Chapter 11

Regression

In previous chapters we introduced the concept of a statistical model. This concept was yet another fundamental concept in statistics. Most statistical analyses start with a stage of setting up a model. When we say that $x_1, x_2, \ldots, x_n$ is a sample from the normal population with a mean μ and the variance σ^2 we already posed a statistical model. In ANOVA the model was

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}.$$

and we were interested in testing and estimating parameters.

Till now we have discussed statistical inferences based only on the sample measurements of a *single* variable. In many experiments two or more variables are observed for each experimental unit.

Regression[1] analyses (and correlation analyses) are dealing with exploring of associations among observable variables. The questions of interests are:

- Whether the variables are related.
- If they are related, what kind of relation they enjoy and how strong is the relationship.
- Whether one variable (maybe of primary interest) can be predicted from the observations on the others.

For example, we might be interested in estimating the relationship between two random variables X and Y, for instance height and weight, income and IQ, age of husband and wife at marriage, length and breadth of Etruscan skulls, etc.

More formally, let the population of interest .....

$$r^2_{adj} = 1 - \frac{n-1}{n-p}\frac{SSE}{SST}.$$

What is usually understood and referred as a regressional line is

$$\hat{\mu_Y}(x) = b_0 + b_1 \cdot x;$$

an estimator of the mean of Y when $X = x$. By b_0 and b_1 we denoted estimators of β_0 and β_1 based on the sample $(x_1, y_1), \ldots, (x_n, y_n)$.

11.0.1 Least Square Estimation

Bode's Law. In 1772 J. E. Bode gave a sample rule for the mean distance of a planet from the sun as a function of the planet order. Let d_n be the distance from the sun to the nth nearest planet. Bode's law predicts that

$$d_n = 4 + 3 \cdot 2^n.$$

[1] The rather curious name *regression* was given to the procedure by British scientist Sir Francis Galton, who analyzed the heights of sons and the average heights of their parents. From his observations, Galton concluded that sons of very tall (or short) parents were generally taller (shorter) than average, but not as tall (short) as their parents. The results were published in 1885 under the title *Regression Toward Mediocrity in Hereditary Stature.* In the course of time the word *regression* became synonym for the statistical study of relationship between two or more variables.

For the first eight planets, it gives: 4, 7, 10, 16, 28, 52, 100, 196 (using $n = -\infty, 0, 1, 2, \ldots, 6$). The seven planets known in 1800 had mean distances of 3.9, 7.2, 10, 15.2, 52, 95, 192 - the units have been chosen so that the distance of the earth to the sun is 10. The law certainly seems to do well, aside from a missing planet 28 units from the sun. This led a group of astronomers to search the heavens at roughly 28 units from the sun. They found a "planet," actually the asteroid series. Clearly the predictive success of Bode's law adds to its believability.

Although difficult and different, the problem of testing "the reality of Bode's law" is not completely intractable. The basic idea is to describe the steps in the data analysis, set up a way of generating data, and see how often a "law" that fits as well can be found. All the steps are difficult, but perhaps not impossible. Quantifying the steps in the data analysis is perhaps exercise. As to generating more data, the simplest approach is to find a new sample (e.g., data from another solar system). Failing this, a simple mathematical generating mechanism may be tried.

Good suggested the following mechanism for testing B: "Bode's law is true" versus $\bar{B}$: "Bode's law is not true."

Under B, the logarithms of the planetary distances are independent Gaussian with mean $\log(a + b \cdot 2^n)$ and common variance. Thus, on a log scale, the data approximately obey Bode's law.

Under $\bar{B}$ the logarithms of the planetary distances are like the ordered lengths between points dropped at random into an interval.

I. J. Good carried out the testing in a Bayesian framework, putting prior distributions on the parameters involved. He concludes that the odds are 30^4 to 1 in favor of Bode's law. B. Efron offers a criticism of Good's formulation and various alternatives. Using different models for B and $\bar{B}$, he concludes that the odds for "the reality of Bode's law" are roughly even. This is a big shift from Good's 30^4 to 1, but it still suggests that Bode's law should be taken seriously. Both authors discuss how similar analyses might be carried out for testing the validity of other simple laws.

11.1 Correlation

11.1.1 Pearson's Coefficient of Correlation

Formula:

$$r = \frac{s_{xy}}{s_x s_y}$$

where $s_{xy} = \frac{1}{n-1}\Sigma_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ is a sample covariation coefficient. and s_x and s_y are sample standard deviations for $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ respectively.

Calculational formulae:

$$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \; \Sigma(Y_i - \bar{Y})^2}}$$

$$r = \frac{\Sigma(X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{(\Sigma X_i^2 - n\bar{X}^2)(\Sigma Y_i^2 - n\bar{Y}^2)}}$$

Corn Yields and Rainfall.[2]

`yield:` yearly corn yield in bushels per acre, in six Corn Belt states (Iowa, Illinois, Nebraska, Missouri, Indiana, and Ohio).

`rain:` rainfall measurements in inches, in the six states. from 1890 to 1927.

1	9.6	24.5	20	12.0	32.3
2	12.9	33.7	21	9.3	34.9
3	9.9	27.9	22	7.7	30.1
4	8.7	27.5	23	11.0	36.9
5	6.8	21.7	24	6.9	26.8
6	12.5	31.9	25	9.5	30.5
7	13.0	36.8	26	16.5	33.3
8	10.1	29.9	27	9.3	29.7
9	10.1	30.2	28	9.4	35.0

[2] M. Ezekiel and K. A. Fox, Methods of Correlation and Re- gression Analysis, p. 212. Copyright 1959, John Wiley and Sons, Inc., New York. Data originally from E. G. Misner, ”Studies of the Relationship of Weather to the Production and Price of Farm Products, I. Corn”, mimeographed publi- cation, Cornell University, March 1928.

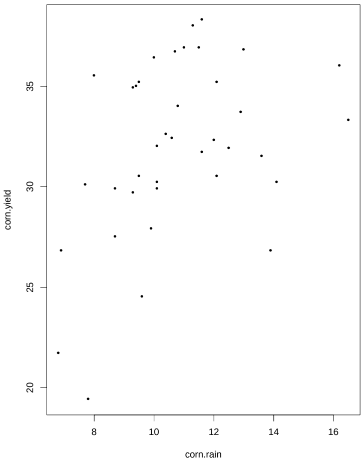

Figure 11.1: Scatterplot corn.rain and corn.yield. The correlation coefficient is 0.403

10	10.1	32.0	29	8.7	29.9
11	10.8	34.0	30	9.5	35.2
12	7.8	19.4	31	11.6	38.3
13	16.2	36.0	32	12.1	35.2
14	14.1	30.2	33	8.0	35.5
15	10.6	32.4	34	10.7	36.7
16	10.0	36.4	35	13.9	26.8
17	11.5	36.9	36	11.3	38.0
18	13.6	31.5	37	11.6	31.7
19	12.1	30.5	38	10.4	32.6

11.1.2 Inferences about ρ

Confidence Intervals

Confidence intervals for ρ can not be obtained directly. First transform: $w = \varphi(r) = r \rightarrow \frac{1}{2}\ln\frac{1+r}{1-r}$.

$(1-\alpha)100\%$ CI for ρ (in terms of w) is:

$$[w_L, w_U] = [w - \frac{z_{1-\alpha/2}}{\sqrt{n-3}}, w + \frac{z_{1-\alpha/2}}{\sqrt{n-3}}]$$

The inverse transformation $r = \varphi^{-1}(w) = \frac{e^{2w}-1}{e^{2w}+1}$ gives $[r_L, r_U]$ where $r_L = \varphi^{-1}(w_L)$ and $r_U = \varphi^{-1}(w_U)$.

Example: If $r = -0.5687$ and $n = 8$ $w = -0.6456$, $w_L = -0.6456 - \frac{1.96}{\sqrt{5}} = -1.522$ and $w_U = -0.6456 + \frac{1.96}{\sqrt{5}} = 0.2309$. In terms of r the confidence interval is:

$$[-0.9091, 0.2269].$$

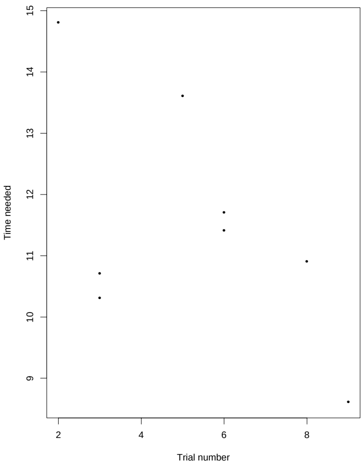

Figure 11.2: `barplot(light)`

Testing $\rho = 0$.
$H_0 : \rho = 0$ vs $H_1 : \rho >, \neq, < 0$.

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

has t distribution with $n-2$ d.f.

Example: The table below gives the number of times the rat has run through a maze (X), and the time it took the rat to run the maze on its last trial (Y).

Rat	Trials (X)	Time (Y)
1	8	10.9
2	9	8.6
3	6	11.4
4	5	13.6
5	3	10.3
6	6	11.7
7	3	10.7
8	2	14.8

(a) Find r
(b) Test the hypothesis that the population correlation coefficient ρ is 0, versus the alternative that is negative.
Solution: $\Sigma X = 42, sumX^2 = 264, \Sigma Y = 92, \Sigma Y^2 = 1084.2, \Sigma XY = 463.8, \bar{X} = 5.25, \bar{Y} = 11.5$.

$$r = \frac{463.8 - 85.2511.5}{\sqrt{(264 - 85.25^2)(1084.2 - 811.5^2)}} = \frac{-19.2}{\sqrt{43.526.2}} = -0.5687.$$

For testing $H_0 : \rho = 0$ vs $H_1 : \rho < 0$ we find

$$t = r\sqrt{\frac{n-2}{1-r^2}} = -1.69.$$

For $\alpha = 0.05$ the critical value is $t_{6,0.05} = -1.943$ and the null hypothesis is not rejected.

Remark: If n were 30, then $t = -3.6588$ and $t_{28,0.05} = -1.7$ and H_0 will be rejected.

Butterflies. The following data were extracted from a larger study by Brower[3] on a speciation in a group of

[3] Brower, L. P. (1959). Speciation in butterflies of the *Papilio glaucus* group. I Morphological Relationships and hybridizations. *Evolution 13, 40-63.*

swallowtail butterflies. Morphological measurements are in millimeters coded × 8. (Y_1 - length of 8th tergile, Y_2 - length of superuncus)

Species	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
Papilio multicaudatus	24	14	21	15	20	17.5	21.5	16.5
	21.5	16	25.5	16	25.5	17.5	28.5	16.5
	23.5	15	22	15.5	22.5	17.5	20.5	19
	21	13.5	19.5	19	26	18	23	17
	21	18	21	17	20.5	16	22.5	15.5
Papilio rutulus	20	11.5	21.5	11	18.5	10	20	11
	19	11	20.5	11	19.5	11	19	10.5
	21.5	11	20	11.5	21.5	10	20.5	12
	20	10.5	21.5	12.5	17.5	12	21	12.5
	21	11.5	21	12	19	10.5	19	11
	18	11.5	21.5	10.5	23	11	22.5	11.5
	19	13	22.5	14	21	12.5	19.5	12.5

The correlation coefficients separately for each species are $r_1 = -0.11$(for *P. multicaudatus*) and $r_2 = 0.176$(for *P. rutulus*)

Test significance of each. Test whether the two correlation coefficients differ significantly.

HINT: The following example may be useful: *Test for the difference between two correlation coefficients.* The test statistic for $H_0 : \rho_1 = \rho_2$ is

$$z = \frac{w_1 - w_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

where $w_i = \frac{1}{2} \ln \frac{1+r_i}{1-r_i}$, $i = 1, 2$ and n_1, n_2 number of pairs in the first, second sample. (You may use the table on page 478 (Iman, Text)) A good approximation for the test cut-points are quantiles of normal distribution.

• The correlation between body weight and wing length in *Drosophila pseudoobscura* was found [4] to be 0.52 in a sample of $n_1 = 39$ at the Grand Canyon and 0.67 in a sample of $n_2 = 20$ at Fagstaff, Arizona.

Grand Canyon: $w_1 = 0.5763$ Flagstaff: $w_2 = 0.8107$

The test statistic for $H_0 : \rho_1 = \rho_2$ is: $z = \frac{0.5763-0.8107}{\sqrt{1/36+1/17}} = -0.7965$.

p-value against two sided hypothesis is $2\Phi(-0.7965) = 0.4257$. Conclusion: Do not reject null hypothesis.

Solution:

Correlation of Y1 and Y2 = -0.112 (for *P. multicaudatus*)

Correlation of Y1 and Y2 = 0.176 (for *P. rutulus*)

$w_1 = -0.1124719, z_2 = 0.1778518. z = \frac{-0.1124719-0.1778518}{\sqrt{1/17+1/25}} = -0.9235328$. p-value against the twosided alternative is $2 * \Phi(-0.9235) = 2 * 0.17786 = 0.3557$.

[4] Sokoloff, A. (1966). Morphological variation in natural and experimental populations of *Drosophila pseudoobscura* and *Drosophila persimilis. Evolution, 20, 49-71.*

Cars. (a) Find r.

Car	Engine Size (in^3)	Miles/gallon
Chevette	98	31
Sentra	98	35
Colt	86	41
Isuzu I Mark	111	27
Mercedes 190D	134	35
Firebird	173	20
VW Rabbit	97	47

Test the hypothesis that $\rho = 0$ vs the alternative that $\rho < 0$ at the significance level $\alpha = 0.01$.

(b) A student has completed the data summary necessary to find r.

$$\Sigma XY = 4739, \Sigma X = 200, \Sigma Y = 287, \Sigma X^2 = 3568, \Sigma Y^2 = 4956, n = 20.$$

How can you tell the student made an error in the data summary calculations? The sums are not related to the part (a) of this problem.

[Ans: $r = 1869/1145.98 > 1$.]

11.1.3 Spearman's Rank Correlation Coefficient

Formula:

$$r_R = 1 - \frac{6\Sigma(R_x - R_y)^2}{n(n^2-1)}$$

When $n > 20$ approximate distribution of r_R is t_{n-2}.

Example:

11.1.4 Problems

1. Let z_X and z_Y be the z scores for samples X and Y. Prove that $r = \frac{\Sigma z_X z_Y}{n-1}$.

2. Prove that $r_{XY} = r_{aX+b,cY-d}$, where a, b, c, and d are constants.

3. Find r

Car	Engine Size (in^3)	Miles/gallon
Chevette	98	31
Sentra	98	35
Colt	86	41
Isuzu I Mark	111	27
Mercedes 190D	134	35
Firebird	173	20
VW Rabbit	97	47

4. A student has completed the data summary necessary to find r.

$$\Sigma XY = 4739, \Sigma X = 200, \Sigma Y = 287, \Sigma X^2 = 3568, \Sigma Y^2 = 4956, n = 20.$$

How can you tell the student made an error in the data summary calculations?
[Ans: $r = 1869/1145.98 > 1$.]

11.2 Regression

Olympic Metric Mile. [5] Table 4 shows to the nearest second the winning times beyond 200 seconds required in the 1500-meter run for men.

Year	Time in seconds minus 200	Year	Time in seconds minus 200
1900	46	1936	28
1904	45	1948	30
1908	43	1952	25
1912	37	1956	21
1920	42	1960	16
1924	34	1964	18
1928	33	1968	15
1932	31	1972	16

Use these data to:
(a) Plot time in seconds as a function of the year.
(b) Fit a line.
(c) Calculate and discuss the residuals.

Meat Research. In meat science research, a *pH* of 6.0 in postmortem muscle is desired before processing begins. In general, the *pH* at time of slaughter is 7.0 to 7.2, which decreases over time. It is not practical to monitor the *pH* decline for each carcass to determine when *pH* 6.0 is attained. Thus, an estimate is needed of the time when *pH* 6.0 is reached.

The estimated times to reach *pH* 6.0 for conventional treatment and for an electrically simulated hot boned (*ESHB*) treatment (Kastner et al., 1980) [6] were studied. The *ESHB* treatment was designed to increase the rate of

[5] Source: Adapted from data in *The World Almanac & Book of Facts*, 1980, edited by G. E. Delury, p. 818. new York: Newspaper Enterprise Association.

[6] Kastner, C. L., et al. (1980). Effects of carcass electrical stimulation and hot boning of selected beef muscles. In *Proceedings of the 26th European Meeting of the Meat Research Workers*, Vol. II, 40.

pH decline, potentially achieving *pH* 6.0 sooner, without affecting the terminal *pH*. Meat products may experience toughening induced by cold if processing begins prior to a *pH* of 6.0 being attained. On the other hand, delaying processing too long after *pH* is attained will diminish the positive benefits achieved from the *ESHB* procedure. A confidence interval about the time required to reach *pH* 6.0 is desired as an interval estimator for when processing can begin.

Postmortem *pH* declines were recorded for the *longissimus dorsi* muscle of 24 steer carcasses, 12 steers in each treatment group. The *pH* for each carcass was measured once at a specified time. Recordings were made at either 1, 2, 4, 6, 8, or 24 hours, where the *pH* of two different carcasses was measured at each hour. The observed data are presented in table below.

To obtain an estimate of the time required to attain *pH* 6.0, a response model is fit to the observed data. Although the two treatment groups each have the same *pH* at slaughter, a rapid change in the *pH* decline occurs after slaughter, which is further altered as *ESHB* treatment is applied. It is difficult for the meat science researcher to obtain data immediately after slaughter to adequately describe this change in the *pH* decline. Therefore, the first observed data point is not necessarily from the time of slaughter.

A typical practice for analyzing data of this type is to utilize a natural logarithm transformation of the independent variable, hour, to "linearize" the data. A simple linear model of the form $pH = \beta_0 + \beta_1(ln(hour)) + \epsilon$ is then fit to the data.

Conventional			*ESHB*		
Animal	Hour	pH	Animal	Hour	pH
1	1	7.02	13	1	7.01
2	1	6.93	14	1	6.96
3	2	6.42	15	2	6.25
4	2	6.51	16	2	6.09
5	4	6.07	17	4	5.77
6	4	5.99	18	4	5.62
7	6	5.59	19	6	5.57
8	6	5.80	20	6	5.39
9	8	5.51	21	8	5.46
10	8	5.36	22	8	5.30
11	24	5.30	23	24	5.37
12	24	5.47	24	24	5.50

Pollution Count. A plant distills liquid air to produce oxygen, nitrogen, and argon. The percentage of impurity in the oxygen is thought to be linearly related to the amount of impurities in the air, as measured by the "pollution count" in parts per million (ppm). Fit a linear regression model to the following data.

Purity(%)	93.3	92.0	92.4	91.7	94.0	94.6	93.6	93.1	93.2	92.9	92.2	91.3	90.1	91.6	91.9
Pollution(ppm)	1.10	1.45	1.36	1.59	1.08	0.75	1.20	0.99	0.83	1.22	1.47	1.81	2.03	1.75	1.68

Brakes, octanes and compression. The brake horsepower developed by an automobile engine on a dynamometer is thought to be a function of the engine speed in revolutions per minute (rpm), the road octane number of the fuel, and the engine compression. An experiment is run in the laboratory and the data that follow are collected. Fit a multiple linear regression model to these data.

Brake Horsepower	rpm	Octane Number	Compression
225	2000	90	100
212	1800	94	95
229	2400	88	110
222	1900	91	96
219	1600	86	100
278	2500	96	110
246	3000	94	98
237	3200	90	100
233	2800	88	105
224	3400	86	97
223	1800	90	100
230	2500	89	104

Sales of air conditioners. An appliance store dealer believes that the weekly sales of air conditioners are dependent upon the average outside temperature during the week. In support of this claim the dealer randomly selects eight weeks and records the average outdoor temperature and the sales of air conditioners during the week. This information is listed in the following chart:

Average Outside Temperature, x	Number of Air Conditioners Sold, y
75	4
83	7
67	2
89	9
95	12
79	6
81	6

(a) Draw a scatter diagram for the data and compute the coefficient of correlation.
(b) Do these data support the dealer's claim?
The MINITAB output for the regression procedure is given below:

```
The regression equation is
sales = - 22.2 + 0.354 temp

Predictor       Coef       Stdev    t-ratio        p
Constant     -22.213       1.704     -13.04    0.000
temp         0.35412     0.02085      16.98    0.000

s = 0.4660      R-sq = 98.3%     R-sq(adj) = 98.0%

Analysis of Variance

SOURCE       DF          SS          MS         F         p
Regression    1      62.628      62.628    288.39     0.000
Error         5       1.086       0.217
Total         6      63.714



     Fit  Stdev.Fit          95% C.I.          95% P.I.
  13.199      0.428    ( 12.098, 14.299)  ( 11.571, 14.826)
```

1. Test the hypothesis $H_0 : \beta_1 = 0.3$ against two sided alternative.
2. Find 98% confidence interval for the unknown population intercept β_0.

3. How many sales do you predict when temperature is 100? Find 99% confidence interval for the mean response when temperature is 100.

4. Discuss the regression output (one paragraph).

World Population Facts. The following is a list of the 1981 number of people per square kilometer of arable land and the per capita Gross National Product (in U.S. dollars) for several countries:[7]

Country	Persons Per Square Kilometer Of Arable Land	Per Capita Gross National Product, y
Egypt	1533	460
Brazil	58	1690
Panama	109	1350
Sweden	223	11,920
Belgium	682	10,890
Israel	321	4180
Italy	326	5240
United States	52	10,820
U.S.S.R	44	4110
China	309	230

(a) Draw a scatter diagram for the above data

(b) Compute the coefficient of correlation.

Primates. The table below contains data on the adult weights and quality of diet for 15 primate species. According to an ecological relationship called the Jarman-Bell Principle, among related species of animals the larger the body weight, the lower the quality of diet. The data in the table were assembled to study this relationship separately by sex, and to examine effects of the species' mating systems on the body weight-diet relationship. Index of dietary quality, derived from the proportions of animal matter (highest quality), plant reproductive parts (middle quality), and plant structural parts (lowest quality) in diet. Higher numbers indicate better quality of diet.[8]

Species	Male weight in kg	Male dietary quality	Female weight in kg	Female dietary quality
Tarsius bancanus	1.20	580	1.10	580
Cebus capucinus	3.80	410	2.70	410
Alouatta seniculus	8.10	351	6.40	351
Ateles belzebuth	7.40	362	7.60	362
Saimiri oerstedii	.89	460	.74	460
Cercopithecus cephus	4.10	406	2.90	426
Cercopithecus neglectus	7.00	392	4.00	392
Cercopithecus nictitans	6.60	394	4.20	366
Cercopithecus pogonias	4.50	428	3.00	429
Cercopithecus ascanius	9.00	416	6.40	443
Cerocebus albigena	9.00	416	6.40	443
Nasalis larvatus	20.40	210	9.98	210
Symphalangus syndactylus	11.10	326	10.30	326
Pongo pygmaeus	69.00	392	37.00	332

11.3 Regression

1. **The planets and their distances from the sun.** Astronomers have long been interested in law-like relations for interplanetary distances. In Table 3 we list the known planets and their true distances, d_n, from the sun. (Here

[7] Source: Statistical Office and Population Division of the United Nations; *Demographic Yearbook, 1980*; Population and Vital Statistics Report.

[8] Gaulin, S. J. C., and L. D. Sailer (1985). "Are females the ecological sex?" *American Anthropologist* 87: 111 - 119.

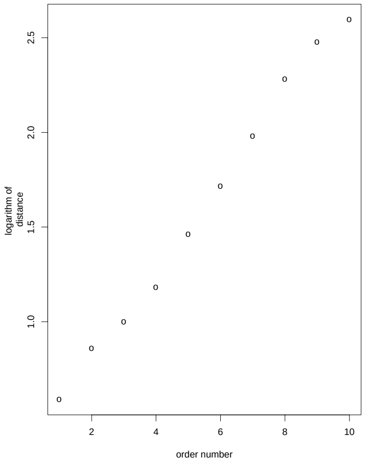

Figure 11.3: Amazing linearity

distances are measured in units of $\frac{1}{10}$ times the Earth's distance from the sun.) If we number planes by their order, moving outward from the sun, then Mercury is 1, Venus 2, and so on.

Planet	n	d_n	$\log_{10} d_n$
Mercury	1	3.87	0. 588
Venus	2	7.23	0.859
Earth	3	10.00	1.000
Mars	4	15.24	1.183
Asteroids	5	29.00	1.462
Jupiter	6	52.03	1.716
Saturn	7	95.46	1.980
Uranus	8	192.0	2.283
Neptune	9	300.9	2.478
Pluto	10	395.0	2.597

```
> num.pla
 [1]  1  2  3  4  5  6  7  8  9 10
> logdist.pla
 [1] 0.588 0.859 1.000 1.183 1.462 1.716 1.980 2.283 2.478 2.597
> Reg(num.pla, logdist.pla)

------------------------------------------------------------------
Source  SS      df      MS      F               p
------------------------------------------------------------------
Regr    4.486   1       4.486   1297.244        0
Error   0.028   8       0.003
Total   4.513   9
------------------------------------------------------------------
s= 0.059         R-squared= 0.994
------------------------------------------------------------------
Regression line is:  y = 0.332 + 0.233 * x
```

```
------------------------------------------------------------------
Parameter       Estim   St.dev  t       p-value
Intercept       0.332   0.04    8.268   0
Slope           0.233   0.006   36.017  0
------------------------------------------------------------------
```

Take of Oxigen. The human body takes in more oxygen when exercising than when it is at rest, and to deliver the oxygen to the muscles, the heart must beat faster. Heart rate is easy to measure but the measurement of oxygen uptake requires elaborate equipment. If oxygen uptake (VO2) can be accurately predicted from heart rate (HR) under a particular set of exercise conditions, then predicted, rather than measured, values can be used for various research purposes. Unfortunately, not all human bodies are the same, and a single equation that works for all people cannot be found. But individuals can be measured for both variables under varying sets of exercise conditions for a short time, and regression equations for predicting oxygen uptake from heart rate can be calculated for each person. The predicted oxygen uptakes can then be used in place of measured uptakes for the same individuals in later experiments. The data for one individual are as follows. [9]

HR	VO2	HR	VO2
94	.473	108	1.403
96	.753	110	1.499
95	.929	113	1.529
95	.939	113	1.599
94	.832	118	1.749
95	.983	115	1.746
94	1.049	121	1.897
104	1.178	127	2.040
104	1.176	13	2.231
106	1.292		

(a) Plot the data. Are there any outliers or unusual points?

(b) Compute the least-squares regression line for predicting oxygen uptake from heart rate for this individual.

(c) Test the null hypothesis that the slope of the regression line is 0.

(d) Since the regressions to be used for prediction, we calculate prediction intervals for future observations. Calculate 95for heart rates of 95 and 110.

(e) From what you have learned in (a), (b), (c), and (d) of this exercise, do you think that the researchers should use predicted VO2 in place of measured VO2 for this individual under similar experimental conditions? Explain your answer.

Volume of RBr. In a chemistry experiment, a volume of RBr is added to a solvent mixture of isopropyl alcohol and water. A reaction then occurs which causes a change in the pH levels of the solution. To monitor the rate of reaction, a pH meter is used. Below are the data obtained:

time	*pH*	[RBr]	time	*pH*	[RBr]
x		(conc. of RBr)	x		(conc. of RBr)
15	3.40	0.0079433	120	2.28	0.0026952
30	2.81	0.0063945	135	2.22	0.0019177
45	2.61	0.0054886	150	2.21	0.0017773
60	2.49	0.0047073	165	2.20	0.0016337
75	2.40	0.0039622	180	2.18	0.0013364
90	2.35	0.0034764	195	2.16	0.0010250
105	2.30	0.0029952	210	2.13	0.0005302

[9] Data provided by Paul Waldsmith from experiments conducted in Don Corrigan's lab at Purdue University, West Lafayette, IN.

Team	Number of Games Won	Team Batting Average
Cleveland	57	0.254
New York	71	0.256
Boston	84	0.269
Toronto	91	0.257
Texas	85	0.270
Detroit	84	0.247
Minnesota	95	0.280
Baltimore	67	0.254
California	81	0.255
Milwaukee	83	0.271
Seattle	83	0.255
Kansas City	82	0.264
Oakland	84	0.248
Chicago	87	0.262

A reaction's rate may be described as either first order or second order. A first order reaction plots a straight line for the graph of "ln[RBr] vs t". A second order reaction plots a straight line for the graph "1/[RBr] vs t".

Basic metabolic rate. A group of researchers is studying basal metabolic rate, i.e. the number of calories your body consumes when you are in a reclined, resting state. They study a group of 15 subjects, both female and male, and measure their heights, weights, and metabolic rates with a gas exchange analysis. Their goal is to see if weight and height correlate with metabolic rate.

bmr	weight	height	bmr	weight	height
1800	70	170	1850	71	175
1725	72	169	1810	73	177
1799	74	172	1847	75	173
1850	76	178	1200	50	155
1344	54	160	1322	55	166
1400	56	161	1386	58	163
1375	63	168	1450	63	170
1433	64	170			

Team's batting average. Is the number of games won by a major league baseball team in a season related to the team's batting average? The accompanying table shows the number of games won and the batting averages for the 14 teams in the American League for the 1991 season. [Team batting average is the ratio of the total number of hits to the total number of "at-bats" for the team.]

(a) Fill in the ANOVA table in MINITAB regression output.

```
MTB > regress 'won' 1 'ba'

 The regression equation is
 won = ______ + ______  ba

 Predictor       Coef       Stdev    t-ratio        p
 Constant      -37.44       _____      -0.54    0.600
 ba             _____       267.0       1.71    0.114

 s = _______        R-sq = _____%     R-sq(adj) = 12.8%

 Analysis of Variance
```

```
SOURCE       DF        SS          MS         F       p
Regression   __    _______       244.94    ______   0.114
Error        __    1011.06       _____
Total        __    _______
```

(b) Test the hypothesis that the slope $\beta_1 = 500$, against the two sided alternative. Take $\alpha = 0.05$.

(c) Find the 95% confidence interval for the intercept β_0.

(d) If the team has BA=0.275, what is the predicted number of games won? (round to the nearest integer)

(e) Give general comments on the model proposed (good predictability, bad predictability, ...).

Low birthweight. The state of Vermont is divided into 10 Health Planning Districts, which correspond roughly to counties. The following data for 1980 represent the percentage of births of babies under 2500 grams (Y), The fertility rate for females younger than 18 and older than 34 years of age (X_1), and the percentage of births to unmarried mother (X_2) for each district. (Both X_1 and X_2 are known to be risk factors for low birthweight.)

District	Y	X_1	X_2
1	6.1	43.0	9.2
2	7.1	55.3	12.0
3	7.4	48.5	10.4
4	6.3	38.8	9.8
5	6.5	46.2	9.8
6	5.7	39.9	7.7
7	6.6	43.1	10.9
8	8.1	48.5	9.5
9	6.3	40.0	11.6
10	6.9	56.7	11.6

Cough syrup. A particular brand of cough syrup comes on $\frac{1}{4}$-liter bottles and the manufacturer recommends that after a bottle is unsealed it be kept under cool conditions. The shelf-life of the cough syrup in question is dependent on the temperature at which it is stored. The quality control laboratory of the manufacturing company has obtained the data in the table below.

Bottle	Shelf-Life	Storage Temperature
Number	(Y, in days)	(X, in $^\circ C$)
1	727	13
2	760	14
3	730	15
4	716	16
5	683	17
6	665	18
7	641	19
8	663	20
9	653	21
10	615	22
11	585	23
12	614	24
13	592	25
14	564	26
15	537	27
16	537	28
17	552	29
18	507	30

Repeating GRE. It is at least part of the folklore that repeated experience with the Graduate Record Examination (GRE) leads to better scores, even without any intervening study. We obtain eight subjects and give them the GRE verbal exam every Saturday morning for 3 weeks. The data follow:

Subject	First score	Second score	Third score
1	550	575	580
2	440	440	470
3	610	630	610
4	650	670	670
5	400	460	450
6	700	680	710
7	490	510	515
8	580	550	590

(a) Write the statistical model for these data,

(b) Run the ANOVA,

(c) What, if anything, would you conclude about practice effects on the GRE?

Vocabulary and spelling scores. An English teacher believes that vocabulary and spelling scores are related. To support this claim, the following data have been collected:

Vocabulary, x	Spelling, y
72	75
89	88
60	72
79	77
64	69
55	70
69	73
65	80
75	81

(a) Draw a scatter diagram for the data and compute the coefficient of correlation.

(b) Is the English teacher's claim justified? (Find r_{XY})

(c) Fill in the ANOVA table.

(d) Test the hypothesis $H_0 : \hat{\beta}_0 =$ against the twosided alternative. Take $\alpha = 0.05$.

(e) Find 95% CI for β_1.

```
The regression equation is
spell = 42.3 + 0.484 vocab

Predictor        Coef       Stdev    t-ratio        p
Constant       42.325       8.903       4.75    0.000
vocab          0.4842      0.1264       3.83    0.006

s = 3.700       R-sq = 67.7%     R-sq(adj) = 63.1%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      201.05      201.05     14.68    0.006
Error         7       95.84       13.69
Total         8      296.89
```

Hemodilution. Clark et al.[10] examined the fat filtration characteristics of a packed polyester-and-wool filter used in the arterial lines during clinical hemodilution. They collected data on the filter's recovery of solids for 10 patients who underwent surgery. The table below shows removal rates of lipids and cholesterol. Fit a regression line to the data, where the cholesterol is the Y variable and lipids are the X variables.

	Removal rates, mg/kg/L $\times 10^{-}2$	
patient	Lipids (X)	Cholesterol (Y)
1	3.81	1.90
2	2.10	1.03
3	0.79	0.44
4	1.99	1.18
5	1.03	0.62
6	2.07	1.29
7	0.74	0.39
8	3.88	2.30
9	1.43	0.93
10	0.41	0.29

Give answers to the following questions using the MINITAB output.

- What test is performed by given p-value. State H_0 and H_1.
- Find the 95% confidence interval for the intercept of the population regression line (β_0).
- Test the hypothesis that the slope of the population regression line is equal to 1 versus the one-sided alternative that it is less than 1.
- Discuss the quality of the proposed linear fit. Predict the cholesterol rate for lipids=1.80 mg/kg/L$\times 10^{-2}$.

[10] Clark, R., Margraf, H., and Beauchamp, R. (1975). Fat and Solid Filtration in Clinical Perfusions. *Surgery,* 77, 216-224.

```
The regression equation is
cholest = 0.0616 + 0.534 lipids

Predictor       Coef       Stdev    t-ratio        p
Constant     0.06163     0.07396       0.83    0.429
lipids       0.53445     0.03423      15.61    0.000

s = 0.1252      R-sq = 96.8%     R-sq(adj) = 96.4%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      3.8175      3.8175    243.71    0.000
Error         8      0.1253      0.0157
Total         9      3.9428
```

Sand Dollars. Pilkey and Hower[11] examined the changes in concentration of Magnesium and Strontium in the tests of a recent echinoid species collected from varying environments over much of its geographical range. They used X-ray technique to analyze the Magnesium and Strontium in specimens of *Dendraster excentricus*, the common Pacific Coast sand dollar, collected from 24 localities between Vancouver Island, British Columbia and Santa Rosalia Bay, Baja California, and calculated the percentage of calcium. Table below shows the mean summer temperature (X) at the 24 locations and the mean percent $MgCO_3$ content (Y) of the specimens collected.

$^\circ C(X)$	$MgCO_3(Y)$	$^\circ C(X)$	$MgCO_3(Y)$	$^\circ C(X)$	$MgCO_3(Y)$
23.0	9.5	18.7	9.0	17.5	9.2
21.0	9.2	20.0	9.4	19.0	9.3
15.3	9.0	14.0	8.5	14.0	9.0
13.7	8.4	13.3	8.8	13.6	8.9
13.1	8.5	13.0	8.7	13.6	8.6
14.2	8.7	13.9	8.5	14.8	9.1
14.2	9.1	13.0	8.0	16.1	8.1
15.9	8.5	13.0	8.4	11.7	8.7

1. Fill-in the ANOVA table.

2. Find the 99% confidence interval for the **mean** percent $MgCO_3$ content if the mean summer temperature is 16.0°C.

```
The regression equation is
MgCO3 = ______ + ______ temp

Predictor       Coef       Stdev    t-ratio        p
Constant      7.3879     _______      22.34    0.000
temp        ________     0.02111       4.33    0.000

s = 0.2976      R-sq = ________    R-sq (adj) = ____________

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression   __      ______     _______     ______    0.000
Error        __      ______     _______
Total        __      3.6096


     Fit  Stdev.Fit          95% C.I.          95% P.I.
  8.8507     0.0621    ( 8.7220, 8.9794)   ( 8.2201, 9.4813)
```

[11] Pilkey, O. and Hower, J. (1960). The Effect of Environment on the Concentration of Skeletal Magnesium and Strontium in *Dendraster*. *J. Geol.* 68, 203-214.

Hint: **Adjusted r^2:** The definition is $1-(1-r^2)\frac{n-1}{n-m-1}$.

Mike's Used Mercedes Prices Data Set.

```
+++++++++++++++++++
USED MERCEDES PRICES  Mike West
+++++++++++++++++++

Data taken from the advertising pages of the Sunday Times on 27th February 1983.
Prices of used Mercedes classified according to type/model of car, age of car
(in six-month units based on date of registration), recorded mileage, and vendor.

Interest lies in explaining price, distinguishing models according to price,
depreciation with age and mileage, possible varying depreciation rates across
models of car, collinearity issues for age and mileage, possible vendor
differences on asking prices, outliers, influential cases, predicting prices

Data columns:
    1. Case number 1,2, .... , n=54
    2. Asking price in thousands of pounds
    3. Type/Model of car
         1=model 500, 2=450, 3=380, 4=280, 5=200
    4. Age of car in six-month units
         (based on advertised registration date)
    5. Recorded mileage (in thousands of miles)
    6. Vendor
    (1,2,3,4 represent different dealerships, 5 represents sale by owner)

  Case  Price Mod  Age  Mile Vend

    1   30.495  1    1     7   1
    2   22.250  1    3    16   1
    3   23.995  1    3     8   2
    4   18.495  1    3    15   3
    5   20.950  1    2    26   5
    6   21.500  1    3    18   5
    7   19.995  1    5    24   1
    8   18.950  1    5    20   4
    9   15.695  2    6    13   1
   10   15.995  2    6    27   2
   11   16.595  2    6    18   3
   12   15.995  2    6    25   3
   13    9.950  2   11    43   5
   14   17.995  3    3    23   1
   15   17.495  3    4    16   1
   16   21.995  3    3    13   2
   17   21.995  3    1     2   2
   18   19.695  3    2     5   3
   19   16.850  3    4    15   5
   20   16.750  3    4    44   5
   21   19.850  3    4    22   5
   22   21.000  3    3     5   4
   23   17.950  3    5    44   4
   24   17.995  4    1     6   2
   25   16.295  4    2    14   1
   26   16.495  4    3     4   2
   27   15.995  4    4    15   2
   28   13.995  4    3    7   2
   29   16.995  4    1   11   3
   30   15.595  4    3   14   3
   31   11.995  4    5   11   3
   32   13.195  4    4   13   3
   33    8.995  4    7   31   5
   34   17.500  4    4   25   5
   35   15.450  4    4   25   4
   36   14.750  4    2   21   4
   37   15.750  4    4   16   4
   38   12.250  4    5   27   4
   39   10.995  5    2   13   1
   40   10.995  5    2   12   1
   41   10.495  5    3   24   1
   42    8.995  5    5   34   1
   43   10.295  5    3    7   3
   44    6.450  5    7   41   4
   45   10.950  5    4   29   4
   46    9.750  5    4   21   4
   47    6.950  5    8   35   4
   48    9.950  5    3   11   4
   49    8.750  5    3   40   4
   50   10.750  5    2   12   5
   51    4.950  5    8   57   5
   52    9.250  5    2   23   5
   53    9.250  5    4   21   5
   54    8.995  5    3   24   5
----------------------------------------------------------------
1
 MTB > read 'merc' c1-c6
      54 ROWS READ

  ROW    C1       C2   C3    C4    C5   C6

    1     1   30.495    1     1     7    1
    2     2   22.250    1     3    16    1
    3     3   23.995    1     3     8    2
    4     4   18.495    1     3    15    3
   .  .  .

 MTB > name c1 'case' c2 'price' c3 'model'  c4 'age' c5 'mileage' c6 'vendor'
 MTB > stepwise 'price' c3-c6

  STEPWISE REGRESSION OF  price   ON  4 PREDICTORS, WITH N =   54

     STEP        1        2
 CONSTANT    25.17    30.22

 model       -2.90    -2.99
 T-RATIO     -9.23   -14.46

 age                  -1.23
 T-RATIO              -8.35

 S            3.20     2.10
 R-SQ        62.07    83.99
  MORE? (YES, NO, SUBCOMMAND, OR HELP)
 SUBC> no
```

```
MTB > #------------------------------------------------------------------
MTB > breg c2 c3-c6

Best Subsets Regression of price

                                          m
                                          i v
                                      m   l e
                                      o   e n
                                      d a a d
              Adj.                    e g g o
Vars   R-sq   R-sq    C-p         s   l e e r

   1   62.1   61.3   70.8    3.2007   X
   1   23.2   21.7  194.5    4.5539       X
   2   84.0   83.4    3.0    2.0999   X X
   2   75.8   74.9   28.9    2.5791   X   X
   3   84.5   83.6    3.3    2.0850   X X X
   3   84.3   83.3    4.2    2.1034   X X   X
   4   84.6   83.4    5.0    2.1002   X X X X


MTB >  regress c2 2 c3 c4;
SUBC>    vif;
SUBC>    dw;
SUBC>    residuals c7.

The regression equation is
price = 30.2 - 2.99 model - 1.23 age

Predictor       Coef       Stdev    t-ratio        p       VIF
Constant     30.2170      0.9814      30.79    0.000
model        -2.9878      0.2067     -14.46    0.000       1.0
age          -1.2335      0.1476      -8.35    0.000       1.0

s = 2.100       R-sq = 84.0%     R-sq(adj) = 83.4%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    2     1179.65      589.82    133.76    0.000
Error        51      224.90        4.41
Total        53     1404.55

SOURCE       DF      SEQ SS
model         1      871.83
age           1      307.82

Unusual Observations
Obs.   model     price       Fit Stdev.Fit  Residual   St.Resid
  1     1.00    30.495    25.996     0.737     4.499      2.29R
  4     1.00    18.495    23.529     0.606    -5.034     -2.50R
 13     2.00     9.950    10.673     1.121    -0.723     -0.41 X
 16     3.00    21.995    17.553     0.330     4.442      2.14R
 34     4.00    17.500    13.332     0.306     4.168      2.01R

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

```
 Durbin-Watson statistic = 1.01

 MTB > plot c7 vs c1

          -
          -  *              *                 *
       3.5+                      **
          -                        *   *
  C7      -                  *        *        * *
          -                *                               *
          -    *          *   ***   **    *              *
       0.0+            *                 * **     *   * *
          -           * **              *    *  *         *    *
          -   *    * *                               * *    *    *
          -       * *                              **         *
          -                                                  *    *
      -3.5+      *                                              *
          -
          -     *
          -
          -
            +---------+---------+---------+---------+---------+------case
            0        10        20        30        40        50
```

Rabbits. The following temperatures (Y) were recorded in a rabbit at various times (X) after being inoculated with rinderpest virus[12]

Time after injection (hrs)	Temperature ($^\circ F$)
24	102.8
32	104.5
48	106.5
56	107.0
72	103.9
80	103.2
96	103.1

Since the linear regression with one predictor (`time`) gives insignificant F statistics we include `time`^{2} (squared time) as a second predictor making the regression quadratic.

MINITAB gives the following output:

```
The regression equation is
 temp = 98.5 + _______ time - 0.00230 time^2

 Predictor       Coef       Stdev    t-ratio        p
 Constant      ______      3.142       ______   0.000
 time          0.2577      0.1179      ______   0.094
 time^2       _________    ______      -2.34    0.079

 s = 1.311       R-sq =  ______    R-sq(adj) = 39.7%

 Analysis of Variance
```

[12] Carter, G. and Mitchell, C. (1958). Methods for adapting the virus of Rinderpest to rabbits. *Science* **128**, *252-253.*

```
SOURCE       DF         SS          MS         F        p
Regression   __      10.239       _____     ______    0.161
Error        __      ______       _____
Total        __      ______
```

1. Fill in the ANOVA output.
2. Discuss the output. Does the proposed regression explain variability in data.

Solution

```
 temp = 98.5 + 0.258 time - 0.00230 time^2

 Predictor       Coef       Stdev    t-ratio        p
 Constant      98.548       3.142      31.37    0.000
 time          0.2577      0.1179       2.19    0.094
 time^2    -0.0022984   0.0009811      -2.34    0.079

 s = 1.311       R-sq = 59.8%     R-sq(adj) = 39.7%

 Analysis of Variance

 SOURCE       DF         SS          MS         F        p
 Regression    2      10.239       5.120      2.98    0.161
 Error         4       6.875       1.719
 Total         6      17.114
```

Rabbits. The following temperatures (Y) were recorded in a rabbit at various times (X) after being inoculated with rinderpest virus[13]

Time after injection (hrs)	Temperature ($^\circ F$)
24	102.8
32	104.5
48	106.5
56	107.0
72	103.9
80	103.2
96	103.1

Since the linear regression with one predictor (`time`) gives insignificant F statistics we include `time`^{2} (squared time) as a second predictor making the regression quadratic. MINITAB gives the following output:

```
The regression equation is
 temp = 98.5 + _______ time - 0.00230 time^2

 Predictor       Coef       Stdev    t-ratio        p
 Constant      ______      3.142      ______    0.000
 time          0.2577     0.1179      ______    0.094
 time^2     _________     ______       -2.34    0.079

 s = 1.311       R-sq =  ______    R-sq(adj) = 39.7%

 Analysis of Variance
```

[13] Carter, G. and Mitchell, C. (1958). Methods for adapting the virus of Rinderpest to rabbits. *Science* **128**, *252-253.*

```
SOURCE       DF          SS          MS         F        p
Regression   __      10.239       _____      ______    0.161
Error        __      ______       _____
Total        __      ______
```

1. Fill in the ANOVA output.
2. Discuss the output. Does the proposed regression explain variability in data.

Potato Leafhopper. Length of developmental period (in days) of potato leafhopper, *Empoasca fabae*, from egg to adult seem to be dependent on the temperature.[14] The original data were weighted means, but for purpose of this analysis we shall consider them as though they were single observed values.

Temp ° F	59.8	67.6	70.0	70.4	74.0	75.3	78.0	80.4	81.4	83.2	88.4	91.4	92.5
Develop (days)	58.1	27.3	26.8	26.3	19.1	19.0	16.5	15.9	14.8	14.2	14.4	14.6	15.3

The MINITAB output is given below.

```
       60+
         -      *
 days    -
         -
         -
       45+
         -
         -
         -
         -
       30+
         -                  *  **
         -
         -
         -                           * *  *
       15+                                    **  *      *    **
         -
           +---------+---------+---------+---------+---------+------temp
        56.0      63.0      70.0      77.0      84.0      91.0

The regression equation is
days = 99.1 - 0.993 temp

Predictor       Coef       Stdev    t-ratio        p
Constant       99.07       17.48       5.67    0.000
temp         -0.9933      0.2228      -4.46    0.001

s = 7.479       R-sq = 64.4%     R-sq(adj) = 61.1%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      1111.5      1111.5      19.87    0.001
Error        11       615.3        55.9
Total        12      1726.7
```

[14] Kouskolekas, C. and Decker, G. (1966). The effect of temperature on the rate of development of the potato leafhopper, *Empoasca fabae* (Homoptera: Cicadelidae) *Ann. Etimol. Soc. Amer.* **59**, *292-298.*

For the temperature of 85°F the following prediction/confidence intervals are obtained:

```
Fit   Stdev.Fit         95% C.I.         95% P.I.
14.64       2.61    (   8.89,   20.39)  (  -2.80,   32.08)
```

1. Discuss the regression output (one paragraph).
2. Find 98% CI for unknown slope β_1.
3. Test the hypothesis that the intercept is equal to 60 against the alternative that it is bigger than 60. Take $\alpha = 0.01$.
4. What is 96% Confidence interval for the mean response (mean number of days) if temperature is 85.

Comet Bennett. Bacos[15] reported observations made on comet Bennett, which are given in the table below.

Heliocentric Dist. (R)	0.685	0.720	0.735	0.750	0.798	0.810	0.828	0.950	0.988	1.210	1.228	1.224	1.267	1.295	1.312	1.330
Reduced Vis. Mag. (H)	1.72	1.80	2.53	2.15	2.81	2.92	2.80	3.82	3.77	5.06	5.21	5.19	5.39	5.61	5.57	5.89

(i) The coefficient of linear correlation, r, between reduced visual magnitude (H) and the heliocentric distance (R) is found to be 0.993. Comment on appropriateness of a linear model. (One or two sentences)

The MINITAB output gives the regression equation $\hat{H} = \hat{\beta}_0 + \hat{\beta}_1 R$.

```
The regression equation is
H = - 2.17 + 6.01 R

Predictor        Coef        Stdev    t-ratio         p
Constant      -2.1734       0.2004     -10.85     0.000
R              6.0145       0.1933      31.11     0.000

s = 0.1864       R-sq = 98.6%      R-sq(adj) = 98.5%

SOURCE        DF           SS          MS          F        p
Regression     1       33.615      33.615     967.97    0.000
Error         14        0.486       0.035
Total         15       34.101
```

(i) Test the hypothesis $H_0 : \beta_1 = 5.5$. against the alternative $H_1 : \beta_1 > 5.5$ at the level of significance $\alpha = 0.05$.

(ii) Find 98%-Confidence interval for the intercept β_0.

(iii) The "fit" for $R = 1.1$ is given in the following part of MINITAB output:

```
   Fit   Stdev.Fit         95% C.I.          95% P.I.
4.4426      0.0499    ( 4.3356, 4.5495)  ( 4.0287, 4.8564)
```

Find a 97% confidence interval for the mean fit. Note that the 95% confidence interval is given: $(4.3356, 4.5495)$.

Taste of Cheese. As cheddar cheese matures a variety of chemical processes take place. The taste of mature cheese is related to the concentration of several chemicals in the final product. In a study of cheddar cheese from LaTrobe Valley of Victoria, Australia, samples of cheese were analyzed for their chemical composition and where subjected to taste tests. The table below presents data[16] for one type of cheese manufacturing process. *Taste* is the response variable of interest. The taste scores where obtained by combining scores from several tasters. Three of the chemicals whose concentrations were measured were *acetic acid, hydrogen sulfide,* and *lactic acid.* For acetic acid and hydrogen sulfide log transformations were taken.

[15] Bacos, G. (1973). Photoelectric observations of comet Bennett. *J. Roy. Astron. Soc. Can., 67, 183-189.*

[16] Data from the experiments of G. T. Lloyd and E. H. Ramshaw, CISRO Food Research, Victoria, Australia. Published in Moore, D. and McCabe, G. *Introduction to the Practice of Statistics,* Freeman, 1989.

Taste	Acetic	H2S	Lactic	Taste	Acetic	H2S	Lactic
12.3	4.54	3.13	0.86	20.9	5.16	5.04	1.53
39.0	5.37	5.44	1.57	47.9	5.76	7.59	1.81
5.6	4.66	3.81	0.99	25.9	5.70	7.60	1.09
37.3	5.89	8.73	1.29	21.9	6.08	7.97	1.78
18.1	4.90	3.85	1.29	21.0	5.24	4.17	1.58
34.9	5.74	6.14	1.68	57.2	6.45	7.91	1.90
0.7	4.48	3.00	1.06	25.9	5.24	4.94	1.30
54.9	6.15	6.75	1.52	40.9	6.37	9.59	1.74
15.9	4.79	3.91	1.16	6.4	5.41	4.70	1.49
18.0	5.25	6.17	1.63	38.9	5.44	9.06	1.99
14.0	4.56	4.95	1.15	15.2	5.30	5.22	1.33
32.0	5.46	9.24	1.44	56.7	5.86	10.20	2.01
16.8	5.37	3.66	1.31	11.6	6.04	3.22	1.46
26.5	6.46	6.92	1.72	.7	5.33	3.91	1.25
13.4	5.80	6.69	1.08	5.5	6.18	4.79	1.25

(1) Fill-in the ANOVA table:

```
The regression equation is
Taste = - 28.8 + _____ Acetic + 3.92 H2S + 19.6 Lactic

Predictor       Coef       Stdev    t-ratio        p
Constant       _____      _____      -1.46    0.155
Acetic         _____      4.450       0.07    0.942
H2S            3.924      1.245       3.15    0.004
Lactic        19.586      8.623       2.27    0.032

s = ____       R-sq = _______%     R-sq(adj) = 61.3%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression  ____      _______      ______    16.30    0.000
Error       ____      2659.7       ______
Total       ____      _______

     Fit  Stdev.Fit          95% C.I.           95% P.I.
   22.13       3.42    (_______, ________) (   0.18,  44.08)
```

(2) Choose the “best” linear model in the problem. Explain your choice.

```
Best Subsets Regression of Taste

                                      A   L
                                      c   a
                                      e   c
                                      t H t
              Adj.                    i 2 i
Vars    R-sq   R-sq    C-p          s   c S c

   1    57.3   55.8    6.0    10.809     X
   1    49.6   47.8   11.8    11.745       X
   2    65.3   62.7    2.0    9.9261     X X
   2    58.4   55.3    7.2    10.865   X X
   3    65.3   61.3    4.0    10.114   X X X
```

Does SAT predict GPA? For years, admission committees from from leading colleges and universities have been responsible for accepting only the best and brightest students to their respective schools. These admission people attempt to predict how a student will perform academically based on their high school experience. One of the main predictors that schools use is SAT score. Eric, Heath and Jeff[17] decided to explore if the SAT scores can be viable predictors of academic success among Duke students. They took as a sample the graduating Duke class of Spring 1995 (n=1414 pairs)

The following is an excerpt from their MINITAB output.

```
MTB > regress 'DUKEGPA' on 1 predictor 'SAT';
SUBC>   predict 1300.

The regression equation is
DUKEGPA = 1.21 + 0.00157 SAT

Predictor     Coef      Stdev      t-ratio        p
Constant     ______    0.1053     _______      0.0000
SAT          ______    0.00008085 _______      0.0000


s=__________    R-sq= __________     R-sq(adj)= 21.0%

Analysis of Variance

SOURCE        DF         SS          MS            F          p
Regression   ____     53.599      _____         _____      0.0000
Error        ____     ______      0.142
Total        1413     ______


Fit          Stdev.Fit     95% CI          95% PI
_______       0.01000     _______        ___________
```

(i) Complete the output.

(ii) Test the hypothesis that the population slope $\beta_1 = 1$ against the alternative $\beta > 1$. Take $\alpha = 0.01$.

(iii) Give 98 % confidence interval for the intercept.

(iv) Give 98 % confidence interval for the mean response for SAT=1300. Compare this interval with the interval in (iii). Why they have different lengths?

HINT: You may find the following relation useful: $s^2_{\text{pred}} = s^2 + s^2_{\text{mean}}$, where s_{pred} and s_{mean} are standard deviations for the predicted fit and the mean fit respectively.

[17] Eric Givner, Heath Mills, and Jeff Laoang: Does SAT predict GPA? STA110E Project, Fall 1995.