

Forty-two people took this exam. The range of the scores went from 16 to 93 with a median of 71. The average score was 67.8 with a standard deviation of 18.1.

1. i) In a survey of 100 randomly selected U. of Minnesota women undergraduates 27 stated that they exercised at least four hours a week. Obtain a 95% confidence interval for the true proportion  $p_1$  of woman students who exercise at least four hours a week.

ii) In a similar survey of 95 U. of Minnesota men undergraduates, 31 stated that they worked out at least four hours a week. If  $p_2$  is the true proportion of U. of Minnesota men students who exercise at least four hours a week find a 95% interval for  $p_1 - p_2$ .

iii) Using the above data find the  $p$ -value or level of significance for testing  $H_0: p_1 = p_2$  against  $K: p_1 < p_2$

**Solution:** i) Since  $z_{.025} = 1.96$  we have

$$\frac{27}{100} \mp 1.96 \sqrt{\frac{27}{100} \frac{73}{100} \frac{1}{100}}$$

ii) In a similar survey of 95 U. of Minnesota men undergraduates, 31 stated that they worked out at least four hours a week. If  $p_2$  is the true proportion of U. of Minnesota men students who exercise at least four hours a week find a 95% interval for  $p_1 - p_2$ .

$$\frac{27}{100} - \frac{31}{95} \mp 1.96 \sqrt{\frac{27}{100} \frac{73}{100} \frac{1}{100} + \frac{31}{95} \frac{64}{95} \frac{1}{95}}$$

iii) Now  $\hat{p}_2 = 31/95$ ,  $\hat{p}_1 = 27/100$  and  $\hat{p}_p = 58/195$  and so

$$\frac{\hat{p}_2 - \hat{p}_1 - 0}{\sqrt{\hat{p}_p(1 - \hat{p}_p)} \sqrt{1/100 + 1/95}} = 0.8599 = z_\alpha$$

and the  $p$ -value =  $\alpha = 0.195$ .

■

2. A hair growth company claims that for bald men who use their product 25% grow lots of hair, 40% grow some hair and 35% grow no hair. Test their claim at level  $\alpha = .01$  if in a random sample of 100 men 20 grew lots of hair, 30 grew some hair and 50 grew no hair.

**Solution:**

Under the null hypothesis the expected number in the three groups are 25, 40 and 35. Since  $P(\chi_2^2 > 9.21) = 0.01$  and

$$\sum \frac{(Observed - Expected)^2}{Expected} = \frac{25}{25} + \frac{100}{40} + \frac{225}{35} = 9.93 > 9.21$$

We reject the claim at level  $\alpha = .01$ .

■

3. A study of gas chromatography, a technique which is used to detect very small amounts of a substance, was carried out. The response variable,  $y$  was the output reading from the gas chromatograph. The measurements were taken on twenty specimens containing different amounts  $x$  of the substance. The purpose of the study was to calibrate the chromatograph by relating the actual amount of the substance to the chromatograph reading. Use the data below and the output from fitting the simple linear regression model  $Y = \beta_0 + \beta_1 x + Z$  to answer the questions on the exam.

```

      x      y
[1,] 0.25  6.55
[2,] 0.25  7.98
[3,] 0.25  6.54
[4,] 0.25  6.37
[5,] 0.25  7.96
[6,] 1.00 29.70
[7,] 1.00 30.00
[8,] 1.00 30.10
[9,] 1.00 29.50
[10,] 1.00 29.10
[11,] 5.00 211.00
[12,] 5.00 204.00
[13,] 5.00 212.00
[14,] 5.00 213.00
[15,] 5.00 205.00
[16,] 20.00 929.00
[17,] 20.00 905.00
[18,] 20.00 922.00
[19,] 20.00 928.00
[20,] 20.00 919.00

```

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.4107      2.6142  -5.512 3.11e-05 ***
x              46.6287      0.2533 184.086 < 2e-16 ***
---

```

Residual standard error: 9.023 on 18 degrees of freedom  
Multiple R-Squared: 0.9995,

```
predict(chrom.lm,data.frame(x=c(2,6,9)),se.fit=T)
```

```
$fit
```

```

      1      2      3
78.84666 265.36137 405.24740

```

```
$se.fit
```

```

      1      2      3
2.325190 2.022677 2.110006

```

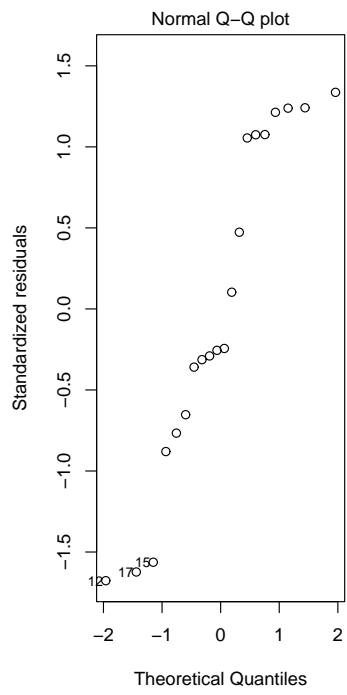
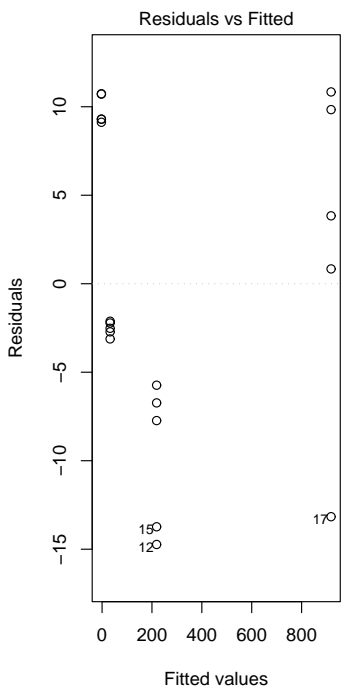
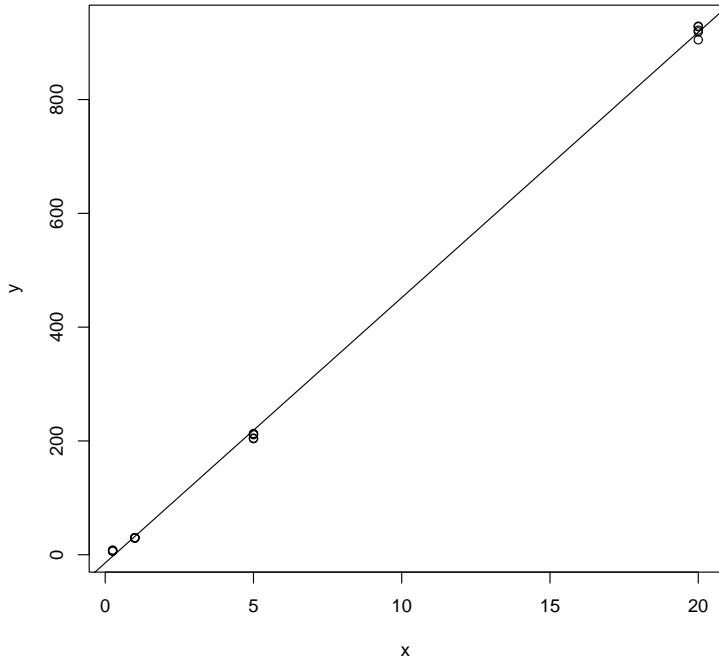
i) Find a 95% confidence interval for the the expected value of  $Y$  when the level of substance  $x = 6$ .

ii) Should one use the fitted model to use  $x$  to predict  $Y$ ? Briefly justify your answer.

**Solution:**

i) Since  $t_{18,0.025} = 2.101$  the interval is

$$265.36 \pm 2.101 \times 2.023$$



ii) With an R-squared of almost one it would seem that the model works very well. But careful examination of the residuals suggests that all is not well. Note there is a pattern in the residuals. They start positive, then turn negative and then positive again. The Normal Q-Q plot is not very linear. Finally a close look at the first plot indicates there is some curvature in the data. One should consider observing some more data at some  $x$  values between 5 and 20.

■

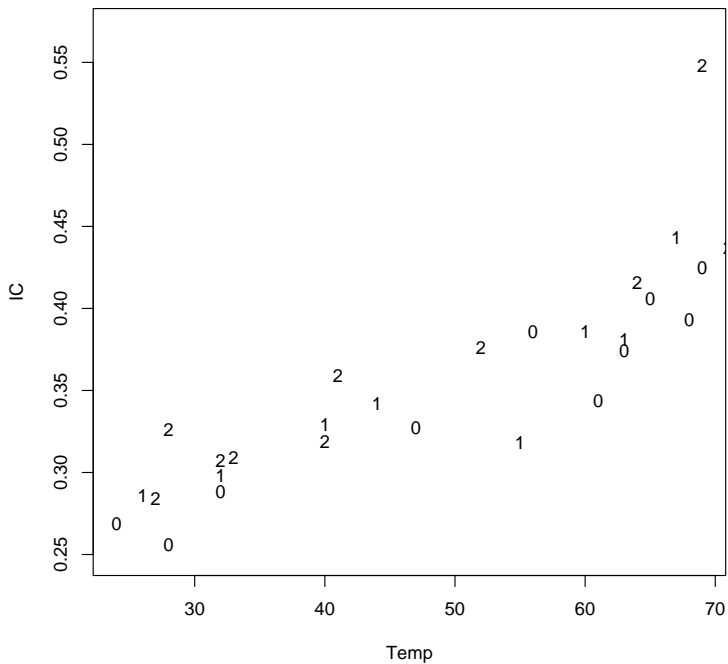


Figure 1: For each of the three years the plot of  $IC$  against  $Temp$ . Note the ten pairs for 1996 are denote by 0's, the ten pairs for 1997 by 1's and the ten pairs for 1998 by 2's.

4. A study was carried out to see if one can predict ice cream consumption using price, income and temperature. For three years Ice cream consumption was measured over 30 four-week periods. For each period the the following data were collected.

- IC: Ice cream consumption in pints per capita
- Pr: Price of ice cream per pint in dollars
- Inc: Average weekly family income in dollars
- Tmp: Mean temperature in degrees F.
- Yr: Year within the study (0 = 1996, 1 = 1997, 2 = 1998)

The model

$$IC = \beta_0 + \beta_{Pr}Pr + \beta_{Inc}Inc + \beta_{Tmp}Tmp + \beta_{Yr}Yr + Z$$

was fit to the data and the usual output is given below. One plot and two additional anova tables when the independent variables were entered in different orders are also given.

```
lm(formula = IC ~ price + income + temp + Year, data = icecream)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.0494466	-0.0128471	-0.0004049	0.0127139	0.0778585

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	0.7119857	0.2521412	2.824	0.00918	**
price	-0.1297962	0.0640002	-2.028	0.05334	.
income	-0.0002232	0.0001849	-1.207	0.23879	
temp	0.0031688	0.0003383	9.367	1.18e-09	***
Year	0.0389074	0.0148564	2.619	0.01477	*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02813 on 25 degrees of freedom

Multiple R-Squared: 0.8424,

F-statistic: 33.41 on 4 and 25 DF, p-value: 1.075e-09

Analysis of Variance Table

Response: IC

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
price	1	0.008459	0.008459	10.6898	0.003132	**
income	1	0.000051	0.000051	0.0644	0.801725	
temp	1	0.091804	0.091804	116.0159	7.015e-11	***
Year	1	0.005427	0.005427	6.8586	0.014773	*
Residuals	25	0.019783	0.000791			

---

## Analysis of Variance Tables

Response: IC

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Year	1	0.008706	0.008706	11.002	0.002786 **
income	1	0.017231	0.017231	21.776	8.845e-05 ***
price	1	0.010369	0.010369	13.103	0.001306 **
temp	1	0.069435	0.069435	87.748	1.179e-09 ***
Residuals	25	0.019783	0.000791		

---

Response: IC

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Temp	1	0.092385	0.092385	116.7508	6.57e-11 ***
Year	1	0.009479	0.009479	11.9784	0.001947 **
Price	1	0.002725	0.002725	3.4432	0.075340 .
Income	1	0.001152	0.001152	1.4564	0.238794
Residuals	25	0.019783	0.000791		

i) In the full model at level  $\alpha = 0.05$  test  $H : \beta_{Tmp} = \beta_{Yr} = 0$  against  $K$ : At least one is not zero.

ii) In the full model at level  $\alpha = 0.05$  test  $H : \beta_{Tmp} = \beta_{Inc} = 0$  against  $K$ : At least one is not zero.

iii) In the model  $IC = \beta_0 + \beta_{Pr}Pr + Z$  at level  $\alpha = 0.05$  test  $H : \beta_{Pr} = 0$  against  $K : \neq 0$ .

iv) Based on this output can you suggest a model which could be useful for predicting ice cream consumption? Briefly justify your answer.

### **Solution:**

i) Using the last anova table (because it is the one where  $Tmp$  and  $Yr$  come last) we find that

$$\frac{(0.091804 + 0.005247)/2}{0.019783/25} = \frac{0.0485255}{0.000791} = 61.35 > 3.38519 = f_{2,25;.05}$$

ii) Cannot do this part since none of the anova tables have these two variables coming last.

iii) We need to use the results from the first anova table. From there we find that

$$SSR(\beta_{Pr}|\beta_0) = 0.008459.$$

Using the rest of this table we can find the residual sum of squares for this model.

$$RSS = .000051 + .091804 + .005427 + .019783 = 0.117065$$

But since

$$\frac{0.008459}{0.117065/28} = \frac{0.008459}{0.004181} = 2.02325 < 4.196 = f_{1,28;.05}$$

we fail to reject  $H$ .

iv) A good model appears to be

$$IC = \beta_0 + \beta_{Tmp}Tmp + \beta_{Yr}Yr + Z$$

From the last anova table we find for this model that

$$SSR(\beta_{Tmp}, \beta_{Yr}|\beta_0) = 0.092385 + 0.009479 = 0.101864$$

and its residual sum of squares is

$$RSS = 0.002725 + .001152 + 0.019783 = 0.02366$$

Hence the Multiple R-Squared for this model is

$$\frac{SSR}{SSR + RSS} = 0.81$$

which almost as good as 0.84 for the full model.

We see from the plot that for each year  $IC$  is approximately a linear function of  $Temp$ . Also there is some evidence of increased consumption over the three year period.

■