Data Analysis Report #1

Philip Booth

Introduction:

Our objective in this report is to investigate whether a relationship exists between gastric activity and metabolism levels in adults. To explore this, we analyzed data from thirty participants (18 females and 12 males), each of whom received two doses of ethanol proportional to their weight. Metabolism was measured by comparing blood alcohol concentrations following both intravenous and oral administration. The key variables in the dataset include the subject ID, sex, metabolism rate, and gastric activity level for each participant. Our primary goal was to compute the average metabolism rate across the sample and assess whether it differed significantly between males and females. Additionally, we evaluated whether sex or gastric activity levels influenced metabolism rates.

Methods:

We utilized three of the four variables from the data set: Sex, Metabolism, and Gastric Activity, which are included in *Table 1*. We generated a new variable before conducting our statistical analysis and modified the existing variables defined in *Table 2.*

Table 1: Shows the variable name and definition for each variable in the data set.

Variable Name	Variable Type	**Definition**
Subject ID	Categorical	The identification number of each Subject
Metabolism	Quantitative	Rate of alcohol metabolized, expressed in millimoles per liter-hour
Gastric Activity	Quantitative	Gastric enzyme activity, expressed in micromoles per minute per gram of tissue
Sex	Categorical	Sex of each subject (Male or Female)

Table 2 shows the variable name and definition for each of the new or modified variables used for data analysis. (N) means new variable and (M) stands for a modified variable.

Variable Name	**Definition**
sexM (N)	Indicator for whether an adult was male (sexM = 1) or not (sexM = 0)
gastric:sexM (N)	Gastric activity in micromoles/min/g of tissue for an adult male

To quantify the overall average rate of metabolism, we took the mean of the metabol column from the sample. Then, to explore if the average rate of metabolism is different between females and males, we took the two means of the metabol variable separately for each sex and conducted a two-sample t-test. To determine if there is a relationship between gastric activity and metabolism, we ran a simple linear regression with metabol as the response variable and gastric as the explanatory variable. Finally, analysis in establishing whether there exists a change in the relationship based on sex was done by running a multi-variable regression with the sex indicator, gastric activity level, and testing the interaction variable between gastric activity and the sex indicator as explanatory variables for predicting the metabolism level as the response variable. All conditions for linear and multi-variable regression were met, so no further changes to the variables were necessary. All these tests and analyses were performed in R.

Results:

Tables 3 and 4 provide descriptive statistics of the data frame. Of the thirty subjects who participated in the research experiment, sixty percent were female, and forty percent were male. *Table 4* shows summary statistics of the two primary variables in the dataset, Gastric Activity and Metabolism. In *Figures 1 and 2*, we visualized data from *Table 3* using histograms. The two figures capture the distribution of the predictor(GA) and response(M) variables.

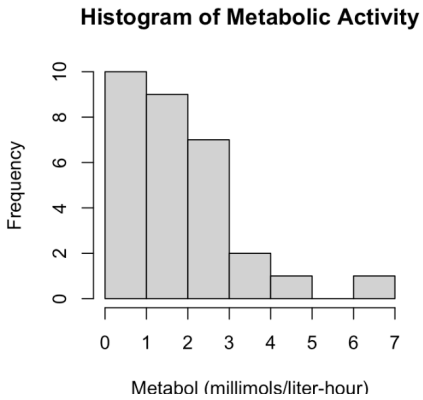
Table 3: shows the proportion of Male/Female subjects in the Sex column of Dataset

Sex	**Proportion**	**Percentage**

Female	18/30	60%
Male	12/30	40%

Table 4: provides summary statistics of the two explanatory and response variable (Gastric vs. Metabol)

Statistic:	**Mean**	**Standard Deviation**	**Median**	**IQR**
Gastric	1.6766667	0.6333727	1.5500	1.0000
Metabol	1.856667	1.460204	1.5500	2.1000



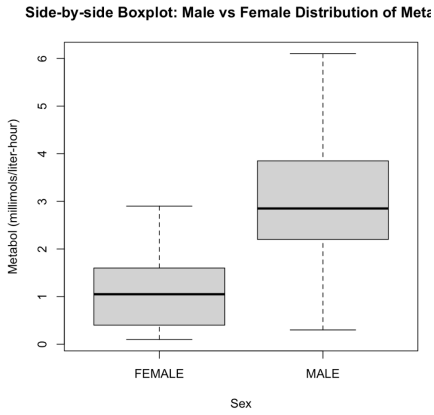


Figure 1. Distribution of gastric activity *Figure 2. Distribution of alcohol metabolism* *Figure 3. Comparison of Metabolic Rates by Sex*

Figure 3 compares the distributions of Metabolism between the Male and Female subjects. After glancing at this figure, we were prompted to run a difference in means test between Metabolism and Sex. After conducting a t-test between these two variables, we calculated t = –4.001 and a p-value of 0.00109, indicating the difference in means between male and female metabolism rates was significant. Moving on, we created a simple linear regression and then ran a t-test for slope. We determined that gastric activity significantly impacts metabolism (p-value: .00199) and calculated an R-squared value of .3951. The equation for the slr is M = -0.5732 + 1.4492(GA), also provided in Figure 4.

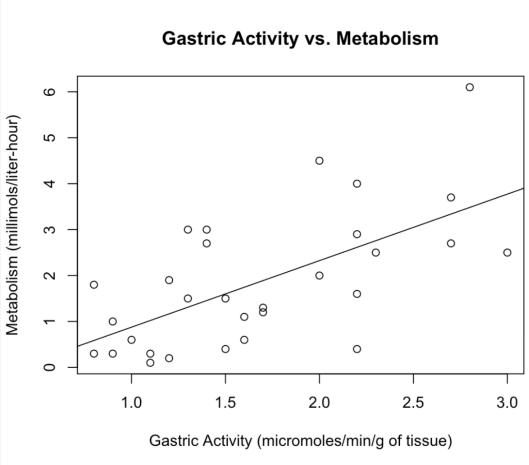


Figure 4: Scatter plot of Gastric Activity (micromoles/min/g of tissue) and Metabolism (millimols/liter-hour). Equation of SLR: Metabolism = -0.5732 + 1.4492(Gastric Activity)

Building on the findings from the previous two tests, we conducted a final MLR to assess whether the indicator variable, sex, significantly influences the model. The results of the MLR suggest that sex impacted the model's intercept (p-value of SexM: 0.000139). The new MLR accounting for sex as an indicator variable model has an R-squared of .8953. The two regression equations produced by the MLR are MetabolismFemale = -0.6823 + 1.1498(GA). MetabolismMale = 0.845 + 1.1498(GA). The overall equation is Metabolism = -0.6823 + 1.1498(GA) + 1.5276(SexM)

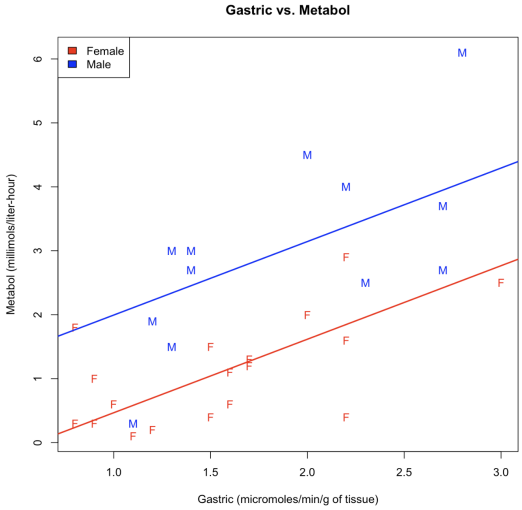


Figure 5: A scatter plot describing the relationship between gastric activity and metabolism while accounting for Male and Female subjects. Equation: Metabolism = -0.6823 + 1.1498(Gastric Activity) + 1.5276(Sex Male).

Lastly, we tested for an MLR for a difference in slopes between Sex by including an interaction term between GA and SexM. After failing the hypothesis test (p-value gastric: SexM = 0.1885), we concluded that the interaction was unnecessary to include in our final model.

Discussion:

We found that alcohol metabolism is significantly associated with both gastric activity and sex. All models presented in our results section yielded meaningful insights. Side-by-side boxplots displaying the distributions of alcohol metabolism for male and female subjects indicate a statistically significant difference confirmed by a t-test (p-value: 0.00109). Identifying this difference early in the analysis facilitated the development of the final multiple linear regression model. To address the primary research question of whether gastric activity was related to metabolism, we created a single linear regression model to examine whether a significant slope existed between the two variables. After concluding that a relationship did exist, we expanded our model to capture Sex as a predictor. Our resulting MLR model suggested that for every one unit increase in the enzyme measured in m/min/g of tissue for Gastric Activity, one can expect a 1.1498 increase in m/liter-hour of Alcohol Metabolism and a flat increase of 1.5276 m/liter-hour of Alcohol Metabolism when the subject is Male. From our final model, we gleaned the rate of alcohol metabolism (1.1498) and that sex changed the relationship between gastric activity and metabolism. The slope coefficient for gastric activity decreased from 1.4492 to 1.1498 in the MLR when Sex was included, indicating that sex changes the relationship between the predictor and response. The indicator variable (SexM) from our MLR suggests that Male subjects are expected to have an increase in 1.5276 millimoles/liter-hour of Alcohol Metabolism compared with Female subjects. After testing for significance in slopes for the two lines by running an MLR on the interaction term(Gastric Activity * Sex), we concluded that a difference in the rate

of metabolism between sexes did not improve the model's precision (p-value: 0.1885). There were a few limitations to this assignment. The distribution of ages across all subjects was not spread, ranging from 20 to 34 years of age, and a side-by-side boxplot between the lower upper halves of age showed no significance. A sample size greater than 30 is preferred; however, the data points were enough to draw significant conclusions across all our tests. In conclusion, we found a significant and positive relationship between gastric activity and metabolism. Our final MLR model suggests that an individual's Sex impacts the baseline/intercept of metabolic rate but does indicate a change in the slope of metabolism.

Appendix:

Descriptive Statistics

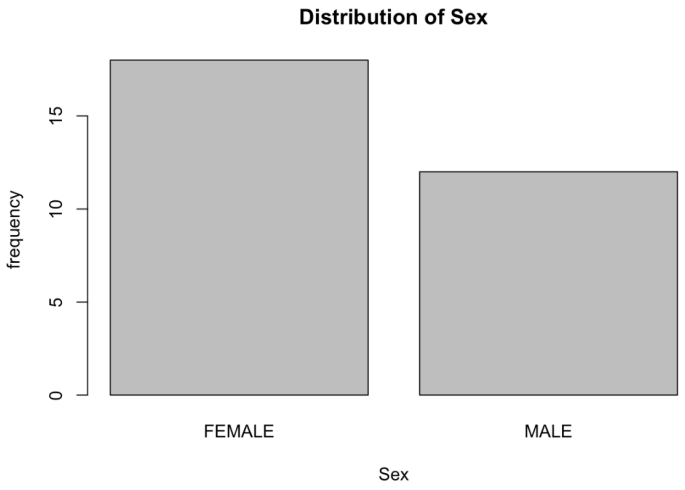
To develop an understanding of the variables in the metabol.csv dataset, we took summary statistics of the two primary quantitative variables using both a table and visualized them in respective histograms. For the primary categorical variable, Sex, we evaluated the distribution of Male and Female participants using a boxplot. The summary table corresponds to *Table 1,* the bar plot of Sex corresponds to *Table 3,* and the two histograms correspond to *Figures 1 & 2*.

```
summary_table <- data.frame(
  Statistic = c("Mean", "Standard Deviation", "Median", "IQR"),
  Gastric = c(mean(df$gastric),sd(df$gastric),median(df$gastric), IQR(df$gastric)),
  Metabol = c(mean(df$metabol),sd(df$metabol),median(df$metabol), IQR(df$metabol)))

t(summary_table)
```

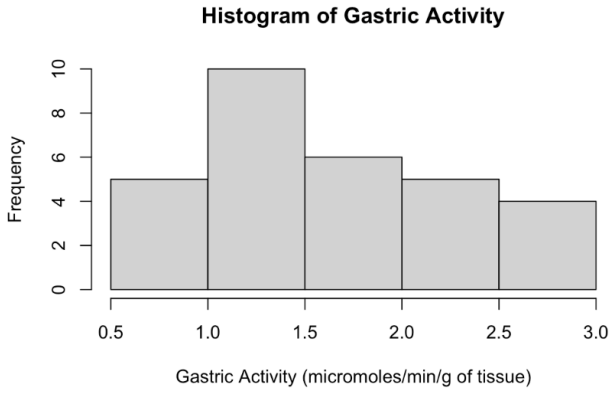
```
##           [,1]        [,2]                [,3]        [,4]
## Statistic "Mean"      "Standard Deviation" "Median"    "IQR"
## Gastric   "1.6766667" "0.6333727"          "1.5500000" "1.0000000"
## Metabol   "1.856667"  "1.460204"           "1.550000"  "2.100000"
```

```
barplot(sex_table,xlab = "Sex", ylab = "frequency", main = "Distribution of Sex")
```

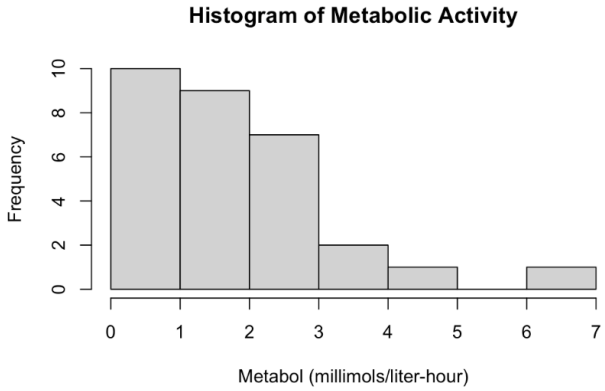


Create histogram distributions of the data

```
hist(x=df$gastric, xlab = 'Gastric Activity (micromoles/min/g of tissue)', main = 'Histogram of Gastric Activit
y')
```




```
hist(x=df$metabol, xlab = 'Metabol (millimols/liter-hour)', main = 'Histogram of Metabolic Activity')
```



Hypothesis Test for difference in means

Parameter: $\mu_1 - \mu_0$

H_0: $\mu_0 = \mu_1$
H_A: $\mu_0 \neq \mu_1$

Conditions:
1) Independent Samples
2) Population is normally distributed or Sample size is greater or equal to 30

```
t.test(df$metabol ~ df$sex)
```

```
##
##  Welch Two Sample t-test
##
## data:  df$metabol by df$sex
## t = -4.001, df = 15.505, p-value = 0.00109
## alternative hypothesis: true difference in means between group FEMALE and group MALE is not equal to 0
## 95 percent confidence interval:
##  -2.8965660 -0.8867673
## sample estimates:
## mean in group FEMALE   mean in group MALE
##             1.100000             2.991667
```

P-value is significant at 0.00109, so we can reject the null hypothesis and assume there is a significant difference in metabolism between Sexes.

Single Linear Regression for T-test

LINES Conditions

To pass LINES, one must assess five conditions. First, the predictor and response variables in the SLR must exhibit a linear relationship. I'm using Figure 6, a scatter plot, to demonstrate the linear relationship between M and GA. The second condition ensures independent groups, which is confirmed since the data was collected with subject ID numbers, guaranteeing no overlap in the sample. The third condition is the normality of residuals. Figure 7's normal Q-Q plot verifies that the residuals from the SLR roughly follow a normal distribution. The fourth condition addresses error variance, meaning the residuals should maintain a consistent spread across values. The RVF plot from Figure 8 indicates that the residuals are adequately scattered throughout the plot. Finally, the data must be collected through simple random sampling, which is confirmed for this experiment. Since all the conditions are satisfied, we can proceed with the sample t-test to assess the regression slope. Using the residual leverage plot in the plot function, we determined that there were no significant outliers in the dataset either. The RVL plot identifies potential outliers that could impact the regression model, using cooks' distance to calculate which points with large residuals significantly influence the model.

Create basic plots of the data

```
plot(df$metabol ~ df$gastric,ylab= "Metabolism (millimols/liter-hour)", xlab='Gastric Activity (micromoles/min/g of tissue)', main='Gastric Activity vs. Metabolism')
```

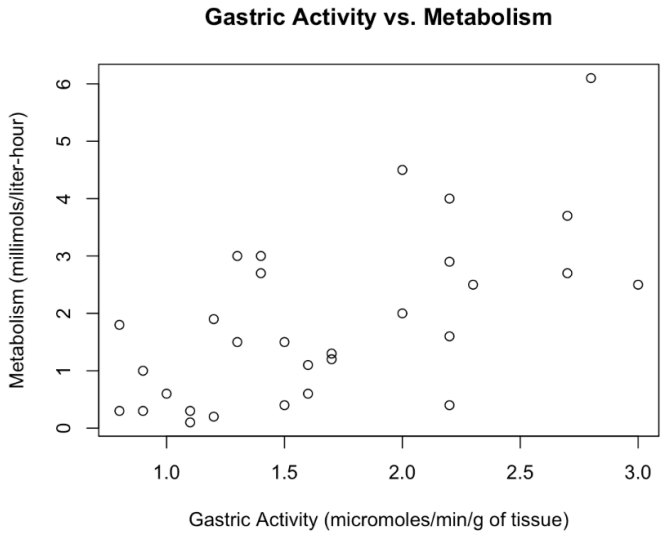


Figure 6: Scatter plot between gastric activity and metabolism

```
slr_model <- lm(df$metabol ~ df$gastric)
```

```
plot(slr_model)
```

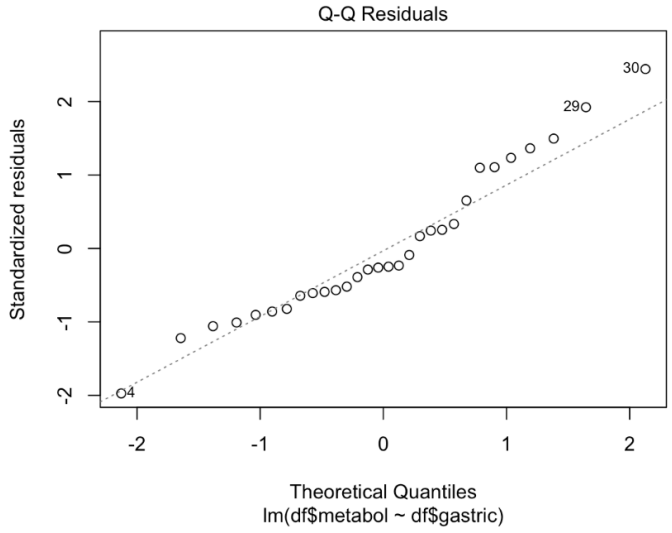


Figure 7: Normal Q-Q Plot of Residuals for Metabolism vs. Gastric Activity

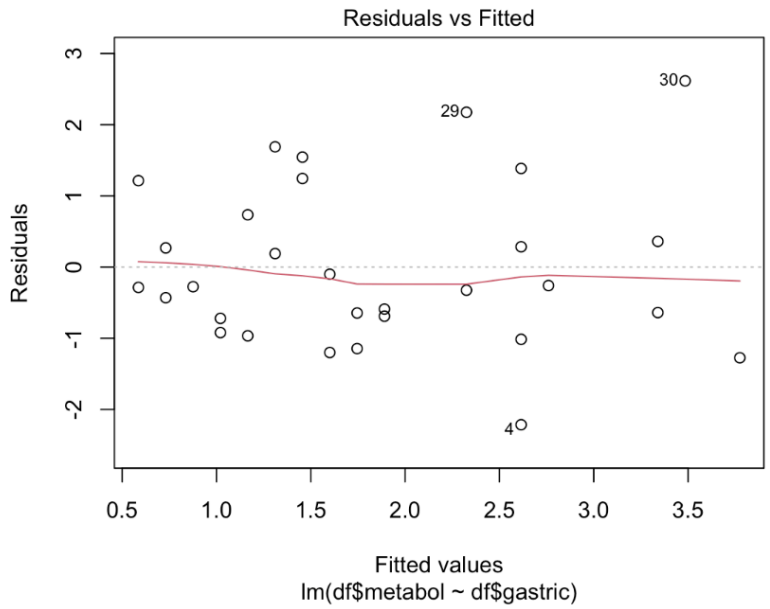


Figure 8. Residual vs Fitted plot for Metabolism vs. Gastric Activity Regression Model

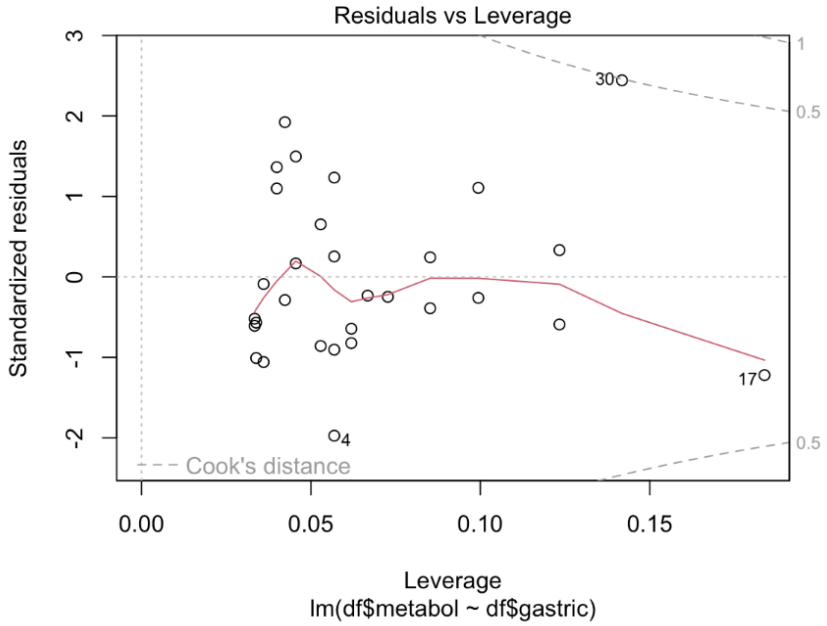


Figure 9. Residuals vs. Leverage Plot for Metabolism vs. Gastric Activity Regression Model

T-test for Slope using SLR

Parameter: β_1

H_0: $\beta_1 = 0$
H_A: $\beta_1 \neq 0$

Conditions: LINES (Passed)

```
slr_model <- lm(df$metabol ~ df$gastric)
summary(slr_model)
```

```
##
## Call:
## lm(formula = df$metabol ~ df$gastric)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2151 -0.7133 -0.2811  0.6407  2.6154
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5732     0.6060  -0.946 0.352358
## df$gastric    1.4492     0.3388   4.277 0.000199 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.156 on 28 degrees of freedom
## Multiple R-squared:  0.3951, Adjusted R-squared:  0.3735
## F-statistic: 18.29 on 1 and 28 DF,  p-value: 0.0001992
```

The test statistic for the slope 4.277 yields a p-value of 0.000199, so we reject the null hypothesis and conclude that the regression model is significant and that the slope is 1.4492. The R-squared value is .3951, which means that 39.51% of the variance in the model can be explained by gastric activity. The equation for the SLR is as follows: Metabolism = -0.5732 + 1.4492(Gastric Activity). The resulting SLR plot corresponds to figure 4:

```
plot(df$metabol ~ df$gastric,ylab= "Metabolism (millimols/liter-hour)", xlab='Gastric Activity (micromoles/min/g
of tissue)', main='Gastric Activity vs. Metabolism')

slr_model <- lm(df$metabol ~ df$gastric)
abline(slr_model)
```

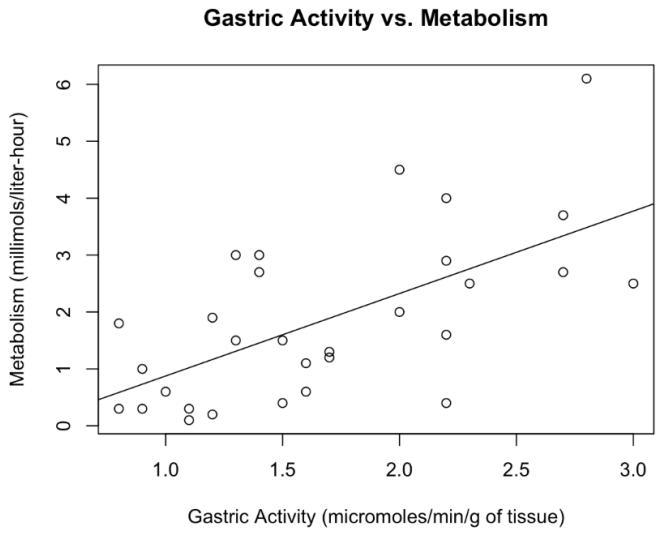


Figure 4: Scatter plot of Gastric Activity (micromoles/min/g of tissue) and Metabolism (millimols/liter-hour). Equation of SLR: Metabolism = -0.5732 + 1.4492(Gastric Activity)

Capturing the impact of Sex on Metabolism vs. Gastric Activity (MLR)

Here, we create a multiple linear regression including the indicator variable SexM into the original SLR to determine whether sex impacts the model.

H_0: β_2 (SexM) has a significant impact on model
H_A: β_2 (SexM) does not has a significant impact on model

Conditions: LINES (Passed)

- ★ Note that in the MLR LINES conditions, rather than check linearity via scatterplot, one must check via the RVF plot, which passed the constant variance condition. Therefore, all LINES conditions hold continuity from SLR → MLR.

```r
#Hypothesis test
model_mlr <- lm(metabol ~ gastric + sexM, data = df)
summary(model_mlr)
```

```
##
## Call:
## lm(formula = metabol ~ gastric + sexM, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81012 -0.49381 -0.05505  0.52307  2.03515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.6823     0.4701  -1.451 0.158244
## gastric       1.1498     0.2710   4.242 0.000232 ***
## sexM          1.5276     0.3445   4.434 0.000139 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8953 on 27 degrees of freedom
## Multiple R-squared:   0.65,  Adjusted R-squared:  0.6241
## F-statistic: 25.07 on 2 and 27 DF,  p-value: 7e-07
```

The test statistic in SexM 4.434 yields a p-value of 0.000139, so we can reject the null hypothesis and conclude that the indicator Sex significantly influences the relationship of the model. The R-squared value is now .65 when including SexM in the model, which can be interpreted that 65% of the variability in the model can be explained by gastric activity and SexM (female = 0,male = 1). The equation for the MLR model is as follows:

Metabolism = -0.6823 + 1.498(Gastric Activity) + 1.5276(SexM)

Here's the code and plot for the two-line MLR regression model corresponding to Figure 5:

```r
plot(df$metabol ~ df$gastric , type = "n",xlab = "Gastric (micromoles/min/g of tissue)", ylab = "Metabol (millimo
ls/liter-hour)",
     main = "Gastric vs. Metabol") +
points(metabol ~ gastric, pch= 'F',  col = "red", data = df[df$sexM =="0",]) +
points(metabol ~ gastric, pch='M',col = "blue", data = df[df$sexM =="1",])
```

```
## integer(0)
```

```r
legend(x = 'topleft', legend= c('Female', 'Male'), fill=c('red','blue'))


#Plot the regression lines for Male vs. Female
model_dummy <- lm(metabol ~ gastric + sexM, data = df)
abline(model_dummy$coeff[1], model_dummy$coeff[2],col="red",lty=1, lwd=2)
abline(model_dummy$coeff[1] + model_dummy$coeff[3], +
       model_dummy$coeff[2],col="blue",lty=1, lwd=2)
```

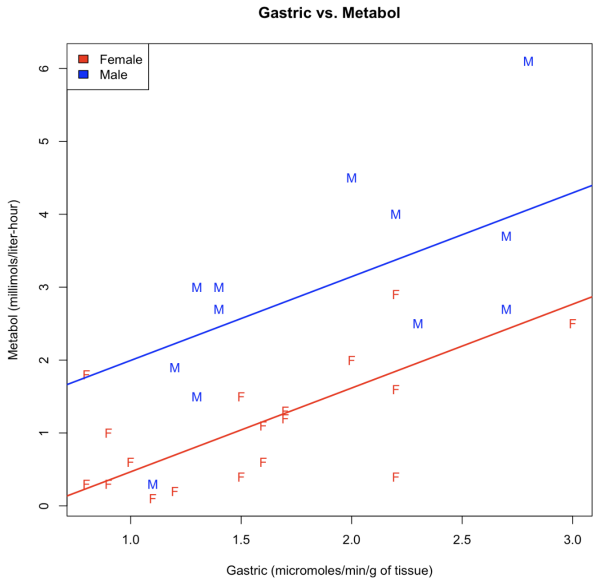


Figure 5: A scatter plot describing the relationship between gastric activity and metabolism while accounting for Male and Female subjects. Equation: Metabolism = -0.6823 + 1.1498(Gastric Activity) + 1.5276(Sex Male).

Evaluating the interaction between Gastric Activity and SexM to determine if it should be included in the MLR model.

After concluding that the indicator variable provided by Sex improved the accuracy of our model, we wanted to see if including a difference in slopes between Sexes would further improve accuracy. We included an interaction term in the new model and evaluated its significance.

Parameter: Interaction term β_3 (Gastric:SexM)

H_0: β_3 (Gastric:SexM) has a significant impact on model
H_A: β_3 (Gastric:SexM) does not has a significant impact on model

Conditions: LINES (passed)

```r
#Hypothesis test
model_mlr <- lm(metabol ~ gastric + sexM + gastric*sexM , data = df)
summary(model_mlr)
```

```
##
## Call:
## lm(formula = metabol ~ gastric + sexM + gastric * sexM, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.59619 -0.60249 -0.04076  0.47590  1.64726
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.1973     0.5860  -0.337   0.7391
## gastric        0.8369     0.3535   2.368   0.0256 *
## sexM           0.2668     0.9932   0.269   0.7904
## gastric:sexM   0.7285     0.5394   1.351   0.1885
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8819 on 26 degrees of freedom
## Multiple R-squared:  0.6729, Adjusted R-squared:  0.6352
## F-statistic: 17.83 on 3 and 26 DF,  p-value: 1.711e-06
```

The test statistic for the interaction term 1.351 yields a p-value of 0.1885, indicating that we cannot reject the null hypothesis and that the rate between gastric activity and metabolism does not significantly differ between sexes.