UNDERGROUND AIR QUALITY USING MULTIPLE LINEAR REGRESSION AND ANOVA

Shaista Syeda
ss7810@rit.edu

Index

Contents

Data Description	1
Overview	1
Description of Variables	1
Dataset	1
Introduction	2
Multiple Linear Regression	2
Claim 1	3
Claim 2	4
Claim 3	6
R - squared	7
ANOVA	8
ANOVA Table	9
Conclusion	10

Dataset Description -

What are the breathing habits of baby birds that live in underground burrows?

Overview -

Some mammals burrow into the ground to live. Scientists have found that the quality of the air in these burrows is not as good as the air aboveground. In fact, some mammals change the way that they breathe in order to accommodate living in the poor air quality conditions underground.

Some researchers (Colby, et al, 1987) wanted to find out if nestling bank swallows, which live in underground burrows, also alter how they breathe. The researchers conducted a randomized experiment on n = 120 nestling bank swallows. In an underground burrow, they varied the percentage of oxygen at four different levels (13%, 15%, 17%, and 19%) and the percentage of carbon dioxide at five different levels (0%, 3%, 4.5%, 6%, and 9%).

Description of variables -

1. Response (y): percentage increase in "minute ventilation," (Vent), i.e., total volume of air breathed per minute (Continuous Variables)
2. Potential predictor (x1): percentage of oxygen (O2) in the air the baby birds breathe (Continuous Variables)
3. Potential predictor (x2): percentage of carbon dioxide (CO2) in the air the baby birds breathe (Continuous Variables)

Total Number of Observations - 120

Type of statistical analysis planned on performing - Multiple Linear Regression and also ANOVA

Questions I will be able to answer through the analysis -

1. Is oxygen related to minute ventilation ?

2. Is carbon dioxide related to minute ventilation ?

3. What is the mean minute ventilation of all nestling bank swallows whose breathing air is 15% oxygen and 5% carbon dioxide?

Dataset - https://github.com/xssti/STATS-Project/blob/main/Babybirds-2.xlsx

Introduction -

The "first-order" model with two quantitative predictors to summarize the data is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

where:

- Y_i is percentage of minute ventilation of nestling bank swallow i
- X_{i1} is percentage of oxygen exposed to nestling bank swallow i
- X_{i2} is percentage of carbon dioxide exposed to nestling bank swallow i

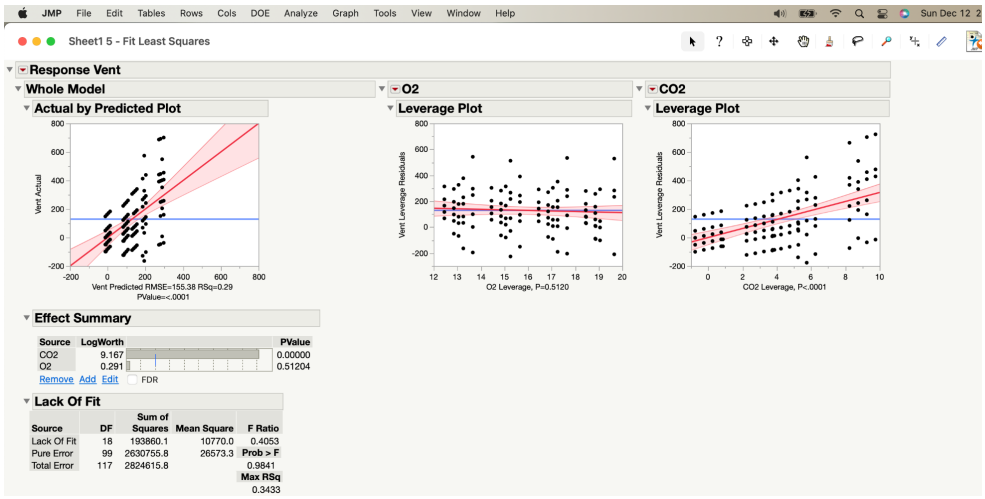
The test for significance is a test to determine if there is a linear relationship between the response and any of the regressor variables

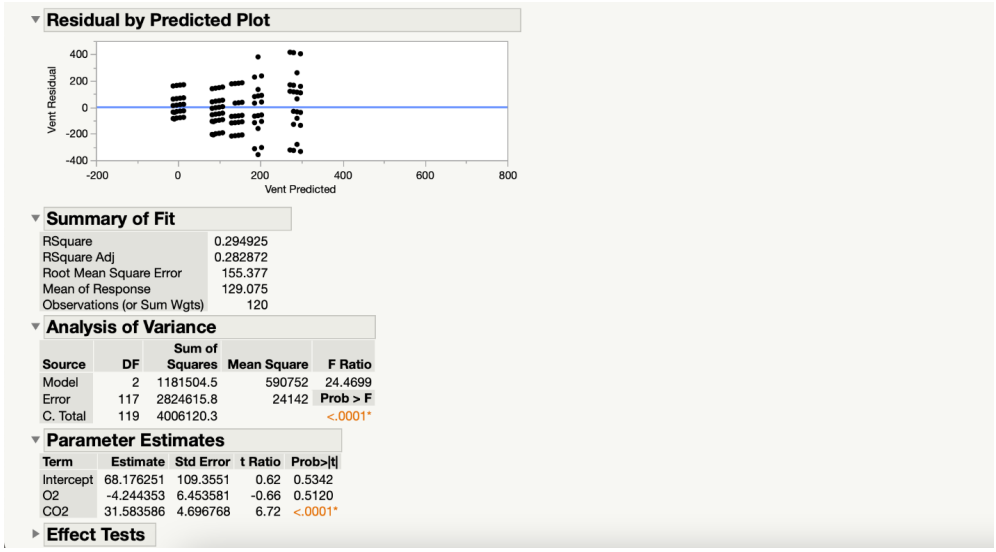
Multiple Linear Regression -

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables.

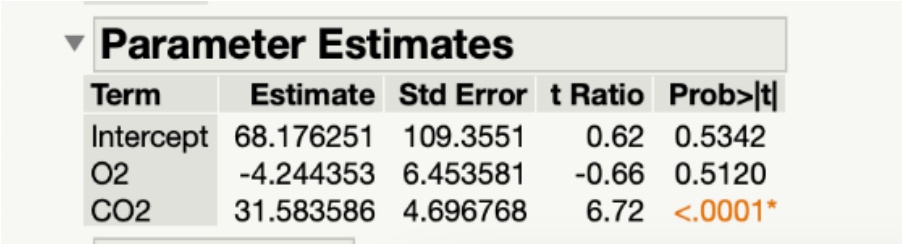
First I will plot a Least Squares plot in JMP and find the Regression line Equation for this dataset.

Plotting Fit Least Squares plot for the dataset-





The Regression equation is



$Y = 68.176 - 4.24X_1 + 31.58X_2$

Claim 1 -

My First Claim is that Minute Ventilation is related to O2 and/or CO2.

To prove this we need to find Test Statistics, critical value, p-value and we will Reject the null hypothesis if

$F_o > F_{\alpha,k,n-k-1}$.

Now Stating the Null and Alternate Hypotheses for this claim

The Null Hypothesis is $H_o: \beta_1 = \beta_2 = 0$

And the alternate hypothesis is $H_1: \beta_i \neq 0$ for at least one i

Summary of Fit

RSquare	0.294925
RSquare Adj	0.282872
Root Mean Square Error	155.377
Mean of Response	129.075
Observations (or Sum Wgts)	120

Analysis of Variance

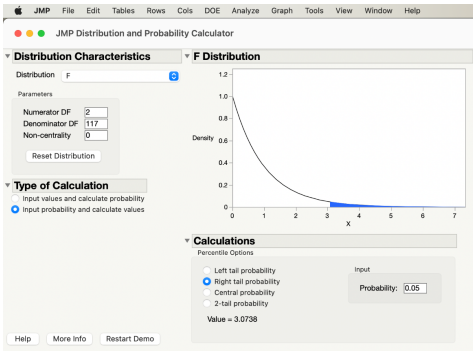
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	1181504.5	590752	24.4699
Error	117	2824615.8	24142	Prob > F
C. Total	119	4006120.3		<.0001*

From the above table,

The value of $F_o = 24.4699$

Calculating the value of $F_{\alpha,k,n-k-1}$

$F_{0.05,2,117} = 3.0738$



The P-value is $P(F_{2,117} > 24.4699) = < 0.0001$

A large value of F_o indicates that regression is significant and as the value of $F_o > F_{0.05,2,117}$, we reject the Null Hypotheses

Hence I can conclude that vent is related to O2 and/or CO2

Claim 2 -

My second claim is that CO2 contributes significantly to the model.

To prove this we need to find Test Statistics, Critical Value, P-value. And reject the Hypotheses $|t_o| > t_{\frac{\alpha}{2},n-k-1}$

The Null and Alternate Hypotheses for this claim are

Null Hypotheses - $H_o: \beta_1 = 0$

Alternate Hypotheses - $H_1: \beta_1 \neq 0$

Test statistics = $t_o = \frac{\beta_j}{s.e.(\beta_j)}$

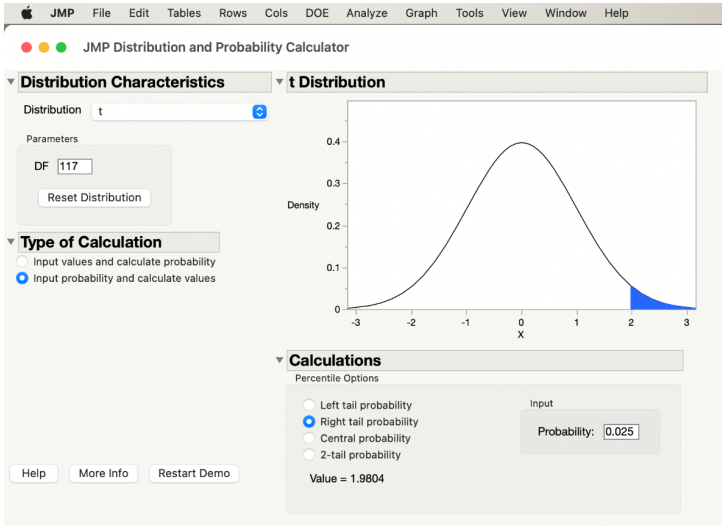
The test statistics value is $t_o = \frac{31.58}{4.6967} = 6.72$

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	68.176251	109.3551	0.62	0.5342
O2	-4.244353	6.453581	-0.66	0.5120
CO2	31.583586	4.696768	6.72	<.0001*

Effect Tests

And the value of $t_{\frac{\alpha}{2},n-k-1}$ is $t_{0.025,117} = 1.9804$



And the P- Value is <0.0001

As $|t_o| > t_{0.025,117}$, we Reject the Null Hypotheses.

Hence I can conclude that CO2 contributes significantly to the model.

Claim 3 -

My third claim is that O2 contributes significantly to the model.

To prove this we need to find Test Statistics, Critical Value, P-value. And reject the Hypotheses $|t_o| > t_{\frac{\alpha}{2},n-k-1}$

The Null and Alternate Hypotheses for this claim are -

Null Hypotheses $\Rightarrow H_o: \beta_1 = 0$

And the Alternate Hypotheses $\Rightarrow H_1: \beta_1 \neq 0$

As we already know the Test statistics value is found by using the formula $t_o = \frac{\beta_j}{s.e.(\beta_j)}$

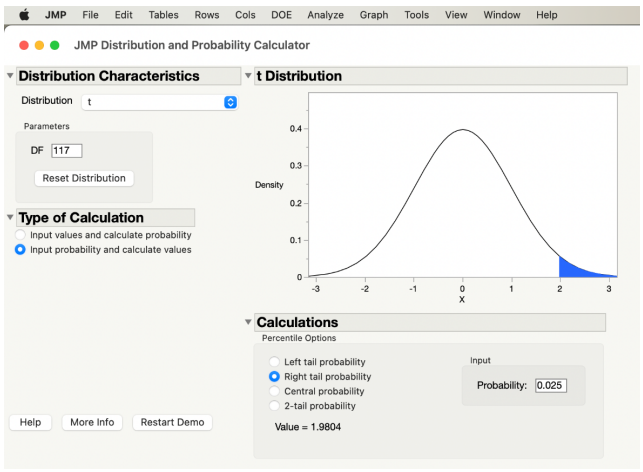
$t_o = \frac{-4.244}{6.4536} = -0.66$

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	68.176251	109.3551	0.62	0.5342
O2	-4.244353	6.453581	-0.66	0.5120
CO2	31.583586	4.696768	6.72	<.0001*

Effect Tests

Finding the value of $t_{\frac{\alpha}{2},n-k-1} \Rightarrow t_{0.025,117} = 1.9804$



P- Value = 0.5120

As $|t_o| < t_{0.025,117}$, we Fail to Reject the Null Hypotheses

Hence I can conclude that there is insufficient evidence to prove that O2 contributes significantly to the model.

R- Squared -

Now Let's understand what R-squared is and give the appropriate conclusion from the obtained R-squared value.

R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable (Minute Ventilation) that's explained by an independent variable or variables (O2 and CO2) in a regression model.

Summary of Fit

RSquare	0.294925
RSquare Adj	0.282872
Root Mean Square Error	155.377
Mean of Response	129.075
Observations (or Sum Wgts)	120

From the above table we can conclude that only 29.49% of the variation in minute ventilation is reduced by taking into account the percentages of oxygen and carbon dioxide.

Consider breathing air is 15% oxygen and 5% carbon dioxide, the mean minute ventilation is found using the Regression equation

$Y = 68.176 - 4.24X_1 + 31.58X_2$

Y = 68.176 - 4.24 (15) + 31.58 (5)

= 68.176 - 63.6 + 157.9

= 162.476

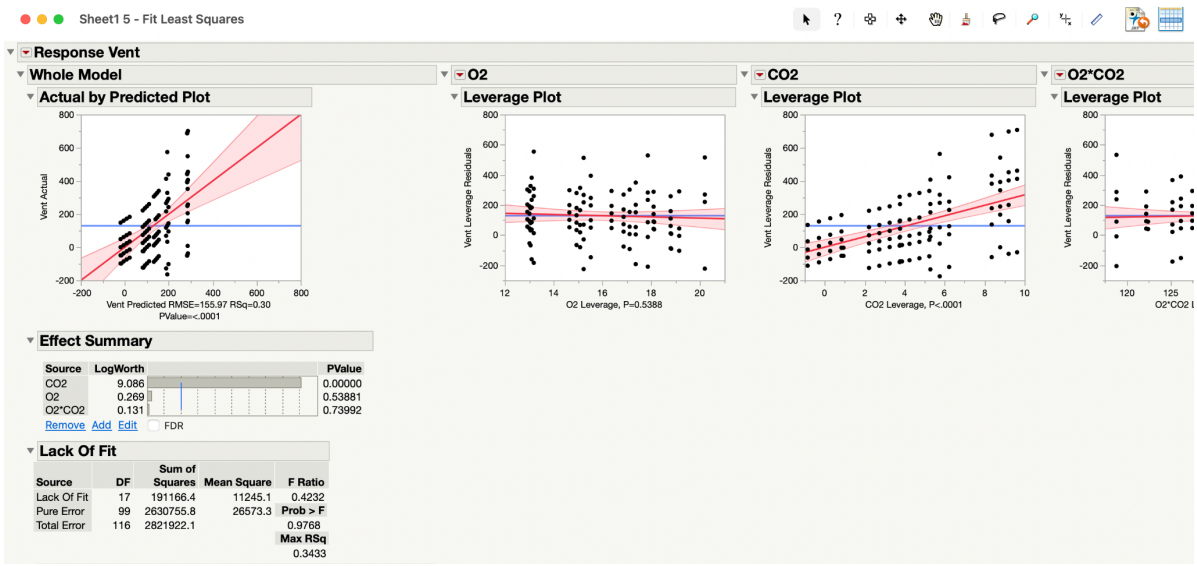
So the Mean minute ventilation whose breathing air is 15% oxygen and 5% carbon dioxide is 162.476.

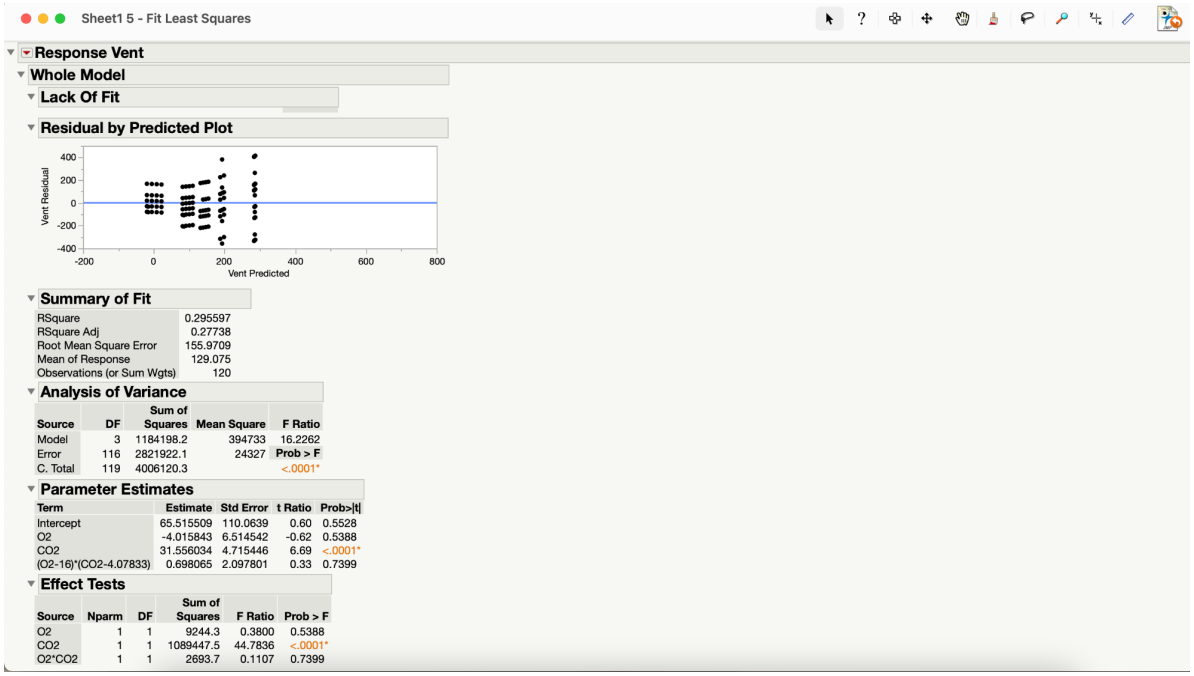
ANOVA -

An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis.

Performing ANOVA test on the dataset -

Plotting Fit Least Squares plot using JMP





Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	1184198.2	394733	16.2262
Error	116	2821922.1	24327	Prob > F
C. Total	119	4006120.3		<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
O2	1	1	9244.3	0.3800	0.5388
CO2	1	1	1089447.5	44.7836	<.0001*
O2*CO2	1	1	2693.7	0.1107	0.7399

Creating the ANOVA table using the information gathered from the above screenshots

ANOVA TABLE -

Source of Variation	Degrees of freedom	Sum of squares	F Test statistics	P-Value
O2	1	9244.3	0.3800	0.5388
CO2	1	1089447.5	44.784	<.0001
Interaction	1	2693.7	0.1107	0.7399
Error	116	2821922.1		
Total	119	3923307.6		

From the ANOVA table I can make the following conclusions and they are

1. Since the p-value for CO2 is very small, we conclude that the main effect of CO2 affects the minute ventilation.
2. But as the p-value for O2 is high, we cannot conclude that the main effect of O2 affects the minute ventilation.

3. Furthermore, since the p -value for the interaction is large, there is no indication of interaction between these factors.

Final Conclusion -

Comparing the conclusions from the Multiple Linear Regression test and ANOVA test we can say that the results obtained from both the tests are similar.