

**Project C: Post-weaning diarrhea I**

**By:**

Name: Menna Abdelsattar ID: 18101590

Name:Omnia Saeed ID: 18102396

Name: Maram Mostafa ID: 18102303

Name: Aya Essam ID: 18100409

Name: Moaz Rizkallah ID: 18102436

Under Supervision by Dr. Mohamed Maysara

Table of Contents

[Introduction 3](#_Toc105235181)

[I Reading the Dataset 4](#_Toc105235182)

[II Descriptive statistics 4](#_Toc105235183)

[III Graphics 5](#_Toc105235184)

[IV Outlier detection 8](#_Toc105235185)

[V Testing for normality/ homoscedasticity 9](#_Toc105235186)

[VI Statistical Inference 14](#_Toc105235187)

[VII Hypothesis testing 16](#_Toc105235188)

[VIII Linear model 23](#_Toc105235189)

[IX Conclusion 28](#_Toc105235190)

[X Contribution 28](#_Toc105235191)

# Introduction

Post-weaning diarrhea (PWD) is a worldwide economically important disease in pigs in piggeries. The disease is characterized by increased mortality, weight loss, retarded growth, increased treatment costs and higher use of antibiotics. Enterotoxigenic Escherichia coli is considered to be the most important cause of the disease.

Currently the disease is often controlled by using antimicrobials, but the emergence of antimicrobial resistance in E. coli urges the need for alternative control strategies. For example, inclusion of additional dietary fiber and reduction of crude protein levels, but also the addition of zinc oxide (ZnO) has been demonstrated to have beneficial effects. However, by 2022 this zinc may no longer be used (EU legislation).

Another strategy is to vaccinate the piglets. In this study we are interested in the effect of vaccination as compared to the addition of ZnO and nutraceuticals (e.g. fibers) to the feed. In particular the following treatments are considered (in the dataset this variables is names Treatment):

**A: normal feed + ZnO**

**B: normal feed + nutraceuticals**

**C: vaccination + high energy/protein in phases 2 and 3 (time periods)**

**D: vaccination + high energy/protein in phases 1, 2 and 3**

**E: vaccination + high energy/protein in phases 1, 2 and 3 + nutraceutics**

We are interested in the following outcomes:

**ADWG0021, ADWG2150, ADWG0050**: average daily weight gain (g/day) in the period between 0 and 21 days post-weaning, between day 21 and day 50 post-weaning and in the period between 0 and 50 days post-weaning, respectively.

**So, we can make some analysis to see that treatments have effect or not.**

# I Reading the Dataset

After reading the data by loading it in R we analyze it and found that the data contains nine columns which are:

1. **Pen** (number of pens which contain 16 piglets in one pen)
2. **Treatment** (contain 5 treatments (A,B,C,D,E), each treatment group consists of 128 piglets (thus 8 pens of 16 piglets))
3. **Feeder** (food)
4. **Sex**  (male, female)
5. **W0** (weight)
6. **P0** (number of piglets in each pen)
7. **ADWG0021** (average daily weight gain (g/day) in the period between 0 and 21 days post-weaning)
8. **ADWG2150** (between day 21 and day 50 post-weaning)
9. **ADWG0050** (period between 0 and 50 days post-weaning)

# II Descriptive statistics

1. Summarize your data, we found that sex is a continuous variable then after calculation convert it into male and female.
2. Calculate the following: mean, median, minimum, maximum, first and third quartile (for each variable), so we get them for continuous variable such as; **W0, P0, ADWG0021, ADWG2150, ADWG0050.**
3. For the categorical variable existing, calculate a frequency table; there is one categorical variable which is sex that contain 20 male and 20 female.
4. Calculate the correlation coefficient (ADWG0021 and ADWG2150) and (ADWG0021and ADWG0050), we used two methods which are; Pearson, Spearman

(Correlation coefficients are indicators of the strength of the linear relationship between two different variables, x and y). So, there is a perfect positive relationship between (ADWG0021 and ADWG2150) and (ADWG0021and ADWG0050) by using two methods; when one variable moves higher or lower, the other variable moves in the same direction with the same magnitude.

# III Graphics

1. Generate a bar chart of a categorical variable for the gender.

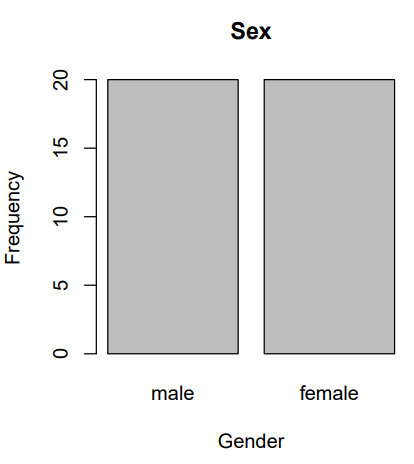


Fig 1: we found that our data contains 20 males and 20 females from 40 pens.

1. Generate a bar chart graph with mean ADWG0021 in males and females

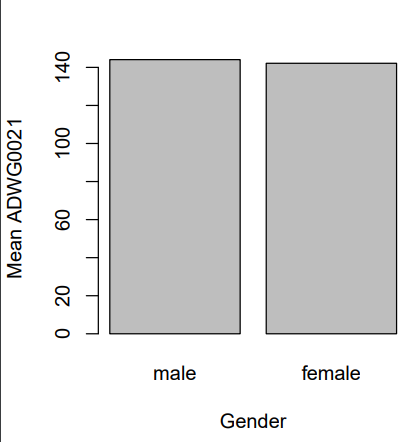


Fig 2: each male and female have the same mean in ADWG0021

1. Make a histogram of a continuous variable: “ADWG2150” as well as “ADWG0021”.

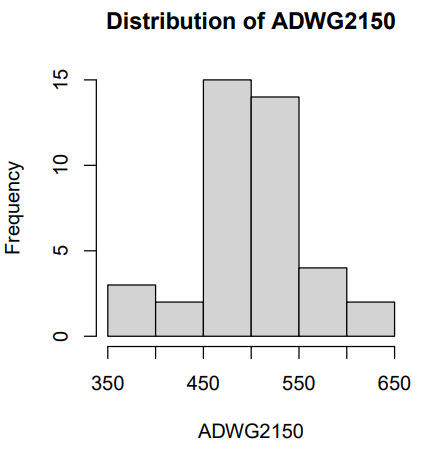


Fig 3: the histogram visualization of ADWG2150 shows that the distribution of data seems to be normal but we will check normality using Shapiro test to check whether the data is really normally distributed or not.

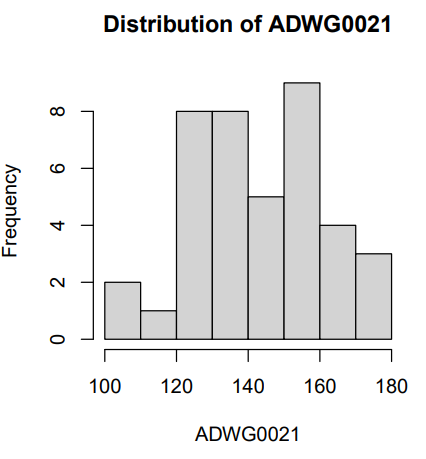


Fig 4: the histogram visualization of ADWG0021 shows that the distribution of data seems to be not normal but we will check normality using Shapiro test to check whether the data is really normally distributed or not.

1. Make a scatterplot of 2 continuous variables ADWG0050 and ADWG0021, and add the regression lines for each gender.

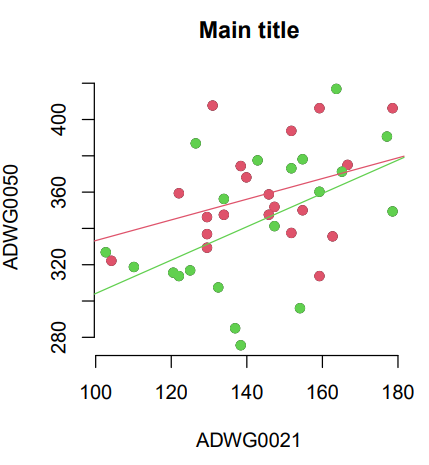


Fig 5: regression of male (green) and female (red), we use the same colors as those used in the scatterplot to differentiate the two regression lines, the distribution of the points suggests a negative relationship between ADWG0050 and ADWG0021 for each gender. Also we put regression attempts to find the line that best fits these points.

1. Make a boxplot of ADWG0021 in and a separate boxplots per Treatment.



Fig 5: we found there are 2 outliers in treatment A and show different variance between the treatments and ADWG0021.

# IV Outlier detection

1. After exploring the data for any existing outliers, we have two outliers in the dataset in treatment A which are 178.5714 and 110.1190.
2. Outliers are values within a dataset that vary greatly from the others—they’re either much larger, or significantly smaller. Outliers may indicate variabilities in a measurement, experimental errors, or a novelty. Outliers may have a negative effect on the result of an analysis; also outliers may have a negative effect on the result of an analysis, but in our data we found only two outliers so it can be effect on our analysis so bad, therefore they must remove them.

# V Testing for normality/ homoscedasticity

1. **We check the normality using:**

**a - Q-Q plot**

First, we started with the categorical data like Treatment and gender column.

1. treatment with all features

A picture containing engineering drawing

Description automatically generated

2- Gender with all features

A picture containing text, map, group, several

Description automatically generated

3- Then, Numeric Columns (W0, ADWG0021, Feeder, ADWG2150, ADWG0050)

Chart

Description automatically generated

**b- Shapiro test:**

In Numeric Columns:

* **Feeder**: p-value = 0.08873
* **W0:** p-value = 0.009664
* **ADWG0021:** p-value = 0.7305
* **ADWG2150:** p-value = 0.9276
* **ADWG0050:** p-value = 0.9086

In Categorical Columns:

* Gender with all features
* Female:
* **Feeder:** p-value = 0.005621
* **w0:** p-value = 0.02526
* **ADWG0021:** p-value = 0.9513
* **ADWG2150**: p-value = 0.9276
* **ADWG0050:** p-value = 0.7312
* Male:
* **Feeder:** p-value = 0.005621
* **w0:** p-value = 0.1768
* **ADWG0021:** p-value = 0.9771
* **ADWG2150**: p-value = 0.849
* **ADWG0050:** p-value = 0.1895
* Treatment with all features
* A:
* **Feeder:** p-value = 0.2738
* **w0:** p-value 0.3348
* **ADWG0021:** p-value = 0.0395
* **ADWG2150**: p-value = 0.6584
* **ADWG0050:** p-value = 0.8976
* B:
* **Feeder:** p-value = 0.2738
* **w0:** p-value = 0.03048
* **ADWG0021:** p-value = 0.8132
* **ADWG2150**: p-value = 0.1098
* **ADWG0050:** p-value = 0.5031
* C:
* **Feeder:** p-value = 0.2738
* **w0:** p-value = 0.6739
* **ADWG0021:** p-value = 0.6954
* **ADWG2150**: p-value = 0.6256
* **ADWG0050:** p-value = 0.5882
* D:
* **Feeder:** p-value = 0.2738
* **w0:** p-value = 0.2784
* **ADWG0021:** p-value = 0.7126
* **ADWG2150**: p-value = 0.2102
* **ADWG0050:** p-value = 0.1702
* E:
* **Feeder:** p-value = 0.2738
* **w0:** p-value = 0.211
* **ADWG0021:** p-value = 0.86
* **ADWG2150**: p-value = 0.5288
* **ADWG0050:** p-value 0.7315

**c- Histogram:**

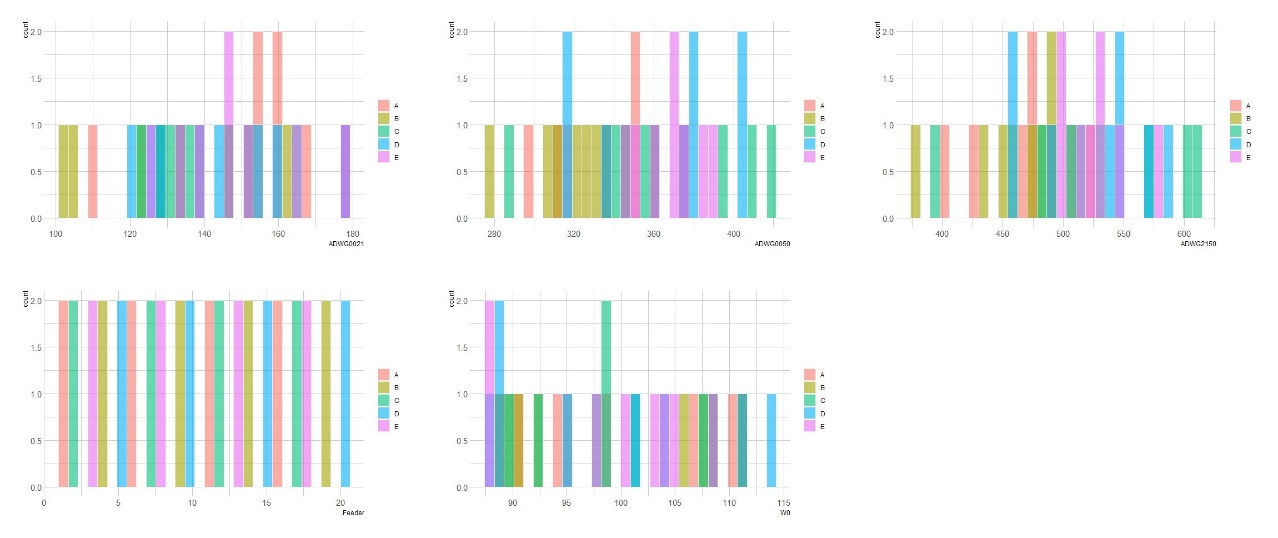
In Numeric Columns:

Chart, histogram

Description automatically generated

In Categorical:

* Treatment with all features



* Gender with all features

A picture containing text, writing implement, stationary, pen

Description automatically generated

**Results of the Normality Tests**

In numerical columns we found that “W0” column is not normal as the p-value of its Shapiro test is 0.009664 < 0.05 so, we have enough evidenced to reject null hypothesis (The Null hypothesis is (H0): assuming that it is normally distributed) in support of the alternative hypothesis (The Alternative hypothesis is (HA): assuming that it not normally distributed).

In categorical columns we found that “Sex” column its Shapiro test with “Feeder” and “W0” columns less than 0.05, so we have enough evidenced to reject null hypothesis (The Null hypothesis is (H0): assuming that it is normally distributed). So “Sex” column is not normally distributed.

In the “Treatment” column its Shapiro test with “ADWG0021” and “W0” columns less than 0.05, so we have enough evidenced to reject null hypothesis (The Null hypothesis is (H0): assuming that it is normally distributed). So “Treatment” column is not normally distributed.

In our Hypothesis we are interested in “ADWG0021” with “Sex” the Shapiro test of these columns in 0.9, and it is greater than 0.05, so it is normally distributed.

**Second, we check the homoscedasticity:**

We use Levene and Bartlett Tests, and we found that all features are Homoscedastic as all results of p-value are greater than 0.05.

Also, we use Boxplot for checking Homoscedasticity.

a- Sex with all features

Chart, box and whisker chart

Description automatically generated

b- Treatment with all features

Chart, box and whisker chart

Description automatically generated

# VI Statistical Inference

* **Calculate the 90%, 95%, 99% confidence interval for the means of ADWG0021per each gender.**

We calculate the confidence interval by this formula;

Diagram

Description automatically generated

Essentially, a calculating a 90 percent confidence interval means that we are 90 percent sure that the true probability falls within the confidence interval range that we create in a standard normal distribution.

Lower Interval: male=139.5012, female=136.6345

Upper Interval: male= 148.6436, female=147.6710

Essentially, a calculating a 95 percent confidence interval means that we are 95 percent sure that the true probability falls within the confidence interval range that we create in a standard normal distribution.

Lower Interval: male= 138.6577, female=135.6162

Upper Interval: male= 149.4871, female=148.6893

Essentially, a calculating a 99 percent confidence interval means that we are 99 percent sure that the true probability falls within the confidence interval range that we create in a standard normal distribution.

Lower Interval: male= 134.9818, female= 131.1788

Upper Interval: male= 153.1630, female= 153.1268

* **How would you describe those inferences and what do you observe in terms of the interval width when request higher confidence (i.e. 99% C.I.)?**

When the confidence interval increases the width increase, so 90 and 95 % confidence interval is narrow comparing to 99% confidence interval.

# VII Hypothesis testing

1. **We hypothesis that ADWG0021is different between male vs female. Assuming normality and homoscedasticity, can you test this hypothesis using statistical hypothesis framework?**

We have the statistical hypothesis testing framework after converting the research question into statistical question which is: is the mean different between male vs female in ADWG0021? ,we have null hypothesis which is the mean of groups male and female are equal and alternative hypothesis which is the mean of groups male and female is different then we compare the p value result to our significance level alpha 0.05 and check results, if results of p value smaller than 0.05 then we have enough evidence to reject the null hypothesis in support of alternative hypothesis and if the results greater than 0.05 then we don't have evidence to reject the null hypothesis, we firstly test normality and homogeneity of variance to know what test we are going to use in the , but here in this case normality and homoscedasticity were assumed so we will use standard two sample t test directly according to these assumptions, here our data is two sided because it is different in general not greater or smaller than a specific value, also data is independent not dependent or paired so paired equal false, the results of the standard two sample t test give a p value equal to 0.7557 which is greater than the significance level alpha 0.05 which means that we do not have enough evidence to reject the null hypothesis in support of alternative hypothesis which means that the mean of the 2 groups males and females is not different from each other, and also means that ADWG0021 is not different between male vs female.

1. **Now we will assess whether the previous test assumptions have been meet for the test**

To test normality, we use QQ plots and shapiro welk test then we have null hypothesis which is the data is normally distributes and alternative hypothesis which is the data is not normally distributed then we compare the p value result to our significance level alpha 0.05 and check results.

Firstly, to test normality of data we used QQ plot and the line has no deviations from the data points so the data of females is normally distributed.

we also tested normality using shapiro welk test of normality and the results of the test gives a p value equal to 0.9513 which is greater than the significance level alpha 0.05, this means that we do not have enough evidence to reject the null hypothesis in support of alternative hypothesis, here the null hypothesis of shapiro test is that the data is normally distributed and the alternative is not normally distributed, so here the data of females is normally distributed. We did the same thing for the data of males, firstly we used QQ plot and the line has no deviations from the data points so the data of males is normally distributed, we also used shapiro wilk test of normality and the results of the test gives a p value equal to 0.9771 which is greater than the significance level alpha 0.05, this means that we do not have enough evidence to reject the null hypothesis in support of alternative hypothesis, here the null hypothesis of shapiro test is that the data is normally distributed and the alternative is not normally distributed, so the data of males is normal.

We tested normality so now we will check homogeneity of variance of data, we will test using levene and F test to double check variance, here the null hypothesis of these tests is homo variance and alternative is hetero variance so we get the p value result and check if smaller than 0.05 so reject null so hetero and if greater than 0.05 so, do not reject null so it is homo.

After normality assessment we tested the homogeneity of variance of the groups males and females by levene test and the result of the test gives a p value equal to 0.3009 which is greater than the significance level alpha 0.05 this means that we do not have enough evidence to reject the null hypothesis in support of alternative hypothesis so the data has homo variance or equal variance, so we will use standard two sample t test not wilsh t test for hetero variance as our data is homo not hetero, also the results of F test gives a p value of 0.4193 which is also greater than 0.05 so we do not have enough evidence to reject null so the data has homo variance.

So the answer is yes, the previous test assumptions have been meet for the test as we tested and the results are that the data is normal and homo in variance.

1. **We hypothesis that ADWG0021is "different" in the group receiving Treatment A (normal feed + ZnO) compared to the Treatment B (normal feed + nutraceuticals). Can you test this hypothesis assuming heteroscedasticity?**

We have the statistical hypothesis testing framework after converting the research question into statistical question which is: Is the mean different between treatments A and B in ADWG0021, then we have the null hypothesis which is mean of treatment A equal to mean of treatment B and alternative hypothesis which is mean of treatment A different from mean of treatment B, then test normality to see what test we going to use it.

We firstly test normality but here heteroscedasticity of variance is assumed so we going to check only normality not variance to know what test we are going to use, then we have the null hypothesis which is mean of treatment A equal to mean of treatment B and alternative hypothesis which is mean of treatment A different from mean of treatment B, then test normality to see what test we going to use.

To test normality we use QQ plots and shapiro welk test then we have null hypothesis which is the data is normally distributes and alternative hypothesis which is the data is not normally distributed then we compare the p value result to our significance level alpha 0.05 and check results , we tested normality of treatment A data we used QQ plot firstly and the line has deviations from the data points so the data of treatment A is not normally distributed, the histogram plot also shows that the data of A is not normal, we also used shapiro wilk test of normality of treatment A and the results of the test gives a p value equal to 0.0395 which is smaller than the significance level alpha 0.05 , this means that we have enough evidence to reject the null hypothesis in support of alternative hypothesis, here the null hypothesis of shapiro test is that the data is normally distributed and the alternative is not normally distributed, so this data of treatment A is not normally distributed.

We did the same thing to treatment B so we tested normality of treatment B data using QQ plot firstly and the line has no deviations from the data points so the data of treatment B is normally distributed, the histogram plot also shows that the data of A is normal.

We also used shapiro wilk test of normality of treatment B and the results of the test gives a p value equal to 0.8132 which is much greater than the significance level alpha 0.05, this means that we do not have enough evidence to reject the null hypothesis in support of alternative hypothesis, here the null hypothesis of shapiro test is that the data is normally distributed and the alternative is not normally distributed, so this data of treatment B is normally distributed.

In our case we have data A not normal and data B normal but when we have at least one of our data not normal so we assume that the data is not normally distributed so here we gonna use man whitny test (welcox test) instead of standart two sample t test assuming non normality of data (non parametric), here we have two sided data.

The results of wilcoxon test gives a p value equal to 0.0312 which is smaller than the significance level alpha 0.05 that means that we have enough evidence to reject the null hypothesis in support of alternative hypothesis , this means that ADWG0021 is "different" in the group receiving Treatment A (normal feed + ZnO) compared to the Treatment B (normal feed + nutraceuticals).

1. **Assess the previous test assumption.**

Now we will test hetero variance of data and see the results, all the results of F test of variance give a p value of 0.9567 for groups A and B which is much greater than the significance level alpha 0.05, this means that we do not have enough evidence to reject the null hypothesis in support of alternative hypothesis which means that the variance of these groups is homo variance or equal variance. This means that the variance is homo in this data not hetero so the answer is No, this do not go with the previous test assumptions.

Now we tested variance and the results were homo and we tested normality before this and the results not normal so we going to use welcox test but make variance equal false.

When we changed the variance homo in code it gives us the same results as we tested before by unequal variance (p value = 0.0312 so reject null so groups are different).

1. **We hypothesis that ADWG0021is different between the different Treatments. Can you perform comparison between the different groups, after assessing the assumptions and performing post-hoc testing (assuming normality and homoscedasticity)?**

Here we have in our data the groups of treatments are categorical and the ADWG0021 is continuous so we going to use anova test.

We have the statistical hypothesis testing framework after converting the research question into statistical question which is: is the mean different between the different Treatments in ADWG0021? , we have null hypothesis which is the mean of the treatment groups are equal and alternative hypothesis which is the mean of treatment groups male and female is different then we compare the p value result to our significance level alpha 0.05 and check results, if results of p value smaller than 0.05 then we have enough evidence to reject the null hypothesis in support of alternative hypothesis and if the results greater than 0.05 then we don't have evidence to reject the null hypothesis, we firstly test normality and homogeneity of variance to know what test we are going to use.

We tested then normality and homo variance to check assumptions are true or not.

QQ plot results show that data of treatment A is not normally distributed as the line but B, C, D and E are normally distributed.

Histogram results show that data of treatment A is not normally distributed but B, C, D and E are normally distributed.

Shapiro test for treatment A gives a p value = 0.0395 which is less than 0.05 so we have enough evidence to reject the null hypothesis in support of alternative hypothesis so the data is not normal.

The results for treatment B gives a p value = 0.8312 which is more than 0.05 so we don’t have enough evidence to reject the null hypothesis in support of alternative hypothesis so the data is normal.

The results for treatment C gives a p value = 0.6954 which is more than 0.05 so we don’t have enough evidence to reject the null hypothesis in support of alternative hypothesis so the data is normal.

The results for treatment D gives a p value = 0.7126 which is more than 0.05 so we don’t have enough evidence to reject the null hypothesis in support of alternative hypothesis so the data is normal.

The results for treatment E gives a p value = 0.86 which is more than 0.05 so we don’t have enough evidence to reject the null hypothesis in support of alternative hypothesis so the data is normal.

Here we have at least one of the data is not normal (data of treatment A) so we going to use kruskal walis test (non-parametric). The results give a p value equal to 0.09972 so we don’t have evidence to reject the null hypothesis so groups are not different in AGDW0021.

Then we checked the variance to see whether it met the assumptions or not, the results of levene test of homogeneity of variance gives a p value 0.8968 which is greater than 0.05 so we do not have enough evidence to reject the null so the data has homo variance.

Now we will assume normality and homoscedasticity as in the question to see whether the assumptions met the test or not so we going to use standard anova test, posthoc is performed after annova using tukey honest test assuming normality and homoscedasticity (includes p value correction to be p adjusted).

The report of anova model has an interpretation that the main effect of Treatment is statistically not significant and large (F(4, 35) = 2.10, p = 0.101; Eta2 = 0.19, 95% CI [0.00, 1.00]).

The results of anova test gives a p value equal to 0.101 which is much greater than the significance level alpha 0.05 so we do not have enough evidence to reject the null hypothesis in support of alternative hypothesis so ADWG0021 is not different between the different Treatments.

After anova we perform posthoc (tukey test). The results of post hoc shows the difference of means between treatments with each other’s (B-A,C-A,D-A,E=A,C-B,D-B,E-B,D-C,E-C,E-D), all the treatments with each others have a confidence interval values indicate that the zero value lies in the interval so this indicates the null hypothesis will not be rejected and the p adjusted value will be too large and if the zero value is not in the interval this will indicate lower p value so we will have evidence to reject the null hypothesis, when the p adjusted value show that all treatments with each other gives a p adjusted value greater than the significance level alpha this means that we do not have enough evidence to reject the null so this means that ADWG0021 is not different.

(The p adjust is the correction of p value performed by posthoc after anova).

The plot of testing posthoc shows that the zero value lies in the middle of confidence interval which results in not rejecting the null and the higher p value, glht and mcp function is done here by dunnet which is same as tuky but here the results gives the linear hypothesis between each pair of treatments and how it can affect the model as the 2 treatments with each other indicate the non-normality of data(they are the reason that the data is not normal), the results shows that the treatments B-A gives the higher p value 0.035 and one star which means that the result is significant as it is lower than alpha 0.05 so we have enough evidence to reject the null hypothesis so not normal data is due to these 2 treatments with each other (B-A). (so we used function ghlt to see the p adjusted value of each pair of treatments with each other to test which mix of treatments cause non normality to the data, so we performed pairwise test with benferroni method, the adjust p value with benferroni method but also results were not significant).

After adjusting p values, treatment B-A gives a p value significant equal to 0.035 which is less than 0.05 so we have enough evidence to reject the null so the mix of treatment B with A is different in ADWG0021.

We also performed tukey and benferoni assuming normality with dunn test and also method bonferroni, the results of tukey test with adjusted p values with benferoni gives a p adjusted value 0.0437 between treatment A and B (A mixed with B) which is lower than 0.05 so we have evidence to reject null so this leads to non-normality of data(also then ADWG0021 is different in A with B after adjusting p value with benferoni) , while the p value of other treatments with each other is greater than 0.05 so we do not have evidence to reject the null so this leads to normality of other treatments (also ADWG0021 is not different between other treatment with each other’s (except A with B)) , also the mean of each pair of treatment is shown in the results.

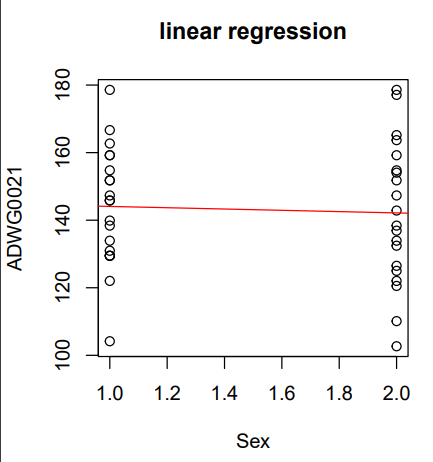
# VIII Linear model

1. **Fit a linear regression to the data and interpret the regression coefficient.**

Firstly we plot the data then we generate our regression model to fit the regression line, then we have the null hypothesis that slope equal zero (y cannot be predicted by x) and alternative hypothesis slope differ from zero (y can be predicted by x) so we test and check results.

Here linear regression is continuous (ADWG0021) with categorical (sex), so convert it to numeric assume male=1, female=2, then make plot and get intercept and slope;

#coefficients a and b (intercept = 144.07 and slope = -1.92), therefore that the intercept is a predicted value of Y when X is zero (unit is the same as in Y), while the slope is the rate of change in y (age) as x changes.



It is obvious that there is no linear relationship between the explanatory variable sex with the response variable ADWG0021 as the sex is categorical (all males equal to one and females equal to 2) and ADVG0021 is continuous so the regression line is horizontal seems to be zero so it means that y (ADWG0021) can't be predicted by x (sex) as there is no linearity between them.

The results of regression models shows that the residuals result is too bad which is 1.23 so the error is too large, also the p value is equal to 0.756 which is much greater than the significance level 0.05 so we do not have enough evidence to reject the null so the slope is equal to zero so y can't be predicted by x (there is no linear relationship between the response variable and the explanatory variable), also the residual standard error (mean square error) equal 19.37 which is too large which is not good and f value is too small 0.09824 which results in the large p value and finally the adjusted R squares is too small -0.02367 which is a not good at all(it explains -0.02 from the variability of y).

Linear regression is between continuous and continuous but here all used in hypothesis were continuous and categorical.

1. **Calculate and interpret a 95% confidence interval of the regression slope.**

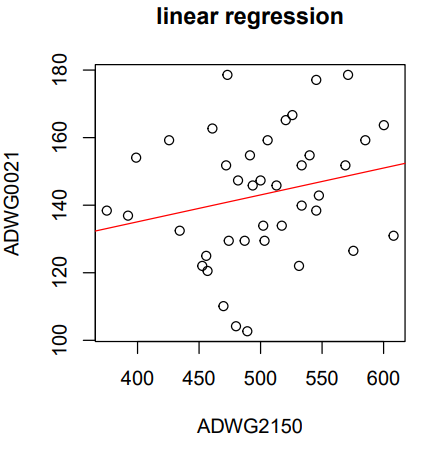
conf int is with the regression model and x axis.

Here we are confident 95 percent that the true population mean falls between 2.5% (-14.31805) and 97.5% (10.47877) of the sampling distribution of sample means

This means that the zero value is in half of the interval which indicates that we do not have enough evidence to reject the null hypothesis and in linear regression the null is that the slope equal zero(no linearity and y can't be predicted by x) so this assures that ADWG0021 can't be predicted by sex.

**Another variables; Now we will perform linear regression but 2 continuous variables as it is supposed to be done like that.**

1. So we will perform ADGW2150 (y axis response variable) with ADWG0021 (x axis explanatory variable). #coefficients a and b (intercept = 103.25181 and slope = 0.07958), therefore that the intercept is a predicted value of Y when X is zero (unit is the same as in Y), while the slope is the rate of change in y (age) as x changes.



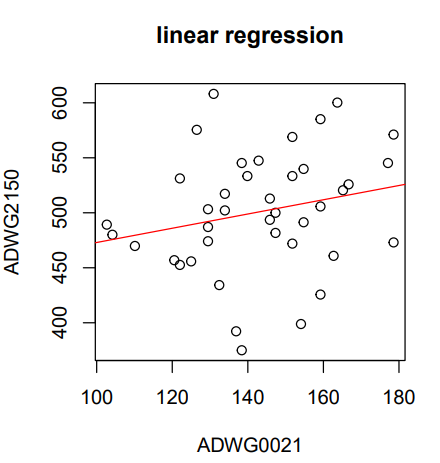
1. The results of linear regression between the 2 continuous variables ADWG0021(y) and ADWG2150(x) gives a bad model as the residuals median is too large(2.849) and the p values is equal to 0.158882 which is much greater than 0.05 so we do not have enough evidence to reject the null hypothesis so y(ADWG0021) can't be predicted by x(ADWG2150) as there is no linearity between them also the regression line when fitted shows that it is horizontally so that means that the slope is close to zero (null hypothesis), also here the adjusted R squared is more accurate than the R squared and the adjusted gives a very small value which is 0.02659(2.5%) which means that the regression model explains(capture) only 2.5% of the total variation of y(ADGW0021) which is not good at all, so there is no linearity.
2. **Calculate and interpret a 95% confidence interval of the regression slope.**

Here we are confident 95 percent that the true population mean falls between 2.5% (-0.03252242) and 97.5% (0.191674) of the sampling distribution of sample means.

This means that the zero value is in half of the interval which indicates that we do not have enough evidence to reject the null hypothesis and in linear regression the null is that the slope equal zero (no linearity and y can't be predicted by x) so this assures that ADWG0021 can't be predicted by sex.

**Now we will perform linear regression the same 2 continuous variables but exchange x with y.**

1. We will perform ADGW2150 (x axis response variable) with ADWG0021 (y axis explanatory variable). #coefficients a and b (intercept = 408.2149 and slope = 0.6477), therefore that the intercept is a predicted value of Y when X is zero (unit is the same as in Y), while the slope is the rate of change in y (age) as x changes.



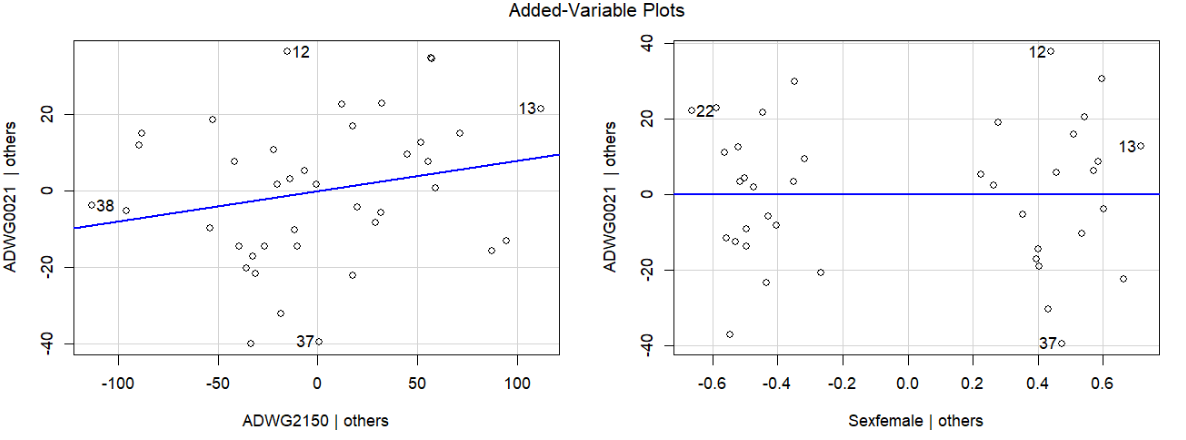
1. The results of linear regression between the 2 continuous variables ADWG0021 (x) and ADWG2150 (y) gives also a bad model as the residuals median is too large (4.802) and the p values is equal to 0.159 which is much greater than 0.05 so we do not have enough evidence to reject the null hypothesis so y (ADWG2150) can't be predicted by x (ADWG0021) as there is no linearity between them also the regression line when fitted shows that it is horizontally so that means that the slope is close to zero (null hypothesis), so that means that the slope is close to zero (null hypothesis), also here the adjusted R squared is more accurate than the R squared and the adjusted gives a very small value which is 0.02659(2.5%) which means that the regression model explains(capture) only 2.5% of the total variation of y(ADGW2150) which is not good at all, so there is no linearity.
2. **Calculate and interpret a 95% confidence interval of the regression slope.**

conf int is with the regression model and x axis.

Here we are confident 95 percent that the true population mean falls between 2.5% (-0.2647333) and 97.5% (1.560231) of the sampling distribution of sample means, this means that the zero value is in half of the interval which indicates that we do not have enough evidence to reject the null hypothesis and in linear regression the null is that the slope equal zero (no linearity and y can't be predicted by x) so this assures that ADWG0021 can't be predicted by sex.

1. **Estimating the average ADWG0021 change for with changing the gender from 1 to 2.**

Now we will perform multiple linear regressions with the 2 continuous variables ADGW0021 and ADGW2150 with the existence of Sex to see whether sex affect or not so, perform model by plotting the response variable (ADWG0021) against the 2 explanatory variables (sex+ADWG2150). #coefficients (intercept = 102.97717 and slope of ADWG2150= 0.07981, slope of Sex = 0.10592), therefore that the intercept is a predicted value of Y when X is zero (unit is the same as in Y), while the slope is the rate of change in y (age) as x changes.



* Produce added variable plots

The results of multiple linear regression between the 3 variables ADWG0021(y) and ADWG2150+Sex gives also a bad model as the residuals median is too large(2.854) and the p value of ADWG2150 is equal to 0.17519 which is much greater than 0.05 so we do not have enough evidence to reject the null hypothesis so y(ADWG0021) can't be predicted by x (ADWG2150) in the existence of Sex as there is no linearity between them also p value of sex equal 0.98652 which is also greater than 0.05 so do not reject the null so sex is not significant and has no effect on the model, also the regression line when fitted shows that it is horizontally so that means that the slope is close to zero (null hypothesis).

Also the also here the adjusted R squared is more accurate than the R squared and the adjusted gives a very small value which is 0.0002851(0.02%) which means that the regression model explains(capture) only 0.02% of the total variation of y(ADGW2150) which is not good at all, so there is no linearity

# IX Conclusion

Finally, we found that the treatment A and B have significant difference from other treatments, so these treatments have effect on the piglets. Also, there is no linear relationship between ADWG0021 and ADWG2150, also between ADWG0021 and Sex (x), so y cannot be predicted by x. when performing multiple linear regression of the two continuous variables ADWG0021 and ADWG2150 with the existence of sex we found that the sex has no effect in the analysis.

# X Contribution

|  |  |
| --- | --- |
| Name | Contribute |
| Moaz Rizkallah | 0. Data reading  1. Descriptive statistics  2. Graphics  3. Outlier detection |
| Aya Essam | 4. Testing for normality/ homoscedasticity |
| Maram Mostafa | 5. Statistical Inference |
| Menna Abdelsattar | 6. Hypothesis testing |
| Omnia Saeed | 7. Linear model |