Worked Example 1:

1. The manager of an oil refinery measures the percentage yield of petroleum spirit y and the specific gravity of crude oil x on seven separate occasions. The data, arranged in order of increasing x-values, are as follows.

x	30.2	32.8	32.9	35.1	42.3	45.5	46.0
y	6.8	10.1	14.3	19.3	10.2	20.0	23.7

(i) Draw a scatter diagram of the data and comment briefly on the suitability of carrying out a simple linear regression analysis on these data.

(5)

(ii) Fit a simple linear regression model $E(Y) = \alpha + \beta x$ to these data, showing details of your calculations.

(6)

(iii) Stating any assumptions you must make and showing details of your calculations, find a 95% confidence interval for β.

(7)

(iv) Give a point prediction for the percentage yield of petroleum spirit when $x = 40$.

(2)

Worked Example 2:

An experimental investigation was made into the heat evolved during the hardening of cement, considered as a function of the chemical composition of the cement. The data recorded were the heat evolved (Y) after 180 days of hardening measured in calories per gram of cement, and the percentages of tricalcium aluminate (X_1) and tricalcium silicate (X_2).

The data were read into a statistical package for analysis. The relevant output follows **at the end of this question**. Use the output to answer the following questions.

(ii) Test the overall regression for significance at the 1% level, and explain the results in terms that a non-statistician would understand.

(4)

(iii) Write down the fitted regression equation of Y on X_1 and X_2 as defined above. Use it to predict the heat evolved during hardening of similar cement with $X_1 = 8$ and $X_2 = 35$.

(5)

(iv) In the output for the fitted model, the p-values for the partial t tests for the regression parameters are missing. Test these parameters for statistical significance at the 0.1% level, quoting the critical value. What do the results imply about the effects of tricalcium aluminate and tricalcium silicate on the heat evolved in hardening?

(4)

(v) Use the value of R^2 to comment on the overall fit of the model.

Regression Analysis: y versus x1 and x2

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	2	2657.9	1328.95	229.53	4.404e-09
Residuals	10	57.9	5.79		
Total	12	2715.8			

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	52.57735	2.28617	23.00	5.46e-10
x1	1.46831	0.12130	12.11	
x2	0.66225	0.04585	14.44	

Residual standard error: 2.406 on 10 degrees of freedom
Multiple R-squared: 0.9787

TABLE 7: PERCENTAGE POINTS OF THE *F* DISTRIBUTION

Upper 1% points

The values in the table are those which a random variable with the F distribution on ν_1 and ν_2 degrees of freedom exceeds with probability 0.01.

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	18	24	∞
1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6106	6192	6235	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.44	99.46	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.05	26.75	26.60	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.08	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.61	9.47	9.02
6	13.74	10.93	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.45	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.21	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.41	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.86	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.46	4.33	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.15	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	3.91	3.78	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.72	3.59	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.56	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.90	3.81	3.67	3.42	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.31	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.21	3.08	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.13	3.00	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.05	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	2.99	2.86	2.42
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.88	2.75	2.31
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.79	2.66	2.21
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.71	2.59	2.13

TABLE 6: PERCENTAGE POINTS OF STUDENT'S t DISTRIBUTION

The values in the table are those which a random variable with Student's t distribution on ν degrees of freedom exceeds with the probability shown.

ν	0.100	0.050	0.025	0.010	0.005	0.001	0.0005
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725

Exercise 1:

With reference the simple linear model and multiple linear model in the above examples and the proof of the mathematics behind, answer (iii) and (iv) with Python script

1. (i) A simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

is to be fitted to some data. What assumptions are usually made about the term representing experimental error (ε_i)?

(2)

(ii) By minimising a suitable function, show that the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1 are given by

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

where $S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y})$ and $S_{xx} = \Sigma(x_i - \bar{x})^2$.

(8)

A clinician recorded the age in years (x) and total cholesterol level of blood (y) for 20 patients suffering from a certain disease. Summary statistics for the data are $\Sigma x_i = 809$, $\Sigma y_i = 68.3$, $S_{xx} = 3630.95$, $S_{xy} = 201.665$, $S_{yy} = 12.9455$.

(iii) Find the equation of the fitted simple linear regression model.

(3)

(iv) Obtain the analysis of variance and hence test the hypothesis that $\beta_1 = 0$.

(7)

Reference:

http://www.hkss.org.hk/images/exam/papers/Past/2015/HC4%202015%20-%20%20HKSS.pdf

Exercise 2:

By making use of Python, translate the following scientific question that can be solved in Python

2. An experiment was carried out to study the variation of the specific heat H in calories per gram of a certain compound with T, its temperature in degrees Celsius. The specific heat was measured twice at each of a series of chosen temperatures, and the results are shown in the following table.

t	50	60	70	80	90	100
h	1.64	1.63	1.67	1.72	1.71	1.71
	1.60	1.65	1.67	1.70	1.72	1.74

You are given that $\sum t = 900$, $\sum h = 20.16$, $\sum t^2 = 71000$, $\sum h^2 = 33.8894$, $\sum th = 1519.9$.

(i) Draw a scatter diagram of the data and comment briefly on the suitability of carrying out a simple linear regression analysis on these data.

(5)

(ii) (a) Fit a simple linear regression model to the data, stating any assumptions made for the purpose of the analysis. Also give a point prediction for the specific heat when $T = 85$.

(5)

Reference:

http://www.hkss.org.hk/images/exam/papers/Past/2008/2008_HC_4_HKSS.pdf

Exercise 3: Under the commercial situation below, formulate multiple linear model by Python and interpret the result to non-statistician

(ii) The data in the following table show the values of price Y (£) for individually patterned Persian carpets of length x_1 (cm) and width x_2 (cm).

y	14	20	37	36	31	42	54	64	38	66	64	77	79	93	119	135
x_1	120	120	120	120	150	150	150	150	180	180	180	180	240	240	240	240
x_2	60	80	100	120	75	100	125	150	90	120	150	180	120	160	200	240

(a) Plot scatter diagrams of price against each of length and width. What do these graphs show?

You should be able to get the following information that support your explanation to other audience

(b) A multiple regression model of price on length and width was fitted to the data given in the table. Edited computer output of the results is as follows.

```
Predictor       Coef     SE Coef         T        P
Constant     -52.671     5.34500     -9.85    0.000
Length       0.32356     0.04250      7.61    0.000
Width        0.44383     0.04012     11.06    0.000

S = 5.32611       R-Sq = 97.9%

Analysis of Variance
Source          DF       SS        MS        F        P
Regression       2  17045.2    8522.6   300.43    0.000
Residual Error  13    368.8      28.4
Total           15  17413.9
```

Interpret these results fully, in terms that a non-statistician would understand. Write down the fitted regression equation of Y on x_1 and x_2, and use it to predict the price of a similar carpet of length 200 cm and width 150 cm. To what extent would you rely on the model to predict the prices of carpets of dimensions outside the sizes observed in the above table (for example, much smaller carpets)?

Reference:

http://www.hkss.org.hk/images/exam/papers/Past/2008/2008_HC_4_HKSS.pdf

Exercise 4:

1. The Devon Motor Racing Grand Prix takes place every five years. Winning average lap speeds (in miles per hour) in the last nine events are shown in the table below.

Year x	1965	1970	1975	1980	1985	1990	1995	2000	2005
Speed y	109	114	116	117	114	127	131	138	141

You are given that

$$\bar{x} = 1985, \quad \sum(x-\bar{x})^2 = 1500, \quad \sum y = 1107, \quad \sum y^2 = 137233, \quad \sum(x-\bar{x})y = 1200.$$

(i) (a) Plot these data and comment on their suitability for simple linear regression analysis. (4)

(b) Fit a simple linear regression model and state its equation. Also compute the total sum of squares and regression sum of squares for this regression, and deduce the error mean square. (6)

(ii) It is later noted that driving conditions in 1985 were affected by a freak thunderstorm which caused partial flooding of the track. The 1985 values were therefore omitted and the regression was recalculated. Results are shown in the computer output below. Compare this analysis with your own results and say with reasons which you regard as the more satisfactory. (3)

```
The regression equation is    y = - 1464 + 0.800 x

Predictor       Coef    SE Coef       T       P
Constant    -1463.87      95.60  -15.31   0.000
x            0.80000    0.04816   16.61   0.000

S = 1.86525   R-Sq = 97.9%   R-Sq(adj) = 97.5%

Analysis of Variance
Source          DF      SS      MS       F       P
Regression       1  960.00  960.00  275.93   0.000
Residual Error   6   20.87    3.48
Total            7  980.87
```

(iii) Use the analysis of part (ii) to obtain point estimates of

(a) the expected winning speed in 1985, (2)

(b) the expected winning speed in 2010, (1)

(c) the time by which a winning speed of 160 mph might be expected. (2)

Mention any reservations you might have about your answers. (2)

Reference:

http://www.hkss.org.hk/images/exam/papers/Past/2009/2009_HC_4_HKSS.pdf

Exercise 5:

2. A certain metal discolours when exposed to air. To protect the metal against discoloration, it is coated with a chemical. In an experiment, coatings of varying thickness, x mm, of the chemical were applied to standard samples of the metal, and the times, t hours, for the metal to discolour were noted. The results are as shown.

x	1.8	3.0	4.0	5.7	7.2	8.4	10.3
t	3.4	5.9	7.0	8.7	9.5	10.4	11.1

(i) You are given that the least squares regression line for these data is

$$t = 3.027 + 0.8617x.$$

Draw a scatter diagram of the data. Plot this regression line on your diagram and comment on the appropriateness of a simple linear regression model for the dependence of t on x.

(5)

(ii) A researcher suggests that the theoretical relationship between t and x should be of the form

$$\exp(t) = Ax^B,$$

where A and B are constants. Show that this relationship may be expressed in the form

$$t = a + b\log x,$$

where a and b are functions of A and B respectively, which you should identify.

(2)

(iii) You are given that

$$\Sigma \log x = 11.2476, \quad \Sigma t = 56, \quad \Sigma(\log x)^2 = 20.3687, \quad \Sigma t \log x = 100.101.$$

Use these results to calculate the least squares regression line of t on $\log x$, and plot this line and the data on a scatter diagram with values of $\log x$ on the horizontal axis.

(6)

(iv) State with reasons which model you prefer. For each of the two models, calculate the predicted value of t when $x = 6$, and comment briefly.

(7)

Reference:

http://www.hkss.org.hk/images/exam/papers/Past/2010/2010_HC_4_HKSS.pdf

Revision Exercise on hypothesis testing:

(b) The amounts of excise duty (in pence per litre) levied on unleaded petrol and diesel in 10 European countries are given in the following table (dated May 2008).

Country	*Unleaded (x)*	*Diesel (y)*
Austria	36	28
Denmark	42	29
Estonia	23	19
Germany	52	37
Greece	26	22
Hungary	30	25
Italy	45	34
Poland	33	23
Spain	31	24
United Kingdom	57	57

You are given that

$$\Sigma x = 375, \quad \Sigma y = 298, \quad \Sigma x^2 = 15\,193, \quad \Sigma y^2 = 9974, \quad \Sigma xy = 12\,191\,.$$

(i) Construct a scatter diagram of y against x and comment on the relationship, if any, between y and x.

(4)

(ii) Calculate the product-moment correlation coefficient for these data, and test at the 1% significance level the null hypothesis that x and y are uncorrelated, against the alternative of a positive correlation.

(4)

(iii) Calculate Spearman's rank correlation coefficient for the above data. Use it to test, at the 1% significance level, the null hypothesis that there is no association between x and y in the underlying population, against the alternative of positive association.

(4)

(iv) Comment on the results of parts (ii) and (iii) and say with a reason which analysis you think is better here. Mention any reservations you may have about this analysis.

(3)

Reference:

http://www.hkss.org.hk/images/exam/papers/Past/2010/2010_HC_4_HKSS.pdf

Revision Exercise on hypothesis testing:

4. The following coded pairs of measurements were taken of the temperature (X) and thrust (Y) of a jet engine while it was being tested under uniform operating conditions.

x	15	20	25	26	30	33	34	35	38	39	41	46	49	52	57
y	1.4	1.2	1.9	1.6	2.5	2.1	2.4	1.5	2.3	2.7	1.8	2.2	2.8	3.4	3.2

You are given that $\sum x = 540$, $\sum x^2 = 21412$, $\sum y = 33.0$, $\sum y^2 = 78.54$, $\sum xy = 1276.6$.

(i) (a) Plot the data on a scatter diagram and comment on the suitability of the Pearson product-moment correlation coefficient, r, as a measure of the association between thrust and temperature.

(6)

(b) Calculate r for these data, and test at the 5% significance level the hypothesis of zero correlation against the alternative that thrust and temperature are positively correlated. State clearly (but do not prove) any formulae that you have used, and list the assumptions you have made in the test.

(6)

(ii) A colleague wishes to test at the 5% significance level the hypothesis of no trend against the alternative of an increasing trend, without assuming that the trend is necessarily linear. State what measure of association he should use, calculate it for the above data and carry out the desired test. Compare the result of this test with your findings in part (i).

(8)

Reference:

http://www.hkss.org.hk/images/exam/papers/Past/2008/2008_HC_4_HKSS.pdf