Final_Jeffrey

Jack Jeffrey

Final

```
# Set working directory and CRAN mirror
setwd("/Users/jackjeffrey/Documents/Poli502_Jeffrey/final502")
options(repos = c(CRAN = "https://cran.rstudio.com/"))

# Install and load required packages
required_packages <- c("dplyr", "ggplot2", "knitr", "kableExtra", "stargazer", "pROC", "tidy

# Check and install missing packages
for (pkg in required_packages) {
  if (!requireNamespace(pkg, quietly = TRUE)) {
    install.packages(pkg)
  }
  library(pkg, character.only = TRUE)
}
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union


Attaching package: 'kableExtra'
```

```
The following object is masked from 'package:dplyr':

    group_rows


Please cite as:

 Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

 R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

Type 'citation("pROC")' for a citation.


Attaching package: 'pROC'

The following objects are masked from 'package:stats':

    cov, smooth, var

-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v stringr   1.5.1
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.1
v readr     2.1.5
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter()          masks stats::filter()
x kableExtra::group_rows() masks dplyr::group_rows()
x dplyr::lag()             masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

#1.

```
# Load the Titanic dataset
td <- read.csv("/Users/jackjeffrey/Documents/Poli502_Jeffrey/Data/titanic2.csv")


# View the structure of the dataset
str(td)
```

```
'data.frame':	1309 obs. of  16 variables:
 $ X          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ survived   : int  1 1 0 0 0 1 1 0 1 0 ...
 $ name       : chr  "Allen, Miss. Elisabeth Walton" "Allison, Master. Hudson Trevor" "Allis
 $ pclass     : int  1 1 1 1 1 1 1 1 1 1 ...
 $ age        : num  29 0.917 2 30 25 ...
 $ child      : chr  "Adult" "Child" "Child" "Adult" ...
 $ old        : int  0 0 0 0 0 0 1 0 1 1 ...
 $ female     : chr  "Female" "Male" "Female" "Male" ...
 $ sibsp      : int  0 1 1 1 1 0 1 0 2 0 ...
 $ parch      : int  0 2 2 2 2 0 0 0 0 0 ...
 $ alone      : int  1 0 0 0 0 1 0 1 0 1 ...
 $ fare       : num  211 152 152 152 152 ...
 $ cherbourg  : int  0 0 0 0 0 0 0 0 0 1 ...
 $ queenstown : int  0 0 0 0 0 0 0 0 0 0 ...
 $ southampton: int  1 1 1 1 1 1 1 1 1 0 ...
 $ port       : chr  "Southampton" "Southampton" "Southampton" "Southampton" ...
```

```
# View the first few rows of the dataset
head(td)
```

```
  X survived                                            name pclass     age
1 1        1                   Allen, Miss. Elisabeth Walton      1 29.0000
2 2        1                  Allison, Master. Hudson Trevor      1  0.9167
3 3        0                    Allison, Miss. Helen Loraine      1  2.0000
4 4        0            Allison, Mr. Hudson Joshua Creighton      1 30.0000
5 5        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels)      1 25.0000
6 6        1                             Anderson, Mr. Harry      1 48.0000
  child old female sibsp parch alone     fare cherbourg queenstown southampton
1 Adult   0 Female     0     0     1 211.3375         0          0           1
2 Child   0   Male     1     2     0 151.5500         0          0           1
3 Child   0 Female     1     2     0 151.5500         0          0           1
4 Adult   0   Male     1     2     0 151.5500         0          0           1
5 Adult   0 Female     1     2     0 151.5500         0          0           1
6 Adult   0   Male     0     0     1  26.5500         0          0           1
         port
1 Southampton
2 Southampton
3 Southampton
4 Southampton
5 Southampton
6 Southampton
```

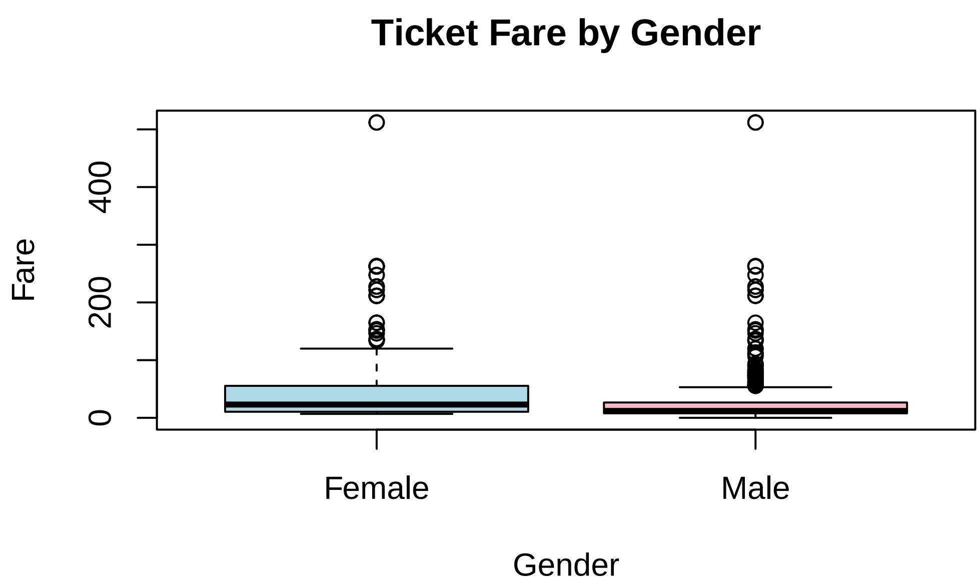
```r
# Check for missing values in the relevant columns
sum(is.na(td$female))  # Missing values in gender
```

```
[1] 0
```

```r
sum(is.na(td$fare))    # Missing values in fare
```

```
[1] 1
```

```r
# Clean dataset by removing rows with NA values in the relevant columns
td_clean <- td[complete.cases(td[, c("survived", "fare", "female", "child")]), ]

# Boxplot to explore the relationship between gender and fare
boxplot(fare ~ female, data = td,
        main = "Ticket Fare by Gender",
        xlab = "Gender",
        ylab = "Fare",
        col = c("lightblue", "lightpink"))
```




```r
# View summary statistics of fare by gender
tapply(td$fare, td$female, summary)
```

```
$Female
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   6.75   10.50   23.00   46.20   55.33  512.00 

$Male
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  0.000   7.876  11.887  26.154  26.550 512.000       1 
```

2.

```r
# Perform a two-sample t-test to compare the fares between female and male passengers
t_test_result <- t.test(fare ~ female, data = td)
t_test_result
```

```
    Welch Two Sample t-test

data:  fare by female
t = 6.1168, df = 701.7, p-value = 1.585e-09
alternative hypothesis: true difference in means between group Female and group Male is not e
95 percent confidence interval:
 13.60960 26.47613
sample estimates:
mean in group Female   mean in group Male 
            46.19668             26.15382 
```

```r
# Interpretation of results
# The p-value of 1.585e-09 is extremely small,
# the difference in means between female and male passengers
# is statistically significant so the null hypothesis can be rejected.
# The confidence interval for the difference in means
# does not contain 0, which confirms a significant difference
# between the fares for females and males.
```

#3a.

```r
# Estimate the first logit model (original fare)
model1 <- glm(survived ~ fare + female + child, data = td_clean, family = "binomial")
model1
```

```
Call:  glm(formula = survived ~ fare + female + child, family = "binomial",
    data = td_clean)

Coefficients:
(Intercept)         fare   femaleMale   childChild
   0.640380     0.009279    -2.362196     0.675223

Degrees of Freedom: 1044 Total (i.e. Null);  1041 Residual
Null Deviance:      1414
Residual Deviance: 1058     AIC: 1066
```

```r
# Clean dataset to remove rows with fare <= 0
td_clean <- td_clean[td_clean$fare > 0, ]

# Estimate the second logit model (log of fare)
model2 <- glm(survived ~ log(fare) + female + child, data = td_clean, family = "binomial")
model2
```

```
Call:  glm(formula = survived ~ log(fare) + female + child, family = "binomial",
    data = td_clean)

Coefficients:
(Intercept)    log(fare)   femaleMale   childChild
    -0.7479       0.5585      -2.3341       0.5597

Degrees of Freedom: 1036 Total (i.e. Null);  1033 Residual
Null Deviance:      1404
Residual Deviance: 1036     AIC: 1044
```

```r
# Display the results using stargazer
stargazer(model1, model2, type = "text", title = "Logit Models of Titanic Passenger Survival'
```

```
Logit Models of Titanic Passenger Survival
==============================================
                      Dependent variable:
                  ----------------------------
                            survived
                       (1)            (2)
```

```
----------------------------------------------
fare                   0.009***
                       (0.002)

log(fare)                            0.559***
                                      (0.082)

femaleMale            -2.362***      -2.334***
                       (0.156)        (0.157)

childChild             0.675***       0.560**
                       (0.235)        (0.234)

Constant               0.640***      -0.748***
                       (0.137)        (0.280)

----------------------------------------------
Observations            1,045          1,037
Log Likelihood         -528.894       -518.122
Akaike Inf. Crit.     1,065.788      1,044.245
==============================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

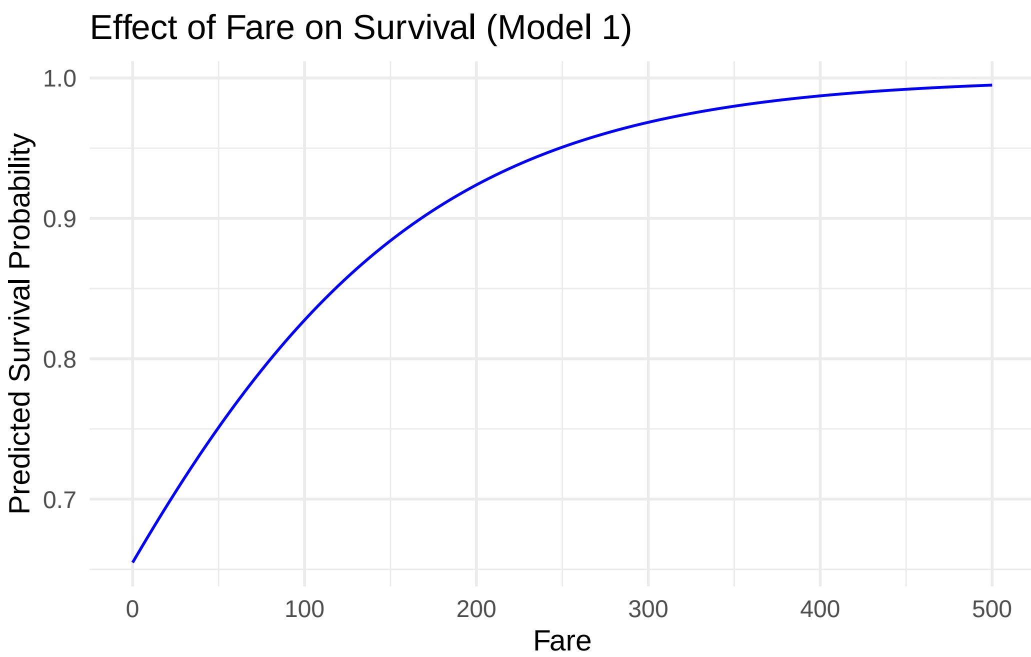
3b.

```
# Based on the model fit statistics,
# Model 2 performs better. This is because it has a
# lower Akaike Information Criterion (AIC) of 976.974 compared
# to Model 1's AIC of 1000.716,
# indicating a better trade-off between model fit and complexity.
```

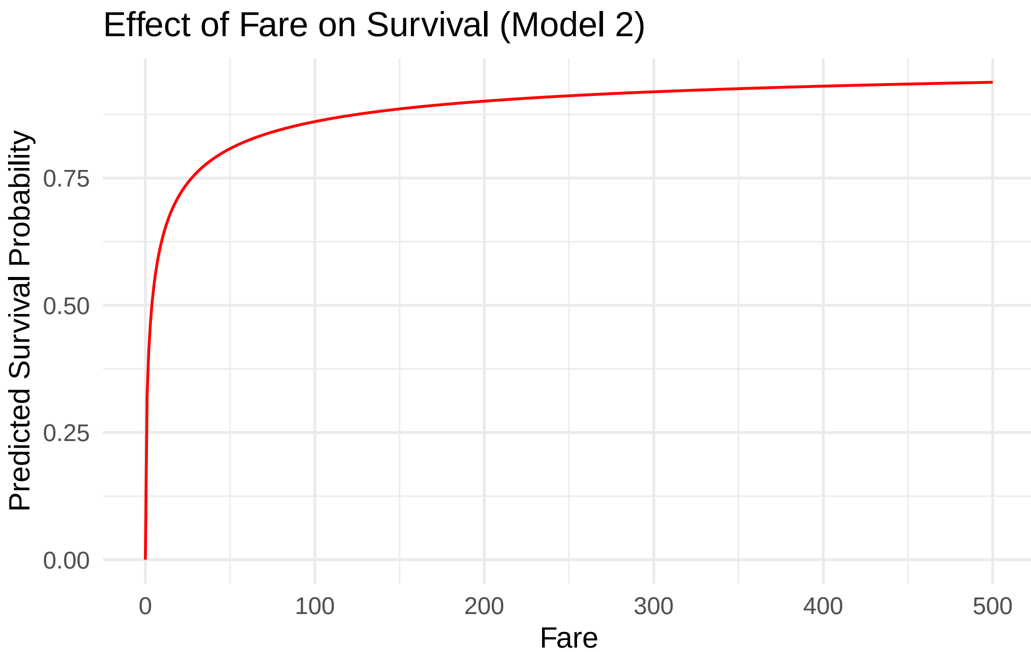
3c.

```
# Calculate the median values of the independent variables (female and child)
median_values <- data.frame(female = "Female", child = "Adult", fare = median(td$fare, na.rm

# Model 1: Using the original fare variable
```

```r
fare_values_1 <- seq(0, 500, by = 1)  # Range for fare
predicted_1 <- predict(model1, newdata = expand.grid(fare = fare_values_1, female = "Female"

# Model 2: Using the log-transformed fare variable
fare_values_2 <- seq(0, 500, by = 1)  # Range for fare
predicted_2 <- predict(model2, newdata = expand.grid(fare = fare_values_2, female = "Female"

# First plot (Model 1)
ggplot(data.frame(fare = fare_values_1, survival_prob = predicted_1), aes(x = fare, y = survi
  geom_line(color = "blue") +
  labs(x = "Fare", y = "Predicted Survival Probability", title = "Effect of Fare on Survival
  theme_minimal()
```




```r
# Plot Effect of fare on passenger survival for Model 2 (log-transformed fare)
ggplot(data.frame(fare = fare_values_2, survival_prob = predicted_2), aes(x = fare, y = survi
  geom_line(color = "red") +
  labs(x = "Fare", y = "Predicted Survival Probability", title = "Effect of Fare on Survival
  theme_minimal()
```



#3d.

```
# Set seed for reproducibility
set.seed(123)

# Remove rows with missing values for fare
td_clean_non_na <- td_clean[!is.na(td_clean$fare), ]

# Create a random 80-20 split
train_id <- sample(1:nrow(td_clean_non_na), nrow(td_clean_non_na) * 0.8)

# Split the data into training and test sets
train_data <- td_clean_non_na[train_id, ]
test_data <- td_clean_non_na[-train_id, ]

# Fit a logit model without the logged fare variable
model_no_log_fare <- glm(survived ~ fare + female + child,
                         data = train_data,
                         family = "binomial")

# Fit a logit model with the logged fare variable
model_with_log_fare <- glm(survived ~ log(fare) + female + child,
                           data = train_data,
                           family = "binomial")
```

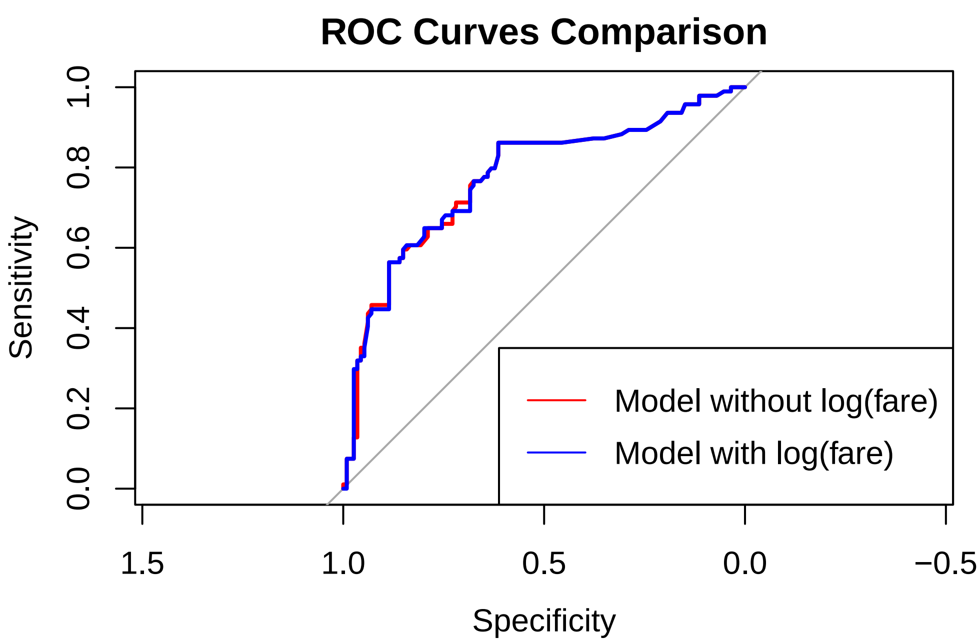
```
# Predict survival probabilities using both models
pred_no_log_fare <- predict(model_no_log_fare, newdata = test_data, type = "response")
pred_with_log_fare <- predict(model_with_log_fare, newdata = test_data, type = "response")

# Generate ROC curve for both models
roc_no_log_fare <- roc(test_data$survived, pred_no_log_fare)
```

```
Setting levels: control = 0, case = 1

Setting direction: controls < cases
```

```
roc_with_log_fare <- roc(test_data$survived, pred_with_log_fare)
```

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```

```
# Plot both ROC curves
plot(roc_no_log_fare, col = "red", main = "ROC Curves Comparison")
plot(roc_with_log_fare, col = "blue", add = TRUE)
legend("bottomright", legend = c("Model without log(fare)", "Model with log(fare)"),
       col = c("red", "blue"), lty = 1)
```



```
# Print AUC values
auc_no_log_fare <- auc(roc_no_log_fare)
auc_with_log_fare <- auc(roc_with_log_fare)

cat("AUC for model without log(fare):", auc_no_log_fare, "\n")
```

```
AUC for model without log(fare): 0.7750093
```

```
cat("AUC for model with log(fare):", auc_with_log_fare, "\n")
```

```
AUC for model with log(fare): 0.7756626
```

3e.

```
# Fit the first logit model (female and child as predictors)
model_female_child <- glm(survived ~ female + child,
                          data = td_clean_non_na,
                          family = "binomial")

# Fit the second logit model (log(fare), female, and child as predictors)
model_log_fare_female_child <- glm(survived ~ log(fare) + female + child,
                                   data = td_clean_non_na,
                                   family = "binomial")

# Generate a stargazer table comparing both models
stargazer(model_female_child, model_log_fare_female_child,
          type = "text",
          title = "Comparison of Logit Models",
          column.labels = c("Model 1: Female + Child", "Model 2: Log(Fare) + Female + Child")
          out = "logit_models_comparison.txt")
```

```
Comparison of Logit Models
==============================================================================
                                      Dependent variable:
                  ------------------------------------------------------------
                                           survived
                  Model 1: Female + Child Model 2: Log(Fare) + Female + Child
```

```
                               (1)                            (2)
-----------------------------------------------------------------------------
log(fare)                                                  0.559***
                                                           (0.082)

femaleMale                  -2.457***                     -2.334***
                             (0.153)                       (0.157)

childChild                  0.660***                       0.560**
                             (0.237)                       (0.234)

Constant                    1.031***                      -0.748***
                             (0.121)                       (0.280)

-----------------------------------------------------------------------------
Observations                  1,037                         1,037
Log Likelihood              -542.342                      -518.122
Akaike Inf. Crit.           1,090.683                     1,044.245
=============================================================================
Note:                                             *p<0.1; **p<0.05; ***p<0.01
```

3f.

```r
# Fit the first logit model (female and child as predictors)
model_female_child <- glm(survived ~ female + child,
                          data = td_clean_non_na,
                          family = "binomial")

# Fit the second logit model (log(fare), female, and child as predictors)
model_log_fare_female_child <- glm(survived ~ log(fare) + female + child,
                                   data = td_clean_non_na,
                                   family = "binomial")

# Fit the second logit model (log(fare), female, and child as predictors)
model_log_fare_female_child <- glm(survived ~ log(fare) + female + child,
                                   data = td_clean_non_na,
                                   family = "binomial")

# Generate a stargazer table comparing both models
stargazer(model_female_child, model_log_fare_female_child,
```

```r
          type = "text",
          title = "Comparison of Logit Models",
          column.labels = c("Model 1: Female + Child", "Model 2: Log(Fare) + Female + Child"),
          out = "logit_models_comparison.txt")
```

```
Comparison of Logit Models
=============================================================================
                                      Dependent variable:
                  -----------------------------------------------------------
                                           survived
                  Model 1: Female + Child Model 2: Log(Fare) + Female + Child
                            (1)                            (2)
-----------------------------------------------------------------------------
log(fare)                                               0.559***
                                                         (0.082)

femaleMale               -2.457***                      -2.334***
                          (0.153)                        (0.157)

childChild               0.660***                        0.560**
                          (0.237)                        (0.234)

Constant                 1.031***                       -0.748***
                          (0.121)                        (0.280)

-----------------------------------------------------------------------------
Observations               1,037                          1,037
Log Likelihood           -542.342                       -518.122
Akaike Inf. Crit.        1,090.683                      1,044.245
=============================================================================
Note:                                             *p<0.1; **p<0.05; ***p<0.01
```

```r
# The second model seems to perform better bases on the lower AIC.
```

3g.

```r
# Get predicted probabilities for both models
pred_female_child <- predict(model_female_child, type = "response")
```

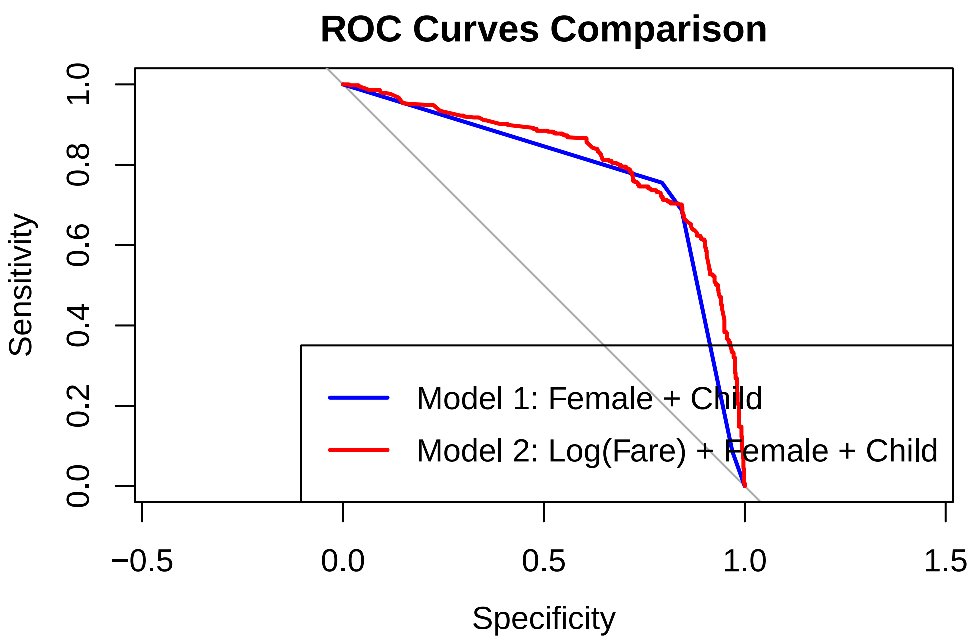
```r
pred_log_fare_female_child <- predict(model_log_fare_female_child, type = "response")

# Generate ROC curves for both models
roc_female_child <- roc(td_clean_non_na$survived, pred_female_child)
```

```
Setting levels: control = 0, case = 1

Setting direction: controls < cases
```

```r
roc_log_fare_female_child <- roc(td_clean_non_na$survived, pred_log_fare_female_child)
```

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```

```r
# Plot ROC curves for both models in one graph
plot(roc_female_child, col = "blue", main = "ROC Curves Comparison",
     lwd = 2, xlim = c(0, 1), ylim = c(0, 1))
lines(roc_log_fare_female_child, col = "red", lwd = 2)
legend("bottomright", legend = c("Model 1: Female + Child", "Model 2: Log(Fare) + Female + Cl
       col = c("blue", "red"), lwd = 2)
```



3h.

```r
# Generate ROC curves for both models (if not already done)
roc_female_child <- roc(td_clean_non_na$survived, pred_female_child)
```

```
Setting levels: control = 0, case = 1

Setting direction: controls < cases
```

```r
roc_log_fare_female_child <- roc(td_clean_non_na$survived, pred_log_fare_female_child)
```

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
```

```r
# Calculate AUC for both models
auc_female_child <- auc(roc_female_child)
auc_log_fare_female_child <- auc(roc_log_fare_female_child)

# Report AUC scores
auc_female_child
```

```
Area under the curve: 0.7824
```

```r
auc_log_fare_female_child
```

```
Area under the curve: 0.8194
```

```r
# The second model controlling for logfare has a higher
# predictive power than the first model.
```

3i.

```
# Based on the ROC curves and AUC scores, the model
# that includes the logged fare variable performs better.
# This is evident from its higher AUC score, which indicates
# that it is more effective in distinguishing between survival
# and non-survival outcomes compared to the model that only
# includes female and child variables. Therefore, the
# inclusion of the logged fare variable enhances
# the model's predictive ability.
```

4.

```
setwd("/Users/jackjeffrey/Documents/Poli502_Jeffrey/Data")
# Load dataset
putnam <- read.csv("putnam.csv")
# Preview data
head(putnam)
```

```
  Region InstPerform CivicCommunity NorthSouth EconModern
1     Ab         7.5              8      South        7.0
2     Ba         7.5              4      South        3.0
3     Cl         1.5              1      South        3.0
4     Cm         2.5              2      South        6.5
5     Em        16.0             18      North       13.0
6     Fr        12.0             17      North       14.5
```

```
# Explore the dataset
str(putnam)  # Get the structure of the data
```

```
'data.frame':	20 obs. of  5 variables:
 $ Region        : chr  "Ab" "Ba" "Cl" "Cm" ...
 $ InstPerform   : num  7.5 7.5 1.5 2.5 16 12 10 11 11 9 ...
 $ CivicCommunity: num  8 4 1 2 18 17 13 16 17 15.5 ...
 $ NorthSouth    : chr  "South" "South" "South" "South" ...
 $ EconModern    : num  7 3 3 6.5 13 14.5 12.5 15.5 19 10.5 ...
```

```
summary(putnam)  # Summary statistics of the dataset
```

```
    Region           InstPerform    CivicCommunity    NorthSouth
 Length:20          Min.   : 1.50   Min.   : 1.000   Length:20
 Class :character   1st Qu.: 6.25   1st Qu.: 3.875   Class :character
 Mode  :character   Median :10.00   Median :15.000   Mode  :character
                    Mean   : 9.15   Mean   :11.350
                    3rd Qu.:11.25   3rd Qu.:16.250
                    Max.   :16.00   Max.   :18.000
  EconModern
 Min.   : 2.50
 1st Qu.: 6.25
 Median :11.75
 Mean   :10.43
 3rd Qu.:14.50
 Max.   :19.00
```

```r
# Check column names
colnames(putnam)
```

```
[1] "Region"         "InstPerform"    "CivicCommunity" "NorthSouth"
[5] "EconModern"
```

```r
# Create a dummy variable for North
putnam$North <- ifelse(putnam$NorthSouth == "North", 1, 0)

# (a)Simple linear regression model
model_a <- lm(InstPerform ~ CivicCommunity, data = putnam)
model_a
```

```

Call:
lm(formula = InstPerform ~ CivicCommunity, data = putnam)

Coefficients:
   (Intercept)  CivicCommunity
        2.7112          0.5673
```

```r
# (b)Fit additive model using dummy variable North
model_b <- lm(InstPerform ~ CivicCommunity + North, data = putnam)
model_b
```

```
Call:
lm(formula = InstPerform ~ CivicCommunity + North, data = putnam)

Coefficients:
   (Intercept)  CivicCommunity           North
       2.69850         0.57094        -0.04781
```

```
# the regression equation would be as follows -
# InstPerform=2.69850+0.57094×CivicCommunity−0.04781×North

# Fit the interactive model
model_c <- lm(InstPerform ~ CivicCommunity * North, data = putnam)
model_c
```

```
Call:
lm(formula = InstPerform ~ CivicCommunity * North, data = putnam)

Coefficients:
         (Intercept)        CivicCommunity                 North
             2.82828               0.54040              -1.19414
CivicCommunity:North
             0.09374
```

```
# Now we need to graph these models to understand the results
```

6.

```
# Interpretation of Results

# For Model 1 it is clear that there is a linear positive relationship,
# indicating that across all regions as Civiv Community index increaes
# Institutional Performance also increases

# For Model 2 both regions held a positive linear relationship,
# however the relationship was stronger in the Northern Region.

# For Model 3 the relationship in both the North and South regions
```

```
# is far more erratic. Neither region shows a clear positive
# relationship and in some cases increased civic community index score
# is leading to reduced institutional performance. This indicates
# there may be other factors at play, such as economic or cultural differences.

# The results of these 3 models and their respective graphs suggest that
# regional distinctions do seem to be affecting the relationship between
# civic community index and institutional performance. This makes regional
# differences spurious in the sense that the positive relationship in
# model 1 was misleading. There may be more complex factors at play
# and we may find more information when we use the variable economic
# modernization instead of civic community index.
```

#7.

```r
# Model a simple linear regression (a)
model_a <- lm(InstPerform ~ EconModern, data = putnam)
summary(model_a)
```

```
Call:
lm(formula = InstPerform ~ EconModern, data = putnam)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3386 -1.7733  0.0086  0.8336  5.5114

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0108     1.3847   2.174 0.043264 *
EconModern    0.5889     0.1200   4.909 0.000113 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.659 on 18 degrees of freedom
Multiple R-squared:  0.5724,    Adjusted R-squared:  0.5487
F-statistic:  24.1 on 1 and 18 DF,  p-value: 0.0001131
```

```r
# Create an additive model
putnam$North <- ifelse(putnam$NorthSouth == "North", 1, 0)
# Additive model with North-South control (Model b)
```

```r
model_b <- lm(InstPerform ~ EconModern + North, data = putnam)
summary(model_b)
```

```

Call:
lm(formula = InstPerform ~ EconModern + North, data = putnam)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6638 -1.1011 -0.2199  1.2497  4.1464 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  5.22207    1.34660   3.878  0.00121 **
EconModern  -0.01941    0.22037  -0.088  0.93083   
North        6.88386    2.22907   3.088  0.00667 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.19 on 17 degrees of freedom
Multiple R-squared:  0.7261,    Adjusted R-squared:  0.6939 
F-statistic: 22.53 on 2 and 17 DF,  p-value: 1.659e-05
```

```r
# Create an interactive model
model_c <- lm(InstPerform ~ EconModern * North, data = putnam)
summary(model_c)
```

```

Call:
lm(formula = InstPerform ~ EconModern * North, data = putnam)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5956 -1.1176 -0.2243  1.3108  4.1278 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)  
(Intercept)       5.05147    2.09263   2.414   0.0281 *
EconModern        0.01471    0.38693   0.038   0.9702  
North             7.30611    4.50649   1.621   0.1245  
EconModern:North -0.05204    0.47786  -0.109   0.9146  
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.256 on 16 degrees of freedom
Multiple R-squared:  0.7263,    Adjusted R-squared:  0.675 
F-statistic: 14.15 on 3 and 16 DF,  p-value: 9.148e-05
```

```r
# Install broom package
library(broom)

# Create objects for each model
model_a_summary <- tidy(model_a)
model_b_summary <- tidy(model_b)
model_c_summary <- tidy(model_c)

# Create a table for regression results

regression_table <- data.frame(
  Model = c("Model a (Simple)", "Model b (Additive)", "Model c (Interactive)"),
  Term = c(model_a_summary$term, model_b_summary$term, model_c_summary$term),
  Estimate = c(model_a_summary$estimate, model_b_summary$estimate, model_c_summary$estimate)
  Std_Error = c(model_a_summary$std.error, model_b_summary$std.error, model_c_summary$std.er
  t_value = c(model_a_summary$statistic, model_b_summary$statistic, model_c_summary$statisti
  p_value = c(model_a_summary$p.value, model_b_summary$p.value, model_c_summary$p.value)
)

# View regression table
regression_table
```

```
                  Model            Term    Estimate Std_Error     t_value
1      Model a (Simple)     (Intercept)  3.01084218 1.3847400  2.17430138
2    Model b (Additive)      EconModern  0.58888804 0.1199651  4.90882696
3 Model c (Interactive)     (Intercept)  5.22206689 1.3465960  3.87797608
4      Model a (Simple)      EconModern -0.01941338 0.2203654 -0.08809629
5    Model b (Additive)           North  6.88386263 2.2290736  3.08821680
6 Model c (Interactive)     (Intercept)  5.05147059 2.0926339  2.41392939
7      Model a (Simple)      EconModern  0.01470588 0.3869282  0.03800675
8    Model b (Additive)           North  7.30610907 4.5064939  1.62124020
9 Model c (Interactive) EconModern:North -0.05204093 0.4778626 -0.10890354
       p_value
1 0.0432639802
2 0.0001131385
```

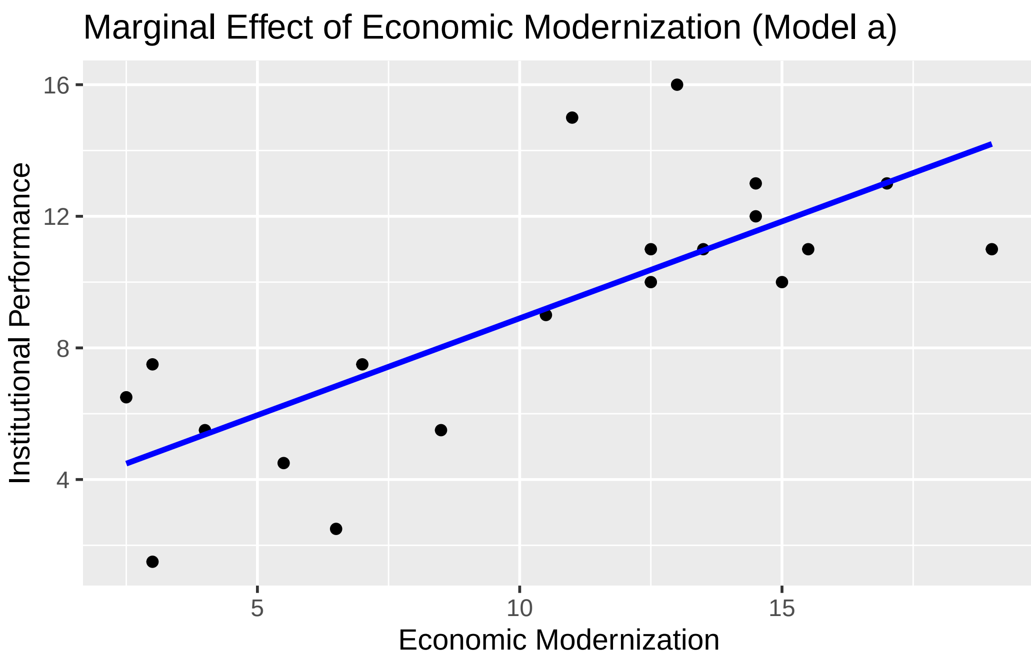
```
3 0.0012080639
4 0.9308296002
5 0.0066708160
6 0.0281279806
7 0.9701524776
8 0.1245037671
9 0.9146331065
```

```
# Interpretation of Results

# Model A: The p-value of 0.9308 indicates that the models results are not
# statistically signicant. The t value of -0.088096 confirms the effect is weak.

# Model B: The p-value of 0.0001 indicates that the effect
# of Economic Modernization on Institutional Performance
# is statistically significant. However, the small effect size
# of 0.5889 suggests that the relationship is weak.
# The coefficient for North is much larger (6.8839), indicating
# that the regional distinction between North and South plays
# a stronger role in explaining Institutional Performance.

# Model C: The p-value for the interaction term
# EconModern:North is 0.9146, which indicates that the
# interaction between Economic Modernization and the North
# region is not statistically significant. The intercept
# for both the North and South regions is statistically significant,
# with p-values of 0.0012 for the North and 0.0281 for the South,
# indicating that Institutional Performance is positively
# influenced by region, but the effect of Economic Modernization
# remains insignificant.
```

8.

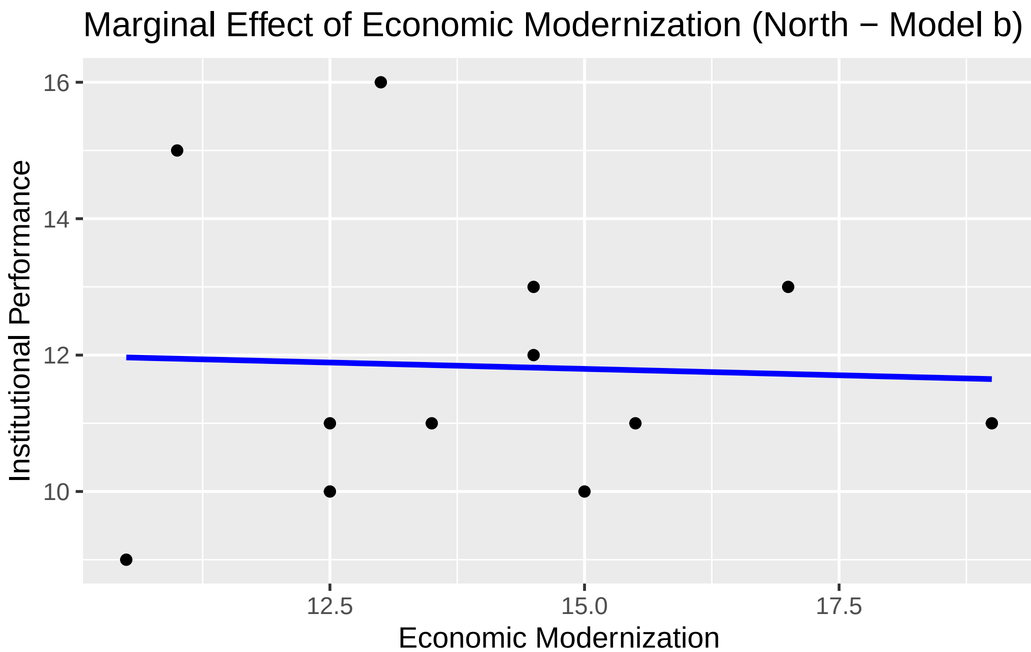
```
# Graph for Model A
ggplot(putnam, aes(x = EconModern, y = InstPerform)) +
  geom_point() +  # Plot the data points
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Regression line
  labs(title = "Marginal Effect of Economic Modernization (Model a)",
       x = "Economic Modernization",
       y = "Institutional Performance")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



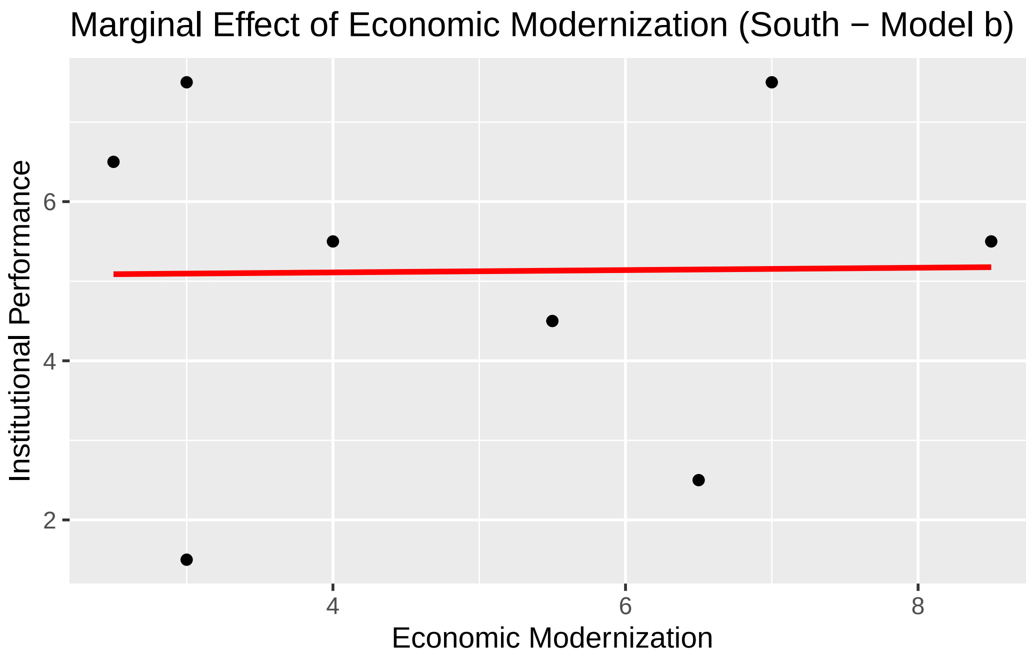

```r
# There is a positive linear relationship between economic modernization
# and institutional performance

# Graph for Model B
# Plot for North region
ggplot(subset(putnam, North == 1), aes(x = EconModern, y = InstPerform)) +
  geom_point() +  # Plot the data points
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Regression line for North
  labs(title = "Marginal Effect of Economic Modernization (North - Model b)",
       x = "Economic Modernization",
       y = "Institutional Performance")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



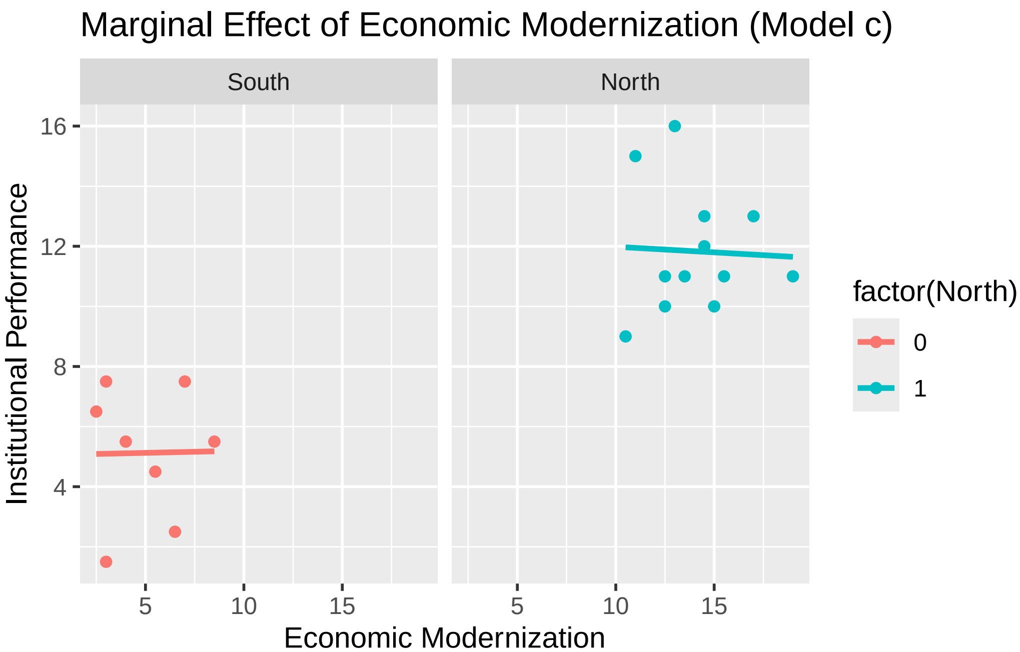

```
# Plot for South region
ggplot(subset(putnam, North == 0), aes(x = EconModern, y = InstPerform)) +
  geom_point() +  # Plot the data points
  geom_smooth(method = "lm", se = FALSE, color = "red") +  # Regression line for South
  labs(title = "Marginal Effect of Economic Modernization (South - Model b)",
       x = "Economic Modernization",
       y = "Institutional Performance")
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
# Based on the graphs for these two models, there does not seem to be
# any real relationship between economic modernization and institutional
# performance in either the north or south regions.

# Graph for Model C
ggplot(putnam, aes(x = EconModern, y = InstPerform, color = factor(North))) +
  geom_point() +  # Plot the data points
  geom_smooth(method = "lm", se = FALSE) +  # Regression line
  facet_wrap(~ North, labeller = labeller(North = c("0" = "South", "1" = "North"))) +  # Two
  labs(
    title = "Marginal Effect of Economic Modernization (Model c)",
    x = "Economic Modernization",
    y = "Institutional Performance"
  )
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
# The results of the side by side graph reinforce the findings from
# model B that there is not a strong relationship
# between our two variables. The North has a much larger intercept
# and even though the relationship is negative, the slope is very slight.
# The South has a smaller y intercept and is positive, but once
# again the slope is very slight making it difficult to draw conlcusions
# from either graph.
```

9.

```
# Interpretation of results from graphs and regression figures


# Model b shows that both economic modernization
# and the North-South distinction have statistically significant
# effects on Institutional Performance,
# the interaction term in Model c does not suggest
# that the relationship between economic modernization
# and institutional performance is drastically altered by region.
# The weak and statistically insignificant results
# in both the regression and graphs suggest that the effect
```

```
# of economic modernization on institutional performance
# is minimal and does not substantially change
# when considering regional distinctions.

# While regional distinctions may influence
# institutional performance independently,
# they do not appear to make the relationship between
# economic Modernization and institutional Performance spurious.

# Essentially the weak relationship between economic modernization
# and institutional performance exists even when
# accounting for the regional distinction
```

#10.

```
# Yi: dependent variable
# Xi: indpendent variable
# a: intercept
# B: Slope
# Ui: Error Term

# Step 1: Write the formula for the slope\
# ( )
#   = Σ[(Xi - mean(X))(Yi - mean(Y))] /
# Σ[(Xi - mean(X))^2]
# Numerator: Covariance between X and Y
# Denominator: Variance of X

# Step 2: Substitute covariance and variance
# into the slope formula
#   = [Σ(Xi * Yi) / n - (mean(X) * mean(Y))] /
# [Σ(Xi^2) / n - mean(X)^2]

# Step 3: Write the formula for the intercept
# ( )
#   = mean(Y) -   * mean(X)

# Step 4: Express the regression equation with
# the derived coefficients
# Yi = (mean(Y) -   * mean(X)) +   * Xi + µi
# Simplified to: Yi =   +  Xi + µi
```