Poli 502 Final Conor Craig

Conor Craig

2023-12-11

Data

```
library(readr)
td <- read.csv ("titanic2.csv")
putnam_data <- read.csv("putnam.csv")
```

Libraries

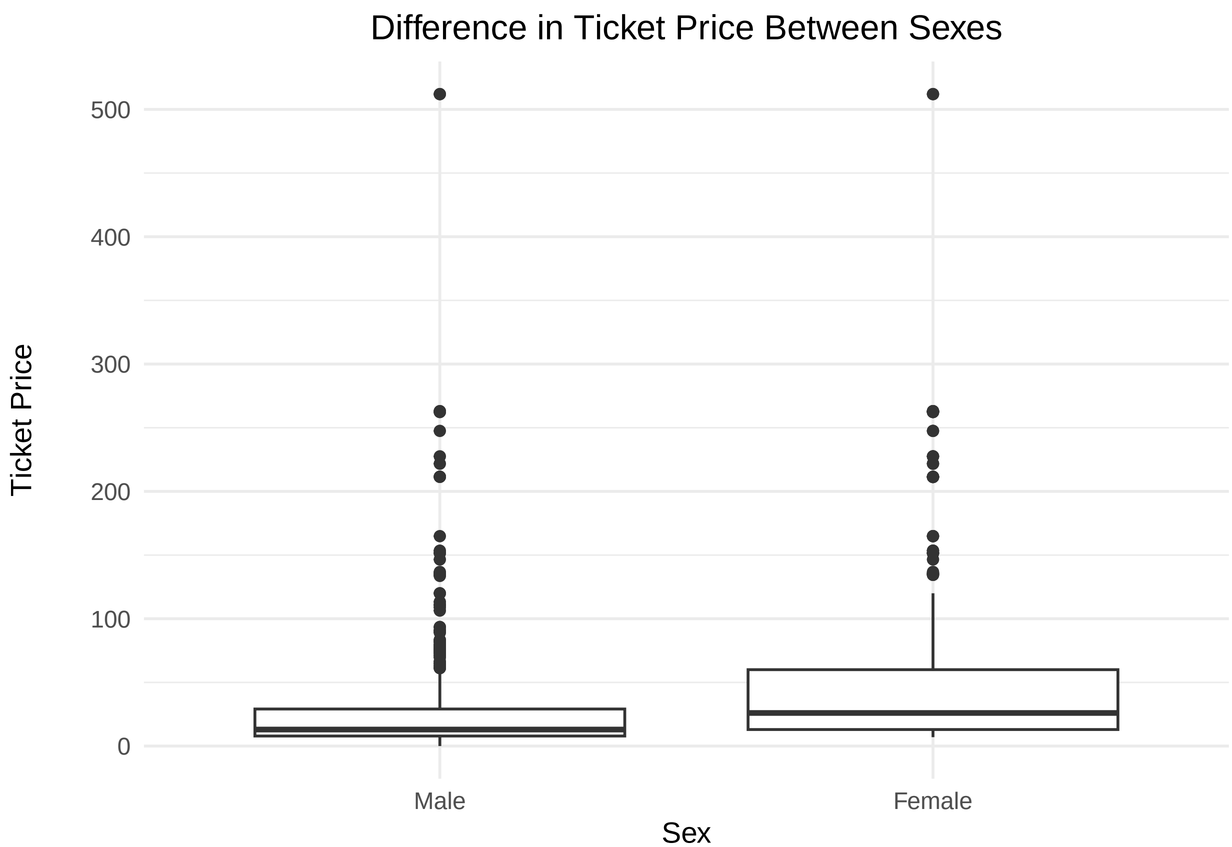
```
library(stargazer)
library(effects)
library(ggplot2)
library(pROC)
library(ROCR)
```

Question 1.

Do females have higher ticket prices than males?

```
# Removing NAs from data
td_no_na <- na.omit(td)
# Creating sex as a numeric binary
td_no_na $ sex <- ifelse (td_no_na$female == "Female", 1, 0)
td_no_na $ sex <- as.factor (td_no_na $ sex)

# Creating plot to see the effect of sex on ticket price
ggplot(td_no_na, aes(x = sex, y = fare)) +
  geom_boxplot() +
  labs(title = "Difference in Ticket Price Between Sexes",
       x = "Sex", y = "Ticket Price") +
  theme_minimal() +
```

```
  theme(plot.title = element_text(hjust = 0.5),
        axis.title.y = element_text(margin = margin(r = 20))) +
  scale_x_discrete(labels = c("Male", "Female"))
```




```
# Women appeared to have paid higher ticket prices. This can be since from the
# different means represented by the line within the box of the box plot.
```

Question 2

Bivariate Statistical Test (Difference of Means Test)

```
# Subseting the data by sex
female_fare_data <- subset(td_no_na, female == "Female")
male_fare_data <- subset (td_no_na, female == "Male")
# pulling out the ticket prices for men and women
female_fare <- female_fare_data$fare
male_fare <- male_fare_data$fare
# Creating and presenting the difference of means test
```

```r
dif_of_mean <- t.test(female_fare, male_fare)
dif_of_mean
```

```
##
##  Welch Two Sample t-test
##
## data:  female_fare and male_fare
## t = 5.5989, df = 550.35, p-value = 3.408e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  14.67275 30.53195
## sample estimates:
## mean of x mean of y
##  51.42835  28.82600
```

```r
# Since one variable (fare) is continuous while the other variable is a binary categor
```

Question 3 A.

Creating a logit model to determine survival

```r
# Regression no log
logit_1 <- glm(survived ~ fare + female + child, data = td_no_na,
               family = binomial)
# Adding the log of fare to the data set
td_no_na $ log_fare <- log(td_no_na$fare)
# removing infinite values to be able to run a regression
td_no_na_log <- td_no_na[!is.infinite(td_no_na$log_fare), ]
# Regression with log_fare
logit_2 <- glm(survived ~ log_fare + female + child, data = td_no_na_log, family = binom
# Stargazer to view results
stargazer(logit_1, logit_2, type = "text")
```

```
##
## ==============================================
##                       Dependent variable:
##                   ----------------------------
##                             survived
##                        (1)            (2)
## ----------------------------------------------
## fare                0.010***
```

```
##                        (0.002)
##
## log_fare                                0.598***
##                                          (0.085)
##
## femaleMale            -2.407***      -2.376***
##                        (0.161)        (0.163)
##
## childChild             0.596**        0.468*
##                        (0.246)        (0.245)
##
## Constant               0.607***      -0.891***
##                        (0.141)        (0.290)
##
## -------------------------------------------------
## Observations             999            991
## Log Likelihood         -496.358      -484.487
## Akaike Inf. Crit.     1,000.716       976.974
## =================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Question 3 B.

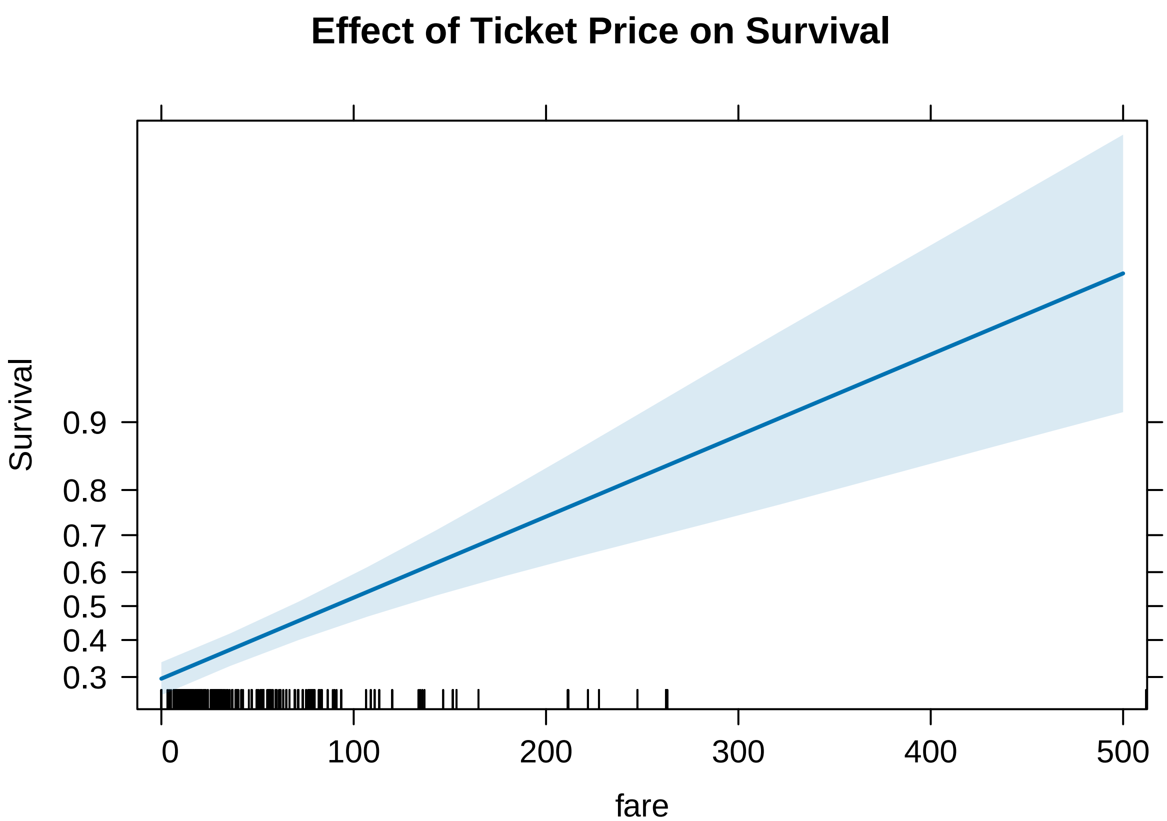
Which model is a better fit?

```
# Model 2 (the logged model) appears to be a better fit. It has a higher log
# likelihood and a lower AIC
```

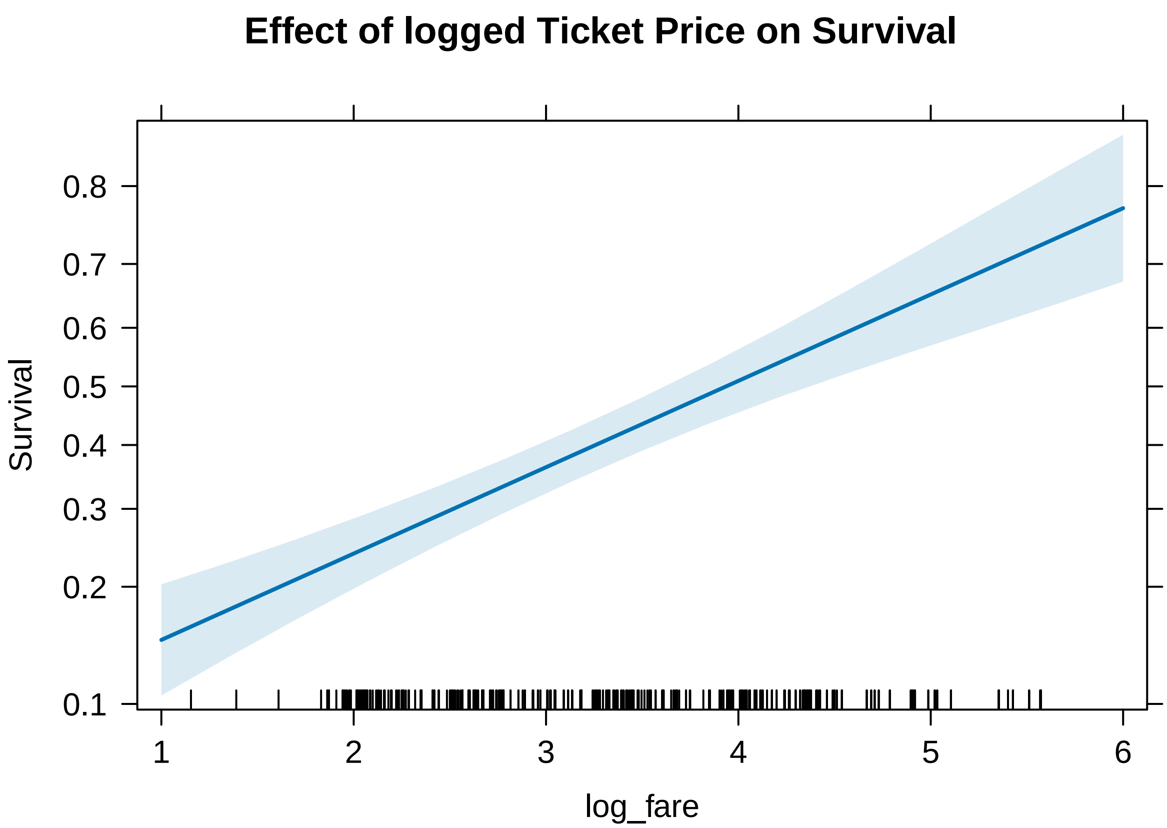
Question 3 C.

Two graphes for the effect of price fare on survival

```
effect_logit_1 <- effect(term = "fare", mod = logit_1)
effect_logit_2 <- effect(term = "log_fare", mod = logit_2)
plot(effect_logit_1,
     main = "Effect of Ticket Price on Survival",
     Xlab = "Ticket Price",
     ylab = "Survival")
```




```
plot(effect_logit_2,
     main = "Effect of logged Ticket Price on Survival",
     Xlab = "Ticket Price",
     ylab = "Survival")
```



Question 3 D.

```
# The first model (the non-logged) is less linear than the second model. While
# both models have point estimate lines that are linear, it's clear that the
# first model has a significantly less linear confidence interval. This means
# that, as the fare price increases, we can't determine if survival increases
# at a slower or faster pace than when fare prices are low. Conversely, the
# second model displays a much more linear effect that survival clearly
# increased as the fare price increased. As stated above, the second model
# outperforms the first. This is seen in the statistics addressed in question
# 3 B
```

Question 3 E.

```
set.seed(123)
training_data_sample <- sample (1: nrow(td_no_na_log), nrow(td_no_na_log) * 0.8)
```

```r
training_data <- td_no_na_log [training_data_sample, ]
testing_data <- td_no_na_log [-training_data_sample, ]

logit_predict_no_log <- glm(survived ~  female + child, data = training_data, family = b

logit_predict_log <- glm(survived ~  female + child + log_fare
                , data = training_data, family=binomial)
```

Question 3 F.

```r
stargazer (logit_predict_no_log, logit_predict_log, type = "text")
```

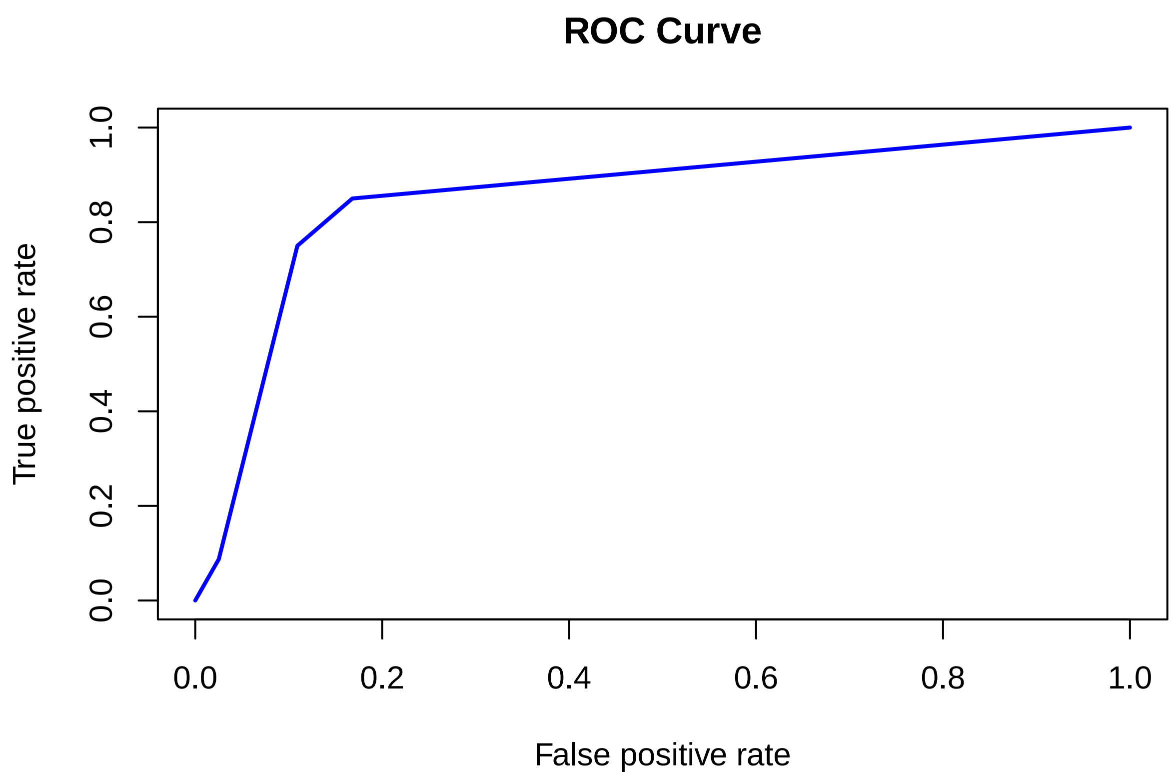
```
##
## ================================================
##                        Dependent variable:
##                     ----------------------------
##                               survived
##                          (1)            (2)
## ------------------------------------------------
## femaleMale            -2.348***      -2.213***
##                        (0.173)        (0.179)
##
## childChild              0.410          0.278
##                        (0.278)        (0.274)
##
## log_fare                             0.645***
##                                       (0.093)
##
## Constant              0.940***       -1.139***
##                        (0.136)        (0.319)
##
## ------------------------------------------------
## Observations             792            792
## Log Likelihood        -420.994       -395.071
## Akaike Inf. Crit.     847.987        798.141
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```r
# the logged model is better with a higher log likelihood and lower AIC
```

Question 3 G.

Code for no log

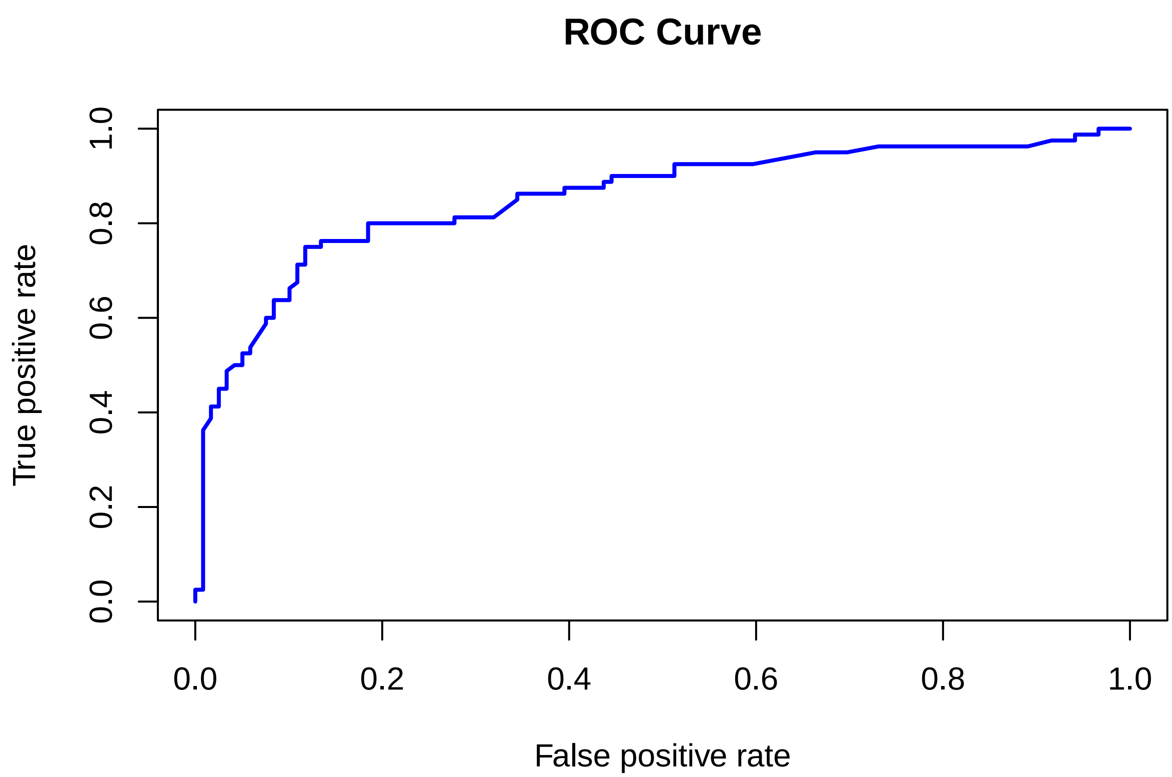
```
# Generate predictions on the testing data
predicted_probs_no_log <- predict(logit_predict_no_log, testing_data, type = "response")

# Create a prediction object
prediction_obj <- prediction(predicted_probs_no_log, testing_data$survived)

# Create a performance object for ROC curve analysis
perf_obj <- performance(prediction_obj, "tpr", "fpr")

# Plot the ROC curve
plot(perf_obj, main = "ROC Curve", col = "blue", lwd = 2)
```



Code for with log

```
# Generate predictions on the testing data
predicted_probs_log <- predict(logit_predict_log, testing_data, type = "response")
```

```r
# Create a prediction object
prediction_obj_log <- prediction(predicted_probs_log, testing_data$survived)

# Create a performance object for ROC curve analysis
perf_obj_log <- performance(prediction_obj_log, "tpr", "fpr")

# Plot the ROC curve
plot(perf_obj_log, main = "ROC Curve", col = "blue", lwd = 2)
```



Question 3 H.

```r
# Calculating AUC score for no log
auc_score_no_log <- performance(prediction_obj, "auc")@y.values[[1]]
auc_score_no_log
```

```
## [1] 0.8528887
```

```r
# Calculating AUC score with log
auc_score_log <- performance(prediction_obj_log, "auc")@y.values[[1]]
auc_score_log
```

```
## [1] 0.8545168
```

Question 3 I.

```r
# Judging first from the ROC graphs, we can see that the second graph is further
# to the right than the first graph. This indicates that the second graph has
# more correct predictions than the first graph. This interpretation is also
# supported by the AUC scores, with the second scoring slightly higher, again
# indicating that it has more correct predictions than the second predictive
# model
```

Moving to Putnam Data

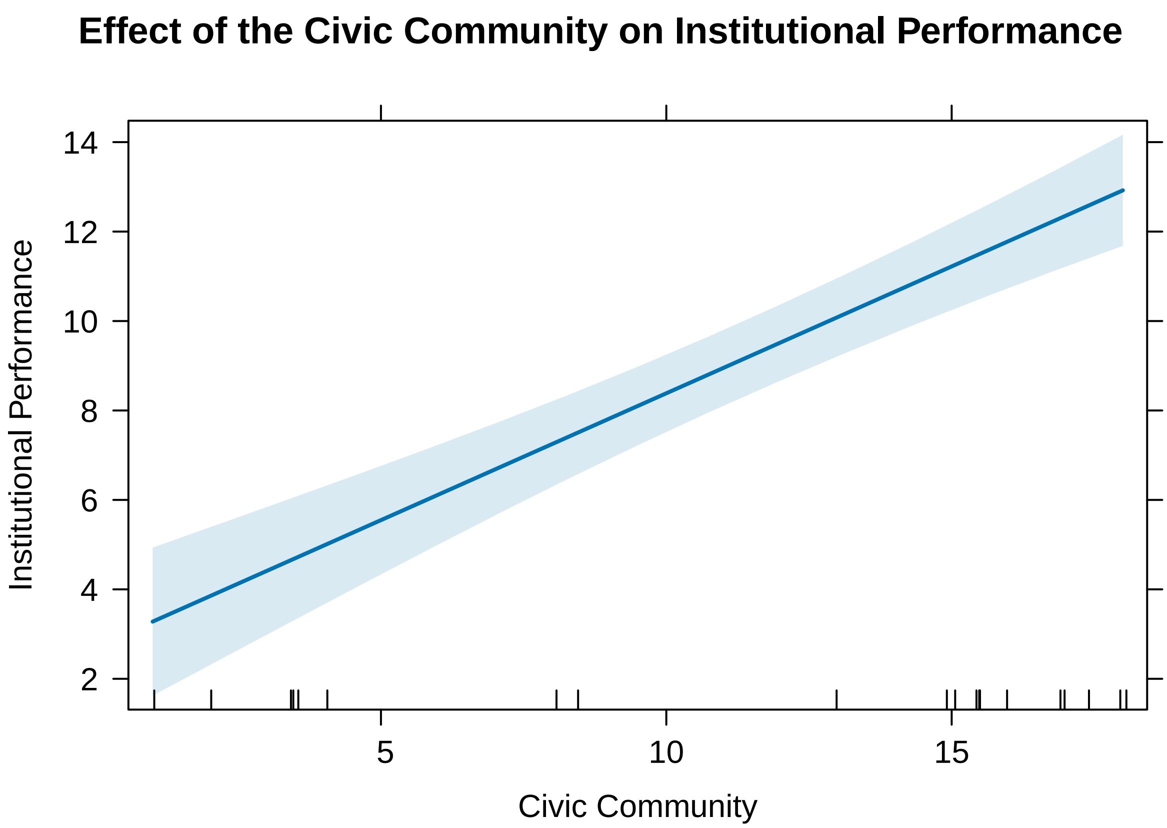
Question 4

```r
# A
put_regress_1 <- lm (InstPerform ~ CivicCommunity, data = putnam_data)
# B
putnam_data $ North <- ifelse(putnam_data$NorthSouth == "North", 1, 0)
putnam_data $ North <- as.factor (putnam_data $ North)

north_data <- subset(putnam_data, North == 1)
south_data <- subset(putnam_data, North == 0)

put_regress_2 <- lm(InstPerform ~ CivicCommunity
                    , data = north_data)

put_regress_2_a <- lm(InstPerform ~ CivicCommunity
                    , data = south_data)
# C
put_regress_3 <- lm (InstPerform ~ CivicCommunity + North +
                      CivicCommunity*North
                    , data = putnam_data)
stargazer (put_regress_1, put_regress_2, put_regress_3,
           type = "text")
```

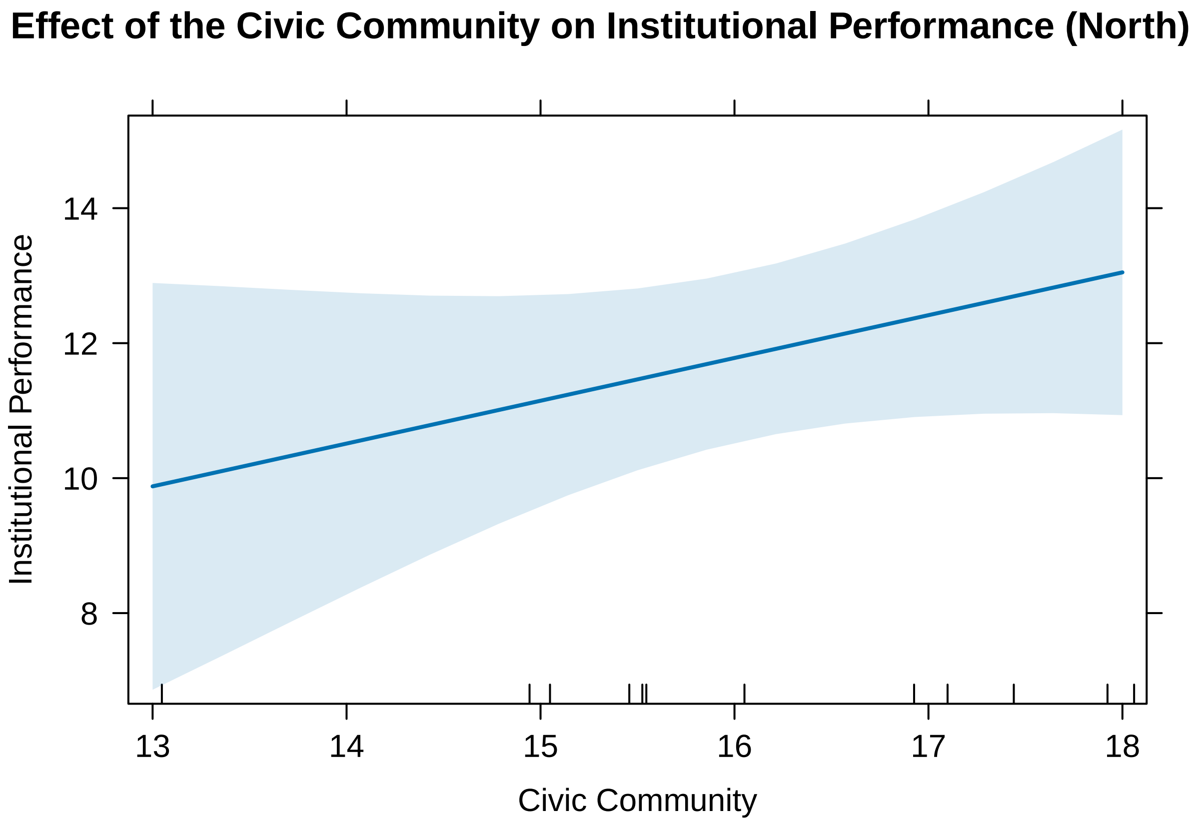
```
##
## ===========================================================================
##                                          Dependent variable:
##                        ----------------------------------------------------
##                                              InstPerform
##                                  (1)                   (2)                   (3)
## ---------------------------------------------------------------------------
## CivicCommunity                0.567***              0.634                0.540*
##                               (0.066)              (0.399)               (0.269)
##
## North1                                                                    -1.194
##                                                                           (6.396)
##
## CivicCommunity:North1                                                      0.094
##                                                                           (0.472)
##
## Constant                      2.711***              1.634                2.828**
##                               (0.844)              (6.440)               (1.326)
##
## ---------------------------------------------------------------------------
## Observations                     20                   12                    20
## R2                             0.806                0.202                 0.807
## Adjusted R2                    0.796                0.122                 0.771
## Residual Std. Error        1.789 (df = 18)      1.951 (df = 10)      1.895 (df = 16)
## F Statistic           74.967*** (df = 1; 18) 2.528 (df = 1; 10) 22.281*** (df = 3; 16
## ===========================================================================
## Note:                                                 *p<0.1; **p<0.05; ***p<0.0
```

Question 5

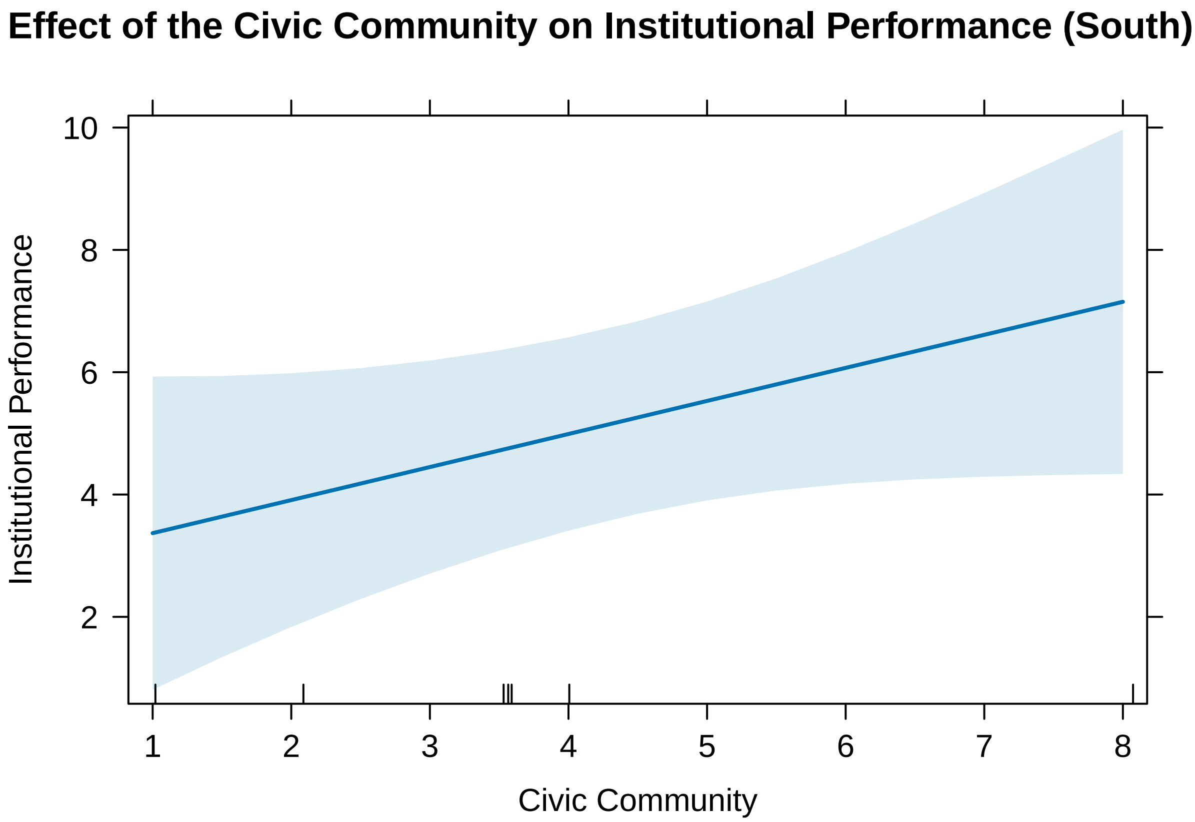
```r
# A
put_effect_1 <- effect(term = "CivicCommunity", mod = put_regress_1)
plot(put_effect_1,
     main = "Effect of the Civic Community on Institutional Performance",
     xlab = "Civic Community",
     ylab = "Institutional Performance")
```



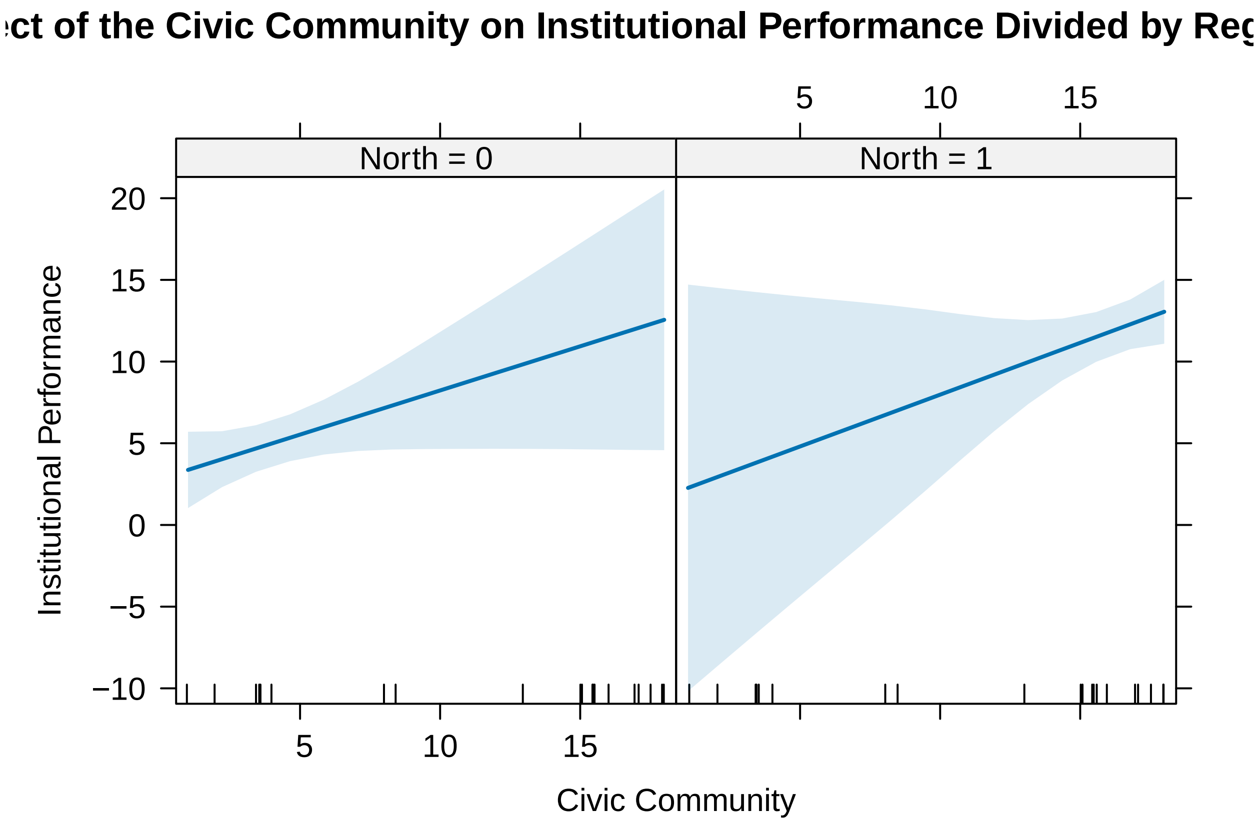

```r
# B
put_effect_2 <- effect(term = "CivicCommunity", mod = put_regress_2)
put_effect_3 <- effect(term = "CivicCommunity", mod = put_regress_2_a)
plot(put_effect_2,
     main = "Effect of the Civic Community on Institutional Performance (North)",
     xlab = "Civic Community",
     ylab = "Institutional Performance")
```




```r
plot(put_effect_3,
     main = "Effect of the Civic Community on Institutional Performance (South)",
     xlab = "Civic Community",
     ylab = "Institutional Performance")
```




```
# C
put_effect_4 <- effect(term = "CivicCommunity:North", mod = put_regress_3)
plot(put_effect_4,
     main = "Effect of the Civic Community on Institutional Performance Divided by Regio
     xlab = "Civic Community",
     ylab = "Institutional Performance")
```



Question 6

No, I would state that the relationship between civic community and institutional performance still appears to be important. Not only is civic community still statistically significant in the third model, but from the graphs subset but region, civic community appears to still have a positive effect on institutional performance; though it should be noted that the confidence intervals are quite large which makes it harder to distinguish North from South.

Question 7

```r
# A
new_put_regress_1 <- lm(InstPerform ~ EconModern, data = putnam_data)

# B
new_put_regress_2 <- lm(InstPerform ~ EconModern
                        , data = north_data)
```

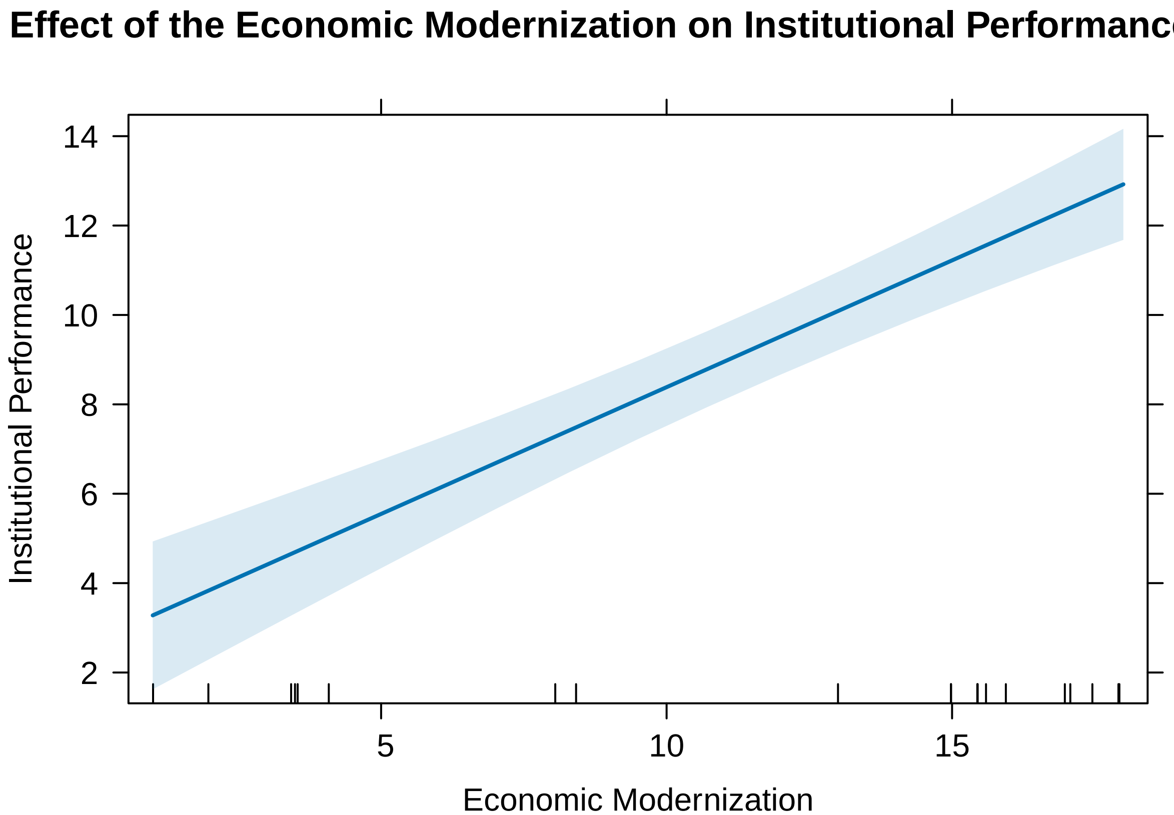
```r
new_put_regress_2 <- lm(InstPerform ~ EconModern
                        , data = south_data)
# C
new_put_regress_3 <- lm(InstPerform ~ EconModern + North +
                          EconModern*North
                        , data = putnam_data)
stargazer(new_put_regress_1, new_put_regress_2, new_put_regress_3, type = "text")
```

```
##
## =====================================================================================
##                                        Dependent variable:
##                      ----------------------------------------------------------------
##                                            InstPerform
##                              (1)                   (2)                   (3)
## -------------------------------------------------------------------------------------
## EconModern                0.589***               0.015                 0.015
##                           (0.120)               (0.407)               (0.387)
##
## North1                                                                 7.306
##                                                                       (4.506)
##
## EconModern:North1                                                     -0.052
##                                                                       (0.478)
##
## Constant                  3.011**               5.051*                5.051**
##                           (1.385)               (2.204)               (2.093)
##
## -------------------------------------------------------------------------------------
## Observations                20                     8                    20
## R2                         0.572                0.0002                 0.726
## Adjusted R2                0.549                -0.166                 0.675
## Residual Std. Error    2.659 (df = 18)      2.376 (df = 6)       2.256 (df = 16)
## F Statistic        24.097*** (df = 1; 18) 0.001 (df = 1; 6) 14.152*** (df = 3; 16)
## =====================================================================================
## Note:                                                    *p<0.1; **p<0.05; ***p<0.01
```

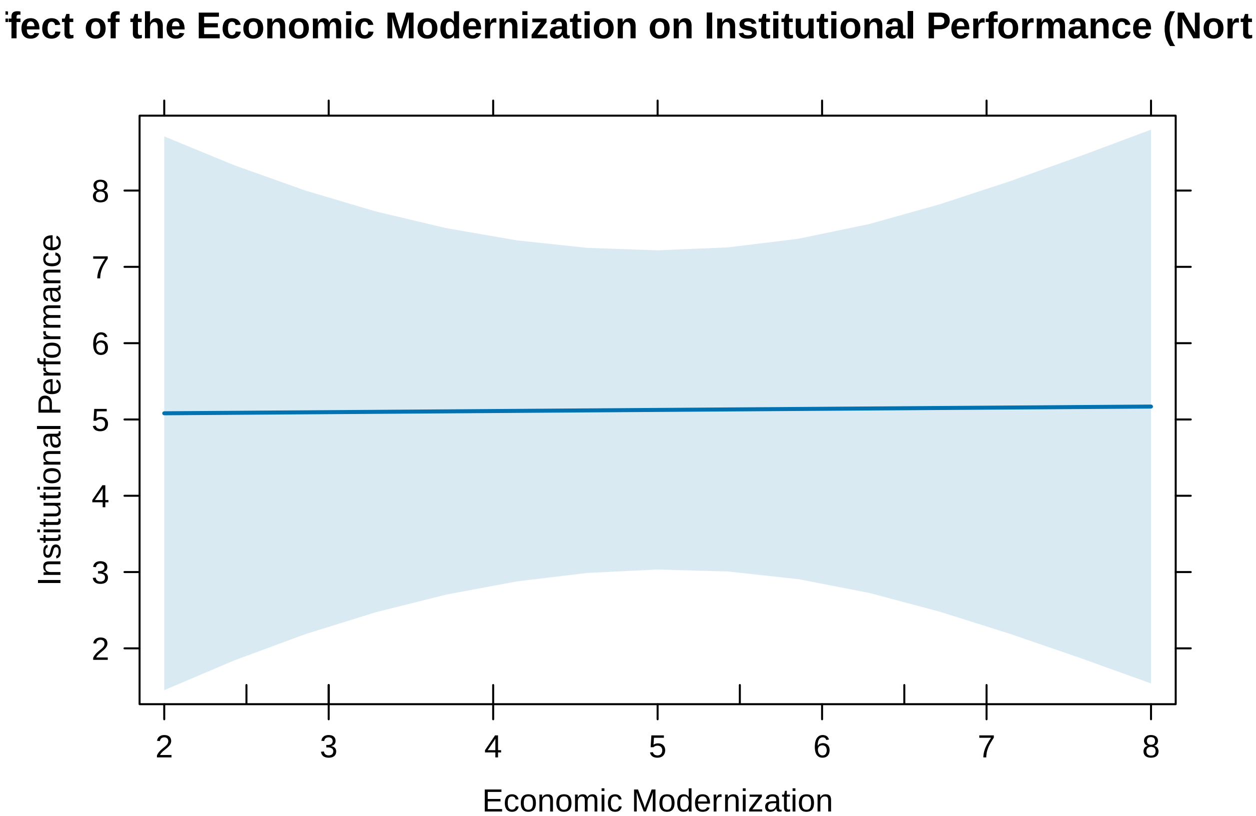
Question 8

```r
# A
new_out_effect_1 <- effect(term = "EconModern", mod = new_put_regress_1)
plot(put_effect_1,
```

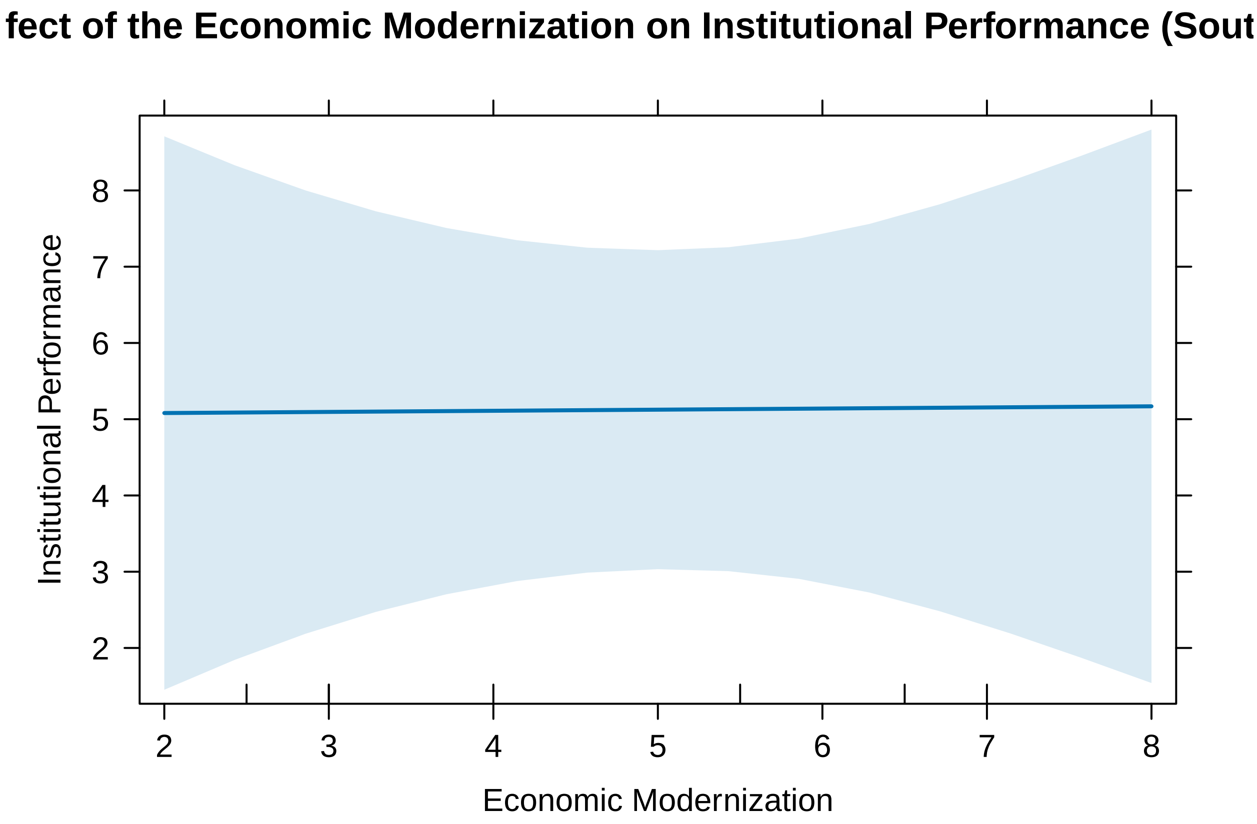
```r
     main = "Effect of the Economic Modernization on Institutional Performance",
     xlab = "Economic Modernization",
     ylab = "Institutional Performance")
```




```r
# B
new_out_effect_2 <- effect(term = "EconModern", mod = new_put_regress_2)
new_out_effect_3 <- effect(term = "EconModern", mod = new_put_regress_2)
plot(new_out_effect_2,
     main = "Effect of the Economic Modernization on Institutional Performance (North)",
     xlab = "Economic Modernization",
     ylab = "Institutional Performance")
```

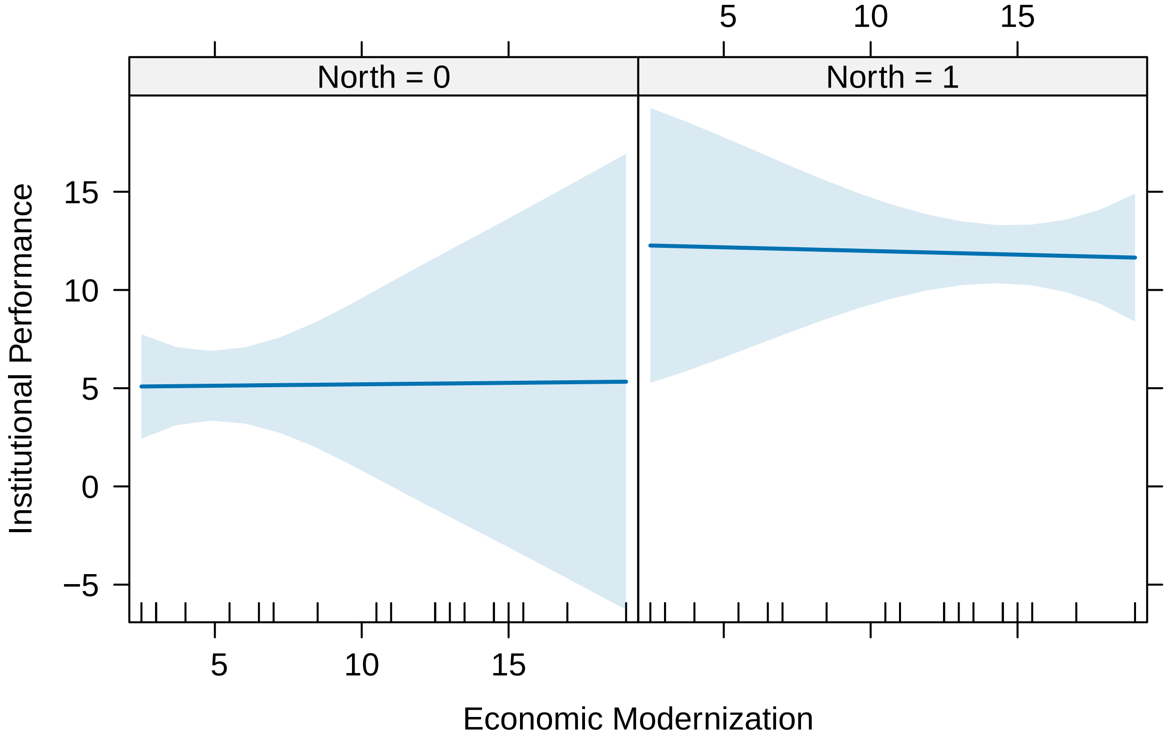



```r
plot(new_out_effect_3,
     main = "Effect of the Economic Modernization on Institutional Performance (South)",
     xlab = "Economic Modernization",
     ylab = "Institutional Performance")
```




```r
# C
new_out_effect_4 <- effect(term = "EconModern:North", mod = new_put_regress_3)
plot(new_out_effect_4,
     main = "Effect of the Economic Modernization on Institutional Performance Divided b
     xlab = "Economic Modernization",
     ylab = "Institutional Performance")
```

f the Economic Modernization on Institutional Performance Divided by



Question 9

```
# Yes, the relationship between institutional performance and economic
# modernization is spuious. This is evident in the last graph particularly.
# The effect of economic modernization on institutional performance is flat
# (null) for both the North and South. This means that the positive effect seen
# in the first model is driven by the difference of institutional performance
# across regions, not across different levels of economic modernization.
```

Question 10

To derive the Othe intercept and slope from the OLS formula, we want to start the equation Yi = alpha +(beta)Xi +ui (u being the error term). The error is the difference between the predicted Y values and the actual Y values. We want this difference to be as small as possible. That is, ui = Yi - Yi(hat). To isolate ui, we rewrite the equation as ui = Yi-(alpha +(beta)Xi). We then sum this, and square this to get positive values. (sum of) (Yi - (alpha +(beta)Xi))^2

We then take the partial derivative set that equal to zero, solving for alpha and beta individually (recall that deriving consists of multiplying the coefficient by the exponent and subtracting the exponent by 1). Then,

alpha = Y(bar) - (beta)X(bar)

The Y and X are now at their mean due both of them being summed up (that is, summing all the Y's and X's respectively), and then being divided by n (the n accompanying alpha out of the summation brackets to the other side of the equation, and getting divided to isolate alpha)

To solve for beta Again, we do a partial derivation and set that equal to zero. This yields the summation of Xi(Yi - alpha - beta*Xi) = 0 we multiple Xi across all time, while also distributing the summation symbol (recall, since alpha and beta are constants, they can get pull out of the summation)

(sum of) XiYi - alpha(sum of)Xi - beta*(sum of) Xi^2 = 0

We then substitude the equation for alpha (alpha = y(bar) - beta*X(bar))

This makes the third term in the question y(bar) - betaX(bar))(sum of)Xi

We then distribute this multiple term to get: y(bar)#(sum of) Xi + betaX(bar))(sum of)Xi

Looking at the whole equation again, we have two terms with beta (the third and fourth terms). We move these to the other side of the equation and then pull out beta (as if it were to be distributed but actually to isolate it)

beta(X(bar)(sum of)Xi - (sum of)X^2) = (sum of)XiYi - Y(bar)(sum of)Xi

From this step, we just divide everything in front of beta to the other side to isolate beta

beta = (sum of)XiYi - Y(bar)(sum of)Xi / (X(bar)*(sum of)Xi - (sum of)X^2)

This is how the slope (and thus the effect) of a variable is estimated on another variable