Ve401 Probabilistic Methods in Engineering

Suggested Solutions to the Sample Final Exam Problems

The following exercises have been compiled from past first midterm exams of Ve401. A first midterm will usually consist of 4-7 such exercises to be completed in 100 minutes. In the actual exam, necessary tables of values of distributions will be provided. You may use all tables in Appendix A of the textbook to solve the sample exercises.

Exercise 1.
A manufacturer of precision measuring instruments claims that the standard deviation in the use of an instrument is not more than 0.00002 inch. An analyst, who is unaware if the claim, uses the instrument eight times and obtains a sample standard deviation of 0.00005 inch.

i) Using $\alpha = 0.01$, is the manufacturer's claim justified?

ii) What is the power of the test if the true standard deviation equals 0.00004 inch?

iii) What is the smallest sample size that can be used to detect a true standard deviation of 0.00004 inch or more with a probability of at least 0.95? Use $\alpha = 0.01$.

(2+1+1 Marks)

Solution. i) We test the hypotheses

$$H_0\colon \sigma \le 0.00002, \qquad\qquad H_1\colon \sigma > 0.00002.$$

at $\alpha = 0.01$. If H_0 is true, the statistic

$$X^2_{n-1} = (n-1)\frac{S^2}{\sigma_0^2}$$

follows a chi-squared distribution with $n-1 = 7$ degrees of freedom. **(1/2 Mark)** The critical value is $\chi^2_{0.01,7} = 18.5$. **(1/2 Mark)** The value of the statistic is

$$x_7^2 = 7 \cdot \frac{25 \cdot 10^{-8}}{4 \cdot 10^{-8}} = 43.75.$$

(1/2 Mark) Since this exceeds the critical value, we can reject H_0 at the 1% level of significance. **(1/2 Mark)** There is evidence that the manufacturer's claim is not justified.

ii) We use the OC curve for the right-tailed chi-squared test. The abscissa parameetr is

$$\lambda = \frac{S}{\sigma_0} = 2$$

and the sample size is $n = 8$. We read off $\beta \approx 0.34$, so the power is approximately $1 - \beta = 0.66$.

iii) Again, we use the OC chart with $\lambda = 2$ and $\beta = 1 - 0.95 = 0.05$. A sample size of $n = 20$ is sufficient to achieve the power stated.

Exercise 2.
The diameters of bolts are known to have a standard deviation of 0.0001 inch. A random sampe of 10 bolts yields an average diameter of 0.2546 inch.

i) Test the hypothesis that the true mean diameter of bolts equals 0.255 inch, using $\alpha = 0.05$.

ii) What size sample would be necessary to detect a true mean bolt diameter of 0.2552 inch or more with a probability of at least 0.90, assuming $\alpha = 0.05$?

(2+2 Marks)

Solution. i) We test $H_0\colon \mu = 0.255$ at $\alpha = 0.05$. We will use the statistic

$$Z = \frac{\overline{X} - 0.255}{\sigma/\sqrt{n}},$$

which follows a standard normal distribution if H_0 is true. For $\alpha = 5\%$, we will reject H_0 if $|Z| > z_{0.025} = 1.96$. Now

$$z = \frac{0.2546 - 0.255}{0.0001/\sqrt{10}} = -12.65.$$

Since $|-4| > 1.96$, we reject H_0, i.e., the true mean diamater is different from 0.255 inch.

ii) Since in our case

$$d = \frac{\mu - \mu_0}{\sigma} = \frac{0.2552 - 0.255}{0.0001} = 2,$$

we can see from the OC curve that in our case $n = 3$ is sufficient.

Exercise 3.

A company wants to test whether a new assembly line procedure increases the physical stress on its workers. It selects eleven workers to work for one day using each of the assembly line procedures. At the end of each day, their pulse frequency is measured:

Procedure 1	X	63	65	71	75	72	75	68	74	62	73	72
Procedure 2	Y	80	78	96	87	88	96	82	83	77	79	71

It is thought that the median pulse frequency is higher in Procedure 2 than in Procedure 1.

i) Formulate H_0 and H_1.

ii) Use the Wilcoxon signed rank test at the 5% level of significance to determine whether you can reject H_0.

iii) Use a paired T-test (formally; the sample size is actually to small for it to give meaningful results) at the 5% level of significance to determine whether you can reject H_0.

(2 + 2 + 2 Marks)

Solution.

i) Denote by M_X the median pulse frequency in procedure 1 and by M_Y the median pulse frequency in procedure 2. Then we have

$$H_0\colon M_Y \le M_X, \qquad\qquad H_1\colon M_Y > M_X.$$

(We are trying to find evidence to support the hypothesis that the median pulse frequency is higher in Procedure 2 than in Procedure 1.)

ii) We calculate $Y - X$:

$Y - X$	17	13	25	12	16	21	14	9	15	6	-1

(1/2 Mark) We can see from the table of $Y - X$ that there is only a single negative value of $Y - X$, which has rank 1. Therefore,

$$W_+ = 2 + 3 + \ldots + 11 = \frac{11 \cdot 12}{2} - 1 = 65, \qquad\qquad |W_-| = 1.$$

(1/2+1/2 Mark) Therefore, $W = \min(W_+, |W|_-) = 1$. **(1/2 Mark)** According to the table for the Wilcoxon signed-rank test, we reject H_0 at the 5% level of significance if $W < 14$, so we can here reject H_0. **(1/2 Mark)**

iii) For the purposes of the t-test, we assume that the medians are equal to the means, i.e., $\mu_X = M_X$, $\mu_Y = M_Y$. If H_0 is true, $\mu_{Y-X} = 0$ and

$$\frac{\hat{\mu}_{Y-X} - \mu_{Y-X}}{\sqrt{S^2_{Y-X}/n}} = \frac{\hat{\mu}_{Y-X}}{\sqrt{S^2_{Y-X}/n}}$$

satisfies the T-distribution with $\gamma = 10$ degrees of freedom. **(1/2 Mark)** By the table for the T-distribution, at $\alpha = 5\%$ level of significance and $\gamma = 10$, the critical value is 1.812.

In our case, the sample mean of $Y - X$ is

$$\hat{\mu}_{Y-X} = \overline{Y - X} = \frac{1}{11}(17 + 13 + 25 + 12 + 16 + 21 + 14 + 9 + 15 + 6 - 1) = 13.36.$$

(1/2 Mark) The sample variance is

$$S^2_{Y-X} = \frac{1}{10}\sum_{i=1}^{11}(Y_i - X_i - \overline{Y - X})^2 = 49.85.$$

(1/2 Mark) We obtain

$$\frac{\hat{\mu}_{Y-X}}{\sqrt{S^2_{Y-X}/n}} = 6.28.$$

Since $6.28 > 1.812$, we can reject H_0.**(1/2 Mark)**

Exercise 4.
In a hardness test, a steel ball is pressed into the material being tested at a standard load. The diameter of the indentation is measured, which is related to the hardness. Two types of steel balls are available, and their performance is compared on 10 randomly selected specimens. The hypothesis that the two steel balls give the same expected hardness measurement is to be tested at a significance level of $\alpha = 0.05$. Each specimen is tested twice, once with each ball. The results are given below:

Specimen	1	2	3	4	5	6	7	8	9	10
Ball x	75	46	57	43	58	38	61	56	64	65
Ball y	52	41	43	47	32	49	52	44	57	60

Use each of the following methods to test the hypothesis

i) A pooled T-test (assume that the variances are unequal).

ii) A Wilcoxon signed rank test.

iii) A paired T-test.

Compare the results obtained by each of the above tests. What assumptions are necessary for the validity of each test? What is your final conclusion regarding the hypothesis?
(2+2+2+3 Marks)

Solution.

i) We first compute the sample means and variances:

$$\overline{x} = 56.3, \qquad \overline{y} = 47.7$$
$$s^2_X = 125.344, \qquad s^2_Y = 67.122$$

(1/2 Mark) The value of the pooled test statistic is

$$T_\gamma = \frac{(\overline{x} - \overline{y})}{\sqrt{s^2_x/10 + s^2_y/10}} = 1.96.$$

(1/2 Mark) The degrees of freedom for this test are

$$\gamma = \frac{\left(s^2_x/10 + s^2_y/10\right)^2}{\frac{(s^2_x/10)^2}{9} + \frac{(s^2_x/10)^2}{10}} = 16.49$$

rounded down to 16. **(1/2 Mark)** Since $t_{0.025,16} = 2.120 > 1.96$, we do not have enough evidence to reject H_0. **(1/2 Mark)**

Assumptions: X, Y both follow a normal distribution. (The sample size is too small for the Central Limit theorem to be applicable.) **(1/2 Mark)**

ii) The Wilcoxon statistics are

```
Diff = SortBy[X - Y, Abs]
{-4, 5, 5, 7, 9, -11, 12, 14, 23, 26}

W+ = 0;
For[i = 1, i ≤ Length[Diff], i++, If[Positive[Diff[[i]]], W+ = W+ + i,]];
W+

48

W- = 0;
For[i = 1, i ≤ Length[Diff], i++, If[Negative[Diff[[i]]], W- = W- + i,]];
W-

7
```

(1/2 Mark) so the test statistic is $W = \min(W_+, |W_-|) = 7$. **(1/2 Mark)** For a two-sided test at $\alpha = 0.05$ the critical value is 8, **(1/2 Mark)** so there is enough evidence to reject H_0. **(1/2 Mark)**

Assumptions: $X - Y$ follows a symmetric distribution so that the mean is equal to the median. **(1/2 Mark)**

iii) For the paired T-test we calculate the sample mean and variance of $D = X - Y$:

$$\overline{d} = 8.6, \qquad s_d^2 = 124.7.$$

(1/2 Mark) We test $H_0 \colon D = 0$. The statistic used is

$$T_9 = \frac{\overline{d}}{\sqrt{s_d^2/10}} = 2.435$$

(1/2 Mark) The critical value is $t_{0.025,9} = 2.262$. **(1/2 Mark)** Since the value of the test statistic exceeds this, we reject H_0. **(1/2 Mark)**

Assumptions: X, Y both follow a normal distribution. (The sample size is too small for the Central Limit theorem to be applicable.) **(1/2 Mark)**

Based on the test results, we can reject the null hypothesis. **(1/2 Mark)** While the pooled test was not powerful enough to do so, by the elimination of extraneous factor through pairing we were able to collect enough evidence to reject H_0. **(1 Mark)**

Exercise 5.

A product developer is interested in reducing the drying time of a primer paint. Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time. From experience, it is known that the standard deviation of drying time is 8 minutes, and this inherent variability should be unaffected by the addition of the new ingredient. Ten specimens are painted with formulation 1, and another 10 specimens are painted with formulation 2; the 20 specimens are painted in random order.

i) The two sample average drying times are $\overline{x}_1 = 121$ minutes and $\overline{x}_2 = 112$ minutes. Perform a significance test to judge the effectiveness of the new ingredient. What is the P-value of the test? What conclusions can you draw about the effectiveness of the new ingredient?

ii) If the true difference in mean drying times is as much as 10 minutes, find the sample sizes required to detect this difference with probability at least 0.90, assuming the hypothesis test is conducted with $\alpha = 0.01$.

(3+3 Marks)

Solution.

i) We test $H_1 \colon \mu_2 < \mu_1$, $H_0 \colon \mu_2 \geq \mu_1$. **(1/2 Mark)** The test statistic

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma^2/n_1 + \sigma^2/n_1}}$$

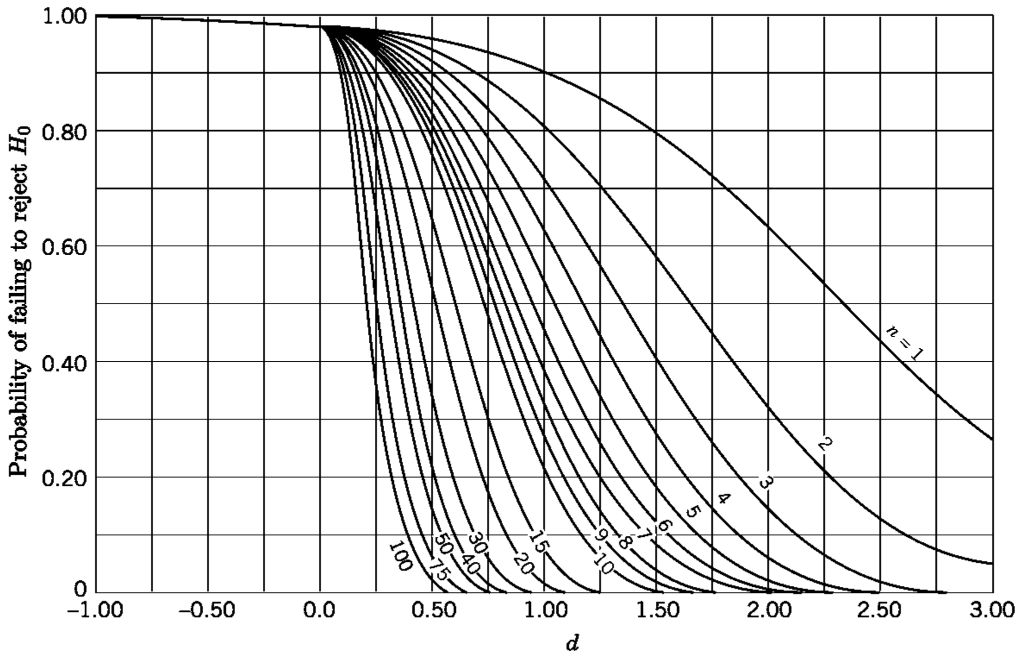
follows (at least approximately) a standard normal distribution. We will reject H_0 if Z is too large to have occurred by chance if H_0 is true. **(1/2 Mark)**

The observed value of Z is

$$z = \frac{121 - 112}{8\sqrt{1/10 + 1/10}} = \frac{9\sqrt{10}}{8\sqrt{2}} = 2.52.$$

(1/2 Mark) The probability of observing this large or a larger result if $\mu_1 = \mu_2$ is $0.5 - 0.4941 = 0.0059$. This is the P-value of the test. **(1 Mark)** Since the P-value is significantly less than 1%, we can conclude that there is evidence that the new drying ingredient reduces the drying time. **(1/2 Mark)**

ii) We use the OC chart for a one-sided test based on the normal distribution:



with $d = (\mu_1 - \mu_2)/\sqrt{\sigma_1^2 + \sigma_2^2} = 10/\sqrt{128} = 10/(8\sqrt{2}) = 0.88$ **(1/2 Mark)** and $\beta = 0.1$. **(1/2 Mark)** This gives a minimum sample size of about $n_1 = n_2 = 17$. **(2 Marks for any number greater than 15 and not greater than 20)**.

Exercise 6.

The second midterm exam of the course Ve401 in Spring 2010 had 25 marks in total. The students taking the exam obtained the following marks:

0, 2, 4.5, 5.5, 8, 8.5, 9, 9.5, 10, 10, 10.5, 10.5, 11, 11.5, 11.5, 12, 12, 12, 12.5, 13, 13, 13.5, 13.5, 14, 14, 14, 14.5, 14.5, 15, 15, 15, 15, 15.5, 15.5, 15.5, 16, 16, 16, 16.5, 16.5, 16.5, 16.5, 16.5, 17, 17.5, 17.5, 17.5, 17.5, 17.5, 18, 18, 18, 18.5, 18.5, 18.5, 18.5, 18.5, 18.5, 18.5, 19, 19, 19, 19, 19, 19, 19, 19, 19.5, 19.5, 20, 20, 20, 20.5, 20.5, 20.5, 21, 21, 21, 21, 21, 21, 21.5, 21.5, 21.5, 21.5, 21.5, 21.5, 21.5, 22, 22, 22, 22.5, 22.5, 22.5, 22.5, 23.5, 23.5, 23.5

Let $S = \{\text{number of marks obtained in the exam}\}$ be the sample space for the the trial "student takes the second midterm exam in Ve401" and consider the random variable $X\colon S \to \mathbb{R}$, $X(s) = 25 - s$. In other words, X gives the difference between the total marks and the marks obtained by a random student.

i) The above data can be used to obtain a random sample of size $n = 98$ from X. Note again that $X(0) = 25$, $X(2) = 23, \ldots, X(23.5) = 1.5$. Plot a stem-and-leaf diagram for the values obtained for X. The stems should have integer units, i.e., there should be 25 stems.

ii) The shape of the stem-and-leaf diagram ressembles that of a chi-squared distribution. Assuming that X follows a chi-squared distribution, find a method-of-moments estimate for the degrees of freedom (rounded to one decimal point).

iii) Use a chi-squared goodness-of-fit test to test the hypotheses

$$H_0\colon X \text{ follows a chi-squared distribution,}$$
$$H_1\colon X \text{ does not follow a chi-squared distribution}$$

at $\alpha = 5\%$. When dividing the positive real axis into categories (intervals), it is not necessary for each interval to have the same expected number of values falling into it. But you should still make sure that the expected numbers are large enough for the chi-squared test to be applicable. If the estimated degrees of freedom are not an integer, interpolate the chi-squared table values linearly to obtain the interval boundaries.

(2+2+4 Marks)

Solution.

i) The stem-and-leaf diagram is shown below

Stem	Leaves
0	
1	555
2	5555
3	0005555555
4	000000555
5	00055
6	000000005555555
7	00055555
8	055555
9	000555
10	000055
11	00055
12	005
13	00055
14	055
15	005
16	05
17	0
18	
19	5
20	5
21	
22	
23	0
24	
25	0

ii) The chi-squared distribution X with γ degrees of freedom is a gamma distribution with $\beta = 2$ and $\alpha = \gamma/2$. Therefore, its expectation is equal to $\mathrm{E}[X] = \alpha\beta = \gamma$. The method-of-moments estimator for n is then simply the sample mean,

$$\hat{\gamma} = \overline{X}.$$

For the given sample, $\hat{\gamma} = 8.3$.

iii) Since we need to use the given chi-squared table, we must choose our categories carefully. Since $n = 98$, we should make sure that at most one interval corresponds to a probability of 5% or less and all categories should have expectations greater than 1. number. We choose the categories in the following way:

Interval boundary	a_1	a_2	a_3	a_4	a_5	a_6	a_7
$P[X < a_i]$	0.025	0.10	0.25	0.50	0.75	0.90	1
E_i	2.45	7.35	14.7	24.5	24.5	14.7	9.8

where E_i is the expected number of values in the interval $[a_{i-1}, a_i]$ One of the expected values is less than five, but this is less than 20% of the six categories. We interpolate the following values for the intervals from the chi-squared table:

γ	0.025	0.10	0.25	0.50	0.75	0.90
8	2.18	3.49	5.07	7.34	10.2	13.4
9	2.70	4.17	5.9	8.34	11.4	14.7
8.3	2.34	3.69	5.31	7.64	10.56	13.79

Therefore, we have the following results

Category	Interval	E_i	O_i
1	$[0, 2.34)$	2.45	3
2	$[2.34, 3.69)$	7.35	14
3	$[3.69, 5.31)$	14.7	12
4	$[5.31, 7.64)$	24.5	25
5	$[7.64, 10.56)$	24.5	18
6	$[10.56, 13.79)$	14.7	13
7	$[13.79, \infty)$	9.8	13

We have the statistic

$$\sum_{i=1}^{7} \frac{(E_i - O_i)^2}{E_i} = 9.612$$

which follows a chi-squared distribution with $7 - 1 - 1 = 5$ degrees of freedom. Since the critical value is $\chi^2_{0.05,5} = 11.1$, there is not enough evidence to reject H_0 at the 5% level of significance. We have no reason to doubt the assumption that the exam marks follow a chi-squared distribution.

Exercise 7.
A study is conducted to test for independence between air quality and air temperature. These data were obtained from records on 200 randomly selected days over the last few years.

	Air Quality		
Temperature	Poor	Fair	Good
Below average	1	3	24
Average	12	28	76
Above average	12	14	30

Do these data indicate an association between these variables? Explain, based on the P-value of the test.
(3 Marks)

Solution. We have the following sums and expected values (in parentheses):

	Air Quality			
Temperature	Poor	Fair	Good	$n_{k\cdot}$
Below average	1 (3.5)	3 (6.3)	24 (18.2)	28
Average	12 (14.5)	28 (26.1)	76 (75.4)	116
Above average	12 (7.0)	14 (12.6)	30 (36.4)	56
$n_{\cdot k}$	25	45	130	200

(1/2 Mark) None of the expected frequencies is smaller than one and only one is smaller than 5. Therefore, we may use our statistical methods. **(1/2 Mark)** The null hypothesis is

$$H_0\colon \text{no association between air quality and temperture.}$$

and the test statistic is

$$\sum_{i,j=1}^{3} \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} = 10.789$$

(1/2 Mark) which follows a χ^2 distribution with $(3-1)(3-1) = 4$ degrees of freedom. **(1/2 Mark)** Since the probability of obtaining a value of less than 9.49 is 95%, the P-value of the test is a little less than 5%. **(1/2 Mark)** Thus we reject H_0. The data indicate an association between air quality and temperature. **(1/2 Mark)**

Exercise 8.

Suppose we have the following data:

x	1.0	1.0	3.3	3.3	4.0	4.0	4.0	5.6	5.6	5.6	6.0	6.0	6.5	6.5
y	1.6	1.8	1.8	2.7	2.6	2.6	2.2	3.5	2.8	2.1	3.4	3.2	3.4	3.9

i) Perform a linear regression for y as a function of x.

ii) Test the model for lack of fit at an $\alpha = 0.05$ level of significance.

(2+3 Marks)

Solution. We first calculate

$$n = 14, \qquad \sum_{i=1}^{n} x_i = 62.4, \qquad \sum_{i=1}^{n} y_i = 37.6,$$
$$\sum_{i=1}^{n} x_i^2 = 322.6, \qquad \sum_{i=1}^{n} y_i^2 = 107.76, \qquad \sum_{i=1}^{n} x_i y_i = 181.94.$$

Then

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\Big(n\sum_{i=1}^{n} x_i^2 - \Big(\sum_{i=1}^{n} x_i\Big)^2\Big) = 44.234$$
$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\Big(n\sum_{i=1}^{n} y_i^2 - \Big(\sum_{i=1}^{n} y_i\Big)^2\Big) = 6.777$$
$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\Big(n\sum_{i=1}^{n} x_i y_i - \Big(\sum_{i=1}^{n} x_i\Big)\Big(\sum_{i=1}^{n} y_i\Big)\Big) = 14.3514$$

i) We assume the model $\mu_{Y|x} = \beta_0 + \beta_1 x$. The estimators for β_0 and β_1 are

$$b_1 = \frac{S_{xy}}{S_{xx}} = 0.324, \qquad b_0 = \overline{y} - b_1\overline{x} = 1.240.$$

It follows that the linear regression line is

$$\mu_{Y|x} = 1.24 + 0.324x,$$

ii) First we note that

$$\text{SSE} = S_{yy} - b_1 S_{xy} = 2.120.$$

Calculating SSE_{pe} is comparatvely simple in this case, since we have two repeated measurements for most values of the regressor:

$$\begin{aligned}
\text{SSE}_{\text{pe}} &= \sum_{i=1}^{4}\sum_{j=1}^{2}(Y_{ij} - \overline{Y}_i)^2 + \sum_{i=5}^{6}\sum_{j=1}^{3}(Y_{ij} - \overline{Y}_i)^2 \\
&= \frac{1}{2}\sum_{i=1}^{4}(Y_{i1} - Y_{i2})^2 + 2(2.6 - 2.467)^2 + (2.2 - 2.467)^2 + 2\cdot 0.7^2 \\
&= 1.66
\end{aligned}$$

Furthermore,

$$\text{SSE}_{\text{lf}} = \text{SSE} - \text{SSE}_{\text{pe}} = 2.12 - 1.66 = 0.46.$$

We test

$$H_0\colon \text{the linear regression model is appropriate,}$$
$$H_1\colon \text{the linear regression model is not appropriate.}$$

using the statistic

$$F_{k-2,n-k} = F_{4,8} = \frac{\text{SSE}_{\text{lf}}/(k-2)}{\text{SSE}_{\text{pe}}/(n-k)} = \frac{8}{4}\cdot\frac{0.46}{1.66} = 0.554 < 1.$$

We reject H_0 if the value of the statistic is too large, indicating a significantly larger lack-of-fit error compared to pure error. However, since the value is actually less than one, we do not reject H_0. (Formally, the P-value of the test is greater than 5% because $P[F_{4,8} > 3.84] = 5\%$.) We conclude there is no evidence to indicate that linear regression is not appropriate.

Exercise 9.

A chemical engineer is investigating the effect of process operating temperature on product yield. The study results in the following data:

Temperature (° C)	100	120	140	160	180
Yield (%)	45	54	66	74	85

Fit a linear regression model and find

i) a 95% confidence interval for the slope;

ii) a 95% confidence interval for the intercept;

iii) a 95% confidence interval for the mean yield at 130° C;

iv) a 95% prediction interval for the yield at 130° C.

(4 × 2 Marks)

Solution. We calculate

$$\begin{aligned} n &= 5, & \sum x_k &= 700, & \sum y_k &= 324, \\ \sum x_k^2 &= 102000, & \sum y_k^2 &= 21998, & \sum x_k y_k &= 47360, \\ \overline{x} &= 10, & S_{xx} &= 4000. \end{aligned}$$

We assume the model $\mu_{Y|x} = \beta_0 + \beta_1 x$. The estimators for β_0 and β_1 are

$$b_1 = \frac{n\sum x_k y_k - \left(\sum x_k\right)\left(\sum y_k\right)}{n\sum x_k^2 - \left(\sum x_k\right)^2} = \frac{1}{2}, \qquad b_0 = \frac{1}{n}\sum y_k - \frac{b_1}{n}\sum x_k = -\frac{26}{5}$$

We need the estimator for the variance,

$$S^2 = \mathrm{SSE}/(n-2) = \mathrm{SSE}/3,$$

where the sum of squares error is

$$\mathrm{SSE} = \sum (y_k - b_0 - b_1 x_k)^2 = 2.8.$$

Thus $s^2 = 0.933$, $s = 0.966$. For a 95% confidence interval we need $t_{0.025} = 3.182$, using a $T_{n-2} = T_3$-distribution. Now 95% confidence intervals for the slope and intercept are given by

$$\begin{aligned} b_1 \pm t_{\alpha/2} s/\sqrt{S_{xx}} &= 0.5 \pm 3.182 \cdot 0.966/\sqrt{4000} = 0.5 \pm 0.049, \\ b_0 \pm t_{\alpha/2} s\sqrt{\sum x_k^2}/\sqrt{nS_{xx}} &= 0.5 \pm \sqrt{102000/5} \cdot 3.182 \cdot 0.966/\sqrt{4000} = 5.2 \pm 6.950, \end{aligned}$$

For the confidence interval for $\mu_{Y|130}$ we need the term

$$s\sqrt{\frac{1}{n} + \frac{(130-\bar{x})^2}{S_{xx}}} = 1.41$$

An confidence interval for $\mu_{Y|130}$ is given by

$$\hat{\mu}_{Y|4} \pm 1.41 = b_0 + b_1 \cdot 130 = 59.8 \pm 1.41.$$

For the prediction interval for $\mu_{Y|130}$ we need the term

$$s\sqrt{1 + \frac{1}{n} + \frac{(130-\bar{x})^2}{S_{xx}}} = 3.40,$$

so

$$\widehat{Y \mid 130} = 59.8 \pm 3.40.$$

Exercise 10.
You have applied a quadratic regression model to a data set of $n = 7$ points. Your computer algebra system tells you that $R^2 = 0.781$. Is the regression significant at the $\alpha = 5\%$ level?
(3 Marks)

Solution. We test

$$H_0\colon \text{regression not significant.}$$

We have $k = 2$ (quadratic model), and we know that

$$\begin{aligned} F_{k,n-k-1} &= \frac{n-k-1}{k}\frac{\mathrm{SSR}}{\mathrm{SSE}} = \frac{n-k-1}{k}\frac{\mathrm{SSR}/S_{yy}}{\mathrm{SSE}/S_{yy}} = \frac{n-k-1}{k}\frac{\mathrm{SSR}/S_{yy}}{1-\mathrm{SSR}/S_{yy}} \\ &= \frac{n-k-1}{k}\frac{R^2}{1-R^2} = \frac{4}{2}\cdot\frac{0.781}{0.219} \\ &= 7.132 = F_{2,4} \end{aligned}$$

According to the table, the critical value the $\alpha = 5\%$ level is 6.944, so we can reject H_0.

Exercise 11.
Consider the following data, which result from an experiment to determine the effect of $x =$ test time in hours at a particular temperature on $y =$ change in oil viscosity.

x	0.50	1.00	1.50	2.00	2.50
y	-0.51	-2.09	-6.03	-9.28	-17.12

i) Fit the model $\mu_{y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$ to the data.

ii) Test whether the regression is significant at a 5% level.

iii) Give a 90% prediction interval for $y \mid 0$.

iv) Use an F-test to test whether a linear model is sufficient at a 5% level of significance; give $\mathrm{SSE}_{\text{full}}$ and $\mathrm{SSE}_{\text{reduced}}$.

In order to solve this exercise without the use of a computer you may use that

$$(X^T X)^{-1} = \begin{pmatrix} 4.6 & -6.6 & 2. \\ -6.6 & 10.6857 & -3.4286 \\ 2. & -3.4286 & 1.1429 \end{pmatrix}$$

where X is the model determination matrix. The entries in the matrix have been rounded; make sure you use all the given decimal places in your calculations, otherwise your results will be off.
(3+3+2+3 Marks)

Solution.

i) From

$$\hat{\beta} = b = (X^T X)^{-1} X^T y.$$

we find

$$b = \begin{pmatrix} 4.6 & -6.6 & 2. \\ -6.6 & 10.6857 & -3.4286 \\ 2. & -3.4286 & 1.1429 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0.5 & 1. & 1.5 & 2. & 2.5 \\ 0.25 & 1. & 2.25 & 4. & 6.25 \end{pmatrix} \begin{pmatrix} -0.51 \\ -2.09 \\ -6.03 \\ -9.28 \\ -17.12 \end{pmatrix} = \begin{pmatrix} -0.798 \\ 2.06361 \\ -3.38477 \end{pmatrix}$$

This gives the regression curve

$$\mu_{Y|x} = -0.798 + 2.064x - 3.385x^2.$$

ii) We test

$$\begin{aligned} &H_0\colon \beta_1 = \beta_2 = \ldots = \beta_p = 0, \\ &H_1\colon \beta_j \neq 0 \quad \text{for at least one } j = 1, \ldots, p. \end{aligned}$$

using the statistic

$$F_{p,n-p-1} = F_{2,2} = \frac{\text{SSR}/2}{\text{SSE}/6} = 3\frac{\text{SSR}}{\text{SSE}}.$$

For $\alpha = 5\%$ the critical region of the statistic is the interval $[19, \infty)$. We have

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} y_i\Big)^2 = 174.782$$

$$\text{SSR} = b_0 \sum_{i=1}^{n} y_i + \sum_{j=1}^{2}\sum_{i=1}^{n} b_j x_i^j y_i - \frac{1}{n}\Big(\sum_{i=1}^{n} y_i\Big)^2 = 173.649$$

so $\text{SSE} = (S_{yy} - \text{SSR}) = 1.133$. Then the value of the statistic is

$$F_{2,6} = \frac{\text{SSR}/2}{\text{SSE}/6} = 3\frac{173.649}{1.133} > 19,$$

so we reject H_0. The regression is significant.

iii) The estimator for the variance is

$$\text{MSE} = \frac{\text{SSE}}{n-p-1} = 0.567.$$

We set

$$x_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

A 90% prediction interval is then given by

$$Y \mid 0 = \hat{Y} \mid 0 \pm t_{\alpha/2,n-p-1}\sqrt{\text{MSE}(1 + x_0^T(X^TX)^{-1}x_0)}.$$

where $t_{\alpha/2,n-p-1} = t_{0.05,2} = 2.920$ and we use the point estimate

$$\hat{Y} \mid 0 = -0.798 + 2.064 \cdot 0 - 3.385 \cdot 0^2 = -0.798.$$

Now $x_0^T(X^TX)^{-1}x_0 = 4.6$, so the prediction interval is

$$Y \mid 0 = \hat{Y} \mid 0 \pm t_{0.05,2}\sqrt{\text{MSE}(1 + x_0^T(X^TX)^{-1}x_0)} = -0.798 \pm 5.20.$$

iv) A (reduced) linear model gives

$$\mu_{Y|x} = 5.117 - 8.082x.$$

with $\text{SSE}_{\text{reduced}} = 3.828$. We have seen above that $\text{SSE}_{\text{full}} = 1.133$. We test

H_0: reduced model is sufficient, $\qquad$ H_1: full model is needed.

at $\alpha = 5\%$. The test statistic

$$F_{p-m,n-p-1} = F_{1,6} = \frac{p-m}{n-p-m}\,\frac{\text{SSE}_{\text{reduced}} - \text{SSE}_{\text{full}}}{\text{SSE}_{\text{full}}} = \frac{1}{5-2-1}\,\frac{\text{SSE}_{\text{reduced}} - \text{SSE}_{\text{full}}}{\text{SSE}_{\text{full}}}$$

follows an F distribution with 1 and 2 degrees of freedom; therefore, the critical interval for rejection of H_0 is $(18.5, \infty)$. Now the statistic takes the value

$$\frac{1}{2}\,\frac{\text{SSE}_{\text{reduced}} - \text{SSE}_{\text{full}}}{\text{SSE}_{\text{full}}} = \frac{1}{2}\,\frac{3.828 - 1.133}{1.133} = 1.18 < 18.5,$$

so we fail to reject H_0 at the stated level of significance. There is no evidence that the linear model is not sufficient.

Exercise 12.
Consider the following data:

x	1	2	3	4	5	6	7
y	8	17	29	34	46	42	52

Fit a model of the form $\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$. You may use that

$$(X^T X)^{-1} = \frac{1}{7}\begin{pmatrix} 17 & -9 & 1 \\ -9 & 67/12 & -2/3 \\ 1 & -2/3 & 1/12 \end{pmatrix},$$

where X is the model determination matrix. What is the value of R^2 for this model?
(3+2 Marks)

Solution. Note first that

$$X^T y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 4 & 9 & 16 & 25 & 36 & 49 \end{pmatrix} \begin{pmatrix} 8 \\ 17 \\ 29 \\ 34 \\ 46 \\ 42 \\ 52 \end{pmatrix} = \begin{pmatrix} 228 \\ 1111 \\ 6091 \end{pmatrix}$$

From

$$\hat{\beta} = b = (X^T X)^{-1} X^T y.$$

we find

$$b = \frac{1}{7}\begin{pmatrix} 17 & -9 & 1 \\ -9 & 67/12 & -2/3 \\ 1 & -2/3 & 1/12 \end{pmatrix} \begin{pmatrix} 228 \\ 1111 \\ 6091 \end{pmatrix} = \begin{pmatrix} -32/7 \\ 155/12 \\ -61/84 \end{pmatrix}$$

This gives the regression curve

$$\mu_{Y|x} = -\frac{32}{7} + \frac{155}{12}x - \frac{61}{84}x^2.$$

To find R^2 we note that

$$S_{yy} = \sum_{i=1}^{7} y_i^2 - \frac{1}{7}\Big(\sum_{i=1}^{7} y_i\Big)^2 = 1507.71$$

$$\begin{aligned} \mathrm{SSR} &= b_0 \sum_{i=1}^{7} y_i + b_1 \sum_{i=1}^{7} x_i y_i + b_2 \sum_{i=1}^{7} x_i^2 y_i - \frac{1}{7}\Big(\sum_{i=1}^{7} y_i\Big)^2 \\ &= \langle X^T y, b\rangle - \frac{1}{7}\Big(\sum_{i=1}^{7} y_i\Big)^2 = 1458.62 \end{aligned}$$

so $R^2 = \mathrm{SSR}/S_{yy} = 0.967$.