

- a) There are 10 pages including this page.
- b) You can use the backs of the question pages for rough work.
- c) There are 50 marks in total. The number of marks available for each part of each question is indicated.
- d) You will not be evaluated on the quality of your writing. Do not waste time writing lengthy answers when concise ones will do just as well.

FAMILY NAME:_____

GIVEN NAME:_____

STUDENT NUMBER:_____

The Story

A gas distribution company is concerned with the health of the gas meters being used by its industrial customers. There are over 20 000 such customers, which is too many to visit to examine each gas meter. So they select a sample of $n=400$ meters from the database and send technicians to visit only these meters. The technicians perform an analysis of each meter, which includes some testing, and record the following data (along with the meter ID)

- **age**: age of meter in months
- **max_kpa**: maximum test pressure in kPa
- **min_kpa**: minimum test pressure in kPa
- **tot_gas**: cumulative amount of gas in m^3 that meter has recorded
- **volts**: result of an electric test run through meter in V
- **rust**: amount of corrosion recorded as “low”, “medium”, “high”.
- **brand**: either brand “A” or “B”.

In this exam you will analyse the dataset in various ways.

1. (20 marks total) First you will have a look at the relationship between `rust` and `volts`. Here is a table that gives sample mean and sample variance for the `volts` readings within each level of `rust`:

Descriptive Statistics: `volts`

Total				
Variable	<code>rust</code>	Count	Mean	Variance
<code>volts</code>	<code>high</code>	93	-5.195	1.031
	<code>medium</code>	197	-5.5652	1.3638
	<code>low</code>	110	-5.788	1.641

A normal quantile plot of the `volts` readings shows a perfect straight line, as do each of normal quantile plots of the `volts` readings within each `rust` group.

Denote by μ_1, μ_2, μ_3 the mean values of `volts` within the “low”, “medium”, and “high” rust groups respectively.

- a) (6 marks) Perform the two-sample t-test for the question “is there any difference in mean `volts` between the “low” and “medium” rust levels?” Include a discussion of any model assumptions you need to make.

- b) **(5 marks)** Given the additional information that the sample variance of the **rust** readings is 1.4011, perform the analysis of variance for the question “are there any differences at all among mean **volts** for the different levels of rust?”. Include a discussion of any model assumptions you need to make.

(Hint: the “MSE” is 1.3629. You need to show how this is calculated to get full marks, but I include this number to let you continue if you get stuck on that point.)

- c) **(5 marks)** Supposing that you had wanted, even in advance of collecting the data, to estimate $\mu_1 - (\mu_2 + \mu_3)/2$. Do so now using a 95% confidence interval.

- d) (**4 marks**) Perform all pairwise comparisons of the mean **volts** readings within different levels of rust, at the experiment-wise error rate of 0.05. (Lots of space here but not because you need it...that's just the way it worked out.)

2. (15 marks) The simple regression model with `volts` as the response and `max_kpa` as the input, or “ x variable”, is fit resulting in the following output with some numbers removed. Recall that the sample variance of the `volts` readings is 1.4011 and the sample size is 400.

The regression equation is

`volts = - 15.8 + ***** max_kpa`

Predictor	Coef	SE Coef	T	P
Constant	-15.761	1.055	-14.94	0.000
max_kpa	*****	0.004506	****	*****

S = ***** R-Sq = 19.13%

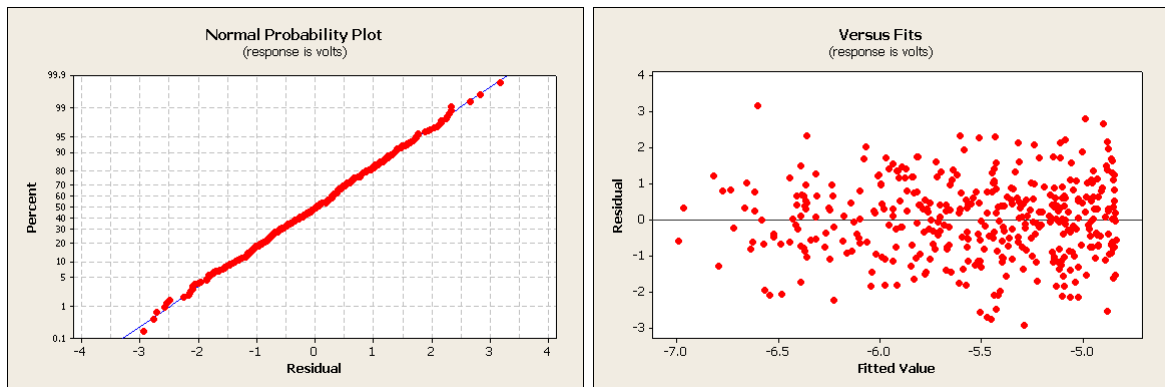
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	***	*****	*****	*****	*****
Residual Error	***	*****	*****		
Total	***	*****			

- a) (5 marks) Show that the correct conclusion to be drawn from the hypothesis test of $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ is that there is “overwhelming evidence against the null hypothesis.”

- b) **(3 marks)** No plot was given of the data. Draw a simple sketch of what the scatterplot of the data would like like given the correct conclusion to be drawn from a) and the fact that R^2 is “only” 19.13%.

- c) **(4 marks)** Here are the normal plot of the residuals and the plot of the residuals versus fitted values. In addition, Minitab gives a list of 27 possible outliers in which the smallest studentized residual is -2.93 and the largest is 3.12. Comment on the extent to which the model assumptions have been satisfied and on the possible existence of outliers.



- d) **(3 marks)** A meter outside the sample of 400 being studied here is chosen at random and found to have a `max_kpa` reading that happens (by sheer luck) to be precisely the sample average of the readings in the dataset. Produce a 95% prediction interval for the `volts` reading for this meter.

3. (10 marks total) The categorical variables `rust` and `brand` are going to be transformed into so-called dummy variables (or “indicator variables” or “0-1 variables”) so that they can be included in a regression model. The new variables called `rust_low`, `rust_medium`, and `brand_A` will be added that take on the values 0 or 1 depending on the actual values of `rust` and `brand` in the dataset according to the following tables:

<code>rust</code>	<code>rust_low</code>	<code>rust_medium</code>	<code>brand</code>	<code>brand_A</code>
low	1	0	A	1
medium	0	1	B	0
high	0	0		

Here are the correlations among the variables that may be included in a regression model in which `volts` remains the output variable:

	<code>age</code>	<code>max_kpa</code>	<code>min_kpa</code>	<code>tot_gas</code>	<code>volts</code>	<code>rust_low</code>	<code>rust_medium</code>
<code>max_kpa</code>	0.077						
<code>min_kpa</code>	-0.032	-0.771					
<code>tot_gas</code>	0.881	0.077	-0.032				
<code>volts</code>	0.070	0.437	-0.488	0.183			
<code>rust_low</code>	-0.311	0.004	-0.009	-0.287	-0.129		
<code>rust_medium</code>	0.039	-0.040	0.004	0.034	-0.021	-0.607	
<code>brand_A</code>	0.010	0.008	0.015	-0.001	-0.031	-0.007	-0.012

- a) **(4 marks)** What would you expect to see if you were to try to include both `max_kpa` and `min_kpa` at the same time in a regression model with output variable `volts`? Include no more than three different possibilities.

- b) **(4 marks)** Recall that the sample variance of the `volts` readings is 1.4011 and the sample size is 400. A possible regression model includes the following input variables: `age`, `max_kpa`, `min_kpa`, `tot_gas`, `rust_low`, `rust_medium`. The R^2 value is 33.86%. Perform the hypothesis test that answers the question “is there any significant linear relationship between the inputs and the output?”
- c) **(2 marks)** If it possible to do so using the table of correlations, state which of the variables from b) would be the most likely candidate to be removed from the model as part of a general stepwise model selection strategy in which variables can be excluded for having high p-values. If it is not possible, state why not.

4. (5 marks total) We'll now consider the variables **rust** and **brand** just by themselves. Here is the contingency table for these two variables with observed cell counts given along with all but one expected cell counts.

Rows: rust	Columns: brand		
	A	B	All
low	33	77	110
	33.55	*****	
medium	59	138	197
	60.09	136.91	
high	30	63	93
	28.36	64.64	
All	122	278	400

a) (2 marks) Provide a 95% confidence interval for the proportion of meters with “low” level of rust, being sure to also comment (very briefly!) on any required assumptions for the calculation to be accurate.

b) (3 marks) Perform the hypothesis test that answers the question “are the rows and columns independent?” being sure to also comment (very briefly!) on any required assumptions for the calculation to be accurate.