This quiz has been regraded; your score was affected.

2020 Quiz 3

Due No due date **Points** 24.5 **Questions** 22 **Time Limit** None

Instructions

This quiz is open from 1pm AEST to 3pm AEST.

You may refer to your lecture notes in answering the quiz.

Answers must be your own work.

Plagiarism declaration

By submitting work for this quiz I hereby declare that I understand the University's policy on academic integrity (https://academicintegrity.unimelb.edu.au/) and that the work submitted is original and solely my work, and that I have not been assisted by any other person (collusion) apart from where the submitted work is for a designated collaborative task, in which case the individual contributions are indicated. I also declare that I have not used any sources without proper acknowledgment (plagiarism). Where the submitted work is a computer program or code, I further declare that any copied code is declared in comments identifying the source at the start of the program or in a header file, that comments inline identify the start and end of the copied code, and that any modifications to code sources elsewhere are commented upon as to the nature of the modification.

Attempt History

	Attempt	Time	Score	Regraded
LATEST	Attempt 1	117 minutes	21.17 out of 24.5	22.17 out of 24.5

Correct answers are no longer available.

Score for this quiz: **22.17** out of 24.5
Submitted Jul 21 at 14:57
This attempt took 117 minutes.

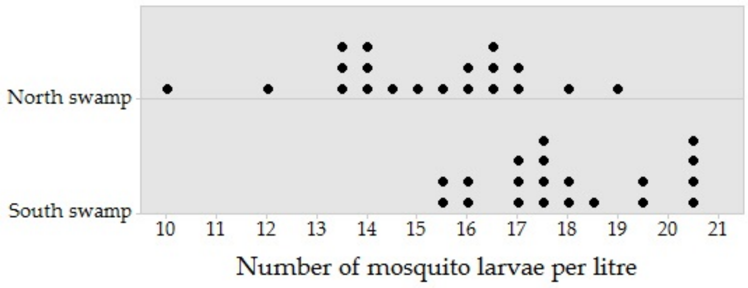
Question 1 1 / 1 pts

Consider samples of mosquito larvae densities for two swamps, where sample densities at a number of points in the two swamps were measured. The densities are the number of larvae per litre.

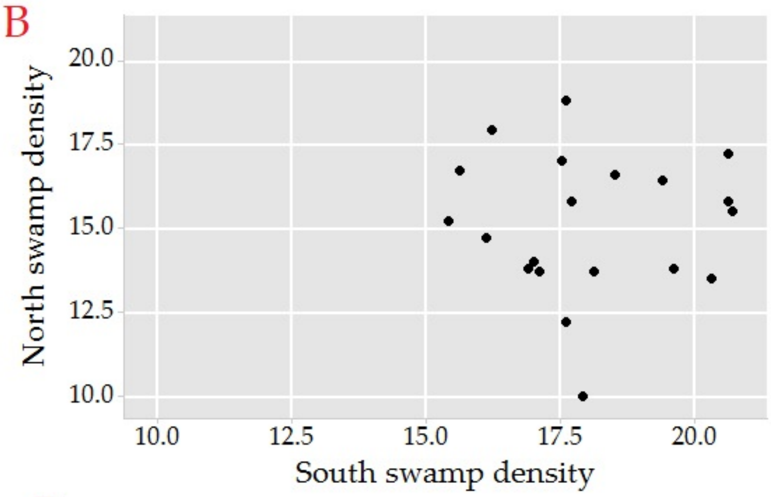
The data are stored as mosquito.mwx, which you should examine.

Which one of the plots below is most suitable for examining the question of differences in density between the two swamps, and representing the data clearly?

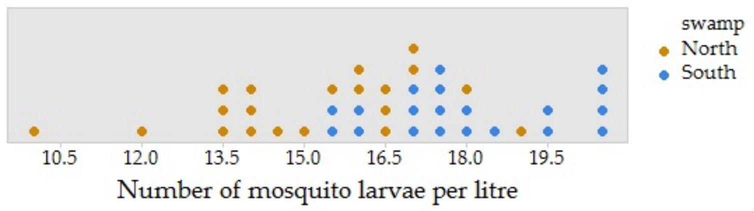
A



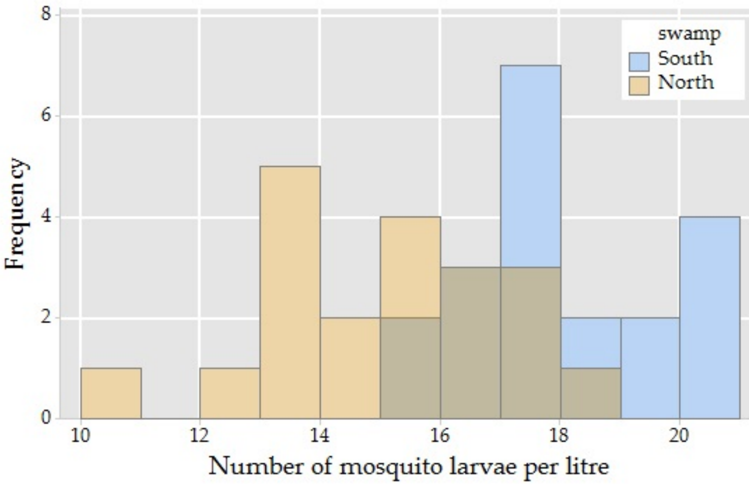
B



C



D



○ C

● A

○ D

○ B

The dot plot on the top left is best. We are interested in the differences between the swamps, and the data from each group is shown most clearly with separate dot plots.

Question 2

Original Score: 0 / 1 pts Regraded Score: 1 / 1 pts

This question has been regraded.

Carry out an appropriate analysis to draw inferences about the mean differences in density between the two swamps (North minus South).

Which one of the following is the best summary of the appropriate analysis?

○ The estimated mean difference was 2.91 larve/litre with a 95% confidence interval (1.69, 4.12); *P* < 0.001 for the test of the null hypothesis of no true difference of means.

○ The estimated mean difference was 2.91 larve/litre with a 95% confidence interval (1.64, 4.17); *P* < 0.001 for the test of the null hypothesis of no true difference of means.

○ The estimated mean difference was -2.91 larve/litre with a 95% confidence interval (-4.17, -1.64); *P* < 0.001 for the test of the null hypothesis of no true difference of means.

● The estimated mean difference was -2.91 larve/litre with a 95% confidence interval (-4.12, -1.69); *P* < 0.001 for the test of the null hypothesis of no true difference of means.

The appropriate analysis is based on two independent sample t-procedures; this produces the results -2.91 larve/litre with a 95% confidence interval (-4.12, -1.69) for estimating the mean difference North minus South.

Question 3

1 / 1 pts

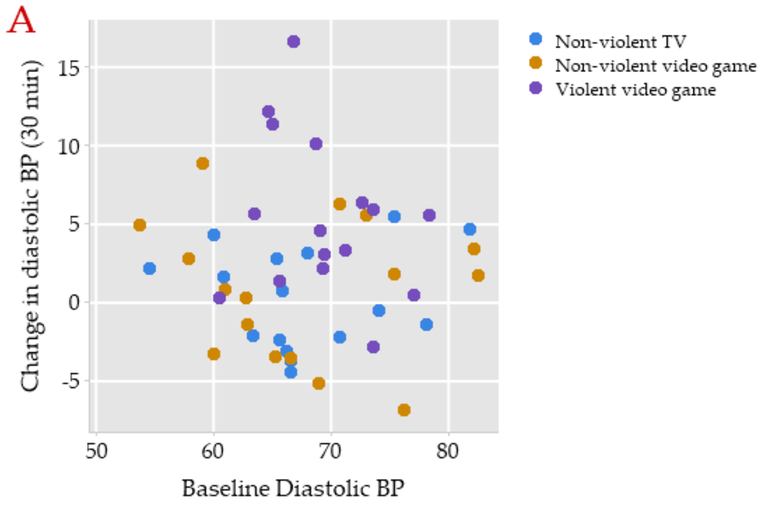
The data stored in videogames.mwx are from an experiment that investigated the effect of activity type on changes in blood pressure.

Blood pressure was measured before commencing the activity and then 30 and 60 minutes into the activity. The change from baseline in diastolic blood pressure was recorded for 30 minutes and 60 minutes into the activity.

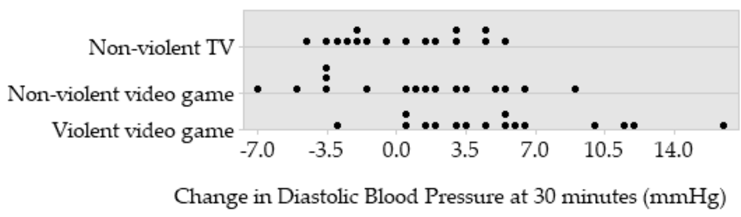
There were three different activity groups: watching non-violent TV, playing a non-violent video game, and playing a violent video game.

Which visual display is most useful for investigating the effect of different activities on the change in diastolic blood pressure after 30 minutes?

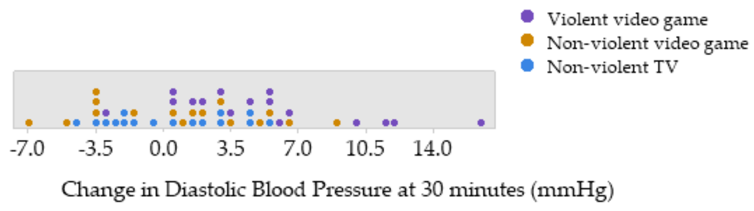
A



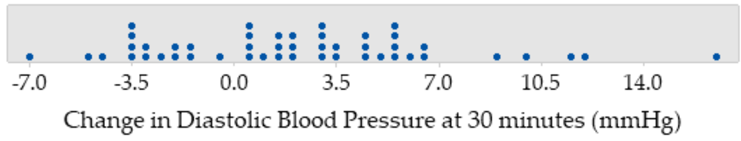
B



C



D



○ C

○ D

○ A

● B

We are interested in making comparisons between types of activity in relation to changes in blood pressure. The top right graph shows the relevant data clearly.

Question 4

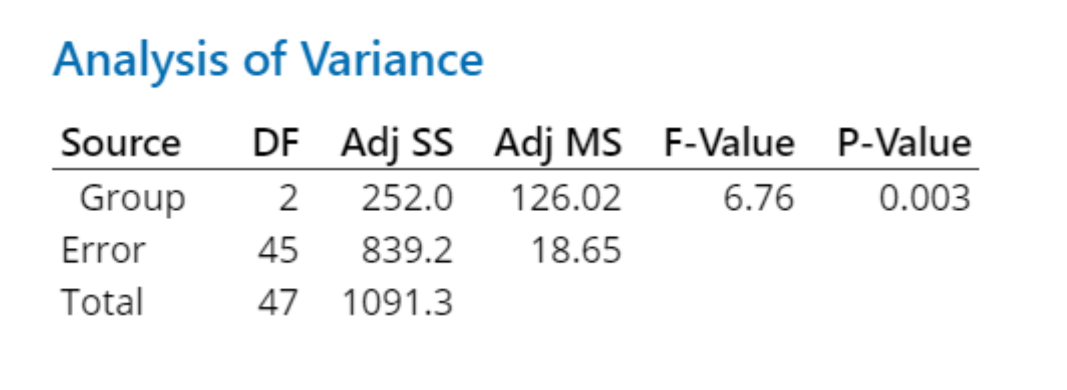
1 / 1 pts

Fit a General Linear Model to test the effects of the 3 activity types on diastolic blood pressure change at 30 minutes.

Report the *P*-value for the test of the effects of activity type on diastolic blood pressure change at 30 minutes to three decimal places.

0.003

The P-value of 0.003 is shown in the output below.



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Group	2	252.0	126.02	6.76	0.003
Error	45	839.2	18.65		
Total	47	1091.3			

Question 5

1 / 1 pts

Consider further your results from the analysis testing the effects of the 3 types of activity on change in diastolic blood pressure.

Which one of the following is a correct report of the test statistic?

○ F($_{2, 47}$) = 6.76

● F($_{2, 45}$) = 6.76

○ F($_{2}$) = 6.76

○ P = 0.003

An appropriate report of the test statistic includes the degrees of freedom for the explanatory variable (group) and error, and the value of the test statistic. From the output in Question 9, the correct report is $F_{(2, 45)} = 6.76$.

Question 6

1 / 1 pts

Which of the following conclusions is most accurate, based on the *P*-value?

○ The null hypothesis of no population mean differences in the change in diastolic blood pressure between the three types of activity is false.

○ The observed pattern of means is consistent with the null hypothesis of no population mean differences in the change in diastolic blood pressure between the three types of activity.

○ The probability that the null hypothesis of no population mean differences in the change in diastolic blood pressure between the three types of activity is false is large (much greater than 0.05).

● The observed pattern of means is not consistent with the null hypothesis of no population mean differences in change in diastolic blood pressure between the three types of activity.

○ The probability that the null hypothesis of no population mean differences in the change in diastolic blood pressure between the three types of activity is true is quite small (much less than 0.05).

○

The null hypothesis of no population mean differences in the change in diastolic blood pressure between the three types of activity is true.

The P-value does not indicate if the null hypothesis is true or false.

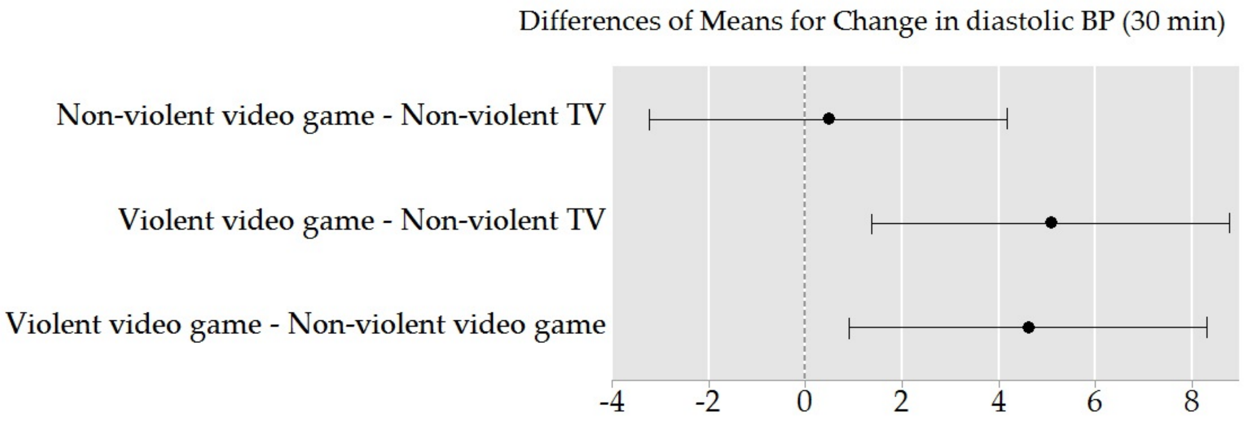
The P-value does not indicate the probability that the null hypothesis is true or false.

Given that the P-value is small, in this case, it suggests that the observed result is not consistent with the null hypothesis.

Question 7

1.5 / 1.5 pts

Consider the graph below, obtained in further analysis of the experiment about the effects of type of activity on change in diastolic blood pressure after 30 minutes.



Which of the following are true? (Tick as many as apply.)

☑

The graph shows Tukey simultaneous 95% confidence intervals, rather than Fisher intervals.

☐

The P-values will be less than 0.05 for all of the pairwise comparisons.

☑

The mean effects on change in diastolic blood pressure primarily relate to the violence of the activity.

□

The mean effects on change in diastolic blood pressure primarily relate to the whether or not video games are being played.

The P-value for the comparison of Non-violent video games and Non-violent TV will be greater than 0.05, as the confidence interval includes zero.

The strongest differences are between Violent and Non-violent activities.

If you carry out the analysis, and check the pairwise comparisons, you can confirm that the confidence intervals on the graph correspond to the Tukey intervals.

Question 8

1 / 1 pts

Consider the data in starch.mwx.

They come from an experiment that investigated preparation methods for denim fabric. Before denim is made into jeans, it is desirable to remove as much starch as possible, to reduce the stiffness of the garments produced.

The experiment investigated three different additives to the washing water:

Alpha amalyze (an enyzme that eats starch)

Caustic soda (a chemical that destroys starch)

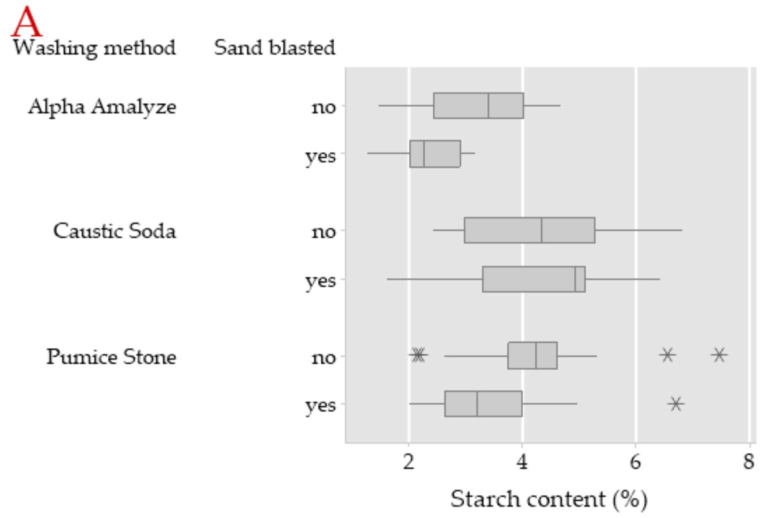
Pumice stone (a physical abrasive)

Additionally, the effect of sand blasting the fabric after washing was investigated.

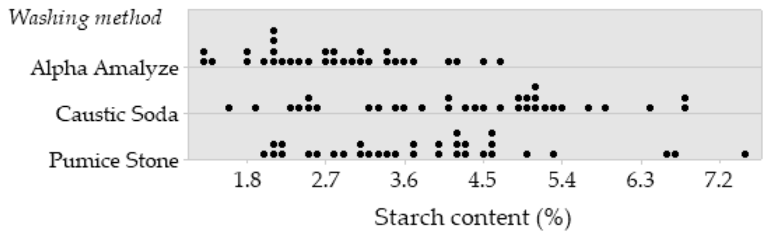
Nearly 100 rolls of fabric were treated with a combination of washing method and sand blasting (yes or no). After treatment, the starch percentage was measured. A low percentage is desirable.

Which of the following graphs is most useful for investigating the question of how to best prepare denim fabric?

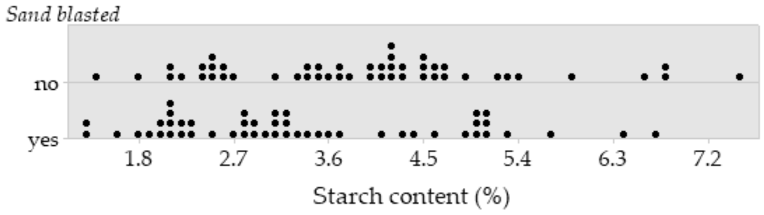
A



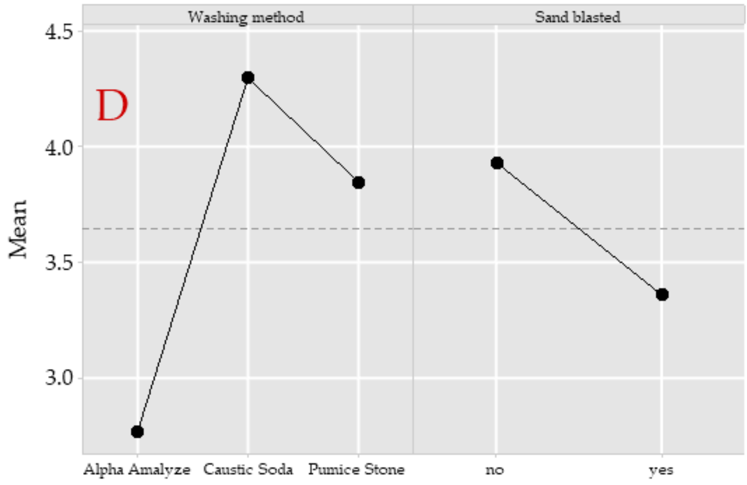
B



C



D



○ D

○ B

○ C

● A

Given that there are two factors of interest (washing method and sand blasting) the interaction is potentially important and of interest. The only graph that represents the interaction is the one containing the boxplots.

Incorrect

Question 9

0 / 1 pts

Fit a General Linear Model to the data to investigate the effects of washing method and use of sand blasting on starch content.

You may assume that the relevant assumptions of the model you are fitting are satisfied.

In the list below, which effects have P-values < 0.05, according to this model? (Tick as many as apply.)

☐ The interaction of washing method and sand blasting.

☑ The null hypothesis.

☐ The alternative hypothesis.

☑ The main effect of washing method.

In this design the standard approach to analysis is to fit a linear model that includes each of the two factors as main effects, and the interaction between them. When this is done, it is found that the interaction term is not statistically significant (F = 1.32, df = 2,92, P = 0.27) but each of the two main effects have small P-values (P < 0.05).

Question 10

1.5 / 1.5 pts

Fit a General Linear Model with main effects but no interaction term to the starch data. Again, assume that the assumptions are satisfied.

Carry out some additional analysis to examine the effects of the various treatments further.

Remember that lower levels of starch are preferable.

Which of the following are correct, based on this model? (Tick as many as apply.)

☐

Caustic soda has the best average outcome of all the washing treatments considered.

☐

The effect of Pumice stone and Caustic soda on the average percentage of starch is the same.

☑

On average, Alpha Amalyze is estimated to be 1% better in terms of starch reduction than Pumice stone

☑

Sand blasting is estimated to reduce the starch content by about 0.6% on average, compared with not using sand blasting.

The mean difference between sand-blasting and no sand-blasting is a reduction of about 0.6%.

Alpha Amalyze is 1% lower than Pumice stone, on average.

Caustic soda is the worst outcome on average of the washing methods; higher mean percentage of starch is worse.

The P-value for a pairwise comparison of caustic soda and pumice stone is > 0.05; however this does not mean that the true effects are the same.

You can examine the pairwise comparisons to confirm these conclusions.

Question 11

1.5 / 1.5 pts

You are asked to make recommendations about the best preparation of denim, based on analysis of these data you have carried out.

Which of the following recommendations are supported by the analysis? (Tick as many as apply.)

☐

Use Alpha Amalyze or Sand blasting but not both as the combination of methods is not important.

☑

Use Alpha Amalyze and Sand blasting, as both reduce average starch content, relative to their comparators.

☐

Use Caustic Soda as it is the best of the washing methods. It doesn't matter whether or not you use sand blasting.

☑ Use Sand blasting as it is better than no Sand blasting.

Both factors show statistically significant differences. The interaction is not at all statistically significant, hence we consider an additive model as suitable here. For Washing method, Alpha Amalyze is better than each of the other two treatments. For Sand blasting, sand blasting is better than no sand blasting. So it is best to use Alpha Amalyze and sand blasting.

Question 12

1 / 1 pts

A study in New Zealand investigated the use of a fungal infection in the reduction of a weed in pasture grazed by farm animals.

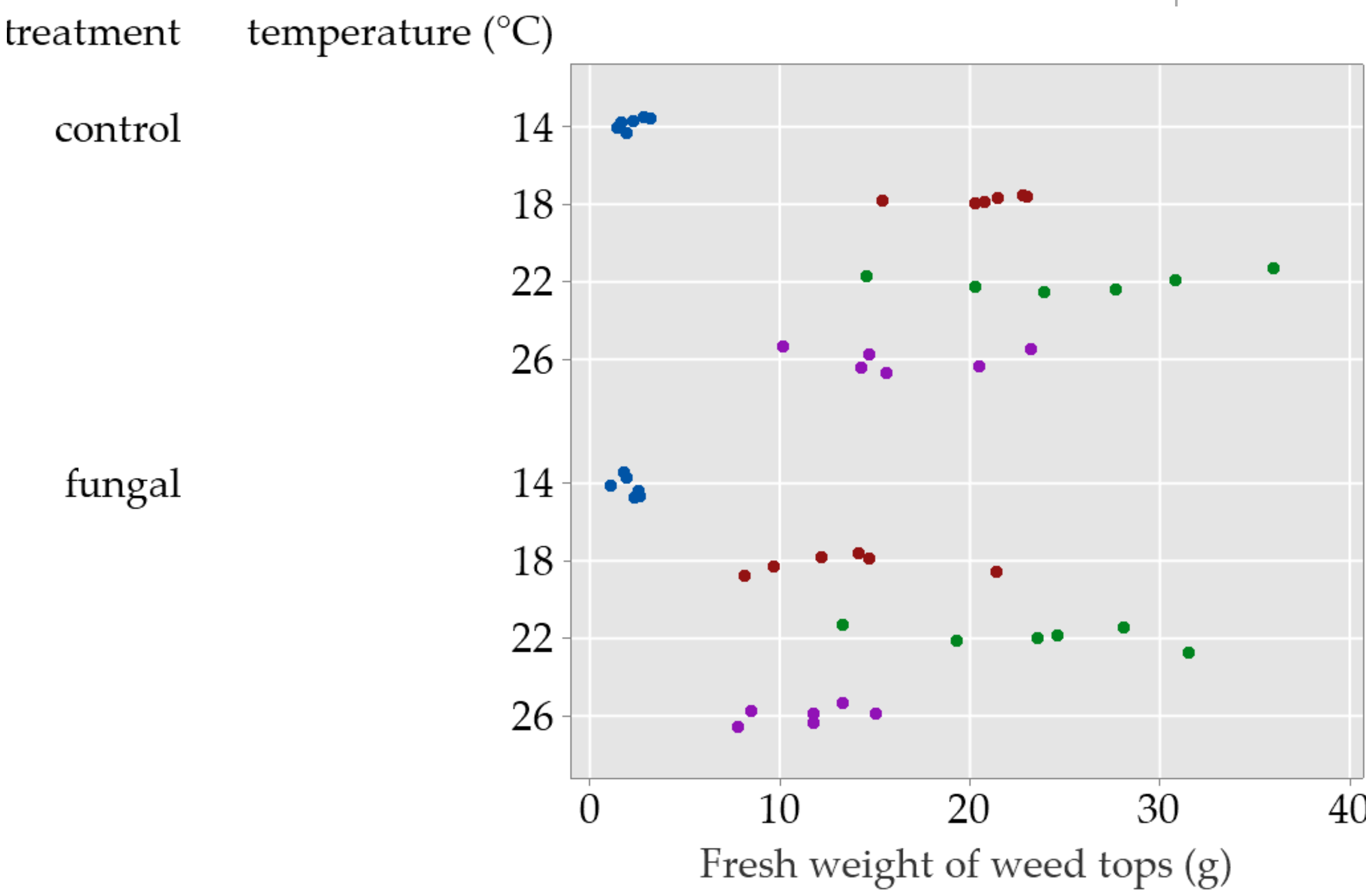
Plants infected with the fungus were compared with control plants at each of four different temperatures. Thus there were eight treatments and six observations were made for each treatment.

Weeds were cut and the data recorded were the fresh weight of the weed tops in grams after a fixed growth period.

The following plot shows the fresh weights by temperature and treatment group (control/fungal infected).

According to the plot, which of the following assumptions, relevant to fitting an appropriate General Linear Model to these data, is likely to be incorrect?

(The data are in fungus.mtw if you wish to examine them. Ignore the fourth column.)



○

Independence of observations in each temperature and treatment combination.

○

Random sampling of weeds from each of the populations underlying the temperature and treatment combinations.

●

Equal variances in the populations underlying the temperature and treatment combinations.

○

Changes in mean weights are caused by changes in temperature and by different treatments.

The variability is much smaller when the temperature is 14 degrees than when the temperature is higher. The assumption of homogeneity of variance is unlikely to be met for these data.

Incorrect

Question 13

0 / 1 pts

Which one or more of the following strategies would be appropriate to investigate to deal with the problem(s) you identified above? (Tick as many as apply.)

☐ Remove outliers.

☑ Leave the group(s) causing the problem out of the analysis.

☐

Use the data as is, given the sample size is relatively large overall.

☑ Investigate transformations of the data.

Removing or changing the outliers substantially alters the data and may lead to the wrong inferences being made. A transformation of the response may lead to a more suitable scale for analysis.

Question 14

1.5 / 1.5 pts

The "Janka hardness" is an important structural property of timber. It essentially measures the resistance of the timber to indentation, abrasion, sawing and so on, and hence is useful when assessing timbers for various purposes (flooring, carving and turning, furniture ...).

The units are kiloNewtons (kN). If you look up timber specifications in catalogues you will often find the Janka hardness of a timber mentioned.

Janka hardness is difficult to measure directly. However, the density of a timber is comparatively easier to measure. Therefore, it is desirable to be able to predict Janka hardness from density.

The Janka hardness (kN) and density (tonnes /m^3) of 36 Australian hardwood specimens are given in janka.mwx.

Obtain a scatterplot of hardness against density. Based on the scatterplot, which of the following are true? (Tick as many as apply.)

□ The relationship between Janka hardness and density is not particularly strong.

□ There is little evidence of a relationship between Janka hardness and density.

☑

The relationship between Janka hardness and density is approximately linear.

☑

The relationship between Janka hardness and density is positive.

There is a clear, strong positive association between Janka hardness and density, and the pattern is roughly linear.

Question 15

1 / 1 pts

Obtain a plot which shows the result of fitting the straight line model:

hi = α + β di + ei

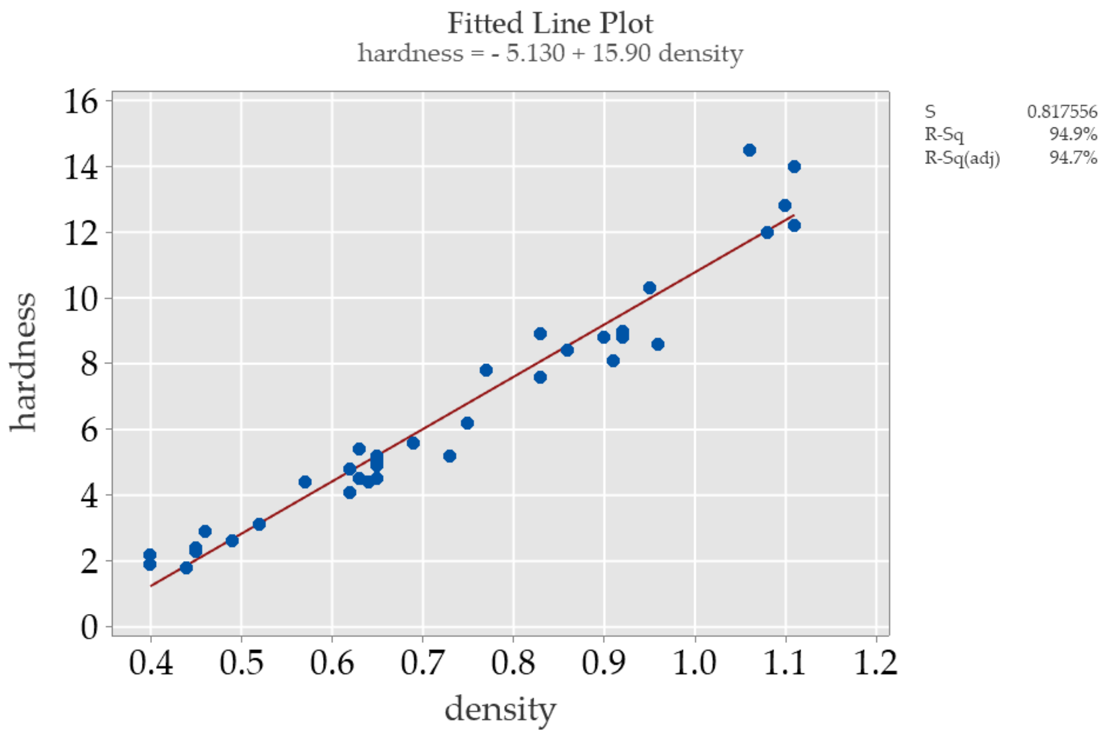
where h = (Janka) hardness and d = density.

Use the commands: Stat > Regression > Fitted line plot.

What is the estimate of β ?

Give your answer to two decimal places.

15.9



Question 16

0.75 / 0.75 pts

Find a 95% confidence interval for β.

Report the lower limit to two decimal places.

14.62

Regression Equation

hardness = -5.130 + 15.904 density

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-5.130	0.484	(-6.113, -4.147)	-10.61	0.000	
density	15.904	0.633	(14.618, 17.190)	25.13	0.000	1.00

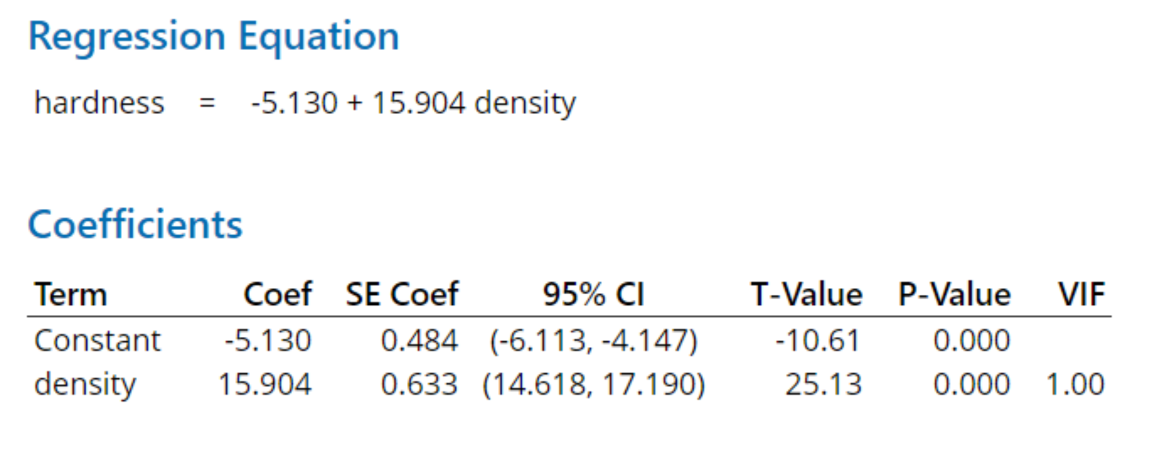
Question 17

0.75 / 0.75 pts

Find a 95% confidence interval for β.

Report the upper limit to two decimal places.

17.19



Regression Equation

hardness = -5.130 + 15.904 density

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-5.130	0.484	(-6.113, -4.147)	-10.61	0.000	
density	15.904	0.633	(14.618, 17.190)	25.13	0.000	1.00

Question 18

1 / 1 pts

Use the results of the regression model to predict the difference between the Janka hardness for two specimens, one with a density of 1.0 and the other a density of 0.5.

One number only is required.

Give your answer to two decimal places.

7.95

The difference will be $15.9 \times (1 - 0.5) = 7.95$; the constants cancel out.

Question 19

1 / 1 pts

A forester with a bit of statistics knowledge looks at your regression output and says:

"The estimate of α is —5.1, and it's significantly different from zero, as well. That's ridiculous: hardness has to be positive! How come you're predicting that for a density of zero, the Janka hardness is —5.1?"

Which one of the following is the most appropriate response to this complaint?

○ That is how the straight line regression line model works. The intercept is negative so the Janka hardness will be negative for zero density.

○ If density is zero, there's no wood, so there's no problem with a negative Janka hardness.

○ That doesn't make sense so the analysis must be wrong.

○ The confidence interval for α includes only negative values, so there's no problem with that result.

● The analysis doesn't apply to very high density or very low density timbers; we wouldn't predict a result for densities close to zero or for zero.

> The regression equation should not be used for prediction using values (well) outside the observed range of x.

Partial

Question 20

0.67 / 1 pts

Consider the data on cheese, contained in cheese.mwx. It consists of 30 observations in which cheese has been tasted, and properties of the specimen tasted were recorded. The response variable is a measure of a quality of the taste, and the explanatory variables in the data are

acetic = log(concentration of acetic acid)
H2S = log(concentration of hydrogen sulphide)
lactic = concentration of lactic acid

The correlations between taste and the three explanatory variables are given below.

		Correlation	**95% CI for ρ**	**P-Value**
acetic	taste	0.550	(0.236, 0.759)	0.002
H2S	taste	0.756	(0.543, 0.877)	0.000
lactic	taste	0.704	(0.461, 0.849)	0.000

You are discussing fitting a multiple linear regression to these data with one of your colleagues. He makes several statements about the data or the proposed analysis. Which would you **disagree** with? (Tick all that apply.)

☑ “The numerical ranges of the explanatory variables are different so you will need to standardise them first, before fitting your regression.”

☑ “It’s a problem that ‘acetic’ and ‘H2S’ are logarithms. That won’t work because it won’t be a linear model.”

□

“The correlations between taste and each of the explanatory variables are all greater than 0.5, and each is very strongly statistically significant. So you’re going to need all of the three variables in your multiple regression model.”

□

“A matrix plot showing all of the pairwise associations between the variables would be a useful exploratory graph.”

Question 21

1.5 / 1.5 pts

Your colleague uses the data to fit some models and reports that the R^2 values are as follows.

[You may assume these calculations are correct.]

R^2	variables in regression
57.12%	H2S
49.59%	lactic
30.20%	acetic
65.17%	H2S and lactic
58.22%	acetic and H2S
52.03%	acetic and lactic
65.18%	H2S, acetic and lactic

Your colleague says: “As I thought, you should include all the variables in the regression, that’s the model with the largest R^2.”

Which of the following are appropriate responses? Tick all that apply.

□

“I could tell that was going to happen from the matrix plot of the variables.”

☑

“By the way, have you done any diagnostic plots of the standardised residuals?”

□

“But the R^2 for the model with all three is only 65% so that leaves too much unexplained variation, I don’t think it will be any use for prediction.”

☑

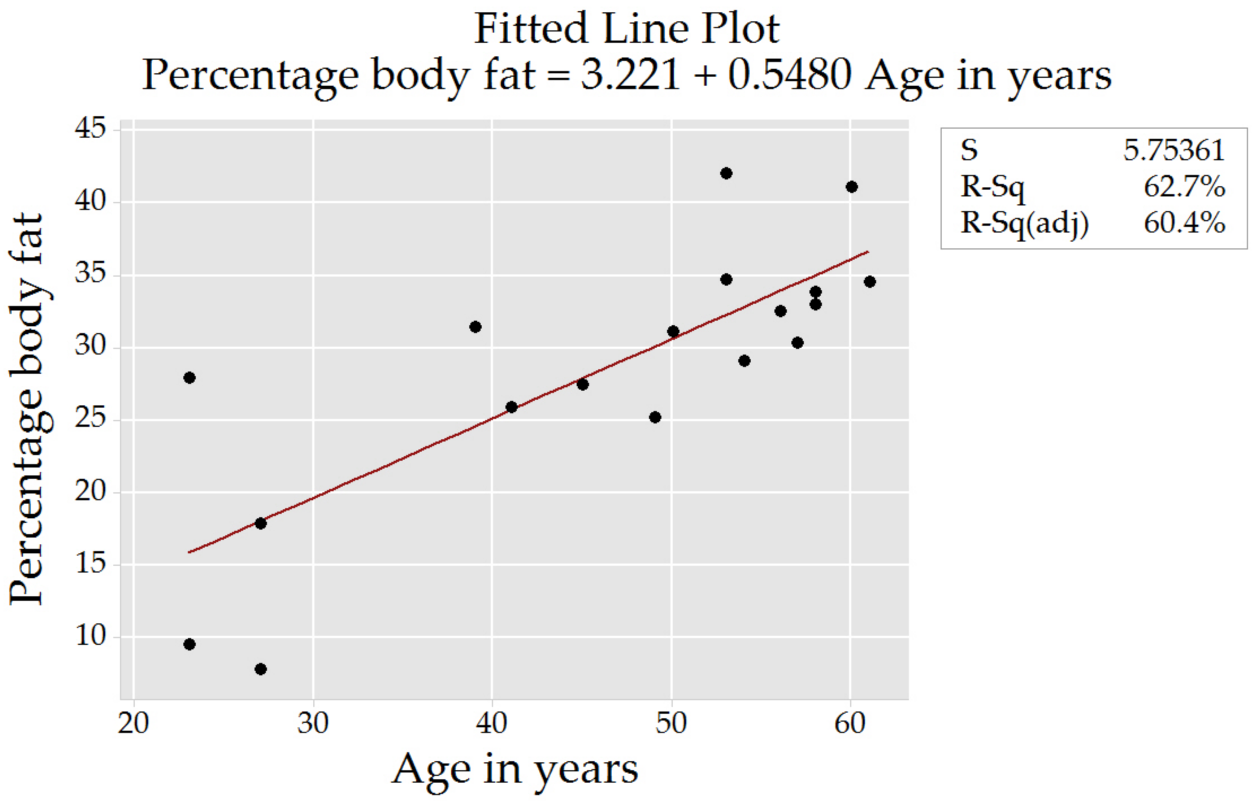
“I would like to see more of the results of analyses, including estimates and confidence intervals of parameters.”

Before deciding which model is appropriate, the residual plots should be examined. The estimates and confidence intervals should also be examined.

Question 22

1.5 / 1.5 pts

The figure below shows the fitted line plot for predicting the percentage of body fat from age in years. The file is: fat.mwx.



According to the analysis above, which of the following are correct? (Tick as many as apply.)

☑ A fifty year old is predicted to have just over 30% body fat.

☑

Body fat is predicted to increase by just over half a percent for each increase of one year in age.

□

A fifty year old is predicted to have 5.5% less body fat than a forty year old.

□

Percentage body fat is predicted to reduce by more than half for each increase of one year in age.

The slope is just over 0.5, so body fat is predicted to increase by just over half a percent for each increase of one year in age.

Quiz Score: **22.17** out of 24.5