Final exam, 12 December 2022

All aids are allowed, except a computer and personal assistance. Restricted use of some computer-like devices (including tablets and smartphones) is permitted under the rules described at the VHM 801 course homepage. The exam consists of three questions that should all be answered. The weights for each of the three questions and also for each subquestion within a question are indicated; these weights total *50 points*. Note that questions, and often also subquestions, can be answered independently of each other. The duration of the exam is 3 hours.

Generally, **statistical models and methods should be specified**, and every statistical analysis should be summarized in a conclusion. Throughout, if you realize that you need more information than is provided to carry out an analysis, specify what information you need, how you would obtain it using statistical software, and how you would use it in the analysis.

Question 1. (*15 points*)

According to superstitions rooted in parts in Christian culture, bad luck may be associated with both the number 13 and with Fridays, and thus in particular with Friday the 13th. A group of British researchers investigated whether they could determine any differences in specific behaviours as well as in "unlucky" accidents on Friday 13th when comparing to the preceding Friday 6th during three years in the beginning of the 1990s. Calendar occurrences of a Friday 13th (and a preceding Friday 6th) within the time period were included in the study, except when any of the two days fell on a holiday; the actual dates can be read off the table in part **c)**.

The outcomes studied were shopping, driving, and accidents in a particular region to the south of London. The numbers of shoppers were determined from records of nine stores (Sainsburys supermarkets) in the area. The numbers of drivers were determined from traffic counts at two highway junctions in the area. The number of accidents were determined from emergency admissions to hospitals for accidents subdivided by type. Extracts of the data for shoppers, drivers and accidents are presented in the subquestions below. Note that all subquestions can be answered independently of each other.

a) (*2 points*)
Explain the study type (e.g., observational or experimental) and the statistical design (e.g., 1-sample or other designs) for comparing the behaviours involved in the outcomes on Friday 6th and Friday 13th. You may discuss this broadly or focus on one of the datasets (of your own choice), e.g. the shoppers data in the Minitab listing on page 3 or the drivers data in the table in part **c)** on page 2. Characterise also the target population for the statistical inference.

b) *Shoppers data* (*5 points*)
The analysis reported in the paper was based on the total number of shoppers per store across all the dates of Friday 6th and Friday 13th. The Minitab listing gives graphical summaries for these two counts as well as the differences between the counts on the 6th

and 13th. Use this information to compute a parametric test to compare the mean total number of shoppers on the two Fridays, and draw conclusions; make sure to include the statistical model your analysis is based on. Include in your interpretation also the magnitude of any difference in the mean number of shoppers on the two dates.

In the paper, the authors used exclusively non-parametric tests because "a normal distribution could not be assumed for any of the data". Discuss (critically) the authors' justification for the use of a non-parametric procedure for this analysis — do you think it is adequate?

c) *Drivers data* (*3 points*)
The table below gives the total traffic flows (that is, vehicle counts) across the two junctions for the five instances of Friday 6th and Friday 13th during the period studied.

Date	Friday 6th	Friday 13th
July 1990	273 258	271 456
September 1991	270 787	267 861
December 1991	244 691	240 364
March 1992	252 924	245 781
November 1992	242 193	240 033

Use a non-parametric test to compare the counts of vehicles on Friday 6th and Friday 13th, in terms of which of the two Fridays in a given month had the higher traffic flow. Make sure to state statistical model as well as your null and alternative hypotheses explicitly, and draw conclusions from your calculated test.

If you are unable to calculate a test from the information provided, explain what extra information you need and how you would use it, or alternatively how you would calculate the test in Minitab (or another statistical software) based on the values in the table. In this case, give enough detail about the method you propose to allow someone with the data at hand to carry out the calculation.

d) *Accidents data I* (*3 points*)
Among the counts of accidents of different types reported were (so-called) transport accidents. A total of 110 transport accidents were reported across all the Fridays; of these, 45 accidents happened on Friday 6th, and 65 accidents happened on Friday 13th. Use these numbers to estimate the probability that a Friday accident happened on the 13th (rather than on the 6th), and compute a 95% confidence interval for this probability. Based on the confidence interval, would you say that transport accidents were equally likely on Friday 13th compared to Friday 6th?

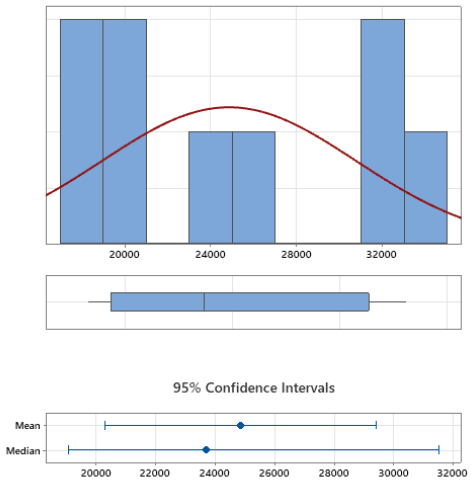
e) *Accidents data II* (*2 points*)
Only four accidents involving animals were reported, but three of them occurred on a Friday 13th. If it was in fact equally likely that such accidents occurred on Friday 6th and Friday 13th dates, what is the probability that most of the four accidents had occurred on a Friday 13th?

Minitab listings and graphical displays for Question 1:

Data

Row	store	shop6th	shop13th	diff13min6
1	Chichester	26564	26832	268
2	Crystal Palace	20611	20964	353
3	Dorking	18278	18371	93
4	East Grinstead	19857	19967	110
5	Epsom	23718	23634	-84
6	Guildford	31031	31194	163
7	Horsham	18843	19216	373
8	Lewisham	31660	32265	605
9	Nine Elms	33111	33322	211

Summary Report for shop6th



Anderson-Darling Normality Test

A-Squared	0.46
P-Value	0.197
Mean	24853
StDev	5908
Variance	34902318
Skewness	0.32418
Kurtosis	-1.84804
N	9
Minimum	18278
1st Quartile	19350
Median	23718
3rd Quartile	31346
Maximum	33111

95% Confidence Interval for Mean

20311	29394

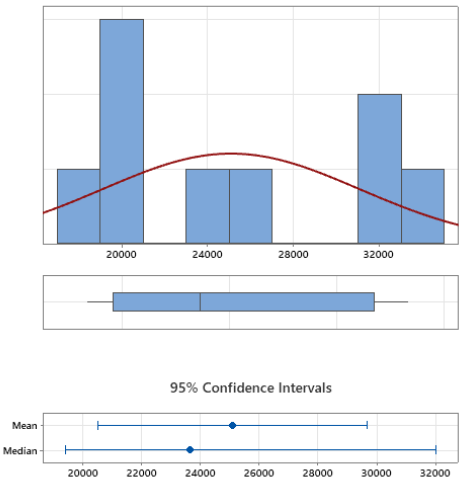
95% Confidence Interval for Median

19074	31517

95% Confidence Interval for StDev

3990	11318

Summary Report for shop13th



Anderson-Darling Normality Test

A-Squared	0.46
P-Value	0.197
Mean	25085
StDev	5964
Variance	35572173
Skewness	0.34120
Kurtosis	-1.84909
N	9
Minimum	18371
1st Quartile	19592
Median	23634
3rd Quartile	31730
Maximum	33322

95% Confidence Interval for Mean

20500	29670

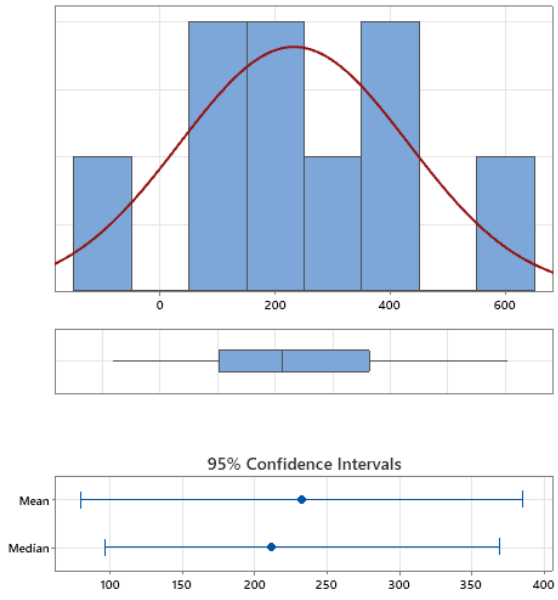
95% Confidence Interval for Median

19387	32021

95% Confidence Interval for StDev

4029	11426

Summary Report for diff13min6



Anderson-Darling Normality Test

A-Squared	0.18
P-Value	0.878
Mean	232.44
StDev	198.18
Variance	39276.03
Skewness	0.406832
Kurtosis	0.728337
N	9
Minimum	-84.00
1st Quartile	101.50
Median	211.00
3rd Quartile	363.00
Maximum	605.00

95% Confidence Interval for Mean

80.11	384.78

95% Confidence Interval for Median

96.87	368.44

95% Confidence Interval for StDev

133.86	379.67

Question 2. (*15 points*)

In an agricultural experiment on the fight against a particular plant disease in barley, six spraying treatments were tested against a control group. The treatments were labeled $1-7$ where 1 is the control. They were applied to plants grown in plots in a field trial. The plots contained variable numbers of plants, for reasons not related to the efficacy of the treatments against the disease. In the later part of the growth season, the plants in each plot were inspected and it was recorded whether any lesions typical for the plant disease were present. The numbers of plants with and without lesions are given for each treatment group in the table below.

Plants	Treatment						
Lesions	1	2	3	4	5	6	7
no	3	10	11	2	9	26	25
yes	17	7	11	11	15	22	29

For example, among the plants subjected to treatment no. 5 there were 9 plants without lesions and 15 plants with lesions. The interest is in identifying treatments that make the plants less susceptible to lesions. The data in the table have been entered into a worksheet with 14 rows and the variables: `plants`, `lesions` and `treatment`.

a) (*6 points*)
The next page shows Minitab listings for 3 analyses of these data. For each of the 3 analyses, identify the type of statistical model/analysis and discuss whether it addresses the question of interest in a meaningful way. A full specification of the statistical models and their assumptions is not required for this part, but your description of the model and analysis method should have enough detail to uniquely identify the method.

b) (*4 points*)
Select the most meaningful model and analysis among those reviewed for part **a)**, and describe and critically review the model's assumptions. Carry out a statistical test to assess any differences between the 7 treatments (including the control), and draw conclusions. If any such differences are seen, compute a measure (or statistic) for each treatment's performance and order the treatments from the worst to the best; if no such difference is seen, compute only a measure of the overall treatment performance.

c) (*5 points*)
The researchers were particularly interested in comparing each of the spraying treatments to the control; this focus was decided prior to collecting and analysing the data. Irrespective of your results in **b)**, continue the analysis by assessing the significance of each of the treatments relative to the control. For this part, an overall (or simultaneous) significance level of at most 5% is desirable — discuss how you can achieve that.
(*Hint:* You may omit the calculation of tests for some treatments against the control if you already (from other information) know what their outcome with regard to signifance will be.)

If you are unable to calculate tests from the information provided, explain what extra information you need and how you would use it, or alternatively how you would calculate the tests in Minitab (or another statistical software) based on the information provided. In this case, be as specific as possible about how you will want to carry out the analysis.

Minitab listings and graphical displays for Question 2:

PLANTS.MTW

Regression Analysis: plants versus treatment

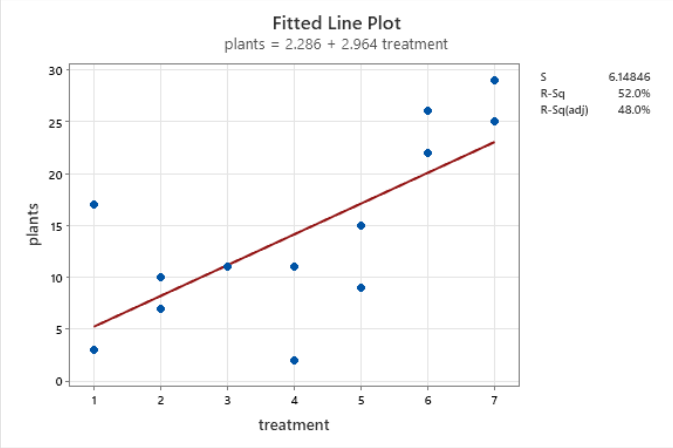
The regression equation is
plants = 2.286 + 2.964 treatment

Model Summary

S	R-sq	R-sq(adj)
6.14846	52.03%	48.03%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	492.071	492.071	13.02	0.004
Error	12	453.643	37.804		
Total	13	945.714			



PLANTS.MTW

One-way ANOVA: plants versus treatment

Factor Information

Factor	Levels	Values
treatment	7	1, 2, 3, 4, 5, 6, 7

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
treatment	6	768.7	128.12	5.07	0.026
Error	7	177.0	25.29		
Total	13	945.7			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
5.02849	81.28%	65.24%	25.14%

Means

treatment	N	Mean	StDev	95% CI
1	2	10.00	9.90	(1.59, 18.41)
2	2	8.50	2.12	(0.09, 16.91)
3	2	11.00	0.00	(2.59, 19.41)
4	2	6.50	6.36	(-1.91, 14.91)
5	2	12.00	4.24	(3.59, 20.41)
6	2	24.00	2.83	(15.59, 32.41)
7	2	27.00	2.83	(18.59, 35.41)

Pooled StDev = 5.02849

PLANTS.MTW

Tabulated Statistics: lesions, treatment

Using frequencies in plants

Rows: lesions Columns: treatment

	1	2	3	4	5	6	7	All
0	3	10	11	2	9	26	25	86
	8.69	7.38	9.56	5.65	10.42	20.85	23.45	
1	17	7	11	11	15	22	29	112
	11.31	9.62	12.44	7.35	13.58	27.15	30.55	
All	20	17	22	13	24	48	54	198

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	15.544	6	0.016
Likelihood Ratio	16.931	6	0.010

Question 3. (*20 points*)

A researcher wanted to study the relationship between the body weight (`w`) and the weight of flight muscles (`f`) in birds, and retrieved values (both recorded in grams) for these variables for 14 species of birds. A list of the data values, together with the values obtained by natural log transformations (`lnw` and `lnf`, respectively), is shown in the Minitab listing on the next page.

a) (*4 points*)
As a first step in exploring the data, graphs were obtained of the two variables plotted against each other. Such graphs are shown on the next page on both original and log-transformed scales. For one of the graphs (of your own choice), describe the relation between the variables, interpret the provided value of the r statistic, and give a statistical assessment of whether the two variables can be considered as independent. Contrast also (briefly) the patterns of points in the two graphs — how are they different?

Because it is easier to measure body weight of birds than the weight of their flight muscles, there was interest in using the data to establish an equation to predict the weight of flight muscles from the body weight.

b) (*4 points*)
The Minitab listings for parts **b)-e)** give selected outputs from four statistical models fitted to the data, each on a separate page. Identify the type of models used and describe how the four models differ. Use the information provided to select one of the four models/analyses, which you think best both matches the stated purpose of analysis and meets the model assumptions. Justify your choice of model/analysis, e.g. by briefly describing why the other models/analyses are of less interest.

c) (*5 points*)
For your chosen model/analysis from **b)**, give estimates and, if possible, also confidence intervals for the model parameters. Explain the meaning of the parameters. Comment also briefly on the strength and statistical significance of the relationship between the variables.

d) (*4 points*)
A species of hummingbirds has a body weight of 2.7 g. Use your selected model from **b)** to predict the weight of the flight muscles for this species. Compare your predicted value with the actual weight of 1.0 g for the flight muscles of this species. Describe how you would use statistical inference for this comparison (calculations are not possible from the information provided). Do you see any issues (or concerns) with this particular prediction?

e) (*3 points*)
One possible relation that has been suggested between the two weights, is that the flight muscles constitute the same proportion of the body weight across different species of birds. With our previous notation of `w` and `f` for the body weight and the weight of flight muscles, respectively, this relation can be described by the equations,

$$\texttt{f} = k \cdot \texttt{w}, \quad \text{or} \quad \ln(\texttt{f}) = \ln(k) + \ln(\texttt{w}).$$

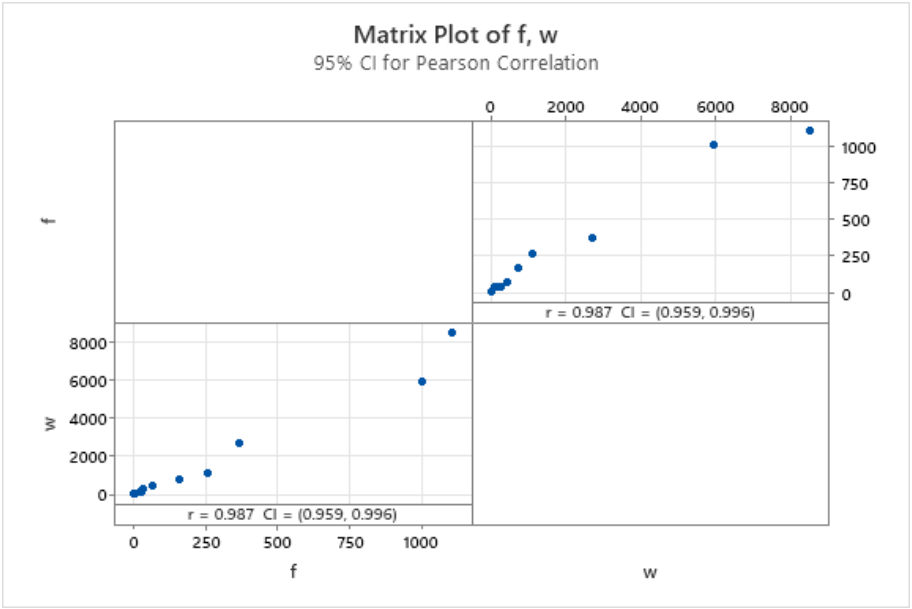
Use the information provided (for any of the models) to assess, by means of a statistical test, whether the data support such a relationship. If you find that the data support the

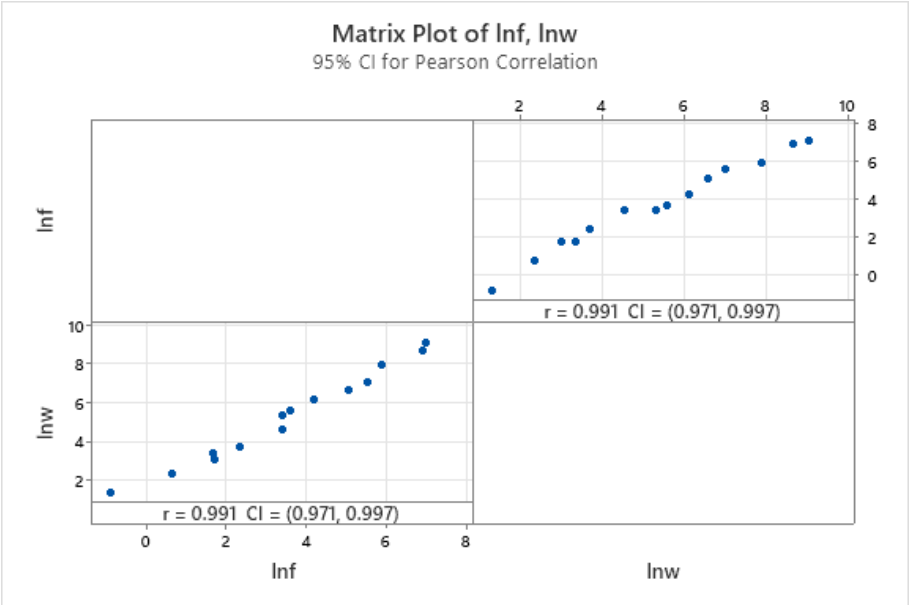
relationship, estimate also the proportionality factor k from the data, and supplement your estimate with a 95% confidence interval.

Minitab listings and graphical displays for Question 3, part a):

Data

Row	w	f	lnw	lnf
1	3.7	0.41	1.31	-0.89
2	10.5	1.97	2.35	0.68
3	20.7	5.64	3.03	1.73
4	28.5	5.37	3.35	1.68
5	41.3	10.50	3.72	2.35
6	95.0	30.00	4.55	3.40
7	209.0	30.00	5.34	3.40
8	270.0	37.00	5.60	3.61
9	455.0	66.00	6.12	4.19
10	728.0	161.00	6.59	5.08
11	1108.0	257.00	7.01	5.55
12	2697.0	369.00	7.90	5.91
13	5943.0	1002.00	8.69	6.91
14	8519.0	1108.00	9.05	7.01





Minitab listings and graphical displays for Question 3, parts b)-e):

BIRDS.MTW

Regression Analysis: w versus f

Regression Equation

w = -91 + 6.938 f

Coefficients

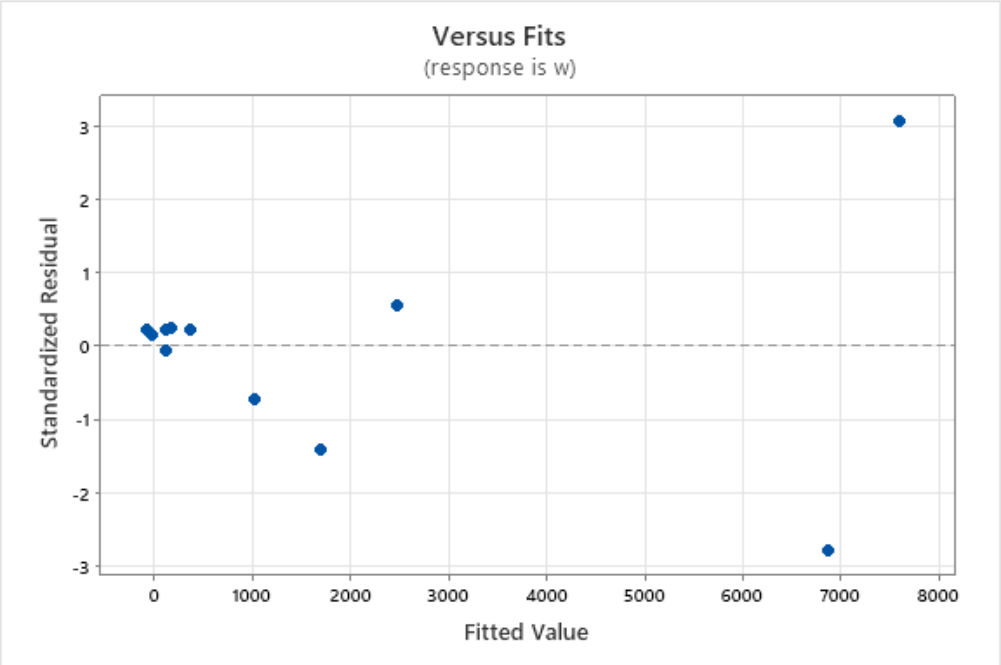
Term	Coef	SE Coef	T-Value	P-Value
Constant	-91	135	-0.67	0.516
f	6.938	0.323	21.50	0.000

Model Summary

S	R-sq	R-sq(adj)
431.569	97.47%	97.26%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	86084423	86084423	462.19	0.000
Error	12	2235020	186252		
Total	13	88319442			



BIRDS.MTW

Regression Analysis: f versus w

Regression Equation

f = 18.3 + 0.14048 w

Coefficients

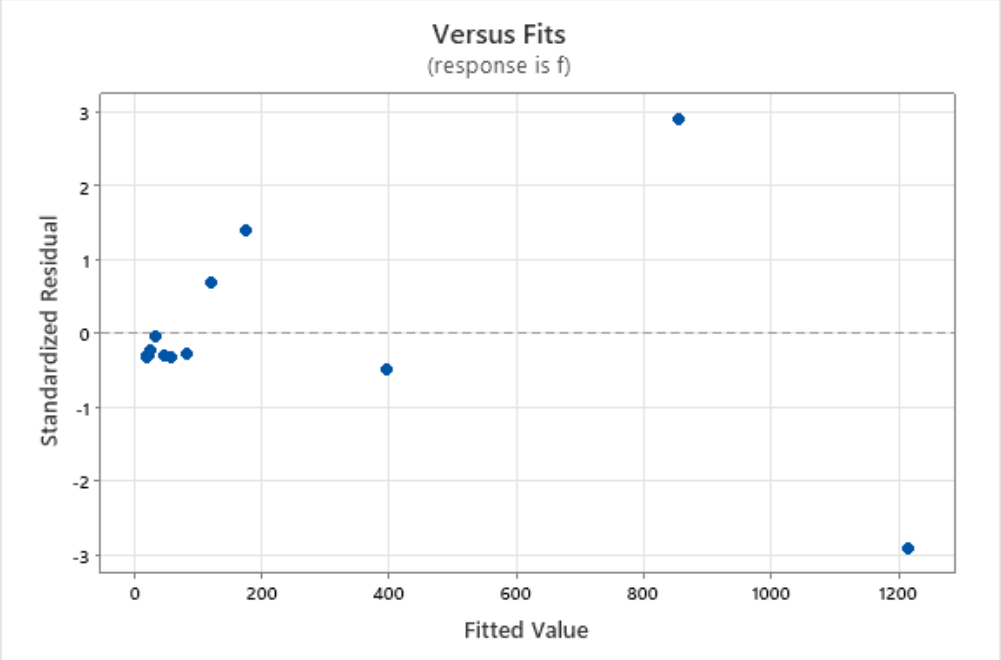
Term	Coef	SE Coef	T-Value	P-Value
Constant	18.3	18.9	0.97	0.352
w	0.14048	0.00653	21.50	0.000

Model Summary

S	R-sq	R-sq(adj)
61.4085	97.47%	97.26%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1742934	1742934	462.19	0.000
Error	12	45252	3771		
Total	13	1788186			



BIRDS.MTW

Regression Analysis: lnw versus lnf

Regression Equation

lnw = 1.684 + 1.0084 lnf

Coefficients

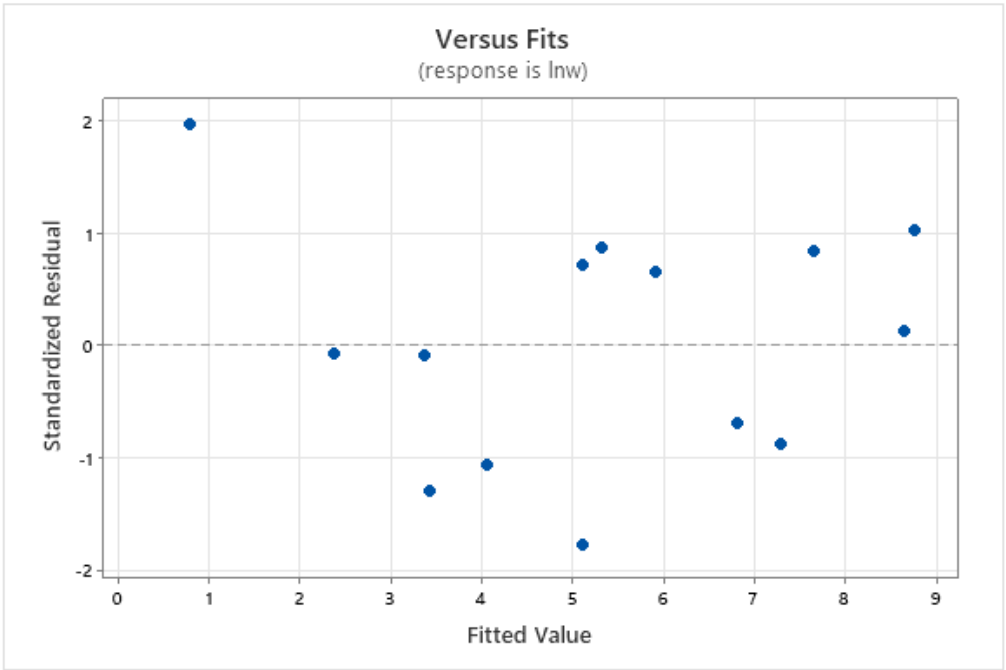
Term	Coef	SE Coef	T-Value	P-Value
Constant	1.684	0.166	10.12	0.000
lnf	1.0084	0.0390	25.86	0.000

Model Summary

S	R-sq	R-sq(adj)
0.330363	98.24%	98.09%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	72.982	72.9820	668.70	0.000
Error	12	1.310	0.1091		
Total	13	74.292			



BIRDS.MTW

Regression Analysis: lnf versus lnw

Regression Equation

lnf = -1.577 + 0.9742 lnw

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	-1.577	0.219	-7.21	0.000
lnw	0.9742	0.0377	25.86	0.000

Model Summary

S	R-sq	R-sq(adj)
0.324709	98.24%	98.09%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	70.505	70.5053	668.70	0.000
Error	12	1.265	0.1054		
Total	13	71.771