confint(fit,"x1",level=0.95)
        2.5 %     97.5 %
x1 -2.288778 0.8887776
```

$\Rightarrow$ Do NOT Reject $H_0: \beta_1 = 0$ at $\alpha = 0.05$

```
> confint(fit,"x2",level=0.95)
        2.5 %    97.5 %
x2 0.3493959 8.450604
```

$\Rightarrow$ Reject $H_0: \beta_2 = 0$ at $\alpha = 0.05$

```
> abline(v=confint(fit,"x1",level=0.95),lty=2)
> abline(h=confint(fit,"x2",level=0.95),lty=2)
```



Reject $H_0: \beta_1 = \beta_2 = 0$ at $\alpha = 0.10$.

Note that the origin $(0, 0)$ lies **outside** the 90% joint confidence region for β_1 and β_2.

g)* Test $H_0 : \beta_0 = \beta_2$ vs. $H_a : \beta_0 \neq \beta_2$ at $\alpha = 0.10$. Find the p-value.

$H_0 : \beta_0 - \beta_2 = 0$ vs. $H_a : \beta_0 - \beta_2 \neq 0$

$X_0^T = [\ 1 \quad 0 \quad -1\ ]$ $\qquad \beta_0 - \beta_2 = X_0^T \boldsymbol{\beta}.$ $\qquad X_0^T \hat{\boldsymbol{\beta}} = -0.7.$

$$X_0^T \mathbf{C}\, X_0 = [1 \quad 0 \quad -1] \cdot \begin{bmatrix} 0.7125 & -0.025 & -0.1375 \\ -0.025 & 0.025 & -0.05 \\ -0.1375 & -0.05 & 0.1625 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = 1.15.$$

$$\hat{\text{Var}}(X_0^T \hat{\boldsymbol{\beta}}) = [X_0^T \mathbf{C}\, X_0]\, s^2 = 1.15 \times 15.28 = 17.572.$$

Test Statistic: $\qquad t = \dfrac{-0.7 - 0}{\sqrt{17.572}} \approx \mathbf{-0.167}.$

$n - p = \mathbf{5}$ d.f.

$$\begin{array}{ccc} -0.167 & > & -0.267 \\ t & > & -t_{0.40}(5) \end{array}$$

left tail > 0.40

p-value = 2 tails. $\qquad$ p-value > **0.80**.

(p-value ≈ 0.8739.)

Do NOT Reject $\mathbf{H_0}$.

```
> x3 = x2 + 1
> x3
[1] 2 6 5 4 2 5 5 3
>
> fite = lm(y ~ x1 + x3 + 0)
> anova(fite,fit)
Analysis of Variance Table

Model 1: y ~ x1 + x3 + 0
Model 2: y ~ x1 + x2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      6 76.826
2      5 76.400  1     0.426 0.0279  0.874
```

h)* Test $H_0 : 2\beta_1 + \beta_2 = 0$ vs. $H_a : 2\beta_1 + \beta_2 \neq 0$ at $\alpha = 0.05$. Find the p-value.

$X_0^T = [\ 0\ \ 2\ \ 1\ ]$ $\qquad 2\beta_1 + \beta_2 = X_0^T \boldsymbol{\beta}$. $\qquad X_0^T \hat{\boldsymbol{\beta}} = 3.0$.

$$X_0^T \mathbf{C} X_0 = [0 \quad 2 \quad 1] \cdot \begin{bmatrix} 0.7125 & -0.025 & -0.1375 \\ -0.025 & 0.025 & -0.05 \\ -0.1375 & -0.05 & 0.1625 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} = 0.0625.$$

$$\hat{\text{Var}}(X_0^T \hat{\boldsymbol{\beta}}) = [X_0^T \mathbf{C} X_0]\, s^2 = 0.0625 \times 15.28 = 0.955.$$

Test Statistic: $\qquad t = \dfrac{3.0 - 0}{\sqrt{0.955}} \approx \mathbf{3.070}$.

$n - p = \mathbf{5}$ d.f.

2.571	<	3.070	<	3.365
$t_{0.025}(5)$	<	t	<	$t_{0.01}(5)$

$0.01 <$ right tail < 0.025

p-value = 2 tails. $\qquad \mathbf{0.02} <$ p-value $< \mathbf{0.05}$.

(p-value ≈ 0.0278.)

Reject H_0.

```
> x4 = x1 - 2*x2
> fitf = lm(y ~ x4)
> anova(fitf,fit)
Analysis of Variance Table

Model 1: y ~ x4
Model 2: y ~ x1 + x2
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1      6 220.4
2      5  76.4  1     144.0 9.4241 0.02779 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```