Regression Analysis for WSDOT Material Applications

WA-RD 143.2

Final Report
February 1988

Washington State Department of Transportation
Planning, Research and Public Transportation Division

in cooperation with the
United States Department of Transportation
Federal Highway Administration

WASHINGTON STATE DEPARTMENT OF TRANSPORTATION

TECHNICAL REPORT STANDARD TITLE PAGE

1. REPORT NO. WA-RD 143.2	2. GOVERNMENT ACCESSION NO.	3. RECIPIENT'S CATALOG NO.

4. TITLE AND SUBTITLE
REGRESSION ANALYSIS FOR WSDOT MATEPIAL APPLICATIONS

5. REPORT DATE
February 1988

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)
Joe P. Mahoney

8. PERFORMING ORGANIZATION REPORT NO.

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Washington State Transportation Center (TRAC)
and the University of Washington
Department of Civil Engineering
Seattle, Washington 98195

10. WORK UNIT NO.

11. CONTRACT OR GRANT NO.
GC 8286, Task 3

12. SPONSORING AGENCY NAME AND ADDRESS
Washington State Department of Transportation
Transportation Building, KF-01
Olympia, Washington 98504

13. TYPE OF REPORT AN PERIOD COVERED
Course notes

14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES
This study was conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.
WSDOT Contract Manager was Mr. Newton C. Jackson.

16. ABSTRACT

This document represents the course notes used in training WSDOT personnel on regression modeling techniques. This work was done in conjunction with the development of pavement performance equations by WSDOT personnel which were based on data in the WSDOT Pavement Management System. This study was funded by the Federal Highway Administration.

17. KEY WORDS
Regression, statistics, materials, pavements, models

18. DISTRIBUTION STATEMENT
No restrictions. This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161

19. SECURITY CLASSIF. (of this report) Unclassified	20. SECURITY CLASSIF. (of this page) Unclassified	21. NO. OF PAGES 71	22. PRICE

REGRESSION ANALYSIS FOR WSDOT MATERIAL APPLICATIONS

by

Joe P. Mahoney
Associate Professor

Washington State Transportation Center (TRAC)
135 More Hall, FX-10
University of Washington
Seattle, Washington 98195

Washington State Department of Transportation
Technical Monitor
Newton Jackson
Pavements Engineer

Final Report

Research Project GC 8286
Task 3

Prepared for

Washington State Transportation Commission
Department of Transportation
and in cooperation with
U.S. Department of Transportation
Federal Highway Administration

February 1988

DISCLAIMER

The contents of this report reflect the views of the author, who is responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Washington State Transportation Commission, Department of Transportation, or the Federal Highway Administration. This report does not constitute a standard, specification, or regulation.

ACKNOWLEDGMENTS

The author gratefully acknowledges the opportunity to develop these notes. This opportunity was created by personnel from the WSDOT Materials Laboratory and the FHWA. Specific individuals who have contributed include Newton C. Jackson, Keith Kay, Dennis M. Crimmins, and Keith W. Anderson of the Washington State Department of Transportation A special thanks is extended to Professor Virgil L. Anderson of Purdue University, who reviewed these notes.

TABLE OF CONTENTS

Section Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

OUTLINE

SECTION 1.0
DEFINITIONS

1. STATISTICAL [1.1]
"Having to do with numbers" or "drawing conclusions from numbers."

2. POPULATION [after Ref. 1.2]
All measurements or counts that are obtainable from all of the objects that process some common characteristic. Example: a "population" of data would be the pavement condition measured on <u>all</u> Interstate highways in Washington state.

3. SAMPLE [1.2]
A set of measurements or counts that constitute a part (or all) of the population.

4. RANDOM SAMPLING [1.2]
A sampling procedure whereby any one measurement in the population is as likely to be included as any other.

5. BIASED SAMPLING [after Ref. 1.2]
A sampling procedure whereby certain individual measurements have a greater chance of being included than others. Example: biased sampling would be taking density measurements only at places on a base course that appeared to be well compacted.

6. MEAN
Average of a group of measurements. The population mean is designated "μ" and a sample mean by "$\bar{x}$."

7. MEDIAN [1.2]
The number, in a set of numbers arranged in ascending order, that divides the set so that half of the numbers are higher and half are lower.

8. RANGE
The largest measurement minus the smallest measurement in a group of data.

9. STANDARD DEVIATION
A measure of variation or dispersion of a group of data. Specifically, the average of the squares of the numerical differences of each measurement (or observation) from the mean. The population standard deviation is designated by "σ" and a sample standard deviation by "s."

10. HISTOGRAM
A graphical form of data presentation. A bar chart that shows in terms of area the relative number of measurements of different classes. The width of the bar represents the class interval, the height represents the number of measurements.

11. VARIABLE [1.3]
A quantity to which any of the values in a given set may be assigned, i.e., something on which measurements are made.

12. CORRELATION
A way to measure the association between two variables.

13. REGRESSION
Goes a step further than correlation. Generates an equation that can be used to predict one variable from another (or others in multiple regression). The predicted variable is the dependent variable and the other variables are called independent variables.

14. SYSTEMATIC SAMPLING [1.2]
Selection of successive observations at uniform intervals in a sequence of time, area, etc. Example: taking pavement deflection measurements every 500 ft. on a project.

15. Examples for the calculation of sample mean, sample standard deviation, range and a histogram are shown in Table 1.1 and Figure 1.1.

Table 1.1. Calculation of Sample Mean, Sample Standard Deviation and Range for Procter Density Data [after Ref. 1.2]

1. Basic data (Procter density) in pcf:

107.5	100.8	107.0	101.5	107.0
112.0	111.4	124.0	103.3	101.3
104.3	109.4	103.5	114.1	98.0
106.0	99.7	110.5	105.0	93.5
101.3	102.5	95.5	94.0	110.1

2. Greek symbol "Σ" indicates a <u>summation</u> calculation is required. To sum the 25 density test results above:

$$\sum_{i=1}^{25} x_i = 107.5 + 112.0 + 104.3 + \ldots + 93.5 + 110.1 = 2,623.2$$

3. Sample mean ($\bar{x}$)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2,623.2}{25} = 104.93 \approx 105 \text{ pcf}$$

4. Sample standard deviation (s)

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{(107.5 - 104.9)^2 + (112.0 - 104.9)^2 + \ldots + (110.5 - 104.9)^2}{25-1}}$$

$$= \sqrt{\frac{1114.27}{25-1}} \quad = 6.8 \text{ pcf}$$

5. Range

Range = largest density - smallest density

$= x_{max} - x_{min}$

$= 124.0 - 93.5 = 30.5$ pcf

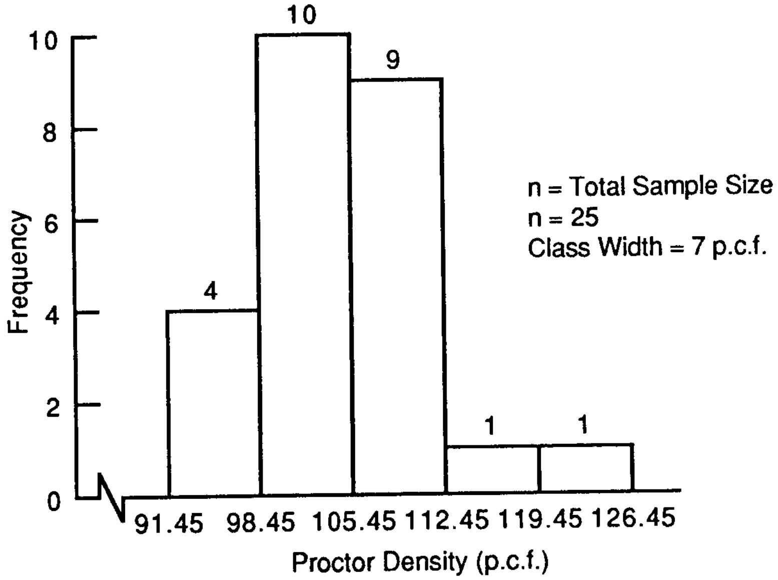


Figure 1.1. Frequency Histogram for Proctor Density Data [after Ref. 1.2]

SECTION 1.0 REFERENCES

1.1 Western Electric Co., Inc., Statistical Quality Control Handbook, Western Electric Co., Inc., AT&T Technologies, Indianapolis, Indiana, May 1985 (11th printing).

1.2 Willenbrock, Jack H., "A Manual for Statistical Quality Control of Highway Construction - Volume I," Federal Highway Administration, National Highway Institute, Washington, D.C., January 1976.

1.3 ASTM, Standards on Precision and Bias for Various Applications, American Society for Testing and Materials, Philadelphia, PA, 1985.

SECTION 2.0
THE NORMAL DISTRIBUTION

1. INTRODUCTION

The normal distribution is a data distribution that can be used to describe many types of measurements in engineering. Basically, a normal distribution is a bell shaped curve. Figure 2.1 illustrates a bell curve, superimposed over a histogram of PCC compressive strength data. Such a distribution is very convenient to use because it is characterized by the mean (μ or $\bar{x}$) and standard deviation (σ or s). As Figure 2.1 shows, most of the strength measurements cluster around the mean ($\bar{x}$ = 4,824 psi), while fewer measurements are near the lowest (3,875 psi) and highest (5,975 psi) strength values.

Since the normal distribution can be defined by the mean and standard deviation, a set of measurements with equal means but differing standard deviations can be illustrated, as shown in Figure 2.2 (subgrade density measurements). In this case, the population mean is equal to 105 pcf and three different normal distributions are shown for population standard deviations of 5, 7 and 9 pcf. If you were a field inspector, which of these three distributions would you prefer?

Figure 2.2 helps to provide an answer to the above question. If the total area under the bell shaped curve is equal to 1.0, then the portion of density tests between 90 and 96 pcf is about 3.5, 8.0 and 11.0 percent for the three standard deviations of 5, 7 and 9 pct, respectively. This suggests that the distribution with σ = 5 pcf is preferable. You will see how to determine these areas later in this section.

Willenbrock [2.1] (slightly modified) helps to explain Figures 2.2 and 2.3

> The theoretical NORMAL DISTRIBUTION extends out infinitely in both directions from a mean of 105 pcf and never quite reaches the horizontal axis...A NORMAL DISTRIBUTION has a total area under the curve of 1.00 (i.e. 100 percent of the data values are represented by the distribution). Since it extends from $-\infty$ to $+\infty$ (minus infinity to plus infinity), it encompasses all of the density results that can occur. The area under the curve within these two limits must therefore be equal to unity (i.e. 1.000 or 100 percent). For all practical purposes, however, most of the data values (actually 99.73 percent) occur between 3 σ limits below 105 pcf and 3 σ limits above 105 pcf.
>
> If the area of each NORMAL DISTRIBUTION is the same (i.e. an area equal to unity, 1.0000), then the distribution shown in Figure 2.2 that has the largest spread (i.e., the largest standard deviation, which occurs in case (c) where σ = 9 pcf) should have the shortest overall height at the average value. Normal distribution (a), on the other hand, has the smallest spread (i.e., σ = 5 pcf), so its horizontal spread is smaller than its distribution (c). Its vertical spread must therefore be larger than its distribution (c).

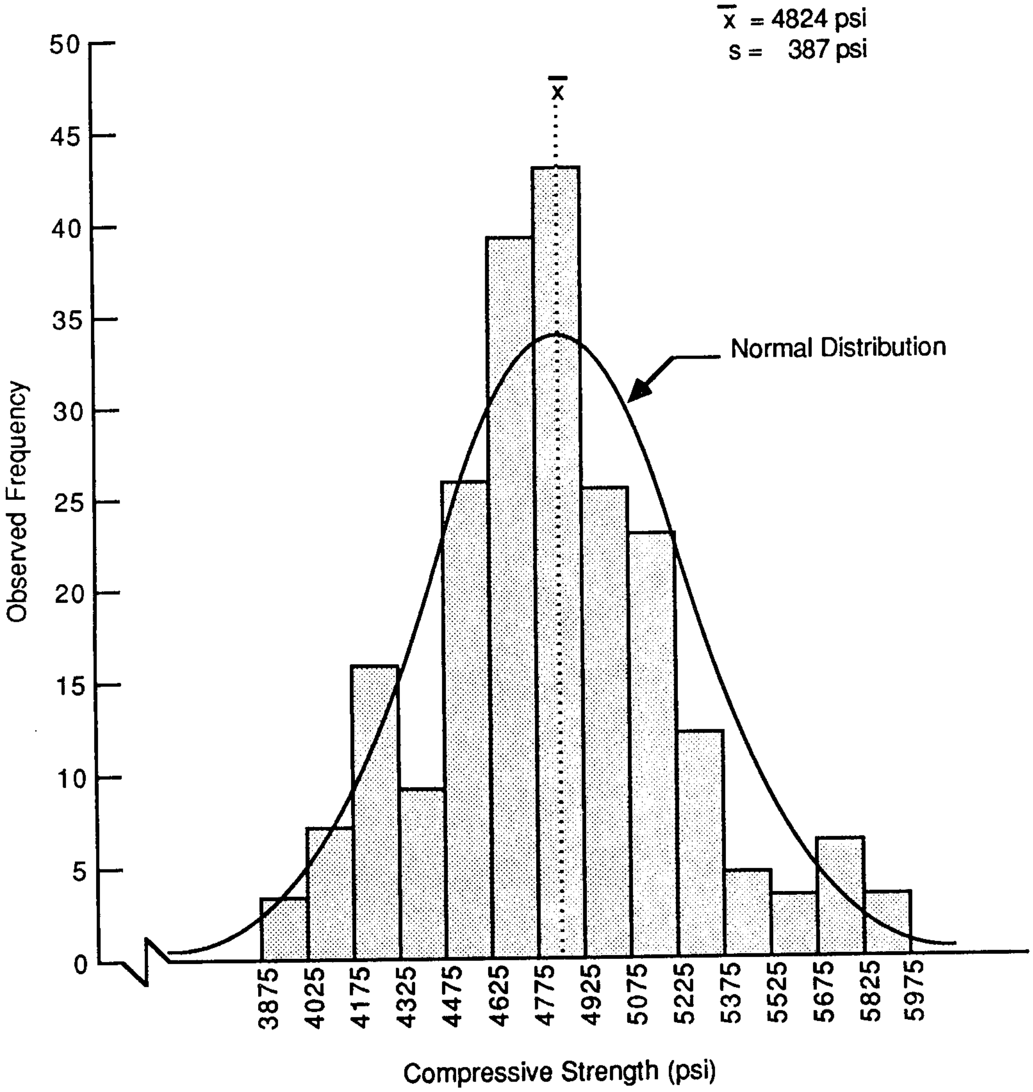


Figure 2.1. Histogram and the Normal Distribution for PCC Compressive Strength Data [after Ref. 2.1]

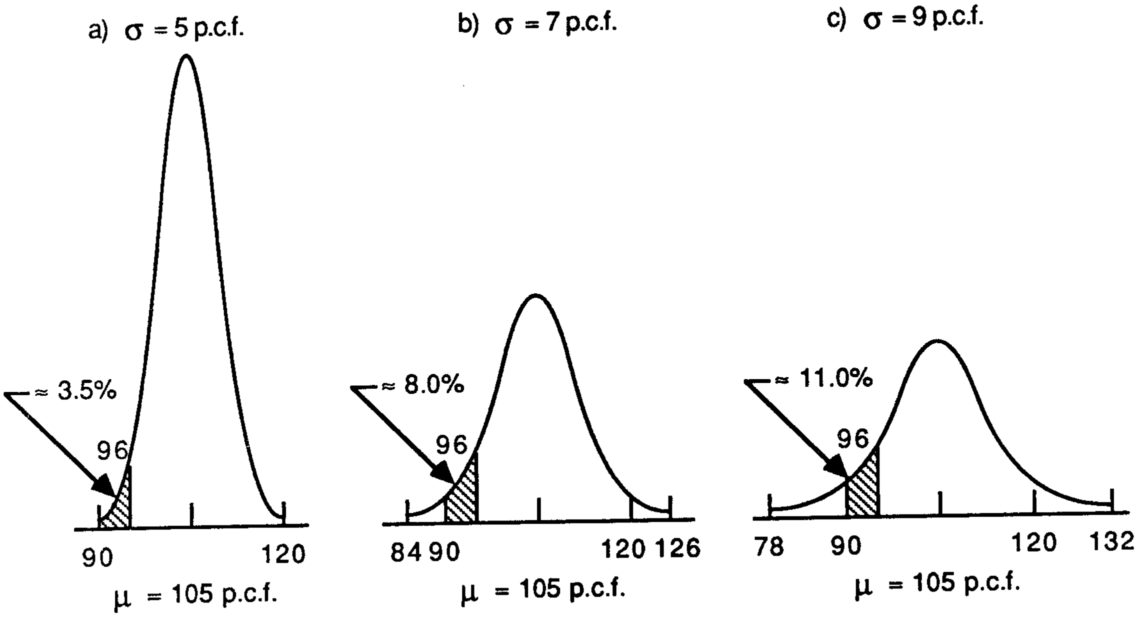


Figure 2.2. Three Normal Distributions for Procter Densities
(Same Means, Different Standard Deviations)
[after Ref. 2.1]

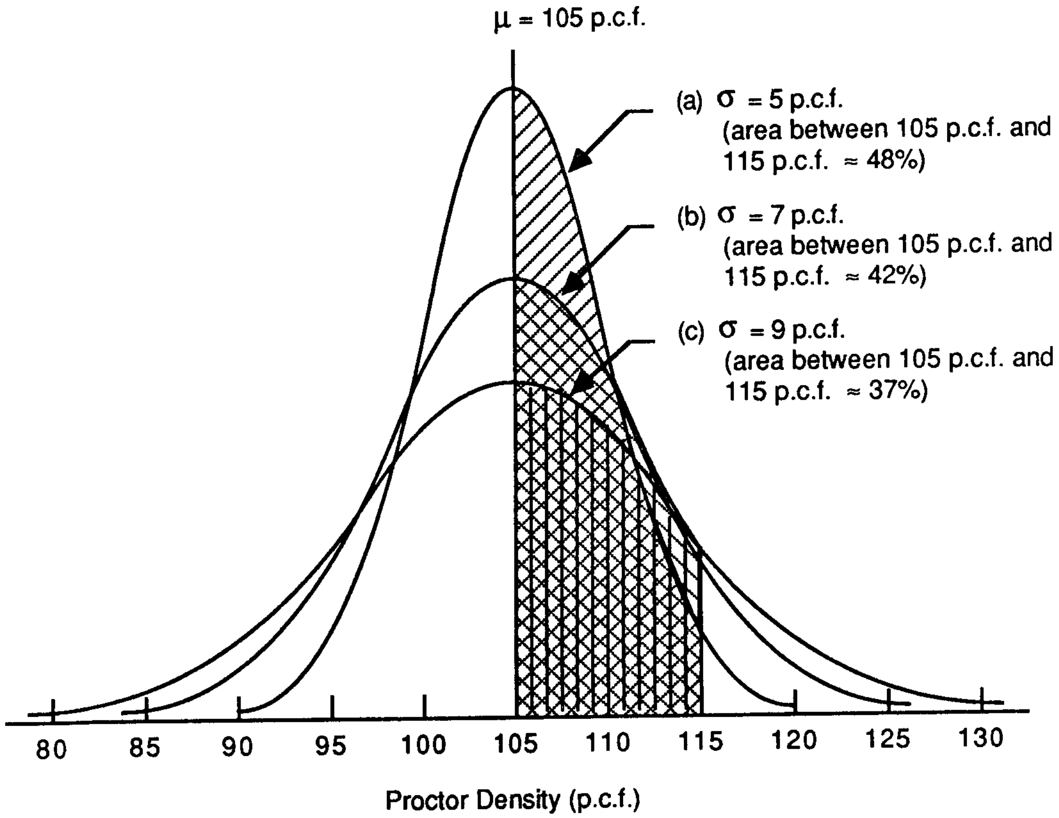


Figure 2.3. Superimposed Normal Distributions [after Ref. 2.1]

A far more important result than those mentioned above is also related to the fact that the area under the curve is equal to 100 percent. Because of this, it can be stated that THE PROBABILITY OF FINDING A DATA VALUE BETWEEN 105 pcf and 115 pcf IS EQUAL TO THE AREA UNDER THE NORMAL DISTRIBUTION BETWEEN 105 pcf AND 115 pcf. For distribution (a), as shown in Figure 2.3 (i.e., $\sigma = 5$ pcf), the area between 105 pcf and 115 pcf represents about 48 percent of the area under the entire distribution.

Figure 2.4 shows two normal distributions with equal population standard deviations ($\sigma = 5$ pcf) but unequal population means ($\mu = 85$ and 105 pcf).

2. NORMAL DISTRIBUTION EQUATION

The height of a normal distribution (y) can be defined by its corresponding value of x (refer to Figure 2.5) by the following equation [after Ref. 2.1]:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x - \mu)^2/2\sigma^2} \qquad \text{Eq. 2.1}$$

where

y = vertical height of a point on the normal distribution,
x = distance along the horizontal axis,
σ = standard deviation of the data distribution,
μ = mean of the data distribution,
e = constant = 2.71828 ... ,
π = constant = 3.14159 ....

To illustrate how Equation 2.1 can be used to determine area under a normal distribution, refer back to the Procter density data (Figure 2.3). Calculate the area under the normal curve between 105 and 115 pcf for a standard deviation of 5 pcf (this is shown in Figure 2.3 to be 0.48 or 48 percent of the total area under the curve). These calculations are

$$y = \frac{1}{(5)\sqrt{2\pi}} e^{-(105 - 105)^2/2(5)^2} = 0.079788$$

and

$$y = \frac{1}{(5)\sqrt{2\pi}} e^{-(115 - 105)^2/2(5)^2} = 0.010798$$

The approximate area under the curve is about 0.45 (or 45 percent), which is close to the "theoretical" value of 48 percent (refer to sketch in Figure 2.6). The significance of this value is that the probability of a density measurement falling within the range of 105 to 115 pcf is about 0.48 (let's use the "theoretical" value).

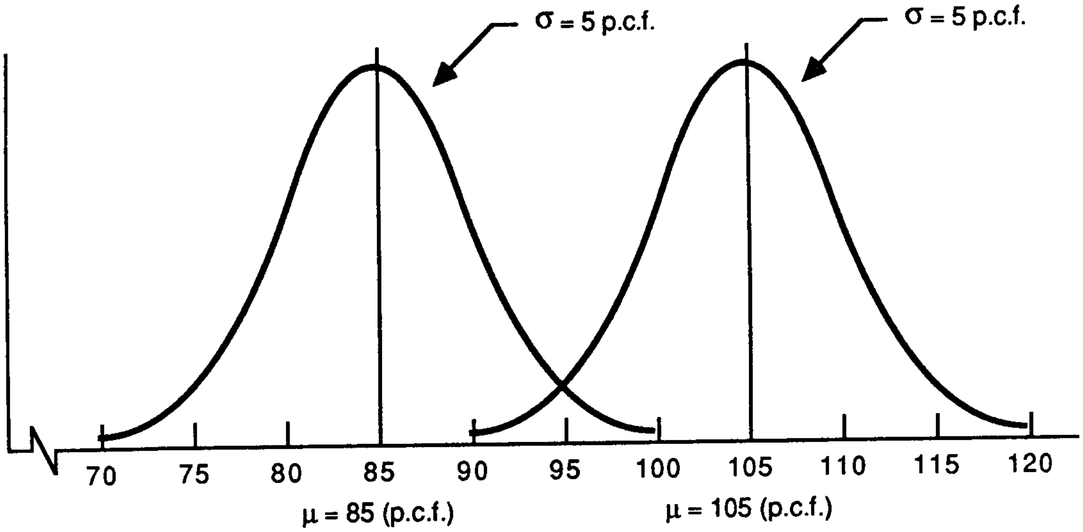


Figure 2.4. Normal Distribution (Different Means, Same Standard Deviations) [after Ref. 2.1]

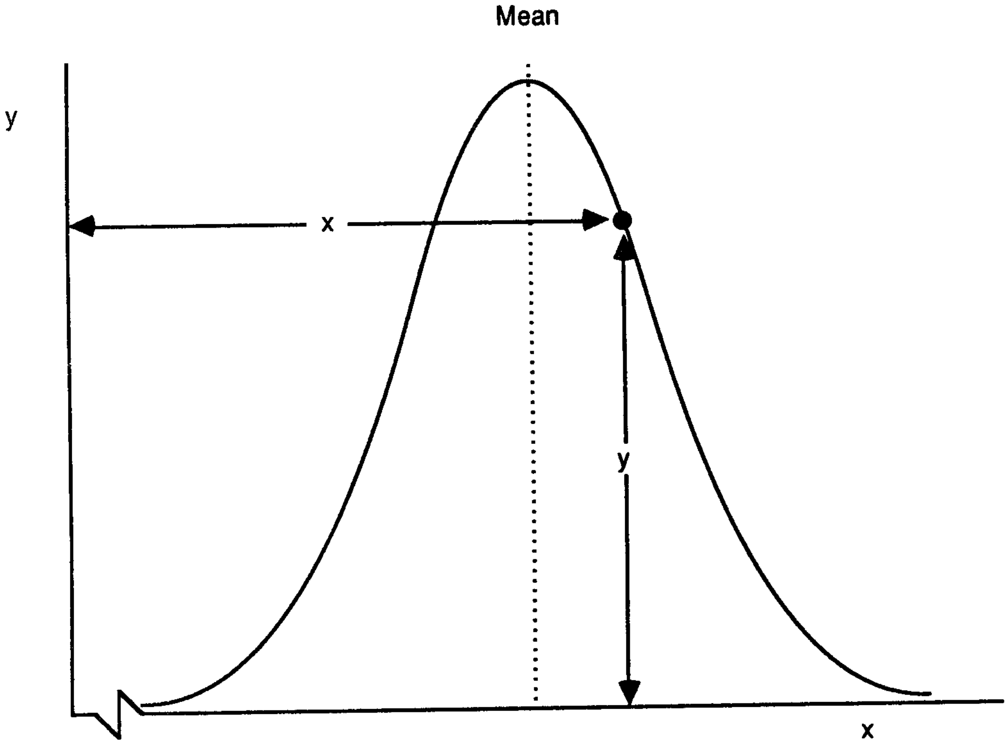


Figure 2.5. Relationship of "y" and "x" Values in the Normal Distribution

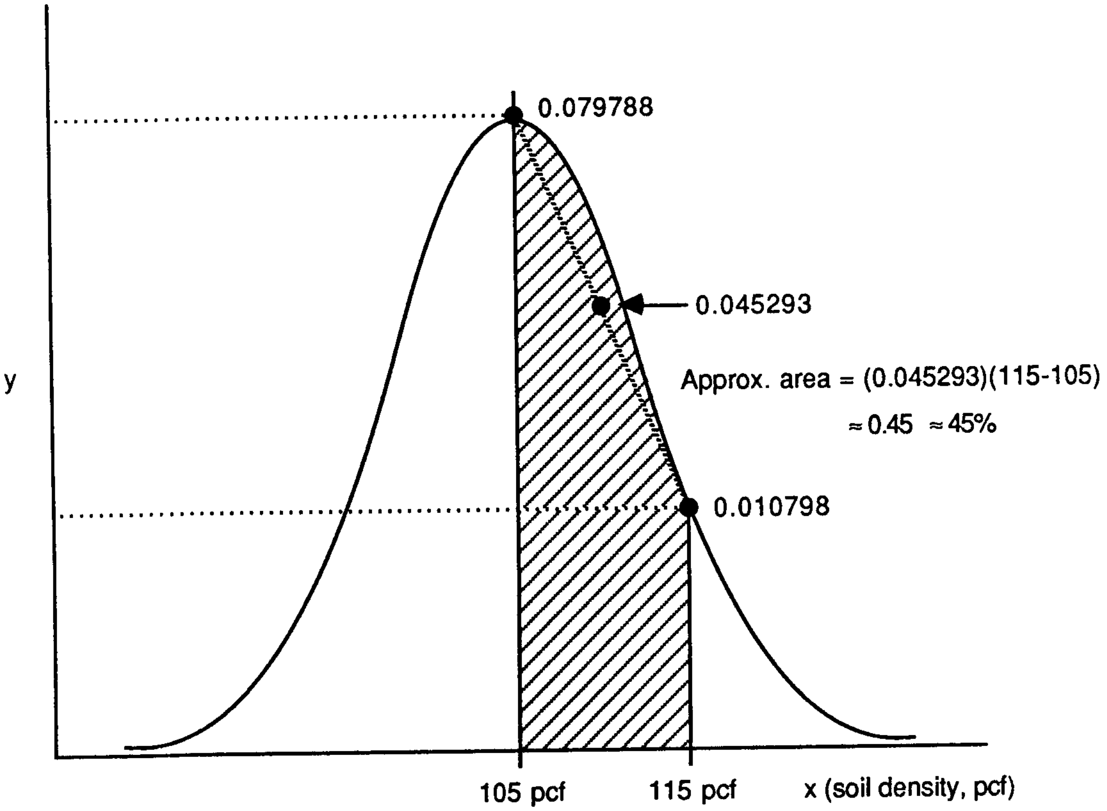


Figure 2.6. Determination of Approximate Area
Under the Normal Distribution

To determine such probabilities in this manner is tedious and time consuming. There is an easier way to determine these probabilities than computing and tabulating y's for various μ's and σ's. To do this, you must convert the normal distribution to a standard normal distribution and define a variable "z," which is:

$$z = \frac{\text{deviation from mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

If you substitute z into Equation 2.1, then following relationship results:

$$y_z = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad \text{Eq. 2.2}$$

where

y_z = vertical height on the standard normal distribution, and
z = as previously defined.

Refer to Figure 2.7, which illustrates this important transformation. Thus, you can see that the probability of having a density test between 105 and 115 pcf is about 47.7 percent (or 34.1 + 13.6 percent). Fortunately, the "z-statistic" has been published in tables to allow for easy computation. Such a table is shown as Table 2.1. You can see that

mean ± 1 standard deviations ≅ 68.2% of area

mean ± 2 standard deviations ≅ 95.4% of area

mean ± 3 standard deviations ≅ 99.8% of area

Recall that all of the area under a normal distribution is 100%.

2.1 Example (more Procter density data)

For normally distributed Procter density data with μ = 105 pcf and σ = 5 pcf, what is the probability the density will be greater than 92 pcf?

First calculate z.

$$z = \frac{92 - 105}{5} = -2.6$$

Now, with z = -2.6, use a cumulative standard normal distribution table (any statistics book will have one, or use Table 2.1) to obtain the appropriate area under the curve that equals 0.0047.

Thus, P(density ≥ 92 pcf) = 1.0000 - 0.0047
= 0.9953 or 99.53 percent

2.2 Example (portland cement concrete strengths)

If a distribution of PCC strength data is $\mu = 5000$ psi and $\sigma = 500$ psi, answer the following questions:

(a) What is the probability that the strength of the PCC will be less than 6,000 psi?

(b) What is the probability it will be less than 5,000 psi?

(c) What is the probability it will be more than 4,000 psi?

(d) What is the probability it will be less than 4,000 psi?

Refer to Figure 2.8 for the results.

2.3 Example (PCC contractor trucks)

The contractor claims that his batch plant can produce PCC mix with

$\mu = 4{,}824$ psi
$\sigma = 387$ psi

Assume that a very unrealistic job specification states that an acceptable PCC must have a compressive strength no lower than 4,700 psi and no higher than 5,000 psi (after seven days of cure).

Question: If the contractor sends 50 truckloads of this mix to the job site, how many of the trucks should be rejected if you know the real, potential compressive strength of each truckload (of course you cannot do this but what the heck)?

Solution: Using z-statistic tables, you find that the total area under the standard normal distribution between 4,700 and 5,000 psi is about 0.30 (recall that the maximum is 1.0000 under the curve).

Thus, approximately 30 percent of the population will be between 4,700 and 5,000 psi. Thus, 1.0000 - 0.30 = 0.70 or about 70 percent of the 50 trucks (i.e., about 35 trucks) should be rejected.

Try to match this solution by using Table 2.1. Hint: start by computing Z_{4700} and Z_{5000}, then use Table 2.1.

2.4 The "t-distribution"

The "t-distribution" is used to test sample means when the population variance (or population standard deviation) is not known (which is usually the case for most of the data you deal with). The "t-statistic" is quite similar to the "z-statistic" but also includes consideration of the sample size (n).

where

$$z = \frac{x - \mu}{\sigma} \qquad \text{Eq. 2.3}$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad \text{Eq. 2.4}$$

This concept will be illustrated in more detail in the next section.

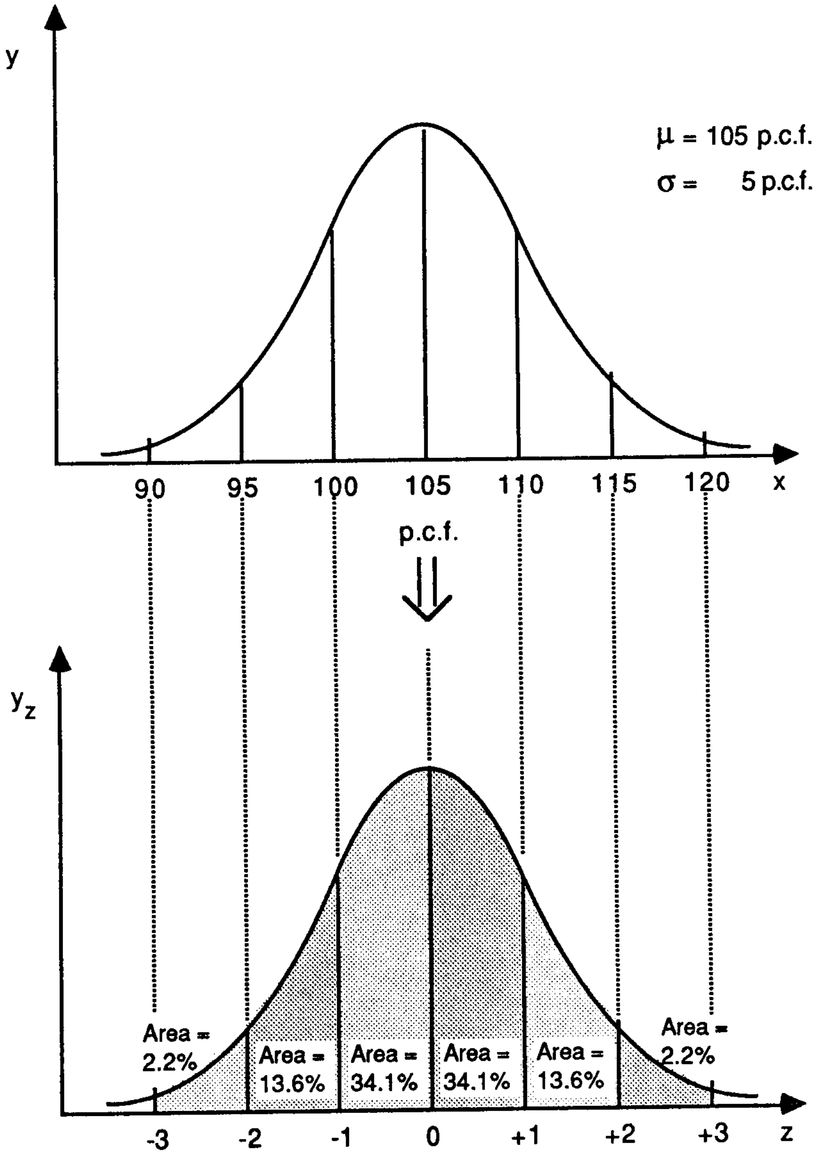
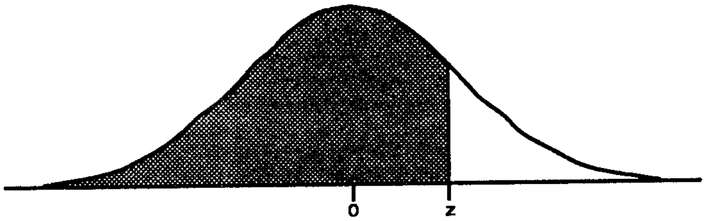


Figure 2.7. The Proctor Density Distributions (Normal and Standard Normal)

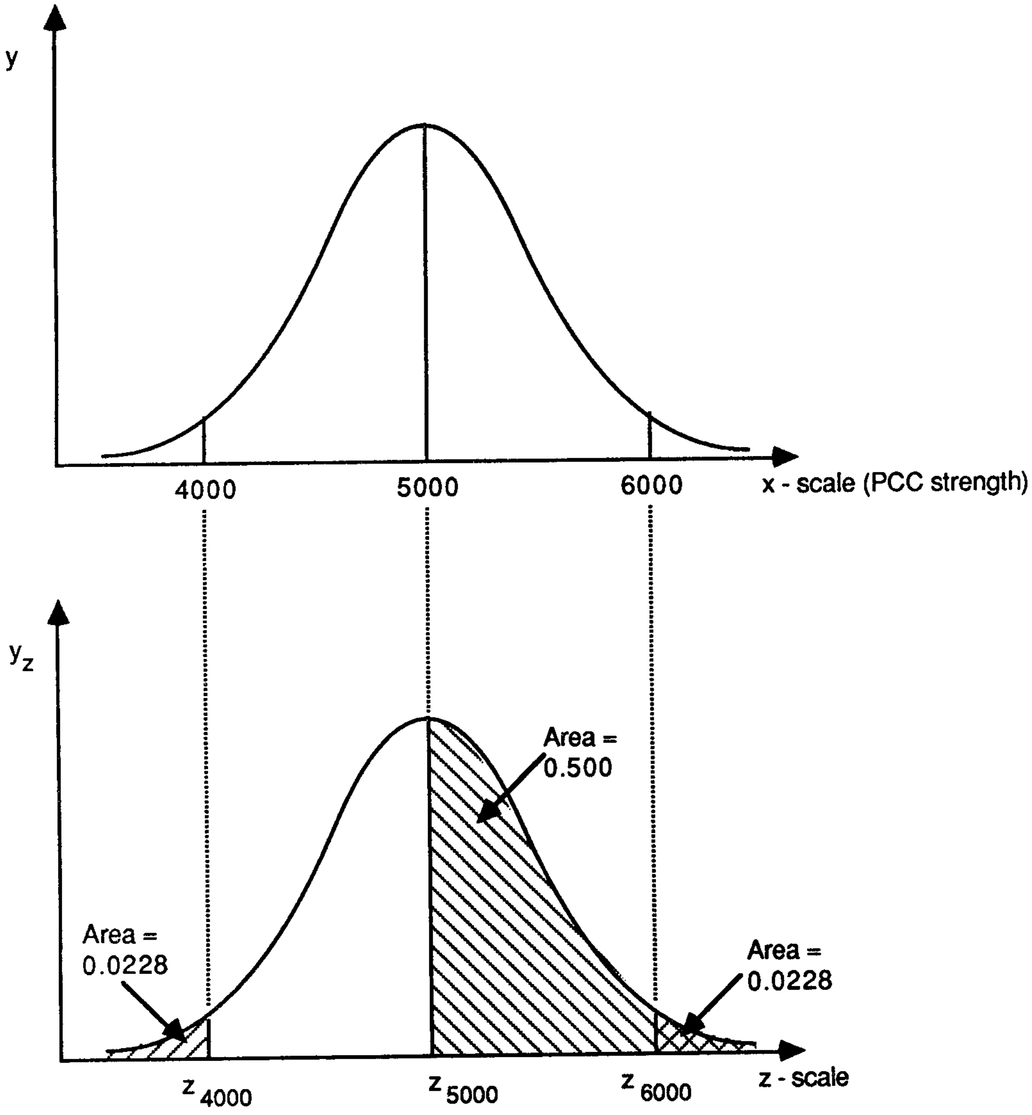
Table 2.1 Normal Distribution Table [from Ref. 2.2]

Normal Distribution



Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-4.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.8	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.7	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483

Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7793	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.7	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.8	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000



From Table 2.1

$z_{4000} = \frac{4000 - 5000}{500} = -2$ (area less than Z = -2 equals 0.0228)

$z_{5000} = \frac{5000 - 5000}{500} = 0$ (area less than Z = 0 equals 0.5000)

$z_{6000} = \frac{6000 - 5000}{500} = +2$ (area less than Z = +2 equals 0.9772)

(a) P (strength ≤ 6000 psi) = 1.000 - 0.0228 = 0.9772 or 97.72%

(b) P (strength ≤ 5000 psi) = 1.000 - 0.5000 = 0.5000 or 50.00%

(c) P (strength ≥ 4000 psi) = 1.000 - 0.0228 = 0.9772 or 97.72%

(d) P (strength ≤ 4000 psi) = 0.0228 or 2.28%

Figure 2.8. PCC Strength Probabilities

SECTION 2.0 REFERENCES

2.1 Willenbrock, Jack H., "A Manual for Statistical Quality Control of Highway Construction - Volume I," Federal Highway Administration, National Highway Institute, Washington, D.C., January 1976.

2.2 Ulberg, Cy, "Highway Research Manual," Draft Final Report, Research Project Y3399, Task 29, Washington State Department of Transportation, Olympia, WA, August 1987.

SECTION 3.0
TESTS OF HYPOTHESES

1. INTRODUCTION

Hypothesis testing is a way in which statistical methods can be used to help in the decision making process. Such testing considers the mean and standard deviation of a group of data, the confidence level (a probability statement) and something about the population being sampled. Hypothesis testing is extremely helpful in performing multiple regression analysis and hence it is important for you to understand the basics.

2. HYPOTHESES

Webster's Seventh New Collegiate Dictionary defines hypothesis as "...a tentative assumption made in order to draw out and test its logical or empirical consequences....an assumption or concession made for the sake of argument..." You can begin to see the problem in explaining hypothesis testing.

There are always two hypotheses for any statistical test [3.1]. These hypotheses are

H_0 = null hypothesis (most important)

H_1 = alternative hypothesis

What is about to be presented is one of the fundamental problems in statistics which is the use of "double negative" statements. Any hypothesis must be tested statistically to be rejected or not rejected (this is a statistical way of accepting something).

The hypotheses (H_0 or H_1) can result in two types of errors if the wrong one is selected, as shown in Table 3.1. The probability of the Type I and II errors is very important, since it determines how carefully you must distinguish between true and false hypotheses.

Table 3.1. Types of Hypothesis Errors

	"The Actual Decision"	
"The Truth"	Reject H_0	Accept H_0
H_0 true	Type I Error (α)	Correct!
H_0 false	Correct!	Type II Error (β)

(This is an area in which statistical "games" can be played, so you need to be very careful). These probabilities are [after Ref. 3.1]

P(Type I error) = α

P(Type II error) = β

The general form for calculating the z-statistic for hypothesis testing is

$$z_{calc} = \frac{\text{(sample mean) - (hypothesized value)}}{\text{standard error}}$$

where

sample mean = $\bar{x}$

hypothesized value = μ (sometimes assumed to = 0 in regression hypothesis testing)

standard error = $\frac{\sigma}{\sqrt{n}}$ = standard deviation of means of random samples of size n from a "parent" population with standard deviation σ. Standard error is sometimes designated $\sigma_{\bar{x}}$.

sample size = n

The same general form applies to the t-statistic for hypothesis testing when none of the population statistics (μ, σ) are known:

$$t_{calc} = \frac{\text{(sample mean) - (hypothesized value)}}{\text{standard error}}$$

where

sample mean = $\bar{x}$

hypothesized value = μ (again, sometimes an <u>assumed</u> or <u>stated</u> value)

standard error = $\frac{s}{\sqrt{n}} = s_{\bar{x}}$

2.1 Example (PCC contractor)

For this example, use the data shown in Figure 2.1. This contractor states that his batch plant has produced a mix in the past of

μ = 4,824 psi

σ = 387 psi

(Since these are population statistics, you can assume that these data were collected over a long period of time)

On the job you take six samples (cylinders) with the result that $\bar{x}$ = 4,549 psi.

Question: Is the contractor correct?

Solution: Assume that the data are normally distributed and use hypothesis testing.

H_0: $\mu = 4{,}824$ psi

H_1: $\mu < 4{,}824$ psi

$$z_{calc} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4549 - 4824}{387/\sqrt{6}} = -1.74$$

$z_{critical} = -1.65$ (for Type I error (or α) = 5%) (Refer to Table 2.1 to determine z.)

since $z_{calc} > z_{critical}$, reject H_0

Thus for your job, you must judge the contractor's claim to be incorrect. Refer to Figure 3.1 for an illustration of this process.

2.2 Example (PCC contractor again) [after Ref 3.2]

The PCC mix contractor claims the following:

PCC mix ≥ 4,000 psi (28-day compressive strength)

You take a random sample of five cylinders and cure them for 28 days (n = 5).

The results:

$\bar{x}$ = 3,740 psi
s = 390 psi
n = 5

Question: If you are willing to accept a 5 percent chance of a Type I error (i.e., rejecting a true H_0), should you believe the contractor?

Solution

H_0: $\mu \geq 4{,}000$ psi (null hypothesis)

H_1: $\mu < 4{,}000$ psi (alternative hypothesis)

$$t_{calc} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{3740 - 4000}{390/\sqrt{5}} = -1.48$$

$t_{critical}$ (@ 5%) < -2.13 (one-tail α = 5% with $\nu = n - 1 = 4$ degrees of freedom)

Therefore, you accept H_0, since

$$t_{calc} = -1.48 < -2.13 = t_{critical}$$

You have no "statistical" reason to doubt the contractor's claim. Refer to Figure 3.2(a) which further illustrates this example. However, note that if the Type I error (rejecting a true H_0) were reduced to a 1 percent chance, then

$$t_{critical}\ (@1\%) > -3.747 \text{ (one-tail } \propto = 1\% \text{ with 4 degrees of freedom)}$$

Thus, you are even more unwilling to accept the alternate hypothesis (H_1) that the contractor's claim was incorrect. Note that the Type I error protects against rejecting a true null hypothesis. In other words, you can select a low Type I error level so that it is difficult to reject the null hypothesis. However, as the Type I error level decreases, the Type II error level increases (not rejecting a false H_o). It is not easy to illustrate the calculation of the Type II error (β), but this example is a good case since a Type I error level of about 11 percent would be needed to reject the null hypothesis.

Often the Type I error is termed the "seller's risk" and the Type II error the "buyer's risk." For the example, the lower the Type I error the lower the risk of the contractor. Correspondingly, the Type II error increases the risk of the DOT accepting PCC of lower than specified quality (again not rejecting a false H_o). Needless to say, a balance between Type I and Type II errors is needed (but not necessarily the same number or value because one error type may be more important than another) in developing statistically based materials "acceptance plans."

At least one illustration of ignoring the Type II error (β) is appropriate The PCC mix contractor example will be used.

Assume that the sample results,

$$\bar{x} = 3{,}740 \text{ psi}$$
$$s = 390 \text{ psi}$$

actually represent the true population (i.e., $\mu = 3{,}740$ psi instead of $\mu = 4{,}000$ psi). First, calculate the value of $\bar{x}$, which corresponds to $t = -2.13$ ($\propto = 5\%$). This value is the one that separated the rejection and acceptance region for H_O ($\mu \geq 4{,}000$ psi). (Refer to Figures 3.2(a) and (b).)

$$\frac{\bar{x} - 4000}{\frac{390}{\sqrt{5}}} = -2.13$$

$$\therefore\ \bar{x} = 3{,}628 \text{ psi}$$

The value of the Type II error (β) is the area under the curve (or distribution) with $\mu = 3{,}740$ psi and $\sigma = 390$ psi but within the acceptance region of the original H_0 distribution ($\mu = 4{,}000$ psi). This is illustrated in Figure 3.2(b).

$$t = \frac{3{,}628 - 3{,}740}{\frac{390}{\sqrt{5}}} = -0.64$$

(Use a t-table, available in most statistics books.)

$$\therefore \beta = P(t \geq -0.64) = 0.78 \text{ (or 78\%)}$$

Therefore, there is a 78 percent chance of accepting a false H_0 ($\mu = 4{,}000$) if the true population mean (μ) = 3,740 psi. You can see that the β will change as the correct population mean changes. Clearly, this level of β is quite high. Thus, the DOT's risk (the "buyer's risk") is too high. A balance in setting $\propto$, β and sample size levels is very important in the proper use of these kinds of statistical tests. This leads to another important area of statistics, generally called "acceptance testing" and "operating characteristic curves," which is not appropriate for discussion in these notes.

2.3 Example (Friction Number data - paired t-test)

The friction number data are for SR-82. The milepost locations indicate that the friction testing was performed on portland cement concrete surfaces only. The friction tests were essentially obtained every one-half mile in all available lanes. The friction number data are shown in Table 3.2. Lane 1 is the "outside" or "curb" lane in all cases.

Table 3.3 presents a few basic statistical measures of the friction number data. This includes mean, standard deviation, the number of data points and the coefficient of variation. The coefficient of variation is a dimensionless number that is the standard deviation divided by the mean multiplied by 100 (to convert to a percent). Stated another way, this value is used to express the standard deviation as a percentage of the mean.

A review of the information in Table 3.3 shows that, in general, the mean (or average) friction number is highest for the "inside" or "median" lanes. This is not surprising since the traffic in the inside lanes is generally lower than that in the "outside" lanes. These differences are about 7 friction numbers. Further, the coefficient of variation is generally higher for the inside lanes. This might suggest that studded tire wear (higher wear exposure in the outside lanes) results in more uniform but lower friction numbers.

Often you may have some idea of what outcome to expect from a comparison of test data (in this case a comparison of friction number data between lanes on SR-82). You may then pose a tentative assumption (or hypothesis) in order to test the outcome. This "test" is done with sample data (friction numbers in this case) and the "t-statistic." For this analysis, the proposed hypothesis (or "null" hypothesis) is that there are no statistically significant differences among the differences between the mean friction numbers of the lanes. The mean friction number differences were based on "paired" observations, i.e., at a specific milepost the difference between any two friction numbers from two separate traffic lanes was calculated. The probability level chosen for this comparison was 95 percent. This implies that there is only a 5 percent chance that the true null hypothesis will be rejected. If the null hypothesis was accepted, then the last column of Table 3.4 would show "no significant difference." This would suggest that there would be only a 5 percent chance of the conclusion being wrong, and one could conclude (for the specific test data used) that there was no real difference in friction numbers for the two lanes compared. On the other hand, if the last column in Table 3.4 showed "significant difference," then the null hypothesis would be rejected. This would result in the conclusion that there was real difference in friction numbers for the specific lanes and route being compared.

A review of the conclusions drawn in Table 3.4 (and the formulas shown in Table 3.5 and the calculations in Table 3.6) suggest that the inside lane has "significantly" higher friction numbers than the other lanes. This indicates that the lower exposure to vehicle traffic (and studded tires) results in higher pavement friction resistance (not a surprising finding). A review of the WSDOT Road Log reveals that this section of SR-82 is 15 years old.

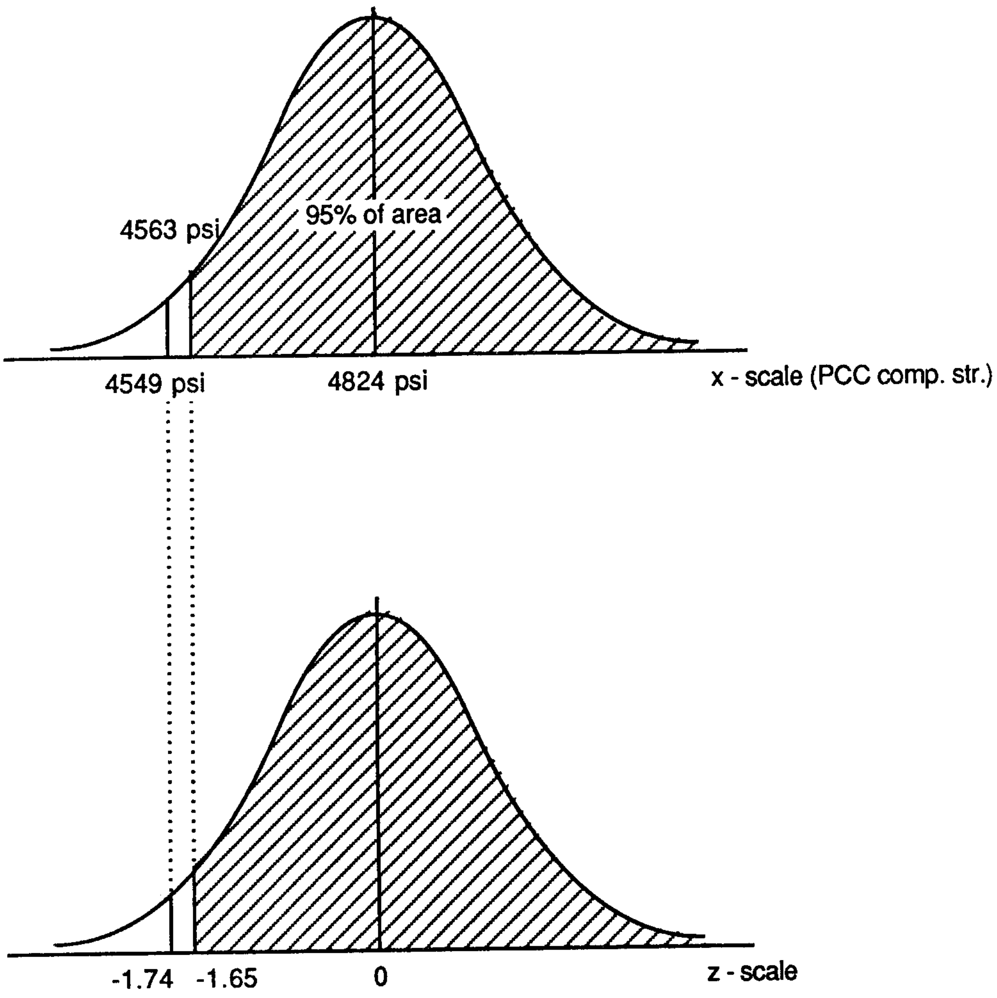


Figure 3.1. Hypothesis Testing with Population Mean and Standard Deviation Known

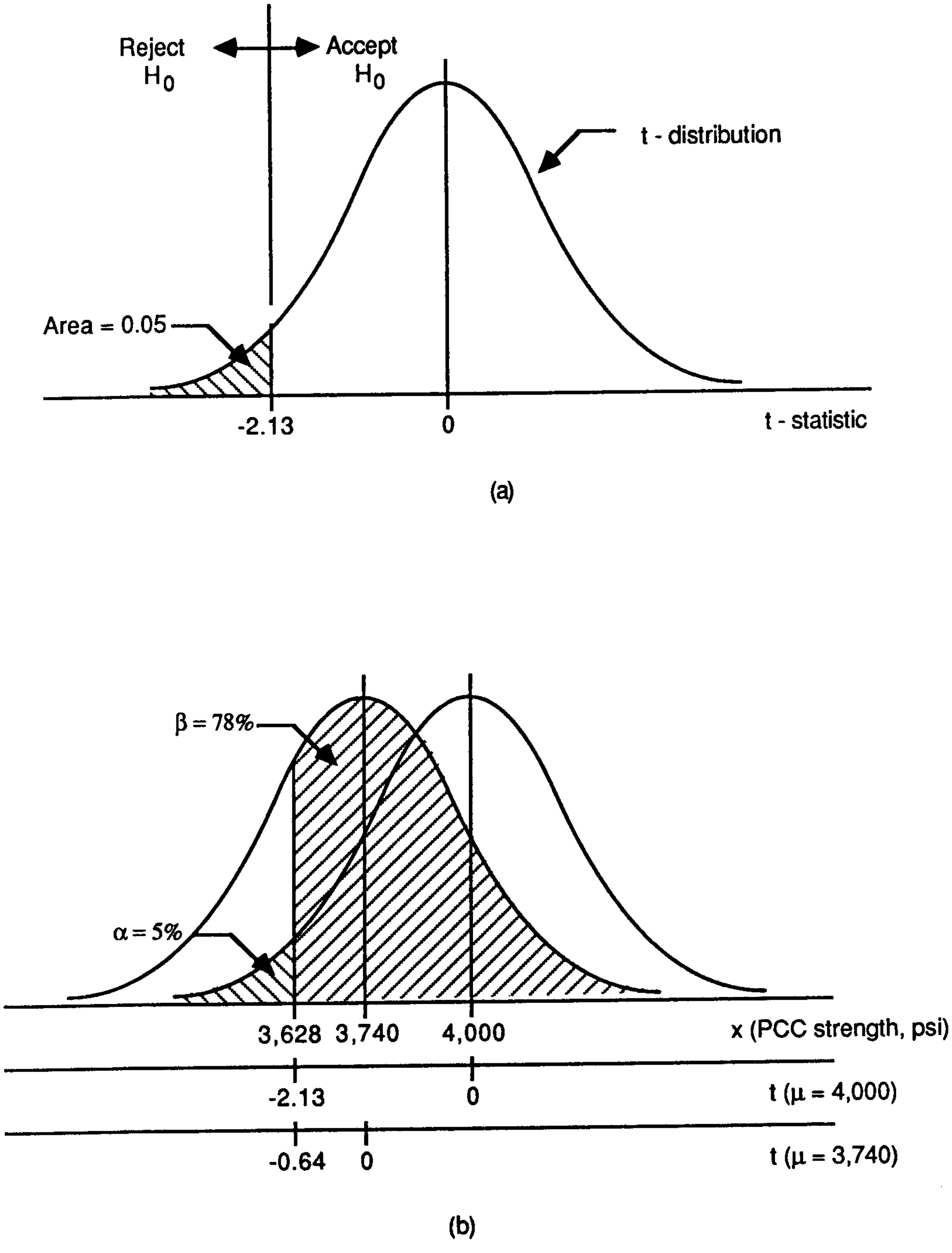


Figure 3.2. t - test Example (PCC Contractor)

Table 3.2. Friction Number Data, SR 82 (MP 4.0 - 14.0)

Mileposts	Friction Numbers			
	Eastbound		Westbound	
	Lane 1	Lane 2	Lane 1	Lane 2
4.0	40.4	50.6	38.5	46.4
4.5	36.2	50.5	41.0	47.2
5.0	38.3	44.3	40.6	47.7
5.5	40.1	46.0	38.7	46.8
6.0	41.3	47.4	39.6	44.4
6.5	42.6	45.9	40.4	47.8
7.0	35.4	43.9	39.5	47.7
7.5	37.9	48.1	41.1	43.3
8.0	41.0	48.6	43.3	55.3
8.5	41.6	45.5	41.7	49.2
9.0	41.0	44.3	38.8	51.7
9.5	40.0	48.5	40.8	46.1
10.0	39.9	50.3	38.6	46.5
10.5	42.9	47.3	43.6	49.4
11.0	41.8	46.2	41.0	50.7
11.5	43.8	46.5	41.7	49.5
12.0	40.9	55.9	42.1	47.0
12.5	41.5	52.9	41.5	53.4
13.0	43.4	50.3	40.8	52.5
13.5	46.3	52.6	40.9	46.4
14.0	40.4	54.0	41.4	46.5

Note: Lane 1: "Outside" or "curb" lane
Lane 2: "Inside" or "median" lane

Table 3.3. Basic Statistics for Friction Number Data, SR 82

Route	Mileposts	Parameter	Lane Direction and Number			
			WB		EB	
			1	2	1	2
SR-82	4.0 - 14.0	Mean ($\bar{x}$) Std. Dev. (s) Data Points (n) Coeff. of Var.	40.7 1.4 21 3.4%	48.4 3.0 21 6.2%	40.8 2.5 21 6.1%	48.6 3.4 21 7.0%

Table 3.4. Results of Hypothesis Testing for Friction Number Data, SR 82

Route	Mileposts	Lane Comparison (Dir: Lane Nos.)	t - statistic		Conclusion (c) (Hypothesis)
			Calculated (a)	Table (b) ($\alpha = 0.05$)	
SR-82	4.0 - 14.0	WB: 1-2 EB: 1-2	-12.587 -9.540	-2.086 -2.086	Significant Difference Significant Difference
	Lane 1 is "Outside" lane of two lanes in one direction. Lane 2 is "Inside" lane of two lanes in one direction.				

Notes:

(a) t-statistic calculated from paired friction number observations for lane directions and numbers shown.

(b) t-statistic for a Type I error of 5% ($\alpha = 0.05$) for n-1 degrees of freedom (number of data points minus one).

(c) Conclusion based on the initial hypothesis that there are no significant differences in friction numbers for the cases shown.

Table 3.5. Formulas Used in Friction Number Statistical Analysis [after Ref. 3.1]

1. Statistical tests reported are "means tests for paired samples."

2. Null hypothesis is $H_o: \delta = \delta_0$

$H_1: \delta \neq \delta_0$

where: $\delta =$ mean difference between paired friction numbers

$\delta_0 =$ 0 (assumes there is no difference between pairs on a population basis)

3. t-statistic

$$t = \frac{(\delta - \delta_0)}{\frac{s_d}{\sqrt{n}}} = \frac{(\bar{d} - 0)}{\frac{s_d}{\sqrt{n}}}$$

where: $\bar{d} = \frac{\Sigma d_i}{n}$

$n =$ sample size

$\Sigma d_i = \Sigma(x_{i1} - x_{i2})$

$x_{i1} =$ individual friction number from Lane 1 (for example)

$x_{i2} =$ individual friction number from Lane 2 (for example)

$$s_d = \left(\frac{\Sigma d_i^2}{n-1} - \frac{n}{n-1} \bar{d}^2\right)^{1/2}$$

Reference 3.1:
Lelank Blank, Statistical Procedures for Engineering, Management, and Science.
McGraw – Hill, 1980, p.384, 630.

Table 3.6. Friction Number Example – Calculation of t-statistics

1. Summation of differences between Lane 1 and Lane 2:

$\Sigma d_i = -162.9$ (eastbound) $\qquad\qquad$ $\Sigma d_i = -159.9$ (westbound)

To illustrate: $\Sigma d_i = (40.4 - 50.6) + (36.2 - 50.5) + \ldots + (40.4 - 54.0) = -162.9$ (for eastbound)

2. Summation of squared differences between Lane 1 and Lane 2:

$\Sigma d_i^2 = 1541.3$ (eastbound) $\qquad\qquad$ $\Sigma d_i^{\,2} = 1371.2$ (westbound)

To illustrate: $\Sigma d_i^2 = (40.4 - 50.6)^2 + (36.2 - 50.5)^2 + \ldots + (40.4 - 54.0)^2 = 1541.3$ (for eastbound)

3. Calculate mean of friction number differences:

$$\bar{d} = \frac{\Sigma d_i}{n} = -\frac{162.9}{21} = -7.757 \quad \text{(eastbound)}$$

$$\bar{d} = \frac{\Sigma d_i}{n} = -\frac{159.9}{21} = -7.614 \quad \text{(westbound)}$$

4. Calculate standard deviation of the differences:

$$s_d = \left[\frac{\Sigma d_i^2}{n-1} - \frac{n}{n-1}\,\bar{d}^2\right]^{1/2}$$

$$s_d = \left[\frac{1541.3}{20} - \left(\frac{21}{20}\right)(-7.757)^2\right]^{1/2} = 3.726 \text{ (eastbound)}$$

$$s_d = \left[\frac{1371.2}{20} - \left(\frac{21}{20}\right)(-7.614)^2\right]^{1/2} = 2.772 \text{ (westbound)}$$

5. Calculate t-statistic:

$$t_{calc} = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}$$

$$t_{calc} = \frac{(-7.757) - (0)}{\frac{3.726}{\sqrt{21}}} = -9.540 \qquad \text{(eastbound)}$$

$$t_{calc} = \frac{(-7.614) - (0)}{\frac{2.772}{\sqrt{21}}} = -12.587 \qquad \text{(westbound)}$$

6. Determine t value

$t_{critical} = \pm 2.086$ (for two tail Type I error = 5% with $v = n - 1 = 20$ degrees of freedom)

(the $t_{critical}$ concept is illustrated in Figure 3.3)

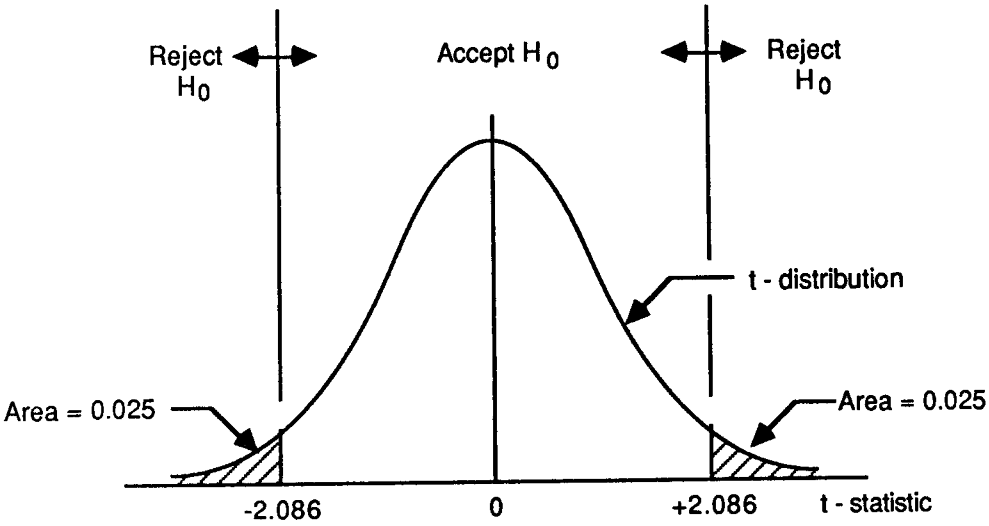


Figure 3.3. t - test Example (SR-82 Friction Numbers)

SECTION 3.0 REFERENCES

3.1 Blank, Leland, Statistical Procedures for Engineering, Management, and Science, McGraw - Hill Book Company, 1980.

3.2 Willenbrock, Jack H., "A Manual for Statistical Quality Control of Highway Construction - Volume I," Federal Highway Administration, National Highway Institute, Washington, D.C., January 1976.

SECTION 4.0
REGRESSION ANALYSIS

1. INTRODUCTION

Recall from SECTION 1.0 that regression analysis can be used to generate an equation to predict one variable from another (or others, which constitutes multiple regression). The predicted variable is the "dependent variable" and the other variables are called "independent variables."

Sir Francis Galton (England) apparently first used the term "regression" in the context of statistics in the late 1800s [4.1]. He was studying the inheritance of human characteristics and noted that offspring tend to "revert" (regress) toward "mediocrity." What he was trying to say was that children's heights, as they grow into adults, tend toward an average or median height.

2. CORRELATION

In statistics, there are several ways two variables can be evaluated so that their collective association can be measured. A common measure of association is correlation. A few of the more significant points about correlation include [after Ref. 4.2] the following.

(a) The correlation coefficient is designated by "r."

(b) The correlation coefficient can range between -1 and +1. If the two variables whose association is being measured are designated as "y" and "x", then the correlation coefficient is positive if an increase in y corresponds to an increase in x.

(c) The correlation coefficient equals 1.0 if all of the y and x values fall on a straight line.

(d) When the correlation coefficient approaches 0.0, then there is little (if any) association between y and x (however, there are exceptions - refer to Figure 10.2 (p. 221) of the MINITAB manual [4.2]).

The basic equation for determining the correlation coefficient is

$$r = \frac{\Sigma\,(x - \bar{x})(y - \bar{y})}{\left[\Sigma\,(x - \bar{x})^2 \Sigma\,(y - \bar{y})^2\right]^{1/2}} \qquad \text{Eq 4.1}$$

The "computation" formula for r is

$$r = \frac{\Sigma xy - \dfrac{\Sigma x \Sigma y}{n}}{\left[\left(\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}\right)\left(\Sigma y^2 - \dfrac{(\Sigma y)^2}{n}\right)\right]^{1/2}} \qquad \text{Eq.4.2}$$

2.1 Example - calculation of the correlation coefficient

An example of correlation will be a comparison of asphalt concrete resilient modulus values obtained in the same laboratory by two different technicians [4.3]. The laboratory data are for a gravel aggregate specimens tested at 77°F. They are shown on Table 4.1.

To simplify the calculations, units of $X10^6$ psi will be used, along with Equation 4.2.

y	x	y^2	x^2	xy
0.204	0.195	0.0416	0.0380	0.0398
0.231	0.207	0.0543	0.0428	0.0478
0.227	0.198	0.0515	0.0392	0.0449
0.228	0.204	0.0520	0.0416	0.0465
0.261	0.229	0.0681	0.0524	0.0598
0.195	0.180	0.0380	0.0324	0.0351
0.225	0.206	0.0506	0.0424	0.0464
0.216	0.202	0.0467	0.0408	0.0436
0.205	0.182	0.0420	0.0331	0.0373
0.232	0.235	0.0538	0.0552	0.0545
0.205	0.186	0.0420	0.0346	0.0381
0.261	0.237	0.0681	0.0562	0.0619
$\Sigma y = 2.690$	$\Sigma x = 2.461$	$\Sigma y^2 = 0.6078$	$\Sigma x^2 = 0.5087$	$\Sigma xy = 0.5557$

After Equation 4.2:

$$r = \frac{0.5557 - \frac{(2.461)(2.690)}{12}}{\left[\left(0.5087 - \frac{(2.461)^2}{12}\right)\left(0.6078 - \frac{(2.690)^2}{12}\right)\right]^{1/2}}$$

$$r = \frac{0.00403}{0.00437} = +0.92$$

Table 4.1. Asphalt Concrete Resilient Modulus Data by Two Technicians - Gravel Aggregate at 77°F

Specimen No.	Resilient Modulus (psi) at 77°F	
	Operator y	Operator x
1	204,000	195,000
2	231,000	207,000
3	227,000	198,000
4	228,000	204,000
5	261,000	229,000
6	195,000	180,000
7	225,000	206,000
8	216,000	202,000
9	205,000	182,000
10	232,000	235,000
11	205,000	186,000
12	261,000	237,000

3. LINEAR REGRESSION WITH ONE INDEPENDENT VARIABLE

3.1 Basic Regression Model

First, examine a basic regression model (or equation, which in this case indicates the same thing):

$$y_i = b_0 + b_1 x_i + \epsilon_i \qquad \text{Eq. 4.3}$$

where

y_i = value of the dependent variable for the ith data point,

x_i = value of the independent variable for the ith data point,

b_0, b_1 = constants (regression parameters),

ϵ_i = random error term, and

$i = 1, 2, 3, ..., n.$

The above model is a simple, linear model. It is simple since there is only one independent variable (x). It is linear since both the parameters (b_0, b_1) and the independent variable (x) are not power functions.

The regression parameters (b_0, b_1) are usually called regression coefficients. The coefficient b_1 is the slope of the regression line and the coefficient b_0 is the intercept of the regression line. This is illustrated in Figure 4.1, which is a plot of Skid Number at 40 mph versus ADT per lane for pavement field data for a select type of limestone rock asphalt surface course. The resulting equation, based on eight data points, is

$$SN_{40} = 56.9 - 0.00666 \text{ (ADT LANE)}$$

The intercept (b_0) at zero ADT per lane is 56.9. This is analogous to a new pavement surfacing that has received no traffic. The slope (b_1) is 0.00666, which means that the Skid Number is reduced by 0.00666 for each increase of 1 ADT per lane (or more understandably, the Skid Number is reduced by about 6.7 for each increase of 1,000 ADT per lane).

3.2 Method of Least Squares

The best relationship (or line) to use to predict some y from x is one that minimizes the differences between the regression line and the actual data. In Figure 4.2 (a), a clear association is shown and one not so clear is shown in Figure 4.2 (b). Thus, Figure 4.2 (a) probably comes closest to minimizing the differences between the line and the plotted y and x data points.

The minimization of the differences between the regression line and the actual data points is illustrated in Figure 4.3, i.e., the differences between the fitted data on the regression line ($\hat{y}_i$) and the actual data points (y_i) are minimized . More specifically, the squared differences are minimized, i.e., a minimum of $\Sigma(\hat{y}_i - y_i)^2$. The squared term results from the derivation, which is based on calculus. From this basic idea come the following equations, which are used to obtain the regression coefficients for a simple, linear regression model:

$$b_1 = \frac{\Sigma\ (x_i - \bar{x})\ (y_i - \bar{y})}{\Sigma\ (x_i - \bar{x})^2} \quad \text{(basic)} \qquad \text{Eq. 4.4}$$

$$b_1 = \frac{\Sigma x_i y_i - \frac{(\Sigma x_i)\ (\Sigma y_i)}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}} \quad \text{(computational)} \qquad \text{Eq. 4.5}$$

$$b_0 = \bar{y} - b_1\ \bar{x} \quad \text{(basic)} \qquad \text{Eq. 4.6}$$

$$b_0 = \frac{1}{n}\ (\Sigma y_i - b_1 \Sigma x_i) \quad \text{(computational)} \qquad \text{Eq. 4.7}$$

3.3 Example - determination of a regression line by the method of least squares

Use the data from Table 4.1 to apply an "adjustment" to the resilient modulus data obtained by Operator "x". In other words, predict the modulus values that would be obtained by Operator "y" from what you know about the results obtained by Operator "x." Determine the appropriate regression line (the basic information needed to determine b_0 and b_1 is contained in SECTION 2.1).

$$b_1 = \text{(from Eq. 4.5)} = \frac{\Sigma x_i y_i - \frac{(\Sigma x_i)\ (\Sigma y_i)}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}$$

$$= \frac{0.5557 - \frac{(2.461)\ (2.690)}{12}}{0.5087 - \frac{(2.461)^2}{12}} = \frac{0.00403}{0.00399}$$

$$= 1.0100$$

$$b_0 = \text{(from Eq. 4.7)} = \frac{1}{n}\ (\Sigma y_i - b_1 \Sigma x_i)$$

$$= \frac{1}{12}\ (2.690 - (1.0100)(2.461))$$

$$= 0.0170$$

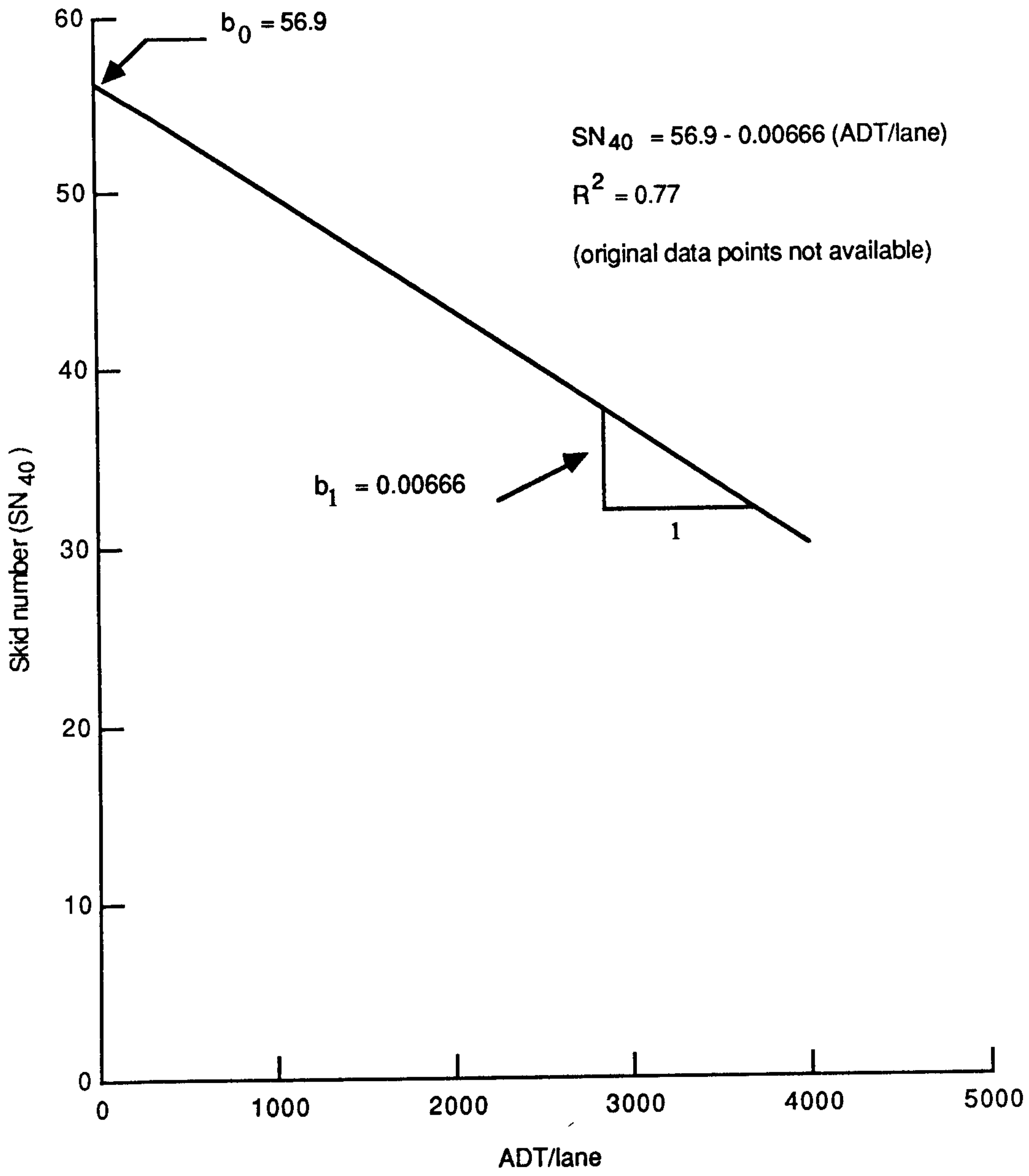
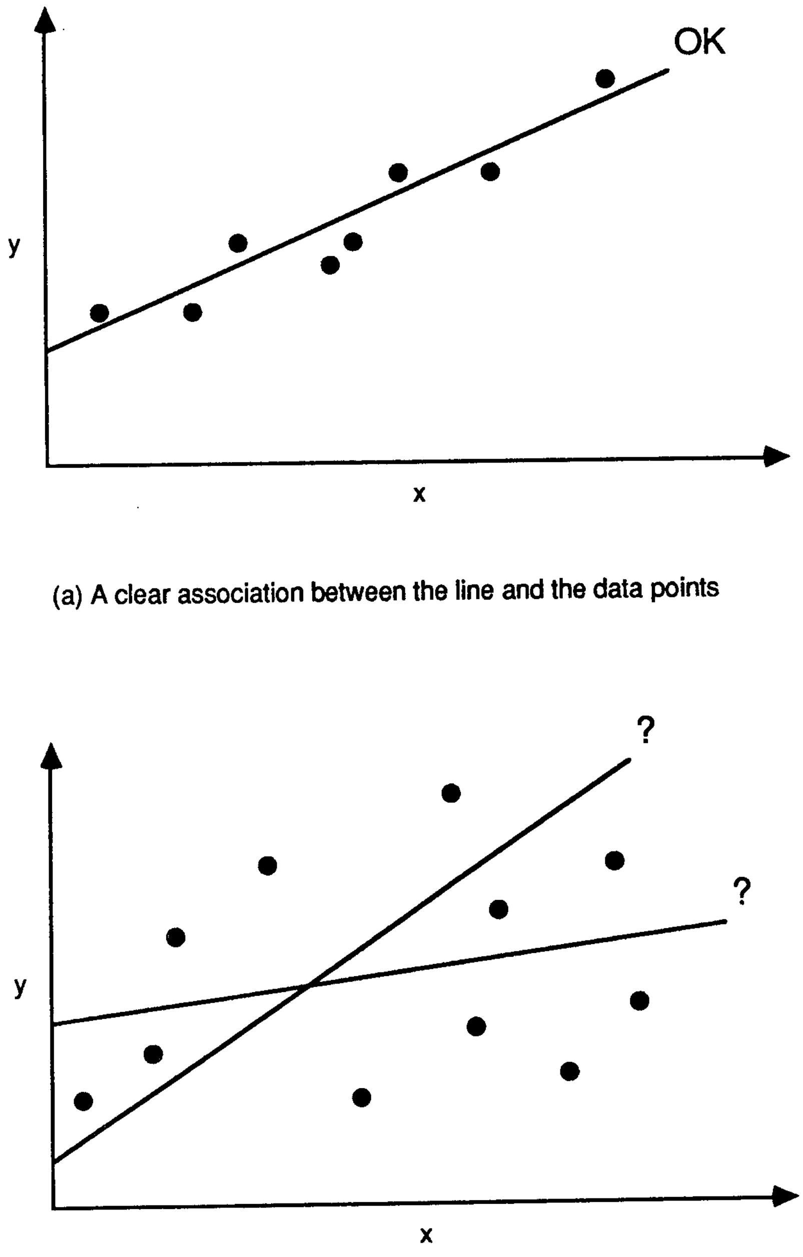


Figure 4.1. Skid Numbers versus ADT per Lane for Limestone Rock Asphalt Surfaces



(a) A clear association between the line and the data points

(b) Unclear associations between lines and the data points

Figure 4.2. Data Fits

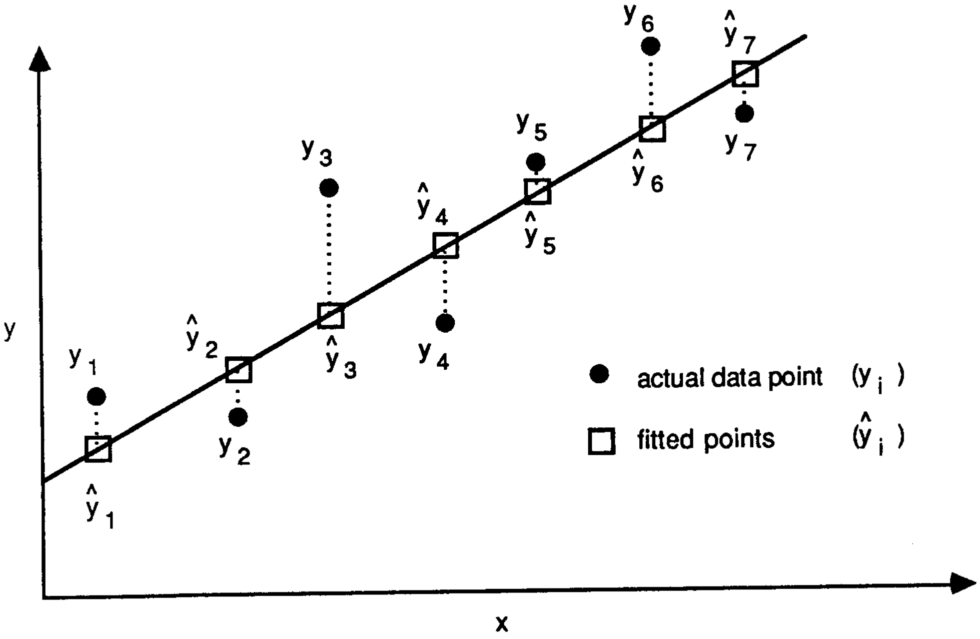


Figure 4.3. Illustration of Minimization of Differences Between the Regression Line and the Data Points

Table 4.2. Actual and Predicted Resilient Modulus Values for Operator Y

Specimen No.	Resilient Modulus (psi) at 77°F		
	Actual Operator y	Predicted Operator y*	Operator x
1	204,000	214,000	195,000
2	231,000	226,000	207,000
3	227,000	217,000	198,000
4	228,000	223,000	204,000
5	261,000	248,000	229,000
6	195,000	199,000	180,000
7	225,000	225,000	206,000
8	216,000	221,000	202,000
9	205,000	201,000	182,000
10	232,000	254,000	235,000
11	205,000	205,000	186,000
12	261,000	256,000	237,000

* Rounded to nearest 1,000 psi

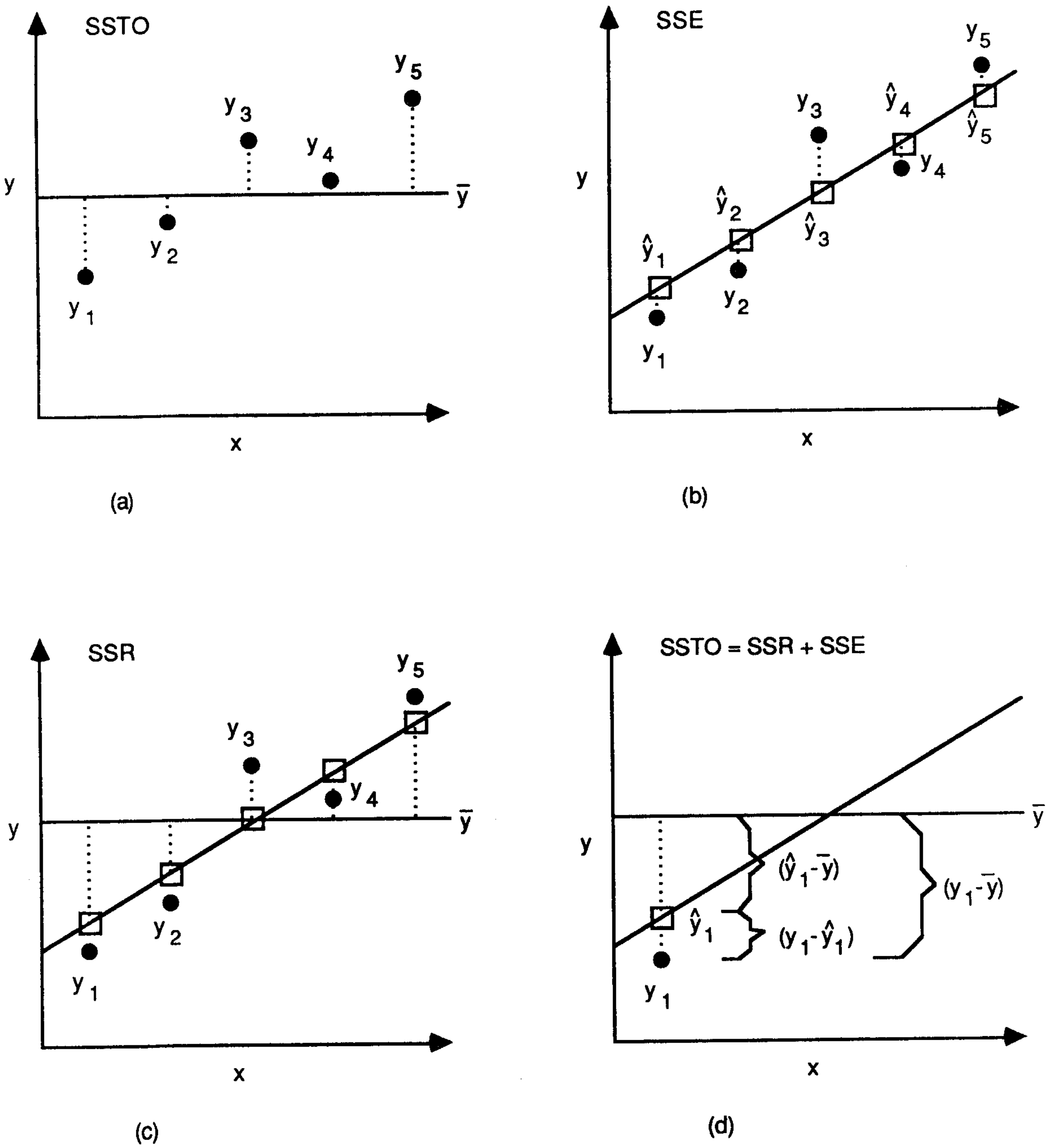


Figure 4.4. Illustration of Deviations Used to Determine Sum of Squares

The following regression equation results:

$$\hat{y} = 0.0170 + 1.0100\,(x)$$

The results of this equation (predicted y) are shown for each of the original "y" and "x" data points in Table 4.2.

3.4 Sum of squares (or the basic information needed to evaluate how "good" a regression line "fits" the data)

3.4.1 Total sum of squares (SSTO) [after Ref. 4.5]

If all y_i data points were identical, then all $y_i = \bar{y}$, which would mean there would be no need for any statistical measure (regression line, etc). However, this is rarely the case with real data. To best evaluate the "fit" of a regression line to actual data, three types of sum of squares measures will be examined, the first of which is illustrated in Figure 4.4 (a) and is denoted the total sum of squares (SSTO):

$$\text{SSTO} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

If SSTO = 0, then all data points must have the same value. The larger the SSTO, the greater the difference between the y_i data points. The calculation of SSTO for the five data points shown in Figure 4.4 (a) is

$$\begin{aligned}\text{SSTO} &= (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2 + (y_4 - \bar{y})^2 + (y_5 - \bar{y})^2 \\ &= \sum_{i=1}^{5} (y_i - \bar{y})^2\end{aligned}$$

Use squared terms, since the sum of deviations of $y_i - \bar{y}$ by definition is zero (which gets us nowhere).

3.4.2 Error sum of squares (SSE)

The "fit" of the regression line to actual data results in some error ("lack-of-fit"), as shown in Figure 4.4 (b). This error is defined as

$$y_i - \hat{y}_i$$

Thus, the error sum of squares (SSE) is defined as

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

or for Figure 4.4 (b), as

$$\text{SSE} = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2 + (y_4 - \hat{y}_4)^2 + (y_5 - \hat{y}_5)^2$$

Stated another way, the SSE is the amount of the sum of squares best explained by the mean ($\bar{y}$) of the dependent variable. The SSE = SSTO when all $\bar{y} = \hat{y}$.

3.4.3 Regression sum of squares (SSR)

Hopefully, for a regression line you wish to develop, the SSTO is much larger than the SSE. The difference is termed the regression sum of squares (SSR):

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

These deviations are illustrated by the dashed lines shown in Figure 4.4 (c). Since the SSR is composed of deviations between the "fitted" regression line and the mean of the data points, the larger the SSR the better the fit of the regression line to the data. Stated another way, the SSR is the amount of the sum of squares explained by the regression equation. For Figure 4.4 (c), SSR is

$$SSR = (\hat{y}_1 - \bar{y})^2 + (\hat{y}_2 - \bar{y})^2 + (\hat{y}_3 - \bar{y})^2 + (\hat{y}_4 - \bar{y})^2 + (\hat{y}_5 - \bar{y})^2$$

3.4.4 Final overview of sum of squares

From the previous sections, we can see that

$y_i - \bar{y}$	=	$\hat{y}_i - \bar{y}$	+	$y_i - \hat{y}_i$
total deviation		deviation of fitted regression value about the mean		deviation around the regression line

or $\Sigma (y_i - \bar{y})^2 = \Sigma (\hat{y}_i - \bar{y})^2 + \Sigma (y_i - \hat{y}_i)^2$

$$SSTO \quad = \quad SSR \quad + \quad SSE$$

3.5 Regression line "goodness-of-fit"

3.5.1 Coefficient of determination (R^2)

The R^2 value explains how much of the total variation in the data is explained by the regression line. Stated another way, the R^2 measures the reduction in the total variation for "y" associated with the use of "x". The $R^2 = 1.0$ when all data points fall on the regression line, as shown in Figure 4.5 (a). The $R^2 = 0$ when the regression line matches the average (or mean) of the data points, as illustrated in Figure 4.5 (b). In other words the mean of the data points is as good a predictor of "y" as any line fit through the data points.

For example, if $R^2 = 0.20$, then the total variation in y is reduced by only 20 percent when x is used (on the other hand $r = \sqrt{R^2} = \sqrt{0.2} = 0.45$).

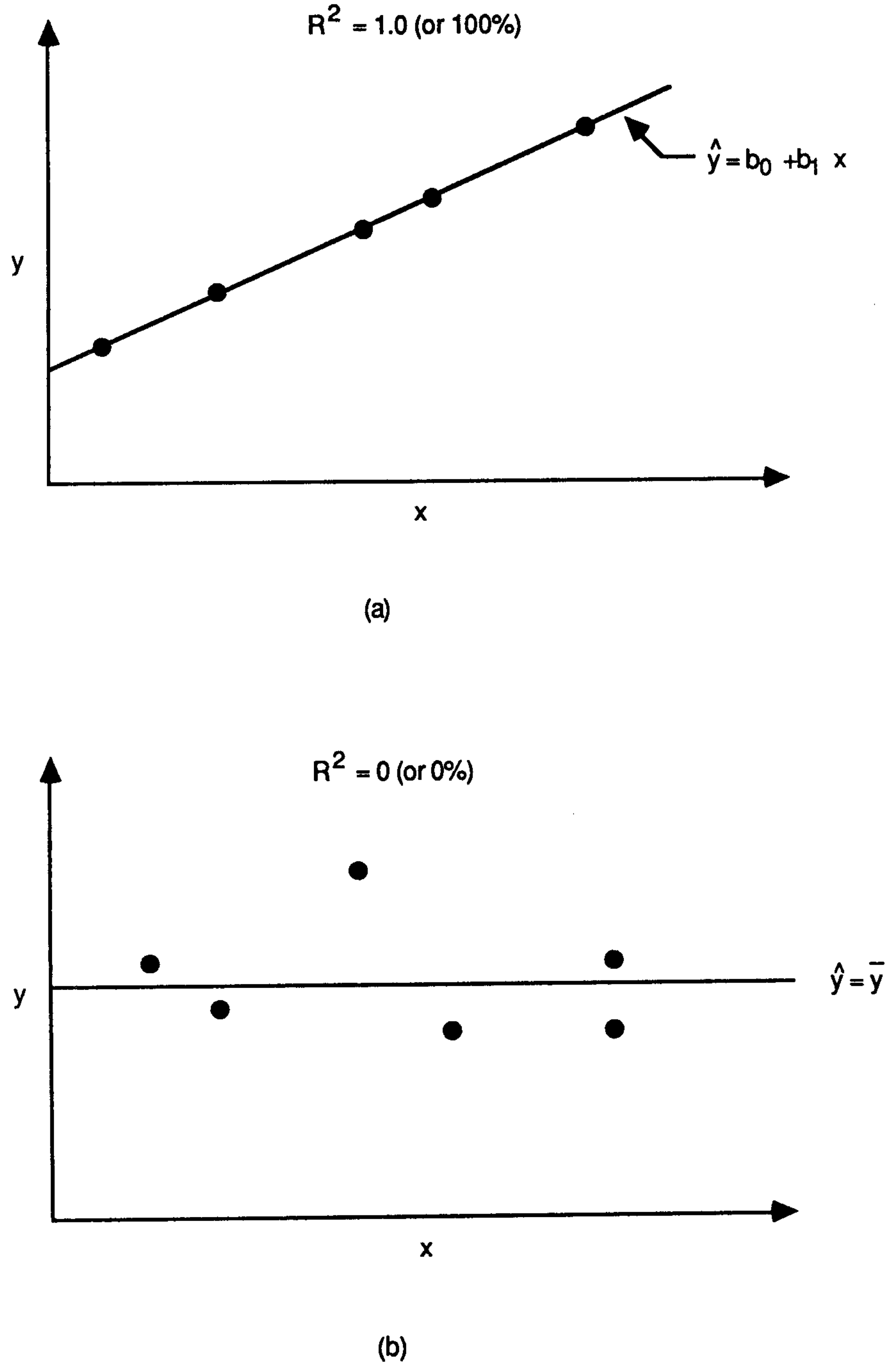
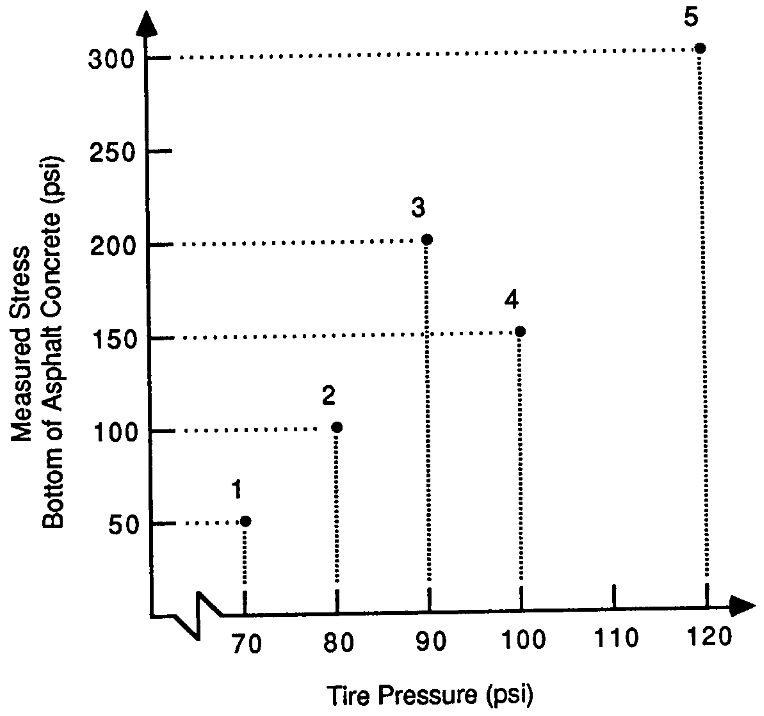


Figure 4.5. Illustrations of the Coefficient of Determination (R^2)



Data Point	y (pavement stress)	x (tire pressure)
1	50	70
2	100	80
3	200	90
4	150	100
5	300	120

Figure 4.6 Example Data - Tire Pressures versus ACP Stress

3.5.2 Mean square error (MSE) and root mean square error (RMSE)

The mean square error is calculated as follows:

$$\text{MSE} = \frac{\text{SSE}}{\text{error degrees of freedom}} = \frac{\text{SSE}}{\text{n-2}}$$

The root mean square error is simply the square root of MSE:

$$\text{RMSE} = \sqrt{\text{MSE}}$$

The RMSE is the standard deviation of the distribution of $\hat{y}$ for a specific x. Stated another way, the RMSE is the standard deviation of the regression line. The larger the RMSE for a specific regression equation, the poorer the associated predictions.

3.6 Another example

An example with five data points is shown in Figure 4.6. Assume that these data resulted from a study of truck tire inflation pressures versus measured horizontal stresses on the bottom of an asphalt concrete surface course. The regression line should more easily estimate the pavement stress for any tire pressure.

(a) Determine the regression coefficients (b_0, b_1)

Data Point	y	x	y^2	x^2	xy
1	50	70	2,500	4,900	3,500
2	100	80	10,000	6,400	8,000
3	200	90	40,000	8,100	18,000
4	150	100	22,500	10,000	15,000
5	300	120	90,000	14,400	36,000
	$\Sigma y = 800$	$\Sigma x = 460$	$\Sigma y^2 = 165{,}000$	$\Sigma x^2 = 43{,}800$	$\Sigma xy = 80{,}500$

$$\bar{y} = \frac{800}{5} = 160 \text{ and } \bar{x} = \frac{460}{5} = 92$$

$$b_1 = \text{(from Eq. 4.5)} = \frac{\Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}$$

$$= \frac{(80{,}500) - \frac{(460)(800)}{5}}{(43{,}800) - \frac{(460)^2}{5}} = \frac{(6{,}900)}{(1{,}480)}$$

$$= 4.662$$

b_0 = (from Eq. 4.7) = $\frac{1}{n}(\Sigma y_i - b_1 \Sigma x_i)$

$= \frac{1}{5}(800 - 4.662(460))$

$= -268.90$

The following regression equation results (and refer to Figure 4.7):

$$\hat{y} = -269 + 4.662\ (x)$$

This results in the following predicted values for "y":

Data Point	Predicted y (pavement stress)	Actual y (pavement stress)	Actual x (tire pressure)
1	57	50	70
2	104	100	80
3	151	200	90
4	197	150	100
5	290	300	120

(b) Total sum of squares (SSTO)

$$\text{SSTO} = (50-160)^2 + (100-160)^2 + (200-160)^2 + (150-160)^2 + (300-160)^2$$

$$= 12{,}100 + 3{,}600 + 1{,}600 + 100 + 19{,}600$$

$$= 37{,}000$$

(c) Error sum of squares (SSE)

$$\text{SSE} = (50-57)^2 + (100-104)^2 + (200-151)^2 + (150-197)^2 + (300-290)^2$$

$$= 49 + 16 + 2401 + 2209 + 100$$

$$= 4{,}775$$

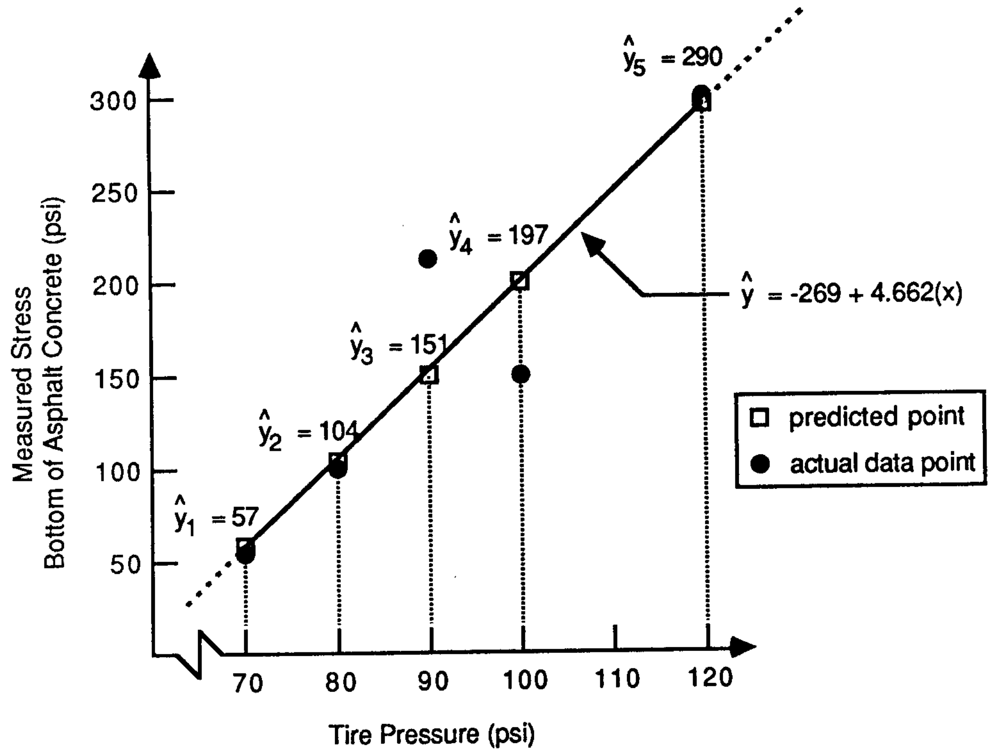


Figure 4.7 Fitted Regression Line for Example Data
(Tire Pressures versus ACP Stress)

(d) Regression sum of squares (SSR)

$$\begin{aligned} SSR &= (57-160)^2 + (104-160)^2 + (151-160)^2 + (197-160)^2 + (290-160)^2 \\ &= 10{,}609 \quad + \quad 3{,}136 \quad + \quad 81 \quad + \quad 1{,}369 \quad + \quad 16{,}900 \\ &= 32{,}095 \end{aligned}$$

(e) Coefficient of determination (R^2)

$$R^2 = \frac{SSR}{SSTO} = \frac{(32{,}095)}{(37{,}000)} = 0.87 \text{ (or 87 percent)}$$

(f) Mean square error (MSE) and root mean square error (RMSE)

$$MSE = \frac{SSE}{df_{error}} = \frac{SSE}{n-2}$$

$$= \frac{(4{,}775)}{3}$$

$$= 1{,}592$$

$$RMSE = \sqrt{1592} = 39.9 \text{ psi}$$

The RMSE is the standard deviation of the distribution of $\hat{y}$ for a fixed x. If you wish to determine the interval estimate of $\hat{y}$ for a given x, as illustrated in Figure 4.8(a), first calculate [after Ref. 4.6]:

$$s_{\hat{y}} = RMSE\left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma(x - \bar{x})^2}\right]^{1/2} \qquad \text{Eq. 4.8}$$

The interval estimate is then

$$\hat{y} \pm ts_{\hat{y}}$$

This interval is the "narrowest" for $\bar{x}$ and gets larger as one moves further away from $\bar{x}$ (either larger or smaller). For the example of pavement stress versus tire pressures:

Data Point	Actual x (tire pressure)	Actual y (pavement stress)	Predicted y ($\hat{y}$)	$(y-\hat{y})^2$	$(x-\bar{x})^2$	Standard Deviation ($s_{\hat{y}}$)*
1	70	50	57	49	484	29.0
2	80	100	104	16	144	21.8
3	90	200	151	2,401	4	18.0
4	100	150	197	2,209	64	19.7
5	120	300	290	100	784	34.1
				4,775	1,480	

* Use Eq. 4.8 and refer to Minitab calculated values in Figure 4.8(b).

If you wish to place a 95% confidence interval estimate for the mean values of $\hat{y}$ for specific x values on the regression line, then

$$t_{table} = 3.182 \text{ for two-tail } \propto = 5\%, \upsilon = n - 2 = 5 - 2 = 3$$

Thus, the confidence intervals are as follows:

Data Point	Predicted y ($\hat{y}$)	Standard Deviation ($s_{\hat{y}}$)	95% Confidence Interval ($\hat{y} \pm ts_{\hat{y}}$)
1	57	29.0	-35, +149
2	104	21.8	+35, +173
3	151	18.0	+94, +208
4	197	19.7	+134, +260
5	290	34.1	+181, +399

This interval estimate is plotted in Figure 4.8(a). It illustrates the uncertainty associated with the regression line, particularly as one moves away from the $\bar{y}$ and $\bar{x}$ values.

The MSE (or RMSE) can also be used for testing whether the regression coefficients (b_0, b_1) are significantly different from zero. This is illustrated for the pavement stress example.

H_0: $b_0 = 0$ (null)
H_1: $b_0 \neq 0$

$$t_{calc} = \frac{b_0 - 0}{s_{bo}}$$

where

$$s_{bo} = \left[MSE \left(\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2} \right) \right]^{1/2}$$

$$= \left[1{,}592 \left(\frac{1}{5} + \frac{92^2}{1{,}480} \right) \right]^{1/2} = 97.1 \text{ psi}$$

$$\therefore\ t_{calc} = \frac{-269 - 0}{97.1} = -2.77$$

If $t_{calc} > t_{table\,(1 - \propto,\, n - 2)}$, conclude H_1

$t_{calc} = -2.77 < -3.182 = t_{table}$

Since $t_{calc} < t_{table}$, you can conclude that b_0 is not different from zero.

For b_1:

H_0: $b_1 = 0$ (null)
H_1: $b_1 \neq 0$ (alternative)

$$t_{calc} = \frac{b_1 - 0}{s_b}$$

where

$$s_b = \left[MSE \left(\frac{1}{\Sigma (x_i - \bar{x})^2} \right) \right]^{1/2}$$

$$= \left[1{,}592 \left(\frac{1}{1{,}480} \right) \right]^{1/2} = 1.04 \text{ psi}$$

$$\therefore t_{calc} = \frac{4.66 - 0}{1.04} = 4.48$$

If $t_{calc} > t_{table\,(1 - \propto,\, n - 2)}$, conclude H_1

$t_{calc} = 4.48 > 3.180 = t_{table}$

Since $t_{calc} > t_{table}$, you can conclude that b_1 is different than zero. Refer to the MINITAB printout (Figure 4.8(b)) to verify these results.

The b_0 value (the intercept) is not statistically significant (at $\alpha = 5\%$) but the b_1 (the slope) is. An inspection of the original data illustrates why this can (and should) occur.

As a rough rule-of-thumb, the t_{calc} values automatically calculated from statistical software (such as Minitab) for the regression b values should equal or exceed the following (for $\propto = 5\%$):

Number of Data Points	t_{calc} (equal to or greater)
3	13
4	4
5-7	3
8-13	2.5
14 or more	2

(g) The Minitab printout for this example is shown as Figure 4.8(b). The results are slightly different because they have been rounded off.

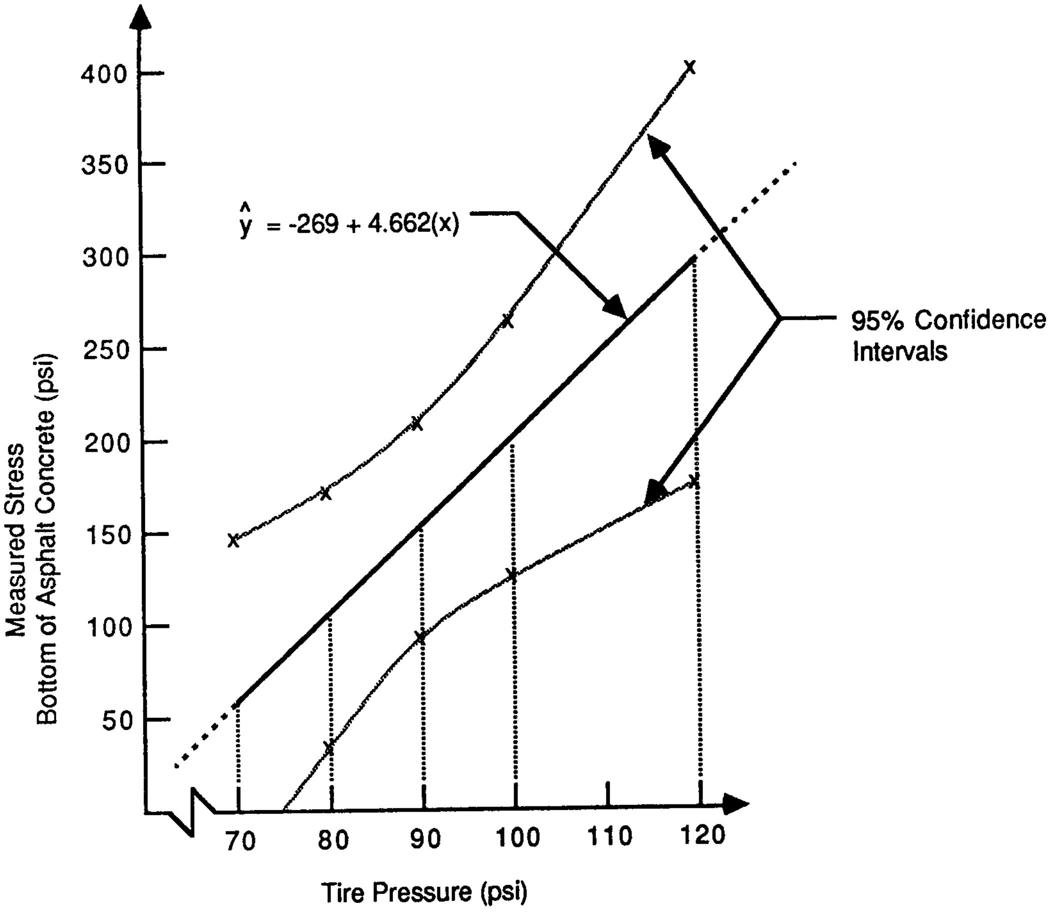
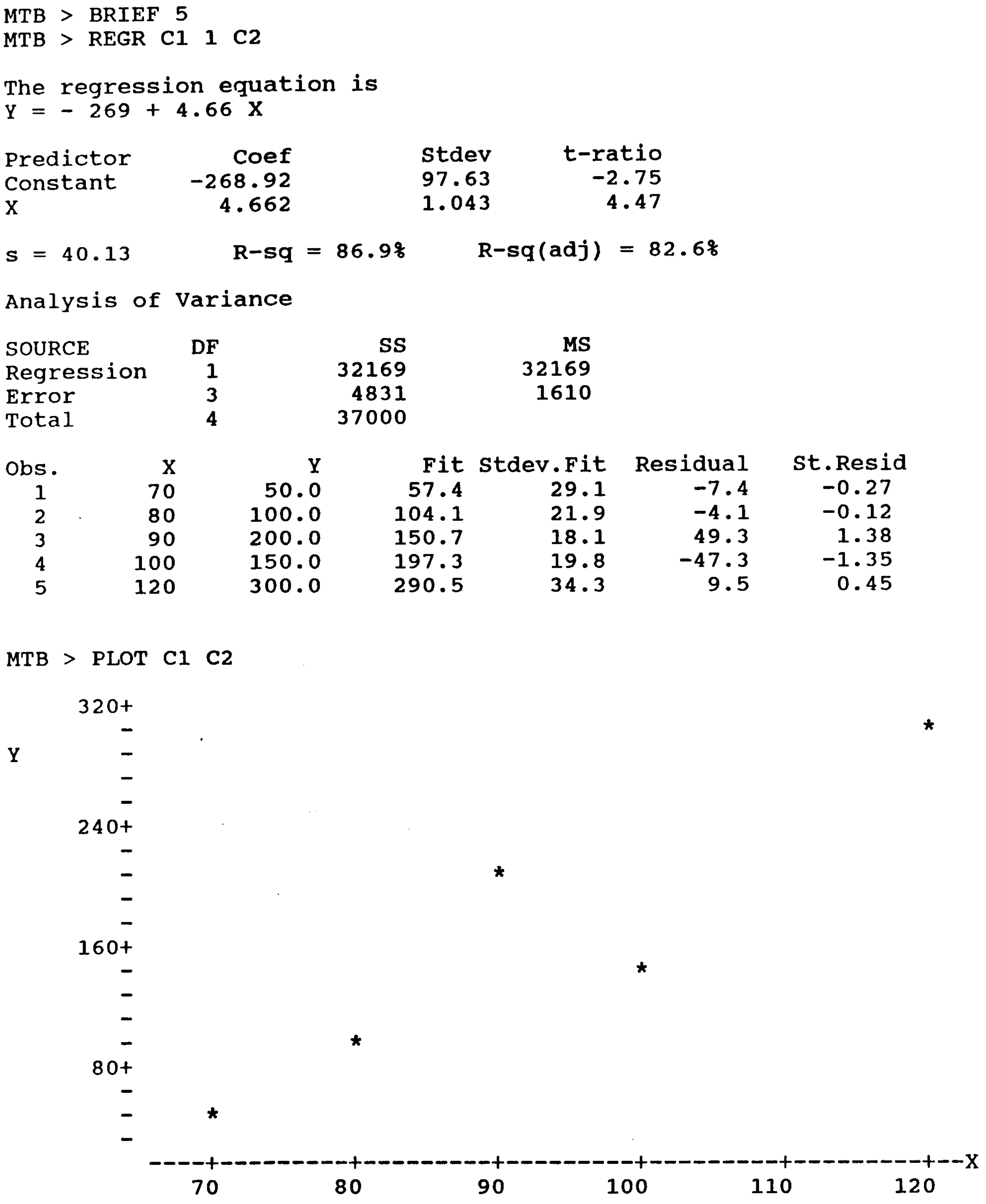


Figure 4.8 (a). Confidence Interval for Regression Line



```
MTB > BRIEF 5
MTB > REGR C1 1 C2

The regression equation is
Y = - 269 + 4.66 X

Predictor        Coef          Stdev      t-ratio
Constant      -268.92          97.63        -2.75
X               4.662          1.043         4.47

s = 40.13        R-sq = 86.9%      R-sq(adj) = 82.6%

Analysis of Variance

SOURCE        DF          SS          MS
Regression     1       32169       32169
Error          3        4831        1610
Total          4       37000

Obs.         X         Y       Fit  Stdev.Fit  Residual    St.Resid
  1         70      50.0      57.4       29.1      -7.4       -0.27
  2         80     100.0     104.1       21.9      -4.1       -0.12
  3         90     200.0     150.7       18.1      49.3        1.38
  4        100     150.0     197.3       19.8     -47.3       -1.35
  5        120     300.0     290.5       34.3       9.5        0.45


MTB > PLOT C1 C2

     320+
        -                                                       *
Y       -
        -
        -
     240+
        -
        -                         *
        -
        -
     160+
        -                                   *
        -
        -
        -               *
      80+
        -
        -     *
        -
          ----+---------+---------+---------+---------+---------+--X
             70        80        90       100       110       120
```

Figure 4.8(b). Minitab Printout of Regression Example

4. TRANSFORMATIONS

4.1 Transformation of variables

Variable transformations can be used for at least three reasons:

(a) to make data distributions more symmetric (or normal),

(b) to simplify a regression equation that relates two or more variables, and

(c) to create a regression equation form that is more "theoretically" correct.

The three most commonly used transformations are:

(a) square root (e.g. $\sqrt{x}$),

(b) logarithm (e.g. $\log_{10}(x)$),

(c) negative reciprocal (e.g. $-1/x$).

The Minitab Handbook (p. 72) is an excellent information source.

4.2 Transformation of equations

Transformations of equations can be quite helpful. For example, a commonly used relationship to describe the stress sensitivity of unstabilized soils is

$$E_R = K_1 (\theta)^{K_2} \qquad \text{Eq. 4.8}$$

(coarse-grained)

or

$$E_R = K_3 (\sigma_d)^{K_4} \qquad \text{Eq. 4.9}$$

(fine-grained)

where

E_R = resilient modulus (psi),

θ = bulk stress (psi),

σ_d = deviator stress (psi), and

K_1, K_2, K_3, K_4 = regression constants.

To obtain the regression constants, a transformation of the entire equation is needed:

$$E_R = K_1\ (\theta)^{K_2}$$

becomes

$$\log E_R = \log K_1^* + K_2\ (\log \theta)$$

and

$$E_R = K_3\ (\sigma_d)^{K_4}$$

becomes

$$\log E_R\ = \log K_3^* + K_4\ (\log \sigma_d)$$

To obtain a regression equation in the form of

$$y = b_0 + b_1(x)$$

then

$$\begin{aligned} y &= \log E_R \\ b_0 &= \log K_1^* &&\qquad (\text{or } \log K_3^*) \\ b_1 &= K_2 &&\qquad (\text{or } K_4) \\ x &= \log \theta &&\qquad (\text{or } \log \sigma_d) \end{aligned}$$

To convert out of $\log_{10}$, convert from

$$\log E_R = \log K_1^* + K_2(\log \theta)$$

to

$$E_R = 10^{K_1^*}\ \theta^{K_2} = K_1\theta^{K_2}$$

$$(\text{letting } 10^{K_1^*} = K_1)$$

Clearly, a similar transformation is appropriate for Equation 4.9.

As a reminder, helpful $\log_{10}$ relationships include the following:

$$\log xy = \log x + \log y$$

$$\log \frac{x}{y} = \log x - \log y$$

$$\log x^n = n \log x$$

$$\log \sqrt[n]{x} = \frac{\log x}{n}$$

4.3 Types of regression models

(a) Linear (refer to Figure 4.9(a))

$$y = b_0 + b_1(x)$$

(b) Exponential (refer to Figure 4.9(b))

$y = b_0(b_1)^x$ (original)

$\log y = \log b_0 + x \log b_1$ (transformed)

(c) Power (refer to Figure 4.9(c))

$y = b_0(x)^{b1}$ (original)

$\log y = \log b_0 + b_1 \log (x)$ (transformed)

(d) Hyperbolic (refer to Figure 4.9(d))

$y = b_0 + b_1 \left(\frac{1}{x}\right)$ (original)

$y = b_0 + b_1 (x^1)$ (transformed)

where $x^1 = \frac{1}{x}$

(e) Polynomial

$y = b_0 + b_1x + b_2x^2 + \ldots + b_kx^k$ (general form)

(i) First degree polynomial (straight line)

$y = b_0 + b_1x$ (refer to Figure 4.10(a))

(ii) Second degree polynomial (parabola)

$y = b_0 + b_1x + b_2x^2$ (refer to Figure 4.10(b))

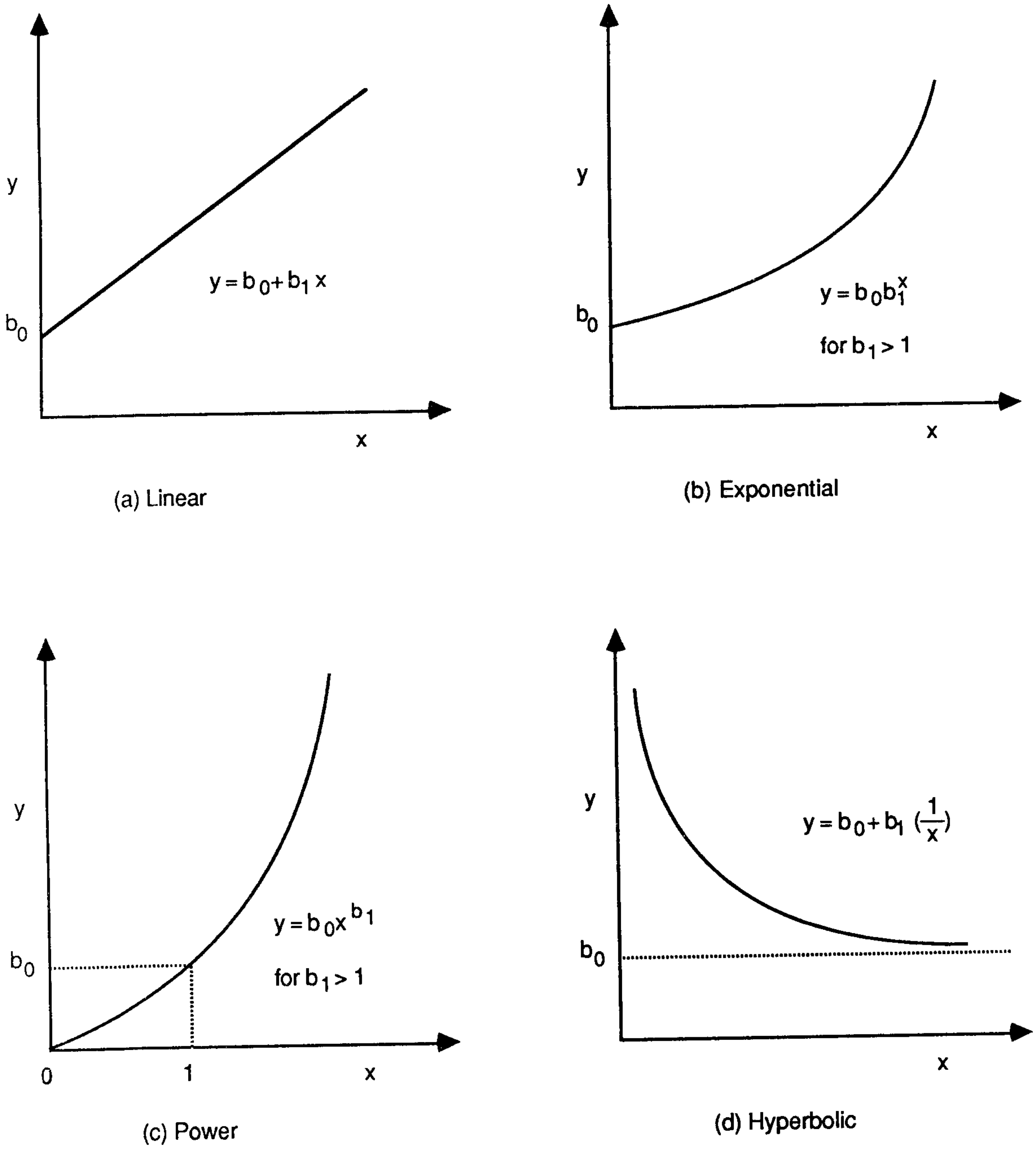
(iii) Third degree polynomial

$y = b_0 + b_1x + b_2x^2 + b_3x^3$ (refer to Figure 4.10(c))

Note: always attempt to obtain the "best fit" with the lowest degree polynomial equation.

5. USE OF REGRESSION EQUATIONS

Any regression equation is only as good as the data used to develop it. Further, such equations should <u>never</u> be used beyond the range of the data from which they were developed. This is particularly important for multiple regression equations.



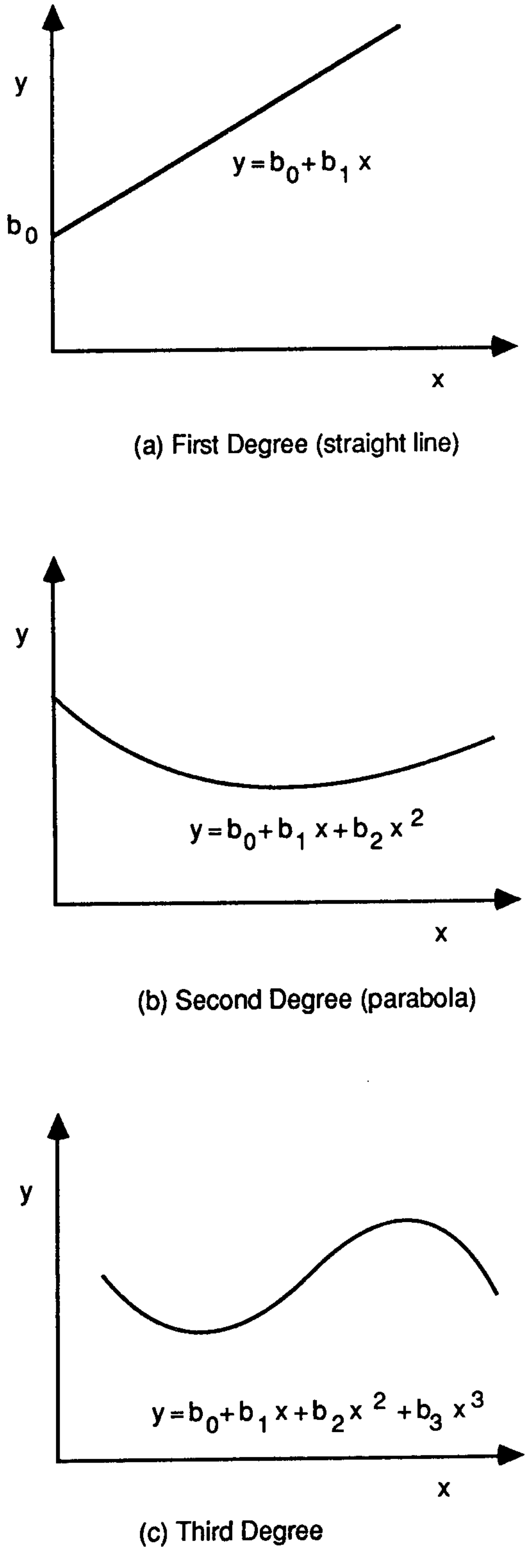
(a) Linear

(b) Exponential

(c) Power

(d) Hyperbolic

Figure 4.9. Types of Regression Models [after Ref. 4.6]



(a) First Degree (straight line)

(b) Second Degree (parabola)

(c) Third Degree

Figure 4.10. Polynomial Regression Models [after Ref. 4.6]

SECTION 4.0 REFERENCES

4.1 Mills, Blake D., "The Origin of the Term Regression as Applied to Plotted Curves," Personal Memorandum, Mechanical Engineering Department, University of Washington, Seattle, WA, September 1979.

4.2 Ryan, Barbara F., Joiner, Brian L., and Ryan, Thomas A., Minitab Handbook, Second Edition, PWS Publishers, Boston, MA, 1985.

4.3 Button, Joe W. and Mahoney, Joe P., "Statistical Summary of Resilient Modulus Measurements," Research Brief No. 2, Texas Transportation Institute, Texas A&M University, College Station, Texas, July 1977.

4.4 Mahoney, Joe P., "A Study of the Skid Performance of Limestone Rock Asphalt Surface Courses," Report submitted to the Institute of Statistics, Texas A&M University, College Station, Texas, May 1978.

4.5 Neter, John, and Wasserman, William, Applied Linear Statistical Models, Richard D. Irwin, Inc., Homewood, Illinois, 1974.

4.6 Blank, Leland, Statistical Procedures for Engineering, Management, and Science, McGraw - Hill Book Company, 1980.

APPENDIX

A BRIEF EXPLANATION OF THE "ANALYSIS OF VARIANCE" SECTION FOR REGRESSION ANALYSIS

APPENDIX

The following numbered sections correspond to the illustration shown as Figure A-1. This figure is part of the Minitab printout for the following regression equation:

$$\text{GRADE} = 7.59 - 0.195(\text{AREA}) - 1.30(\text{CORN}) + 0.912(\text{WHEEL}) - 0.456(\text{SHAPE}) + 25.9\ (\text{RATIO})$$

Definitions of the independent variables are not important for the purposes of this example.

1. The measure of total variation.

   total sum of squares = SSTO = $\Sigma\ (y_i - \bar{y})^2$

   where: y_i = individual data points, and
   $\bar{y}$ = mean of all data points.

2. The measure of uncertainity or error sum of squares.

   SSE = $\Sigma\ (y_i - \hat{y})^2$

   where: y_i = individual data points, and
   $\hat{y}_i$ = fitted data points from the regression equation.

3. Explained variation or regression sum of squares

   SSR = $\Sigma\ (\hat{y}_i - \bar{y})^2$

   This is the sum of squares of the differences between the fitted values ($\hat{y}_i$) and the mean of the values ($\bar{y}$).

4. Sum of squares

   $$\Sigma\ (y_i - \bar{y})^2 = \Sigma\ (\hat{y}_i - \bar{y})^2 + \Sigma\ (y_i - \hat{y}_i)^2$$

   SSTO = SSR + SSE

5. Coefficient of determination (R^2)

   $$R^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}} = \frac{\text{SSR}}{\text{SSTO}}$$

   The R^2 value explains how much of the total variation in the data is explained by the regression equation.

6. Regression degrees of freedom

The regression sum of squares has the nunber of independent variables (x's) minus one degree of freedom. For example, in Figure A-1 the equation has the general form

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5$$

This "model" has six estimated parameters (b's), so the degrees of freedom = 6 - 1 = 5.

7. Total degrees of freedom

The total sum of squares has the total number of data points minus one degree of freedom. For the regression equation in Figure A-1, the total number of data points used to develop the equation was 170.

8. Error degrees of freedom

The error sum of squares is the difference between total and regression degrees of freedom, since degrees of freedom are additive. For our example, the error degrees of freedom = 169 - 5 = 164.

9. Mean square error (MSE) and root mean square error (RMSE)

The mean square error is calculated as follows:

$$\text{MSE} = \frac{\text{SSE}}{\text{df}_{\text{error}}}$$

The MSE is an estimate of σ^2 for the regression model or

$$\text{E (MSE)} = \sigma^2 = \text{variance}$$

The above equation reads as follows: the expected value of MSE is equal to σ^2.

The root mean square error (RMSE) = $\sqrt{\text{MSE}}$. This term is the standard deviation of the regression model, i.e., it is the standard deviation of y (dependent variable) for any x (independent variable).

Additionally, the "error" terms are sometimes called "residuals."

10. F value

This value can be calculated for the example as follows:

$$F_{\text{calc}} = \frac{\text{MSR}}{\text{MSE}} = \frac{94.218}{4.098} = 22.99$$

This calculation is testing whether there is a relation between the dependent variable, y, and the set of independent variables, $x_1, ..., x_{p-1}$, that is, which of the following alternatives to choose:

H_0: $\beta_1 = \beta_2 = ... = \beta_{p-1} = 0$ (null hypothesis)
H_1: not all β_k (k = 1, ..., p-1) equal 0 (alternative hypothesis)

where p-1 = number of independent variables

The decision rule to control the Type I error (rejecting a true hypothesis) at α is as follows:

If $F_{calc} \leq F(1 - \alpha, p - 1, n - p)$, conclude H_0
If $F_{calc} > F(1 - \alpha, p - 1, n - p)$, conclude H_1

Note: The existence of a regression relation, by itself, does not assure that useful predictions can be made by using it.

11. t-statistic

The t-statistic is used for testing whether the corresponding beta (regression coefficients) are zero. For example

$$t_{calc} = \frac{\beta_k}{s(b_k)} = \frac{-0.19549}{0.04418} = -4.43 \quad \text{(refer to "AREA" coefficient)}$$

$$\text{where } s(b_k) = \sqrt{\frac{(MSE)}{\Sigma x_k^2 - \frac{(\Sigma x_k)^2}{n}}}$$

The t-statistic tests the following hypotheses:

H_0: $\beta_K = 0$ (null)
H_1: $\beta_K \neq 0$ (alternative)

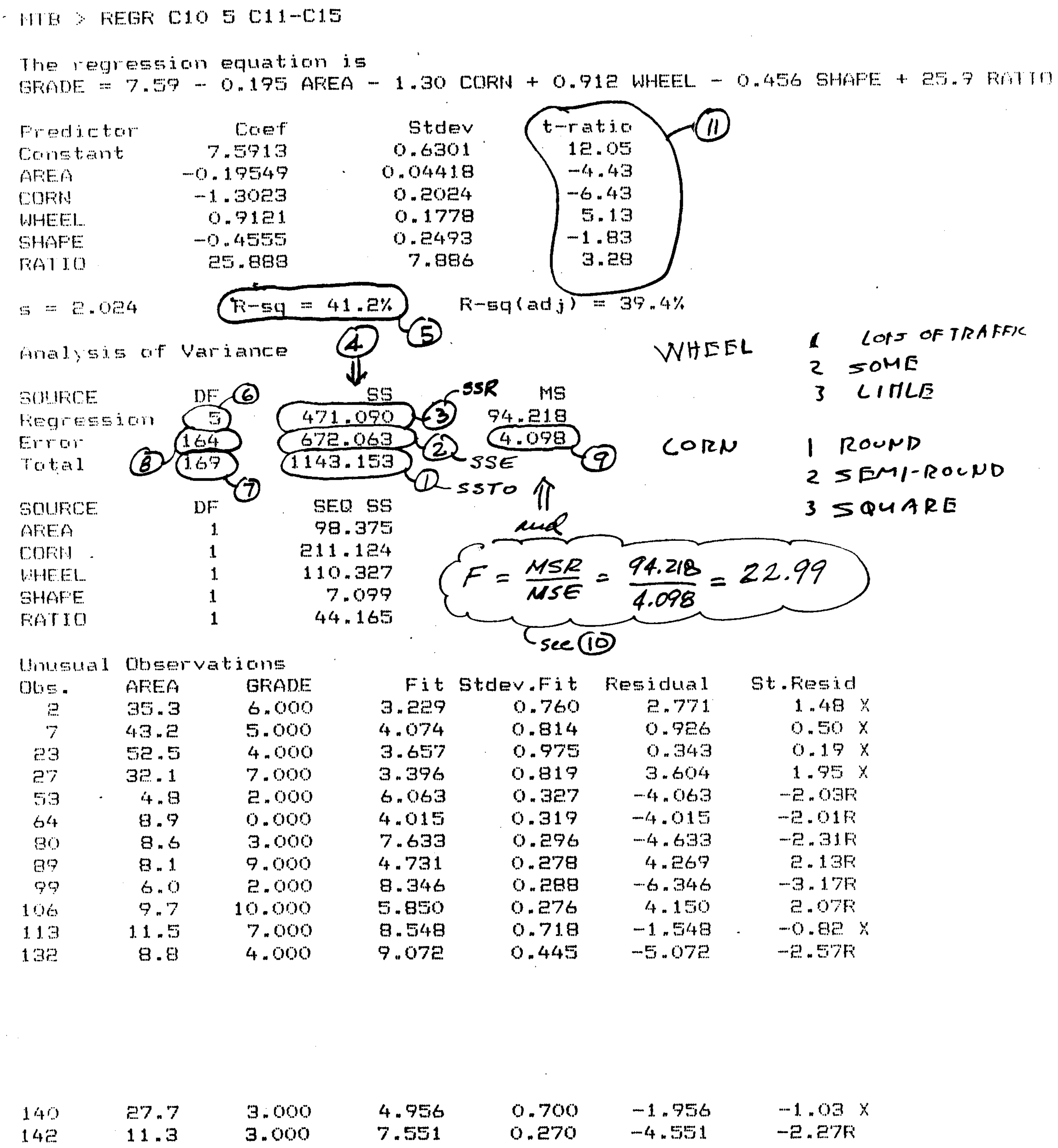
If $t_{calc} \leq t_{table(1 - \alpha/2, n - p)}$, conclude H_0

If $t_{calc} > t_{table(1 - \alpha/2, n - p)}$, conclude H_1

For example, use "AREA" coefficient:

$t_{calc} = -4.43$
and if we set a = 0.05, p = 6 (total β's including intercept), n = 170, $\therefore$ n - p = 164
$\therefore t_{table(1 - 0.05/2, 164)} = 1.96$

Since $t_{calc} > t_{table}$, you can conclude that the "AREA" is a viable variable. Any t_{calc} less than ~ 2.0 is probably "weak."



```
MTB > REGR C10 5 C11-C15

The regression equation is
GRADE = 7.59 - 0.195 AREA - 1.30 CORN + 0.912 WHEEL - 0.456 SHAPE + 25.9 RATIO

Predictor       Coef        Stdev      t-ratio
Constant      7.5913       0.6301        12.05
AREA        -0.19549      0.04418        -4.43
CORN         -1.3023       0.2024        -6.43
WHEEL         0.9121       0.1778         5.13
SHAPE        -0.4555       0.2493        -1.83
RATIO         25.888        7.886         3.28

s = 2.024      R-sq = 41.2%      R-sq(adj) = 39.4%

Analysis of Variance

SOURCE        DF          SS          MS
Regression     5     471.090      94.218
Error        164     672.063       4.098
Total        169    1143.153

SOURCE        DF      SEQ SS
AREA           1      98.375
CORN           1     211.124
WHEEL          1     110.327
SHAPE          1       7.099
RATIO          1      44.165

Unusual Observations
Obs.    AREA     GRADE       Fit  Stdev.Fit   Residual    St.Resid
  2     35.3     6.000     3.229      0.760      2.771       1.48 X
  7     43.2     5.000     4.074      0.814      0.926       0.50 X
 23     52.5     4.000     3.657      0.975      0.343       0.19 X
 27     32.1     7.000     3.396      0.819      3.604       1.95 X
 53      4.8     2.000     6.063      0.327     -4.063      -2.03R
 64      8.9     0.000     4.015      0.319     -4.015      -2.01R
 80      8.6     3.000     7.633      0.296     -4.633      -2.31R
 89      8.1     9.000     4.731      0.278      4.269       2.13R
 99      6.0     2.000     8.346      0.288     -6.346      -3.17R
106      9.7    10.000     5.850      0.276      4.150       2.07R
113     11.5     7.000     8.548      0.718     -1.548      -0.82 X
132      8.8     4.000     9.072      0.445     -5.072      -2.57R




140     27.7     3.000     4.956      0.700     -1.956      -1.03 X
142     11.3     3.000     7.551      0.270     -4.551      -2.27R
```

(11) t-ratio

(5) R-sq = 41.2%

(4) SS

(6) DF

(3) SSR 471.090

(2) SSE 672.063

(1) SSTO 1143.153

(9) 4.098

(8) 164

(7) 169

and

$$F = \frac{MSR}{MSE} = \frac{94.218}{4.098} = 22.99$$

see (10)

WHEEL 1 LOTS OF TRAFFIC
2 SOME
3 LITTLE

CORN 1 ROUND
2 SEMI-ROUND
3 SQUARE

Figure A-1. Minitab Prinout