# Simple Linear Model Report - FA3

Far Eastern University
APM1205: Linear Model
Mr. Teodolfo Bonitez
March 4, 2024

ROSALES, Frances Aneth C.

# FA3

Frances Aneth Rosales

2024-03-04

## Load The Data

```
loandata <- read_excel("C:\\Users\\asus\\Documents\\ALL FEU FILES\\FEU FOLDER 6\\LINEA
R MODEL\\FA3\\Loan_Data.xlsx")

loandata

## # A tibble: 614 × 13

##    Loan_ID  Gender Married Dependents Education  Self_…¹ Appli…² Coapp…³ LoanA…⁴

##    <chr>    <chr>  <chr>   <chr>      <chr>      <chr>     <dbl>   <dbl>   <dbl>

##  1 LP001002 Male   No      0          Graduate   No         5849       0      NA

##  2 LP001003 Male   Yes     1          Graduate   No         4583    1508     128

##  3 LP001005 Male   Yes     0          Graduate   Yes        3000       0      66

##  4 LP001006 Male   Yes     0          Not Gradu… No         2583    2358     120

##  5 LP001008 Male   No      0          Graduate   No         6000       0     141

##  6 LP001011 Male   Yes     2          Graduate   Yes        5417    4196     267

##  7 LP001013 Male   Yes     0          Not Gradu… No         2333    1516      95

##  8 LP001014 Male   Yes     3+         Graduate   No         3036    2504     158

##  9 LP001018 Male   Yes     2          Graduate   No         4006    1526     168

## 10 LP001020 Male   Yes     1          Graduate   No        12841   10968     349

View(loandata)
```

# Loan Amount vs Total Income

In this field, we created a new column for the Total Income of applicant which includes "Applicant Income" and Coapplicant Income".

```
loandata$Total_Income<- loandata$ApplicantIncome + loandata$CoapplicantIncome


loan_vs_total_income <- lm(LoanAmount ~ Total_Income, data = loandata)
summary(loan_vs_total_income)
## Call:
## lm(formula = LoanAmount ~ Total_Income, data = loandata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -391.83  -27.55   -6.75   20.99  396.19
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.872e+01  4.047e+00   21.92   <2e-16 ***
## Total_Income 8.186e-03  4.214e-04   19.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.89 on 590 degrees of freedom
##   (22 observations deleted due to missingness)
## Multiple R-squared:  0.3902, Adjusted R-squared:  0.3891
## F-statistic: 377.5 on 1 and 590 DF,  p-value: < 2.2e-16
```

***Result***

$$Yi=\beta 0+\beta 1Xi+\varepsilon i$$

**Formula: LoanAmount= (88.72) + Total Income (0.008186)+ ε** , as we compare the Loan Amount and applicant's Total Income.

Upon analyzing, the coefficient for Total_Income is (0.008186) indicates that, for a one-unit increase in applicant's Total_Income, the LoanAmount is expected to increase by 0.008186 units.

Additionally, the estimated intercept $\beta 0$ =(88.72) represents the estimated LoanAmount when the Total_Income is zero. However, in reality, the concept of having zero income for applicants loaning in bank would be unrealistic since customer's income should also be the requirement for loaning.

Lastly, we can interpret the *p-value: < 2.2e-16* which is considered extremely small. Since the p-value is smaller than significance thresholds like 0.05, which indicates that there is strong evidence against the null hypothesis.

# Loan Amount vs Married Status

```
loan_vs_Married <- lm(LoanAmount ~ Married, data = loandata)
summary_married<- summary(loan_vs_Married)
summary_married
##
## Call:
## lm(formula = LoanAmount ~ Married, data = loandata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -138.75  -45.85  -18.88   20.00  544.25
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  128.883      5.911  21.804  < 2e-16 ***
## MarriedYes    26.867      7.327   3.667 0.000268 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.84 on 588 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.02236,    Adjusted R-squared:  0.02069
## F-statistic: 13.45 on 1 and 588 DF,  p-value: 0.0002678
```

***Result***

$$Yi=\beta 0+\beta 1Xi+\varepsilon i$$

**Formula: LoanAmount= (128.883) + Married(Yes) (26.867)+ ε** , as we compare the Loan Amount and applicant's status of marriage.

The estimated intercept *β0* =(128.883) represents the estimated LoanAmount when the applicant is not married.

Additionally, upon analyzing, the coefficient for Married applicant is (26.867) indicates that, on average, married applicants have an expected LoanAmount approximately 26.867 units higher than unmarried applicants.

Lastly, we can interpret the *p-value: 0.0002678* which is also considered extremely small. The fact that the p-value is below significance thresholds like 0.05 indicates that there is an evidence once more that the null hypothesis is not supported.

# Loan Amount vs Self-Employed

```
loan_vs_Self_Employed <- lm(LoanAmount ~ Self_Employed, data = loandata)
summary(loan_vs_Self_Employed)
##
## Call:
## lm(formula = LoanAmount ~ Self_Employed, data = loandata)
##
## Residuals:
##      Min       1Q  Median       3Q     Max
## -147.00  -44.00  -17.75   19.00  558.25
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       141.749      3.844  36.872  < 2e-16 ***
## Self_EmployedYes   30.251     10.245   2.953  0.00328 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.4 on 559 degrees of freedom
##   (53 observations deleted due to missingness)
## Multiple R-squared:  0.01536,    Adjusted R-squared:  0.0136
## F-statistic:  8.72 on 1 and 559 DF,  p-value: 0.00328
```

***Result***

$$Yi=\beta 0+\beta 1Xi+\varepsilon i$$

**Formula: LoanAmount= (141.749) + Self-Employed(Yes) (30.251)+ ε** , as we compare variables Loan Amount and applicant's being Self-Employed.
Upon analyzing, the coefficient for being Self-Employed is (30.251) indicates that,self-employed applicants have an expected Loan Amount approximately 30.251 units higher than non-self-employed applicants.
Additionally, the estimated intercept $\beta 0$ =(141.749) represents the estimated LoanAmount when the our independent variable Self_Employed is "No" (Self_EmployedNo). In other words, for applicants who are not self-employed, the expected LoanAmount is 141.749.
Lastly, we can interpret the *p-value: 0.00328* < 0.05, thus, shows we can go against the null hypothesis.

# Loan Amount vs Gender

```
loan_vs_Gender <- lm(LoanAmount ~ Gender, data = loandata)


summary(loan_vs_Gender)
##
## Call:
## lm(formula = LoanAmount ~ Gender, data = loandata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -132.27  -45.27  -18.27   22.73  500.73
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  126.697      7.870  16.099   <2e-16 ***
## GenderMale    22.569      8.735   2.584     0.01 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.16 on 577 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.01144,    Adjusted R-squared:  0.009724
## F-statistic: 6.676 on 1 and 577 DF,  p-value: 0.01002
```

***Result***

$$Yi=\beta 0+\beta 1Xi+\varepsilon i$$

**Formula: LoanAmount= (126.697) + Gender(Male) (22.569)+ ε** , as we compare the Loan Amount and applicant's Gender as Male.
The estimated intercept *β0* =(126.697) represents the estimated LoanAmount when the applicant is Female. Additionally, upon analyzing, the coefficient for Male Applicants applicant is (22.569) indicates that, male applicants have an expected Loan Amount approximately 22.569 units higher than female applicants. Lastly, we can interpret the *p-value: 0.01002* which is also considered extremely small. Since the p-value< 0.05, thus, shows that there is go against the null hypothesis.

# Loan Amount vs Property Area

```
loan_vs_Property_Area <- lm(LoanAmount ~ Property_Area, data = loandata)
summary(loan_vs_Property_Area)
##
## Call:
## lm(formula = LoanAmount ~ Property_Area, data = loandata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -133.20  -45.50  -20.26   19.74  557.80
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               152.260      6.511  23.385   <2e-16 ***
## Property_AreaSemiurban     -6.756      8.635  -0.782    0.434
## Property_AreaUrban        -10.061      8.988  -1.119    0.263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.64 on 589 degrees of freedom
##   (22 observations deleted due to missingness)
## Multiple R-squared:  0.002193,   Adjusted R-squared:  -0.001195
## F-statistic: 0.6473 on 2 and 589 DF,  p-value: 0.5238
```

***Result***

$$Yi=\beta 0+\beta 1Xi+\varepsilon i$$

**Formula:**
**LoanAmount= (152.260) + Property_AreaSemiurban (-6.756)+ Property_AreaUrban(-10.061)+ε** , as we compare the Loan Amount and applicant's Property Area.
The estimated intercept β0 =(152.260) represents the estimated LoanAmount when the categorical variable Property_Area is "Rural". Therefore, for applicants from rural areas, the expected LoanAmount is 152.260. Additionally, upon analyzing, the coefficient for Applicants in Semiurban Applicants indicates that the LoanAmount is expected to be lower by 6.756 units for applicants compared to Rural areas, while Urban Applicants expected to be lower by 10.061 loan amount compared to Rural areas.
Lastly, we can interpret p-value which is 0.5238 which is unfortunately, 0. 5238 > 0.05, thus we cannot reject the null hypothesis.

# Summary

**List of least p-value**

- Total Income (p-value: < 2.2e-16):
- Married (p-value: 0.0002678):
- Self-Employed (p-value: 0.00328):
- Gender (p-value: 0.01002):
- Property Area (p-value: 0.5238):

**CONCLUSION**

Total Income, followed by Gender, Self-Employment status, and Marital Status (Married), seems to be the most significant variable in predicting Loan Amount based on the p-values. Property Area's greater p-value suggests that it is less significant.

## Final Equation:

$$LoanAmount=\beta 0+\beta 1\times \text{Total Income}+\beta 2\times \text{Married (Yes)}+\beta 3\times \text{Self-Employed (Yes)}+\beta 4\times \text{Gender (Male)}+\beta 5\times \text{Property Area (Semiurban)}+\beta 6\times \text{Property Area (Urban)}+\varepsilon$$

# Loan Amount vs Total Income (loglm)

```
library(MASS)

contingency_table_loan_vs_Total_Income <- table(loandata$LoanAmount, loandata$Total_In
come)


loglm_model_loan_vs_Total_Income <- loglm(Freq ~ ., data = as.data.frame(contingency_t
able_loan_vs_Total_Income))


df_Total_Income <- loglm_model_loan_vs_Total_Income$df

lrt_Total_Income <- loglm_model_loan_vs_Total_Income$lrt

pearson_value_Total_Income <- loglm_model_loan_vs_Total_Income$pearson



p_value_Total_Income <- 1 - pchisq(lrt_Total_Income, df = df_Total_Income)

print("Statistics:")

## [1] "Statistics:"

result_p_Total_Income <- data.frame(

  Test = c("Likelihood Ratio Test", "Pearson Test"),

  P_Value = c(p_value_Total_Income, pearson_value_Total_Income)

)

print(result_p_Total_Income)

##                    Test P_Value

## 1 Likelihood Ratio Test       1

## 2          Pearson Test     NaN
```

The log-linear model examining the relationship between Loan Amount and Total Income a **p-value of 1.00** suggests that the test did not find evidence to disclose the null hypothesis, indicating **insufficient of significance** in our data. It is important to recognize that the continuous traits of the variables make Pearson’s chi-square inapplicable in this scenario.

# Loan Amount vs Self-Employed (loglm)

```
contingency_table_loan_vs_Self_Employed <- table(loandata$LoanAmount, loandata$Self_Em
ployed)


loglm_model_loan_vs_Self_Employed <- loglm(Freq ~ ., data = as.data.frame(contingency_
table_loan_vs_Self_Employed))


df_Total_Self_Employed <- loglm_model_loan_vs_Self_Employed$df

lrt_Total_Self_Employed <- loglm_model_loan_vs_Self_Employed$lrt

pearson_value_Self_Employed <- loglm_model_loan_vs_Self_Employed$pearson




p_value_Total_Self_Employed <- 1 - pchisq(lrt_Total_Self_Employed, df = df_Total_Self_
Employed)

print("Statistics:")

## [1] "Statistics:"

result_p_Total_Self_Employed <- data.frame(

  Test = c("Likelihood Ratio Test", "Pearson Test"),

  P_Value = c(p_value_Total_Self_Employed, pearson_value_Self_Employed)

)

print(result_p_Total_Self_Employed)

##                    Test   P_Value

## 1 Likelihood Ratio Test 0.7686985

## 2          Pearson Test       NaN
```

The likelihood between Loan Amount and Self_Employed yields a **p-value of 0.7687**, indicating that there is **no correlation between these two variables**. Therefore, based on this test, Self_Employed status does not have a strong significant impact on Loan Amount in the dataset.

# Loan Amount vs Gender (loglm)

```
contingency_table_loan_vs_Gender <- table(loandata$LoanAmount, loandata$Gender)


loglm_model_loan_vs_Gender <- loglm(Freq ~ ., data = as.data.frame(contingency_table_l
oan_vs_Gender))


df_gender <- loglm_model_loan_vs_Gender$df

lrt_gender <- loglm_model_loan_vs_Gender$lrt

pearson_gender <- loglm_model_loan_vs_Gender$pearson




p_value_gender <- 1 - pchisq(lrt_gender, df = df_gender)

print("Statistics:")

## [1] "Statistics:"

result_p_value_gender <- data.frame(

  Test = c("Likelihood Ratio Test", "Pearson Test"),

  P_Value = c(p_value_gender, pearson_gender)

)

print(result_p_value_gender)

##                    Test   P_Value

## 1 Likelihood Ratio Test 0.2800676

## 2          Pearson Test       NaN
```

The likelihood between Loan Amount and Gender accumulate **p-value of 0.2801**. This suggests that there is **no significant association between Gender and Loan Amount** in the dataset. Therefore, based on this test, Gender does not have a significant impact on Loan Amount.

# Loan Amount vs Property Area (loglm)

```
contingency_table_loan_vs_Property_Area <- table(loandata$LoanAmount, loandata$Propert
y_Area)


loglm_model_loan_vs_Property_Area <- loglm(Freq ~ ., data = as.data.frame(contingency_
table_loan_vs_Property_Area))



df_Total_Property_Area <- loglm_model_loan_vs_Property_Area$df
lrt_Total_Property_Area <- loglm_model_loan_vs_Property_Area$lrt
pearson_value_Property_Area <- loglm_model_loan_vs_Property_Area$pearson




p_value_Total_Property_Area <- 1 - pchisq(lrt_Total_Property_Area, df = df_Total_Prope
rty_Area)


p_pearson_Total_Property_Area <- 1 - pchisq(pearson_value_Property_Area, df = df_Total
_Property_Area)


print("Statistics:")
## [1] "Statistics:"
result_p_Total_Property_Area <- data.frame(
  Test = c("Likelihood Ratio Test", "Pearson Test"),
  P_Value = c(p_value_Total_Property_Area, p_pearson_Total_Property_Area)
)
print(result_p_Total_Property_Area)
##                    Test     P_Value
## 1 Likelihood Ratio Test 0.001137601
## 2          Pearson Test 0.288010874
```

The likelihood ratio between Loan Amount and Property Area results in a **p-value of 0.0011**, indicating a statistically significant association between these variables. This suggests that **Property Area has a significant impact on Loan Amount** in the dataset. However, the Pearson chi-square test does not show a significant association.

# Loan Amount vs Married Status(loglm)

```
contingency_table_loan_vs_Married <- table(loandata$LoanAmount, loandata$Married)


loglm_model_loan_vs_Total_Married <- loglm(Freq ~ ., data = as.data.frame(contingency_
table_loan_vs_Married))


df_Total_Married <- loglm_model_loan_vs_Total_Married$df

lrt_Total_Married <- loglm_model_loan_vs_Total_Married$lrt

pearson_Total_Married <- loglm_model_loan_vs_Total_Married$pearson



p_value_Total_Married <- 1 - pchisq(lrt_Total_Married, df = df_Total_Married)


p_pearson_Total_Married <- 1 - pchisq(pearson_Total_Married, df = df_Total_Married)


print("Statistics:")

## [1] "Statistics:"

result_p_value_Total_Married <- data.frame(

  Test = c("Likelihood Ratio Test", "Pearson Test"),

  P_Value = c(p_value_Total_Married, p_pearson_Total_Married)

)

print(result_p_value_Total_Married)

##                    Test     P_Value

## 1 Likelihood Ratio Test 0.004475452

## 2          Pearson Test 0.389058983
```

The likelihood ratio between Loan Amount and Marital Status (Married) yields a **p-value of 0.0045**, indicating a statistically significant association between variables. This suggests that **Marital Status has a significant impact on Loan Amount** in the dataset. However, the Pearson chi-square test does not show a significant association again.

# SUMMARY

```
library(dplyr)
New_result_p_value_Total_Married <- result_p_value_Total_Married %>%
  mutate(Variable = "Married Variables")

New_result_p_Total_Property_Area <- result_p_Total_Property_Area %>%
  mutate(Variable = "Property Area Variables")



New_result_p_value_gender <- result_p_value_gender %>%
  mutate(Variable = "Gender Variables")



New_result_p_Total_Self_Employed <- result_p_Total_Self_Employed %>%
  mutate(Variable = "Self Employes Variables")

New_result_p_Total_Income <- result_p_Total_Income %>%
  mutate(Variable = "Total Income Variables")



combined_data <- bind_rows(
  New_result_p_Total_Property_Area,
  New_result_p_value_gender,
  New_result_p_Total_Self_Employed,
  New_result_p_Total_Income,
  New_result_p_value_Total_Married
)

Summary_Table <- combined_data %>%
  select(Variable, Test, P_Value) %>%
  spread(Test, P_Value)



Sorted_Summary_Table <- Summary_Table %>% arrange(`Likelihood Ratio Test`)
print("List of Most Accuracy for  Linear Model (loglm)")
```

```
## [1] "List of Most Accuracy for  Linear Model (loglm)"
print(Sorted_Summary_Table)



##                   Variable Likelihood Ratio Test Pearson Test
## 1 Property Area Variables           0.001137601    0.2880109
## 2       Married Variables           0.004475452    0.3890590
## 3        Gender Variables           0.280067625          NaN
## 4 Self Employes Variables           0.768698460          NaN
## 5  Total Income Variables           1.000000000          NaN
```

## List of Most Accuracy for Linear Model (loglm)

| | Variable | Likelihood Ratio Test | Pearson Test |
|---|---|---|---|
| β1 | Property Area Variables | 0.001137601 | 0.2880109 |
| β2 | Married Variables | 0.004475452 | 0.3890590 |
| β3 | Gender Variables | 0.280067625 | NaN |
| β4 | Self Employes Variables | 0.768698460 | NaN |
| β5 | Total Income Variables | 1.000000000 | NaN |

In summary, based on the likelihood ratio test, Property Area Variables and Married Variables appear to be more significant in the log-linear model compared to Gender, Self-Employment status, and Total Income Variables. The Pearson test provides additional information, and for some variables, it didn't yield valid p-values.

### Final Equation: Likelihood Ratio Test;

$$\text{LoanAmount}=\beta 0+\beta 1\times \text{LRT for Property Area Variables}+\beta 2 \times \text{LRT for Married Variables}+\beta 3\times \text{LRT for Gender Variables}+\beta 4 \times \text{LRT for Self-Employed Variables}+\beta 5\times \text{LRT for Total Income Variables}+\varepsilon$$