STAT 420 Fall 2012

Homework #6

(due Friday, October 19, by 3:00 p.m.)

1. Chemists often use ion-sensitive electrodes (ISEs)to measure the ion concentration of aqueous solutions. These devices measure the migration of the charge of these ions and give a reading in millivolts (mV). A standard curve is produced by measuring known concentrations (in ppm) and fitting a line to the millivolt data. The table on the right gives the concentrations in ppm and the voltage in mV for calcium ISE.

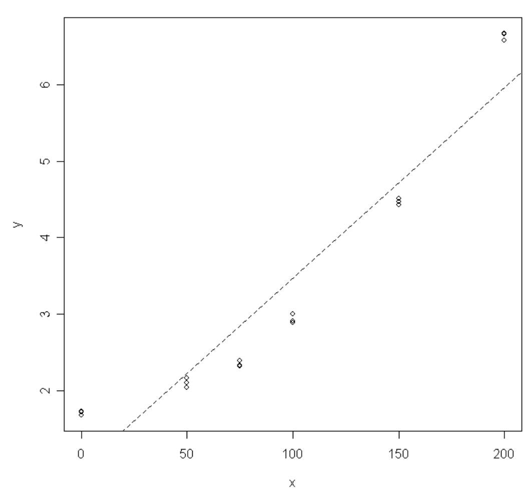
ppm	mV
0	1.72
0	1.68
0	1.74
50	2.04
50	2.11
50	2.17
75	2.40
75	2.32
75	2.33
100	2.91
100	3.00
100	2.89
150	4.47
150	4.51
150	4.43
200	6.67
200	6.66
200	6.57

The data are also stored in Hw06_1.dat

a) Plot the points mV (y) versus ppm (x). Does linear model seem to be appropriate here?

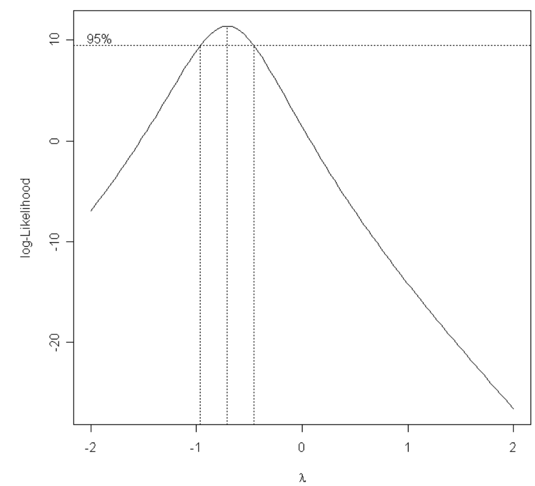
```
> plot(x,y)
> abline(lm(y~x)$coefficients,lty=2)
```

Linear model does NOT seem to be appropriate here.



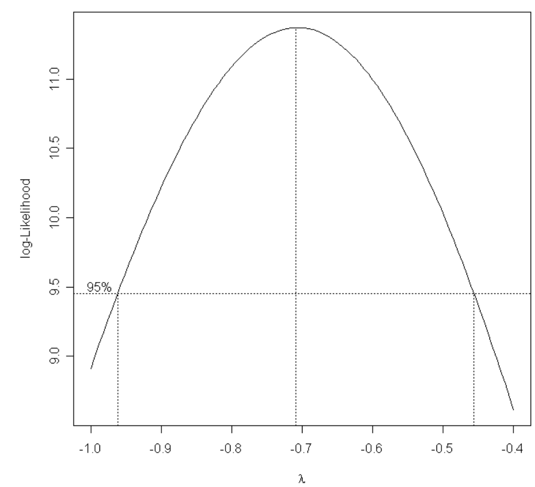
b) Use the Box-Cox method to determine the best transformation on the response variable mV.

```
> library(MASS)
> boxcox(fit,plotit=T)
```




```
> boxcox(lm(y~x),plotit=T,lambda=seq(-1.0,-0.4,by=0.01))
```

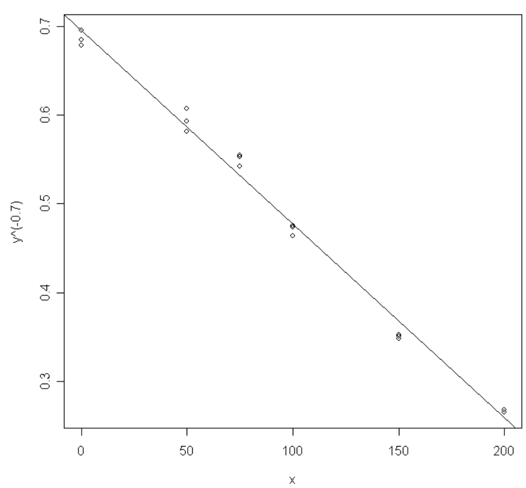
$\lambda \approx -0.7$ seems to give the best transformation of the response variable.



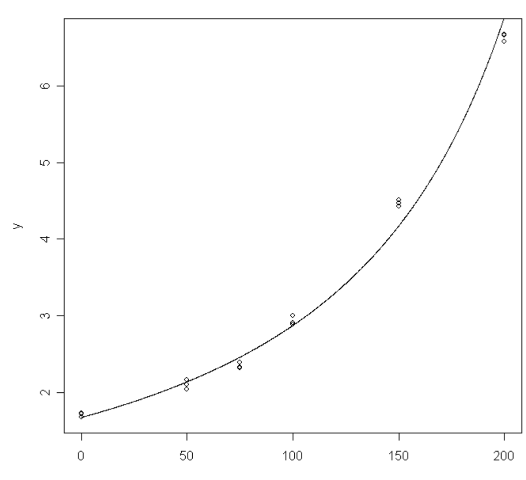

```
> fit1 = lm(y^(-0.7) ~ x)
> summary(fit1)

Call:
lm(formula = y^(-0.7) ~ x)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0194470 -0.0131026 -0.0007467  0.0085100  0.0230990 
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.956e-01  6.003e-03  115.87   <2e-16 ***
x           -2.185e-03  5.179e-05  -42.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01433 on 16 degrees of freedom
Multiple R-squared: 0.9911,     Adjusted R-squared: 0.9905
F-statistic:  1780 on 1 and 16 DF,  p-value: < 2.2e-16
```

```
> plot(x,y^(-0.7))
> abline(fit1$coefficients)
```

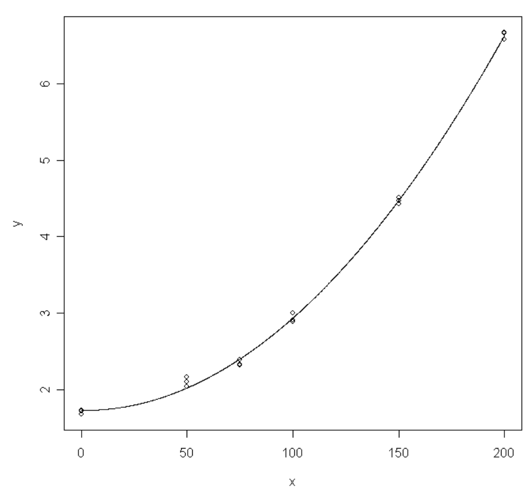



```
> xx = seq(0,200,by=0.1)
> yy = (fit1$coefficients[1]+fit1$coefficients[2]*xx)^(1/(-0.7))
> plot(x,y)
> lines(xx,yy)
```



c) In part (a), a linear model does not seem appropriate. Fit a quadratic model. Does it seem to provide a better fit?

```
> fit2 = lm(y ~ x + I(x^2))
> summary(fit2)

Call:
lm(formula = y ~ x + I(x^2))

Residuals:
      Min        1Q    Median        3Q       Max 
-0.085354 -0.047679 -0.004113  0.035984  0.143329 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.735e+00  3.442e-02  50.410  < 2e-16 ***
x           -3.772e-04  7.688e-04  -0.491    0.631    
I(x^2)       1.242e-04  3.605e-06  34.452 1.07e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.06341 on 15 degrees of freedom
Multiple R-squared: 0.9988,     Adjusted R-squared: 0.9987 
F-statistic:  6500 on 2 and 15 DF,  p-value: < 2.2e-16 


> yy2 = fit2$coefficients[1] + fit2$coefficients[2]*xx +
+ fit2$coefficients[3]*xx^2
> par(mfrow=c(1,1))
> plot(x,y)
> lines(xx,yy2)
```



2.2 The dataset `uswages` is drawn as a sample from the Current Population Survey in 1988. Fit a model with weekly wages as the response and years of education and experience as predictors. Report and give a simple interpretation to the regression coefficient for years of education. Now fit the same model but with logged weekly wages. Give an interpretation to the regression coefficient for years of education. Which interpretation is more natural?

```
> library(faraway)
> data(uswages)
```

The data are also stored in uswages.csv

The log rule: if the values of a variable range over more than one order of magnitude and the variable is strictly positive, then replacing the variable by its logarithm is likely to be helpful.

```
> library(faraway)
> data(uswages)
> attach(uswages)
>
> fit1 = lm(wage ~ educ + exper)
> summary(fit1)

Call:
lm(formula = wage ~ educ + exper)

Residuals:
     Min       1Q   Median       3Q      Max
-1018.23  -237.86   -50.87   149.88  7228.61

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -242.7994    50.6816  -4.791 1.78e-06 ***
educ          51.1753     3.3419  15.313  < 2e-16 ***
exper          9.7748     0.7506  13.023  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 427.9 on 1997 degrees of freedom
Multiple R-Squared: 0.1351,     Adjusted R-squared: 0.1343
F-statistic:   156 on 2 and 1997 DF,  p-value: < 2.2e-16
```

The fitted regression function is

$$\text{wage} = -242.7994 + 51.1753 * \text{educ} + 9.7748 * \text{exper}.$$

The regression coefficient for years of education is 51.1753. We would expect weekly wages to increase by 51.1753 on average for every 1-year increase of years of education with experience fixed.

```
> fit2 = lm(log(wage)~educ+exper)
> summary(fit2)

Call:
lm(formula = log(wage) ~ educ + exper)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7533 -0.3495  0.1068  0.4381  3.5699 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.650319   0.078354   59.35   <2e-16 ***
educ        0.090506   0.005167   17.52   <2e-16 ***
exper       0.018079   0.001160   15.58   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.6615 on 1997 degrees of freedom
Multiple R-Squared: 0.1749,     Adjusted R-squared: 0.174 
F-statistic: 211.6 on 2 and 1997 DF,  p-value: < 2.2e-16
```

The fitted regression function is

$$\ln(\text{wage}) = 4.650319 + 0.090506 * \text{educ} + 0.018079 * \text{exper}.$$

That is,

$$\text{wage} = e^{4.650319} \cdot e^{0.090506 * \text{educ}} \cdot e^{0.018079 * \text{exper}}.$$

We would expect weekly wages to increase $e^{0.090506} = 1.094728$ times (“on average”) [that is, by 9.473%] for every 1-year increase of years of education with experience fixed.

The second model makes more sense since the first model allows wage to have negative values, while the wage cannot be negative.

```
> min(wage)
[1] 50.39
> max(wage)
[1] 7716.05
```

The values of wage range over more than one order of magnitude, and the variable is strictly positive, it would be better to replace the variable by its logarithm ("the log rule"). Therefore, the second model is more natural.

3. Data set `mammals` contains the average body weight in kg (x) and the average brain weight in g (y) for 62 species of land mammals.

```
> library(MASS)
> data(mammals)
```

The data are also stored in mammals.csv

Researchers such as Sprent (1972) and Gould (1996) have noted that the following relationship seems to work well:

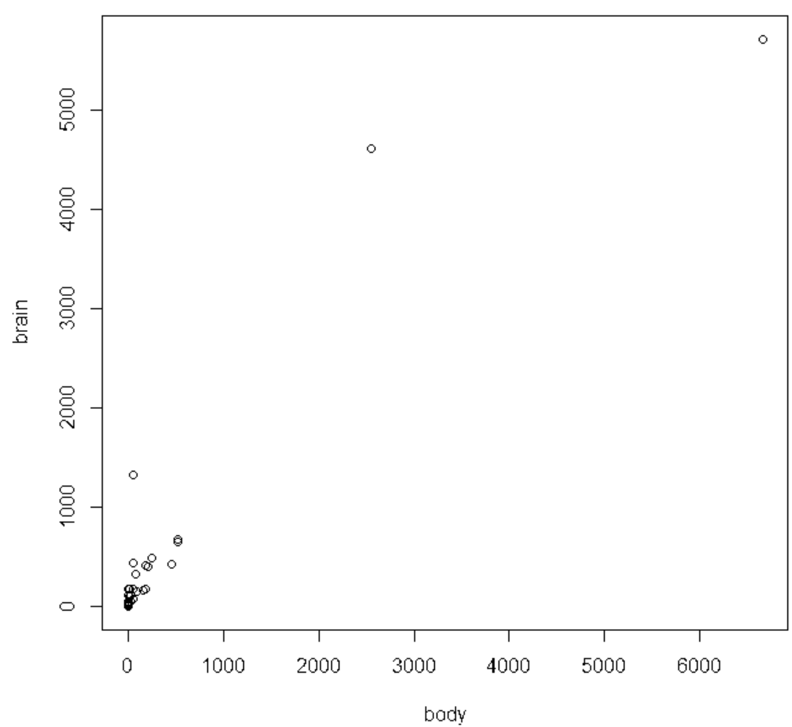
$$\text{brain weight} = \gamma_0 (\text{body weight})^{\beta_1} (\varepsilon).$$

This model asserts that brain weight is proportional to body weight raised to the β_1 power, with a multiplicative error ε. Obviously, this model can be linearized if we take the logarithm of both x and y. That is,

$$\log(\text{brain weight}) = \log(\gamma_0) + \beta_1 \log(\text{body weight}) + \log(\varepsilon).$$

a) Plot the average brain weight (y) vs. the average body weight (x).

```
> attach(mammals)
> plot(body, brain)
```

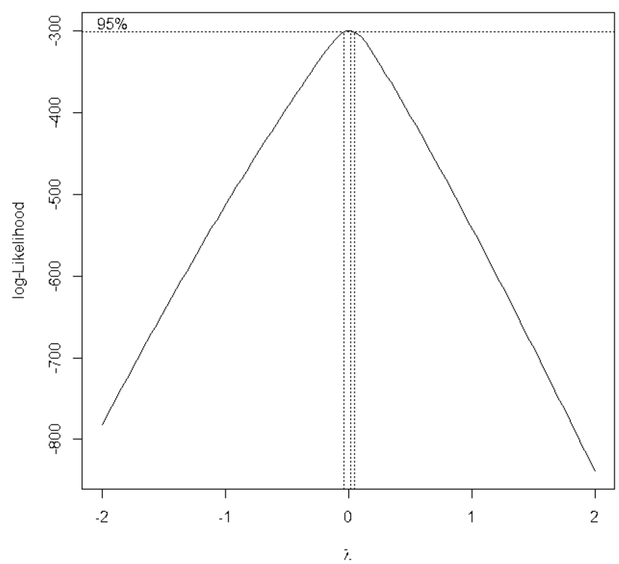


The log rule: if the values of a variable range over more than one order of magnitude and the variable is strictly positive, then replacing the variable by its logarithm is likely to be helpful.

Since the body weights do range over more than one order of magnitude and are strictly positive, we will use $\log(\text{body weight})$ as our predictor. Use the Box-Cox method to verify that $\log(\text{brain weight})$ is a "recommended" transformation of the response variable. That is, verify that $\lambda = 0$ is among the "recommended" values of λ (☺ include printout ☺) when considering

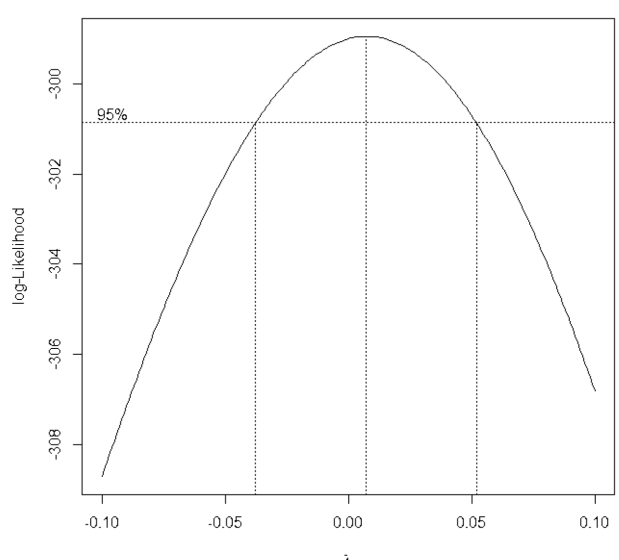
$$g_\lambda(y) = \beta_0 + \beta_1 \log(\text{body weight}) + \varepsilon.$$

```
> lbody = log(body)
> fit1 = lm(brain ~ lbody)
> boxcox(fit1, plotit=T)
```




```
> boxcox(fit1, lambda=seq(-0.10,0.10,by=0.001), plotit=T)
```

$\lambda = 0$ is among the "recommended" values of λ, pretty close to the "optimal" λ.



b) Plot log (brain weight) vs. log (body weight). Does linear relationship seem to be appropriate here? Fit the model

$$\log(\text{brain weight}) = \beta_0 + \beta_1 \log(\text{body weight}) + \varepsilon.$$

and use it to predict the average brain weight of a Siberian tiger (average body weight 227 kg). Construct a 95% prediction interval.

```
> fit2 = lm(log(brain) ~ lbody)
> summary(fit2)

Call:
lm(formula = log(brain) ~ lbody)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.71550 -0.49228 -0.06162  0.43597  1.94829 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.13479    0.09604   22.23   <2e-16 ***
lbody        0.75169    0.02846   26.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-squared: 0.9208,     Adjusted R-squared: 0.9195 
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

$$\log(\text{brain weight}) = 2.13479 + 0.75169 \log(\text{body weight}).$$

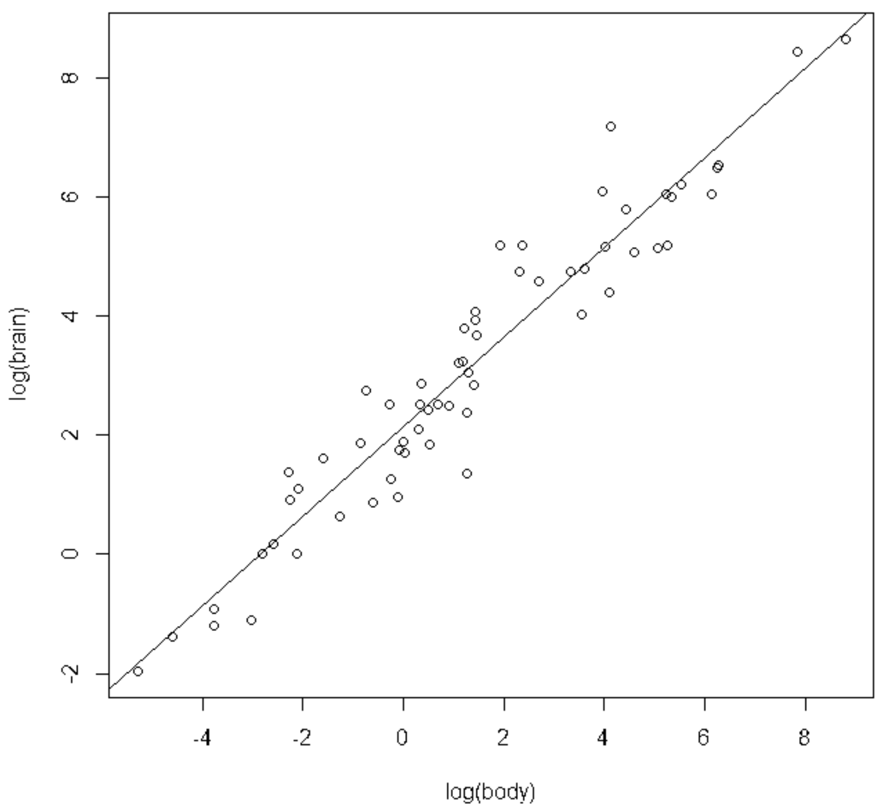
$$\text{brain weight} = 8.45527 \, (\text{body weight})^{0.75169}.$$

```
> predict.lm(fit2, data.frame(lbody=log(227)), 
interval=c("prediction"))
       fit      lwr     upr
1 6.212647 4.793485 7.63181
> 
> exp(6.212647)
[1] 499.0204
> 
> exp(4.793485)
[1] 120.7214
> exp(7.63181)
[1] 2062.78
```

Prediction of the average brain weight of a Siberian tiger = **499** g.

95% prediction interval (120.72 , 2062.78)

```
> plot(log(body), log(brain))
> abline(fit2$coefficients)
```



4. Can a corporation's annual profit be predicted from information about the company's chief executive officer (CEO)? *Forbes* (May, 1999) presented data on company profit (y), (in \$ millions), CEO's annual income (x_1) (in \$ thousands), and percentage of the company's stock owned by the CEO (x_2).

Company	Profit, y	CEO	Income, x_1	Stock, x_2
Gap	824.5	Drexler	3,743	1.71%
Intel	6,068.0	Grove	52,598	.13
Gateway 2000	346.4	Waitt	855	43.93
HJ Heinz	746.9	O'Reilly	2,916	1.63
Conseco	630.7	Hilbert	124,579	3.64
Citicorp	5,807.0	Reed	6,200	.22
Cisco Systems	1,362.3	Chambers	560	.06
General Electric	9,296.0	Welch	40,626	.03
America Online	254.0	Case	26,917	.54
Computer Associates	570.0	Wang	10,614	3.79
Lockheed Martin	1,001.0	Augustine	2,533	.01
Bear Stearns	538.6	Cayne	23,215	3.44

Source: "Compensation Fit for a King," *Forbes*, May 1999.

The data are stored in Hw06_4.csv

```
> Hw06_4
               Company      y       CEO     x1    x2
1                  Gap  824.5   Drexler   3743  1.71
2                Intel 6068.0     Grove  52598  0.13
3         Gateway 2000  346.4     Waitt    855 43.93
4             HJ Heinz  746.9  O'Reilly   2916  1.63
5              Conseco  630.7   Hilbert 124579  3.64
6             Citicorp 5807.0      Reed   6200  0.22
7        Cisco Systems 1362.3  Chambers    560  0.06
8     General Electric 9296.0     Welch  40626  0.03
9       America Online  254.0      Case  26917  0.54
10 Computer Associates  570.0      Wang  10614  3.79
11     Lockheed Martin 1001.0 Augustine   2533  0.01
12        Bear Stearns  538.6     Cayne  23215  3.44
```

a) Fit the interaction model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Give the least squares prediction equation and determine whether the overall model is statistically useful for predicting company profit at $\alpha = 0.10$.

```
> attach(Hw06_4)
> fit = lm(y ~ x1 + x2 + I(x1*x2))
> summary(fit)

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

Residuals:
    Min       1Q   Median       3Q      Max
-3674.4   -621.1   -476.8    175.8   3938.4

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1160.50587  983.14706   1.180   0.2717
x1             0.12176    0.04234   2.876   0.0206 *
x2             6.02726   61.19247   0.098   0.9240
I(x1 * x2)    -0.03528    0.01168  -3.021   0.0165 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2311 on 8 degrees of freedom
Multiple R-squared: 0.5704,     Adjusted R-squared: 0.4093
F-statistic: 3.541 on 3 and 8 DF,  p-value: 0.0678
```

$$\hat{Y} = 1160.50587 + 0.12176\, x_1 + 6.02726\, x_2 - 0.03528\, x_1 x_2$$

$H_0: \beta_1 = \beta_2 = \beta_3 = 0.$ $\qquad \alpha = 0.10.$

Test Statistic $F = 3.541$, 3 and 8 degrees of freedom. $\qquad F_{0.10}(3, 8) = 2.92.$

p-value $= 0.0678 < 0.10 = \alpha.$ $\qquad$ **Reject H_0**: $\beta_1 = \beta_2 = \beta_3 = 0$ at $\alpha = 0.10.$

The overall model is statistically useful for predicting company profit at $\alpha = 0.10$.

b) Is there evidence to indicate that CEO income x_1 and stock percentage x_2 interact? Use $\alpha = 0.05$.

$H_0: \beta_3 = 0.$ $\qquad \alpha = 0.05.$

Test Statistic $t = -3.021$, 8 degrees of freedom. $\qquad \pm t_{0.025}(8) = \pm 2.306.$

p-value = $0.0165 < 0.05 = \alpha$. $\qquad$ **Reject $\mathbf{H_0}$**$: \beta_3 = 0$ at $\alpha = 0.05$.

There is evidence $(\text{at } \alpha = 0.05)$ that CEO income x_1 and stock percentage x_2 interact.

c) Based on the least squares estimates of the β parameters, give the estimate of the change in profit for every one thousand dollar increase in a CEO's income when CEO owns 2% of the company's stock.

$(\beta_1 + \beta_3 x_2)$ represents the change in $E(Y)$ for every 1-unit increase in x_1, holding x_2 fixed.

$0.12176 - 0.03528\, x_2 \;=\; 0.12176 - 0.03528 \cdot 2 \;=\;$ \$0.0512 million $=$ **\$51,200**.

We estimate that the company profit would increase by \$51,200 (on average) for every one thousand dollar increase in a CEO's income when CEO owns 2% of the company's stock.

5. Suppose the interaction model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

was fit to $n = 20$ data points, and the following results were obtained:

```
> sum( lm( y ~ 1 )$residuals^2 )
[1]  57
> sum( lm( y ~ x1 )$residuals^2 )
[1]  40
> sum( lm( y ~ x2 )$residuals^2 )
[1]  45
> sum( lm( y ~ x1 + x2 )$residuals^2 )
[1]  36
> sum( lm( y ~ x1 + x2 + I(x1*x2) )$residuals^2 )
[1]  30
> lm( y ~ x1 + x2 + I(x1*x2) )$coefficients
(Intercept)   x1   x2   I(x1 * x2)
         10    5   -2            3
```

a) Perform the significance of the regression test at $\alpha = 0.05$.

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ vs H_1 : at least one of $\beta_1, \beta_2, \beta_3$ is not zero.

Null model: $Y = \beta_0 + \varepsilon$

$SSResid_{Null} = 57$ $SSResid_{Full} = 30$

ANOVA table:

Source	SS	DF	MS	F
Regression (Diff.)	27	3	9	4.8
Residuals (Full)	30	16	1.875	
Total (Null)	57	19		

$F_{0.05}(3, 16) = 3.24$ **Reject $\mathbf{H_0}$** at $\alpha = 0.05$.

b) Do x_1 and x_2 interact? Perform the appropriate test at $\alpha = 0.05$.

$H_0: \beta_3 = 0$ vs $H_1: \beta_3 \neq 0.$

Null model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

$\text{SSResid}_{\text{Null}} = 36$ $\text{SSResid}_{\text{Full}} = 30$

ANOVA table:

Source	SS	DF	MS	F
Diff.	6	1	6	3.2
Full	30	16	1.875	
Null	36	17		

$F_{0.05}(1, 16) = 4.49$ **Do NOT Reject $\mathbf{H_0}$** at $\alpha = 0.05$

c) Is there sufficient evidence to indicate that x_2 contributes information for the prediction of y? Perform the appropriate test at $\alpha = 0.05$. What is the p-value of this test?

$H_0: \beta_2 = \beta_3 = 0$ vs H_1: at least one of β_2, β_3 is not zero.

Null model: $Y = \beta_0 + \beta_1 x_1 + \varepsilon$

$\text{SSResid}_{\text{Null}} = 40$ $\text{SSResid}_{\text{Full}} = 30$

ANOVA table:

Source	SS	DF	MS	F
Diff.	10	2	5	2.6667
Full	30	16	1.875	
Null	40	18		

$F_{0.05}(2, 16) = 3.63$ **Do NOT Reject H_0** at $\alpha = 0.05$

$F_{0.10}(2, 16) = 2.67$ p-value $\approx$ **0.10**

d) Estimate the change in $E(Y)$ for every 1-unit increase in x_1, when $x_2 = 2$.

$(\beta_1 + \beta_3 x_2)$ represents the change in $E(Y)$ for every 1-unit increase in x_1, holding x_2 fixed.

$\hat{\beta}_1 + \hat{\beta}_3 \times 2 = 5 + 3 \times 2 = \mathbf{11}$.

e) Estimate the change in $E(Y)$ for every 1-unit increase in x_2, when $x_1 = 3$.

$(\beta_2 + \beta_3 x_1)$ represents the change in $E(Y)$ for every 1-unit increase in x_2, holding x_1 fixed.

$\hat{\beta}_2 + \hat{\beta}_3 \times 5 = -2 + 3 \times 3 = \mathbf{7}$.