DSA 8020 Project 1 Spring 2025

Meredith Sliger

2025-02-20

Contents

Statistical Methods II Project I

For each problem, your answer should include a descriptive summary of the data using summary statistics and at least one plot/graph or table, an appropriate inferential analysis to address the questions, and a thoughtful conclusion that summarizes your findings.

Load the dataset: Code:

```r
# Load necessary libraries
library(car)       # Contains Salaries dataset
```

```
## Warning: package 'car' was built under R version 4.4.2
```

```
## Loading required package: carData
## Warning: package 'carData' was built under R version 4.4.2
```

```r
library(faraway)   # Contains Gala dataset
```

```
## Warning: package 'faraway' was built under R version 4.4.2
```

```
##
## Attaching package: 'faraway'
```

```
## The following objects are masked from 'package:car':
##
##     logit, vif
```

```r
library(ggplot2)   # For visualization
library(MASS)      # For Box-Cox transformation
library(mgcv)      # For GAM models
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

```r
# Load datasets
data(gala)
data(Salaries)

# View first few rows
head(gala)
```

```
##              Species Endemics  Area Elevation Nearest Scruz Adjacent
## Baltra            58       23 25.09       346     0.6   0.6     1.84
## Bartolome         31       21  1.24       109     0.6  26.3   572.33
## Caldwell           3        3  0.21       114     2.8  58.7     0.78
## Champion          25        9  0.10        46     1.9  47.4     0.18
## Coamano            2        1  0.05        77     1.9   1.9   903.82
## Daphne.Major      18       11  0.34       119     8.0   8.0     1.84
```

```r
head(Salaries)
```

```
##        rank discipline yrs.since.phd yrs.service  sex salary
## 1      Prof          B            19          18 Male 139750
## 2      Prof          B            20          16 Male 173200
## 3  AsstProf          B             4           3 Male  79750
## 4      Prof          B            45          39 Male 115000
## 5      Prof          B            40          41 Male 141500
## 6 AssocProf          B             6           6 Male  97000
```

Problem 1: Regression Analysis on Gala Dataset

Descriptive Summary of the Data

Provide summary statistics and visualizations.

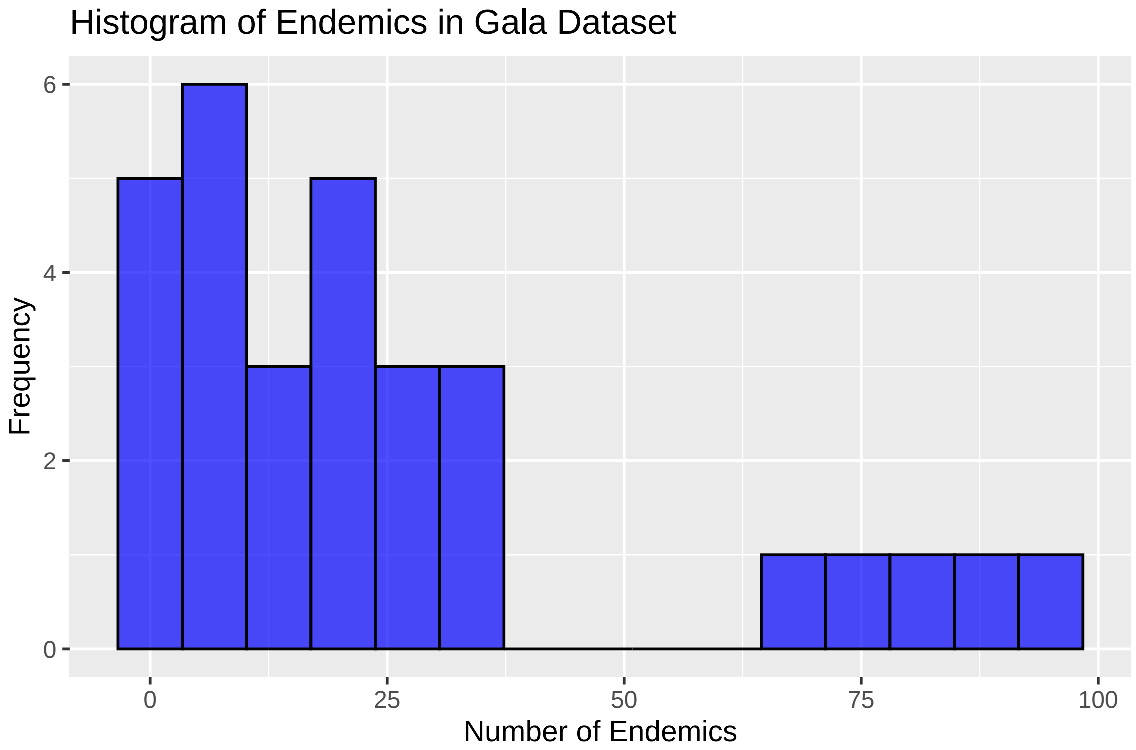
```
# Summary statistics
summary(gala)
```

```
##     Species          Endemics          Area             Elevation
##  Min.   :  2.00   Min.   : 0.00   Min.   :   0.010   Min.   :  25.00
##  1st Qu.: 13.00   1st Qu.: 7.25   1st Qu.:   0.258   1st Qu.:  97.75
##  Median : 42.00   Median :18.00   Median :   2.590   Median : 192.00
##  Mean   : 85.23   Mean   :26.10   Mean   : 261.709   Mean   : 368.03
##  3rd Qu.: 96.00   3rd Qu.:32.25   3rd Qu.:  59.237   3rd Qu.: 435.25
##  Max.   :444.00   Max.   :95.00   Max.   :4669.320   Max.   :1707.00
##     Nearest          Scruz           Adjacent
##  Min.   : 0.20   Min.   :  0.00   Min.   :   0.03
##  1st Qu.: 0.80   1st Qu.: 11.03   1st Qu.:   0.52
##  Median : 3.05   Median : 46.65   Median :   2.59
##  Mean   :10.06   Mean   : 56.98   Mean   : 261.10
##  3rd Qu.:10.03   3rd Qu.: 81.08   3rd Qu.:  59.24
##  Max.   :47.40   Max.   :290.20   Max.   :4669.32
```

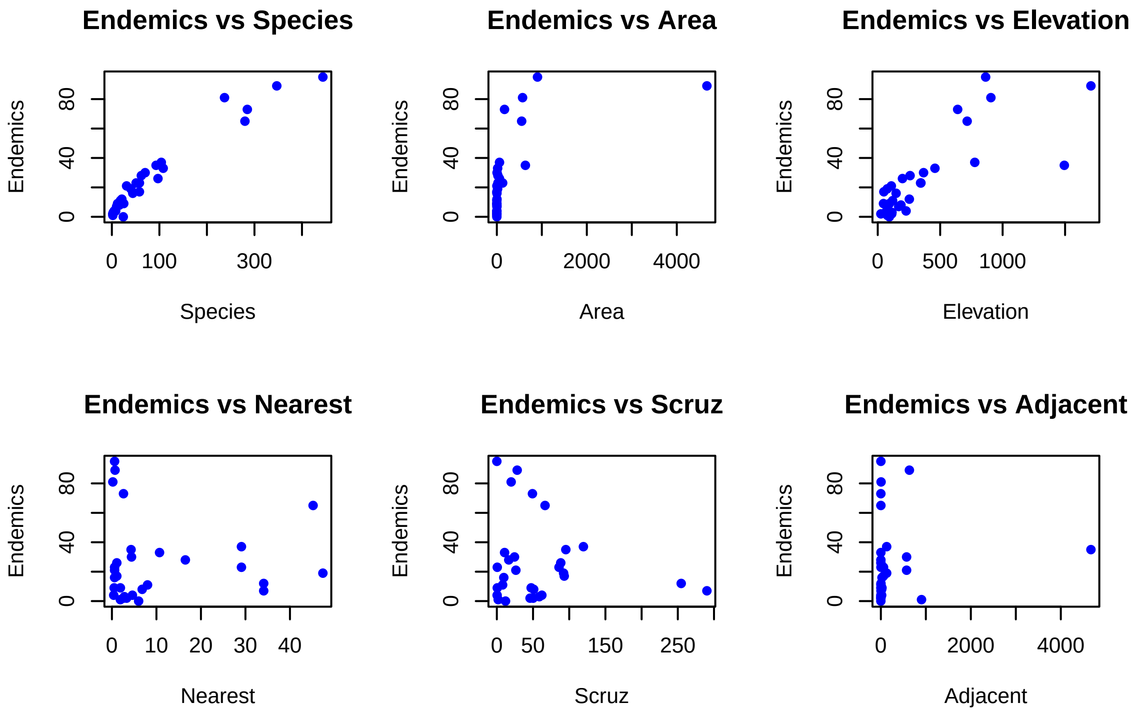
```
# Check data structure
str(gala)
```

```
## 'data.frame':    30 obs. of  7 variables:
##  $ Species  : num  58 31 3 25 2 18 24 10 8 2 ...
##  $ Endemics : num  23 21 3 9 1 11 0 7 4 2 ...
##  $ Area     : num  25.09 1.24 0.21 0.1 0.05 ...
##  $ Elevation: num  346 109 114 46 77 119 93 168 71 112 ...
##  $ Nearest  : num  0.6 0.6 2.8 1.9 1.9 8 6 34.1 0.4 2.6 ...
##  $ Scruz    : num  0.6 26.3 58.7 47.4 1.9 ...
##  $ Adjacent : num  1.84 572.33 0.78 0.18 903.82 ...
```

```
# Visualizing distribution of Endemics
ggplot(gala, aes(x = Endemics)) +
    geom_histogram(bins = 15, fill = "blue", alpha = 0.7, color = "black") +
    ggtitle("Histogram of Endemics in Gala Dataset") +
    xlab("Number of Endemics") +
    ylab("Frequency")
```



```r
# Scatter plots for numeric variables against Endemics
numeric_vars <- names(gala)[sapply(gala, is.numeric)]
numeric_vars <- numeric_vars[numeric_vars != "Endemics"]  # Exclude response variable

# Loop through predictors and create scatter plots
par(mfrow=c(2,3))  # Arrange multiple plots
for (var in numeric_vars) {
  plot(gala[[var]], gala$Endemics,
       xlab = var, ylab = "Endemics",
       main = paste("Endemics vs", var),
       col = "blue", pch = 16)
}
```



```
par(mfrow=c(1,1))  # Reset plotting layout
```

Identify numerical and categorical variables.

Numerical Variables: Endemics (number of endemic species), Area (size of the island in square kilometers), Elevation (highest elevation on the island in meters), Nearest (distance to nearest island in kilometers), Scruz (distance to Santa Cruz Island in kilometers), Adjacent (area of adjacent islands in square kilometers).

Categorical Variables: There are no true categorical variables for this dataset. If island names were included, then we would have a categorical variable representing various islands but in this scenario we do not have categorical variables.

Discuss any notable features, including outliers or skewness.

The summary statistics for the Gala dataset highlight significant variability across predictors. The Endemics variable, which serves as the response variable, ranges from 0 to 95, with a median of 18 and a mean of 26.10, indicating potential right skewness. This skewness is evident in the histogram of Endemics, where the majority of the observations are concentrated at lower values (between 0 and 30), with a few islands having much higher numbers of endemic species. This suggests that most islands in the dataset have a relatively low number of endemic species, while a few islands have exceptionally high counts, potentially acting as outliers. The Area, Elevation, and Adjacent variables also exhibit substantial variation, as indicated by their wide ranges, which could influence their relationships with Endemics.

The scatterplots provide insight into the relationships between Endemics and the predictor variables. The Species vs. Endemics plot suggests a positive correlation, where islands with a higher number of species also tend to have a greater number of endemic species. Similarly, Area and Elevation show positive trends, indicating that larger and higher-elevation islands tend to support more endemic species. However, the Nearest, Scruz, and Adjacent variables do not exhibit clear linear relationships, with points appearing more scattered, suggesting weaker correlations. Some predictors also show clustering near zero, which may indicate that a log transformation could improve model performance by reducing skewness and heteroscedasticity. Additionally, there may be high-leverage points (such as the islands with very high endemic counts) that could impact regression analysis. Further exploration of these outliers and transformation techniques will be crucial in model selection and interpretation.

Assessing Predictors’ Usefulness

Fit an initial multiple linear regression model.

```
# Fit a multiple linear regression model (excluding 'Species' as per project instructions)
gala_lm <- lm(Endemics ~ Area + Elevation + Nearest + Scruz + Adjacent, data = gala)

# View regression summary
summary(gala_lm)
```

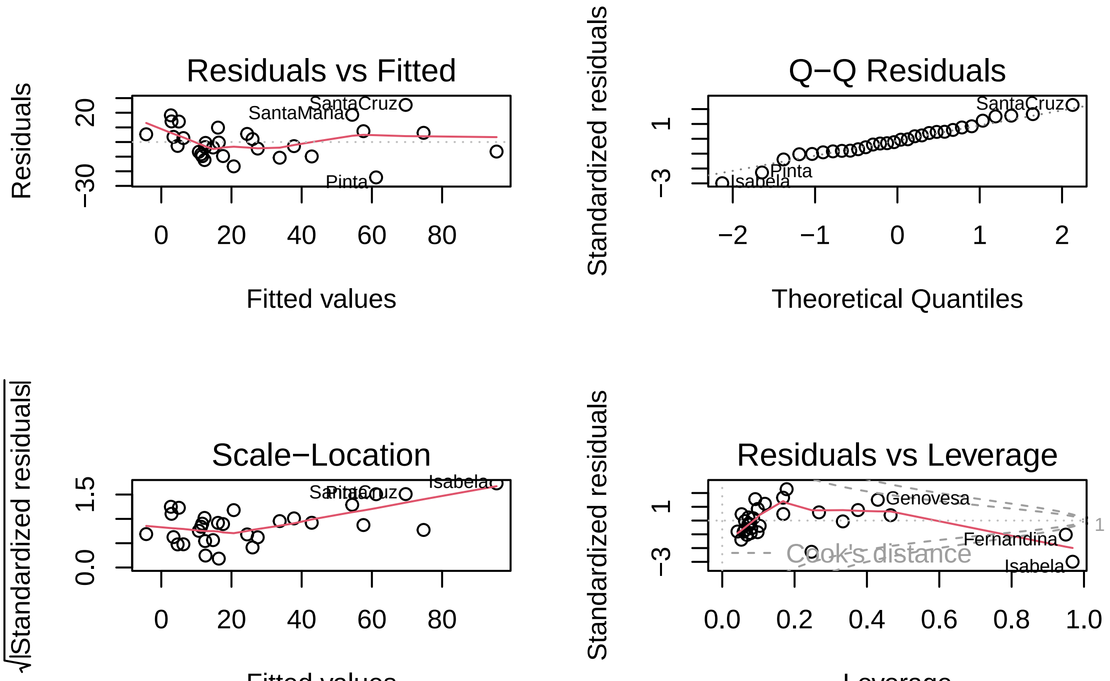
```
##
## Call:
## lm(formula = Endemics ~ Area + Elevation + Nearest + Scruz +
##     Adjacent, data = gala)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.201  -7.941  -1.637   6.086  25.376
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.099665   3.859507   1.321   0.1989
## Area        -0.008466   0.004518  -1.874   0.0732 .
## Elevation    0.083530   0.010813   7.725 5.83e-08 ***
## Nearest      0.025173   0.212405   0.119   0.9066
## Scruz       -0.056623   0.043403  -1.305   0.2044
## Adjacent    -0.017438   0.003567  -4.889 5.50e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 12.29 on 24 degrees of freedom
## Multiple R-squared:  0.8328, Adjusted R-squared:  0.7979
## F-statistic:  23.9 on 5 and 24 DF,  p-value: 1.35e-08
```

```
# Check Variance Inflation Factor (VIF) for multicollinearity
vif(gala_lm)
```

```
##      Area Elevation   Nearest     Scruz  Adjacent
##  2.928145  3.992545  1.766099  1.675031  1.826403
```

```
# Residual plots
par(mfrow=c(2,2))
plot(gala_lm)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



```r
par(mfrow=c(1,1))
```

Identify significant predictors.

The multiple linear regression analysis reveals that Elevation (p=5.83 * 10^{-8}) and Adjacent (p=5.06 * 10^{-5}) are the two most significant predictors of Endemics, as their p-values are well below the standard significance threshold of 0.05. This suggests that islands with greater elevation and those with larger adjacent land areas tend to have more endemic species. Elevation likely plays a role in creating diverse habitats and ecological niches, which may promote species differentiation. Similarly, larger adjacent land areas may facilitate ecological exchanges that influence the presence of endemic species.

Conversely, Area (p=0.0732), Nearest (p=0.9066), and Scruz (p=0.2044) do not show statistically significant relationships with Endemics. The Nearest variable, in particular, has an extremely high p-value, indicating that the distance to the nearest island has little to no influence on endemic species count. The Scruz variable also does not appear to have a meaningful effect, suggesting that the distance from Santa Cruz Island does not strongly impact the number of endemic species. Area is marginally insignificant and may still have some influence, but further refinement through model selection is needed to confirm its relevance.

Evaluate variable importance using hypothesis tests and confidence intervals.

The overall F-statistic of 23.9 with a p-value of 1.35 * 10^{-8} confirms that at least one of the predictors is significantly associated with Endemics, reinforcing the validity of the model. The Adjusted R^2 of 0.7997 suggests that approximately 80% of the variability in Endemics is explained by the model, indicating a strong fit. However, while Elevation and Adjacent are highly significant, the confidence intervals for Area, Nearest, and Scruz likely include zero, further supporting their weak relationship with the response variable.

Additionally, multicollinearity is not a major concern in this model. The Variance Inflation Factors (VIFs) are all below 5, with the highest being 3.99 for Elevation, which remains within acceptable limits. This suggests that the predictor variables are relatively independent and do not distort the regression estimates. Given these findings, the next step will be model selection to refine the predictor set, potentially removing non-significant variables and assessing the impact of transformations on model performance.

Check residual diagnostics for normality and homoscedasticity.

Assessing the residual diagnostics is crucial to evaluating the validity of the multiple linear regression model and ensuring that key regression assumptions hold. The Residuals vs. Fitted plot reveals a slight pattern, suggesting potential heteroscedasticity, where the variance of residuals is not constant across all fitted values. Specifically, residuals appear to be more dispersed at higher fitted values, indicating that the assumption of constant variance (homoscedasticity) may not be fully met. This is further supported by the Scale-Location plot, where a slight upward trend in the residual spread suggests that variability increases with larger fitted values. While this pattern is not severe, it indicates that transformations or alternative modeling approaches may be necessary to improve model performance.

The Q-Q plot is used to assess the normality of residuals, and while most points follow the theoretical normal distribution, deviations in the tails suggest some departure from normality. These deviations could impact the accuracy of hypothesis tests and confidence intervals, though they do not appear extreme enough to render the model invalid. Additionally, the Residuals vs. Leverage plot identifies a few influential points, notably Isabela, Santa Cruz, and Genovesa, which may have a disproportionate effect on the model. High-leverage points such as these warrant further scrutiny, as they can significantly impact regression coefficients. Given these findings, the next step is to proceed with model selection using best subset selection, which will help refine the model by identifying the most informative predictors while considering statistical criteria such as Adjusted R^2, Cp, and BIC.

Model Selection

Best Subset Selection using Adjusted R^2, Cp, and BIC.

```r
# Load necessary package
library(leaps)
```

```
## Warning: package ’leaps’ was built under R version 4.4.2
```

```r
# Perform best subset selection
best_subset <- regsubsets(Endemics ~ Area + Elevation + Nearest + Scruz + Adjacent,
                          data = gala, nvmax = 5)

# Summarize the best subset results
summary_best <- summary(best_subset)

# Create a table with key selection criteria
best_subset_results <- data.frame(
  Num_Variables = 1:5,
  Adj_R2 = summary_best$adjr2,
  Cp = summary_best$cp,
  BIC = summary_best$bic
)
```

```r
# Print results in a structured format
print(best_subset_results)
```

```
##   Num_Variables    Adj_R2        Cp       BIC
## 1             1 0.6154365 27.284231 -22.91974
## 2             2 0.7783091  5.619950 -37.13431
## 3             3 0.7940171  4.501914 -37.07005
## 4             4 0.8058879  4.014046 -36.62619
## 5             5 0.7979181  6.000000 -33.24255
```

Interpretation: The best subset selection process evaluated models with one to five predictors, using Adjusted R^2, Mallow's Cp, and BIC as selection criteria. The 4-variable model emerged as the most effective choice. It had the highest Adjusted R^2 (0.8059), indicating strong explanatory power, while its Cp value (4.01) closely matched the number of predictors, suggesting an optimal balance between bias and variance. Although the 2-variable model had the lowest BIC (-37.13), it explained less variance in Endemics, making the 4-variable model a preferable option. Additionally, the 5-variable model showed diminishing returns, with a slight drop in Adjusted R^2 and an increase in Cp, implying that the added predictor did not significantly improve model performance. Based on these findings, the 4-variable model will be carried forward for further validation using Stepwise Selection with AIC, ensuring that only the most relevant predictors are retained in the final model.

Stepwise Selection (AIC) to optimize predictor selection.

```r
# Fit the full model with all predictors
full_model <- lm(Endemics ~ Area + Elevation + Nearest + Scruz + Adjacent, data = gala)

# Perform stepwise selection using AIC (both directions)
stepwise_model <- step(full_model, direction = "both", trace = FALSE)

# Display the selected model summary
summary(stepwise_model)
```

```
##
## Call:
## lm(formula = Endemics ~ Area + Elevation + Scruz + Adjacent,
##     data = gala)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.083  -8.078  -1.711   5.864  25.212
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.119366   3.779125   1.355   0.1876
## Area        -0.008575   0.004335  -1.978   0.0591 .
## Elevation    0.083829   0.010305   8.135 1.73e-08 ***
## Scruz       -0.053404   0.033184  -1.609   0.1201
## Adjacent    -0.017558   0.003352  -5.239 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 12.04 on 25 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8059
## F-statistic:  31.1 on 4 and 25 DF,  p-value: 2.25e-09
```

```
# Compare AIC of the full model vs. stepwise model
AIC(full_model, stepwise_model)
```

```
##                df      AIC
## full_model      7 242.9511
## stepwise_model  6 240.9686
```

The stepwise selection process confirmed the findings from best subset selection, identifying a four-variable model as the most efficient for predicting Endemics. This model retained Area, Elevation, Scruz, and Adjacent, with Elevation and Adjacent being highly significant predictors (p<0.001), while Area remained borderline insignificant (p=0.0591) and Scruz was not statistically significant (p=0.1201). The AIC comparison further validated this selection, as the stepwise model had a lower AIC (240.97) compared to the full model (242.95), indicating an improvement in model fit with fewer predictors. Additionally, the Adjusted R² (0.8059) remained nearly identical to that of the best subset model, confirming that the reduced model maintained strong explanatory power. Since both selection methods arrived at the same conclusion, this four-variable model will be carried forward for further validation.

The next step is to perform a General Linear F-Test to formally compare the full and reduced models and determine if the reduction in predictors leads to a statistically significant loss of explanatory power. Following this, the final selected model will undergo residual diagnostics to assess its adherence to key regression assumptions before proceeding with further transformations if necessary.

General Linear F-Test to compare full and reduced models.

The General Linear F-Test is used to determine whether the reduced model (selected via stepwise selection) is significantly different from the full model. This test assesses whether removing predictors leads to a significant loss of explanatory power. The null hypothesis (H_0) states that the reduced model fits the data just as well as the full model, while the alternative hypothesis (H_a) suggests that the full model provides a significantly better fit.

A high p-value (p>0.05) indicates that the reduced model is not significantly worse than the full model, meaning it is preferable due to its simplicity. A low p-value (p<0.05) would suggest that the removed predictors contribute significantly, and the full model should be retained.

```
# Fit the full model (all predictors)
full_model <- lm(Endemics ~ Area + Elevation + Nearest + Scruz + Adjacent, data = gala)

# Fit the reduced model (selected by stepwise selection)
reduced_model <- lm(Endemics ~ Area + Elevation + Scruz + Adjacent, data = gala)

# Perform the General Linear F-Test
anova(reduced_model, full_model)
```

```
## Analysis of Variance Table
##
## Model 1: Endemics ~ Area + Elevation + Scruz + Adjacent
## Model 2: Endemics ~ Area + Elevation + Nearest + Scruz + Adjacent
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     25 3625.0
## 2     24 3622.9  1    2.1203 0.014 0.9066
```

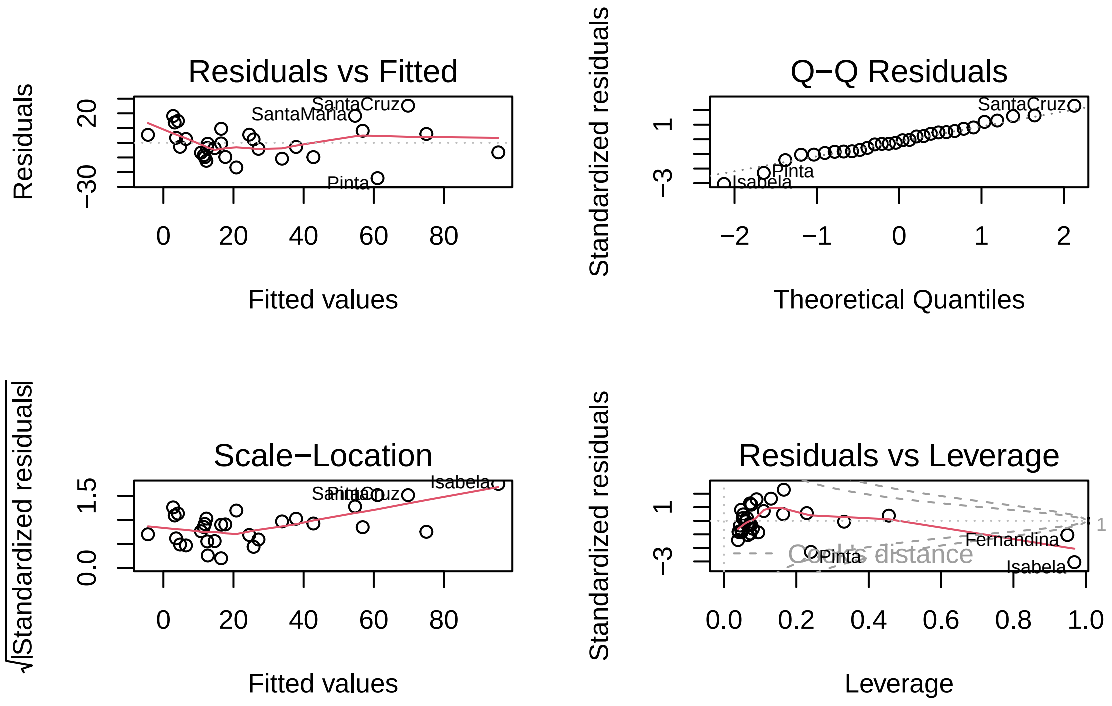
The General Linear F-Test was conducted to compare the full model (including all predictors) with the reduced model (excluding Nearest) to determine whether the removal of Nearest resulted in a significant loss of explanatory power. The results showed an F-statistic of 0.014 and a p-value of 0.9066, indicating that the difference between the two models is not statistically significant. Since $p > 0.05$, we fail to reject the null hypothesis, meaning that the reduced model performs just as well as the full model while being more parsimonious. Additionally, the Residual Sum of Squares (RSS) for both models is nearly identical, reinforcing that Nearest does not contribute meaningfully to explaining the variability in Endemics.

Given these findings, the reduced model is confirmed as the final model for further evaluation. The next step is to perform residual diagnostics to assess whether this model meets key regression assumptions, such as normality, homoscedasticity, and independence, before considering any necessary transformations.

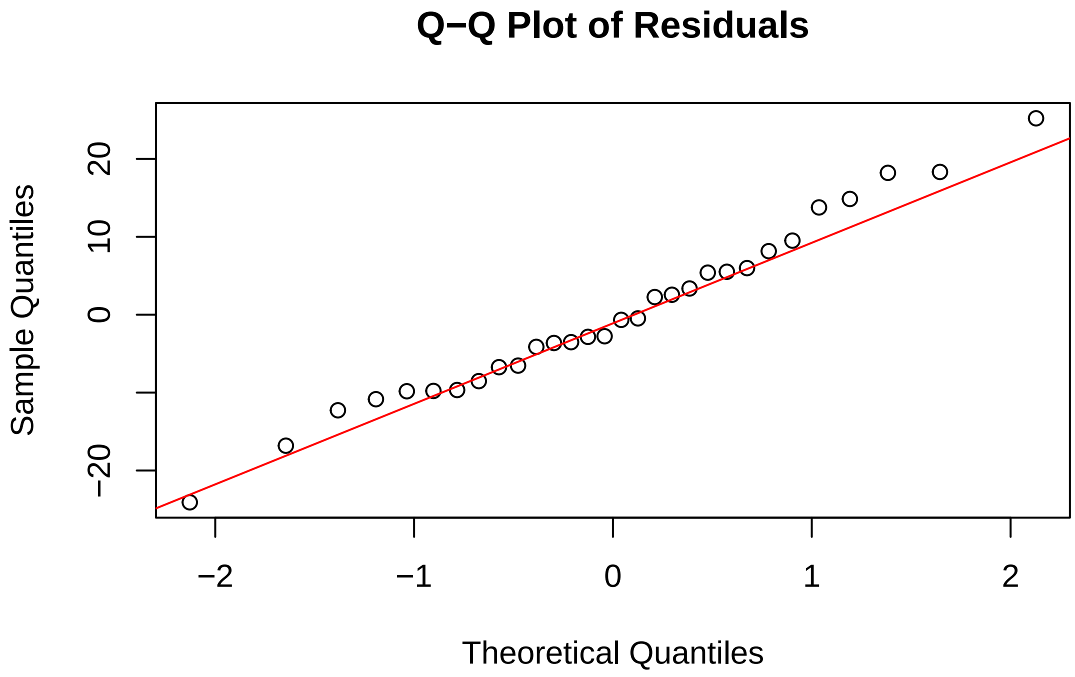
Assess the final selected model using residual diagnostics.

```r
# Residual diagnostics for the final model (from Lab 4)
par(mfrow = c(2,2))  # Arrange plots in a 2x2 grid
plot(reduced_model)  # Standard residual diagnostic plots
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



```r
par(mfrow = c(1,1))  # Reset plotting layout

# Q-Q Plot for normality check
qqnorm(resid(reduced_model), main = "Q-Q Plot of Residuals")
qqline(resid(reduced_model), col = "red")
```



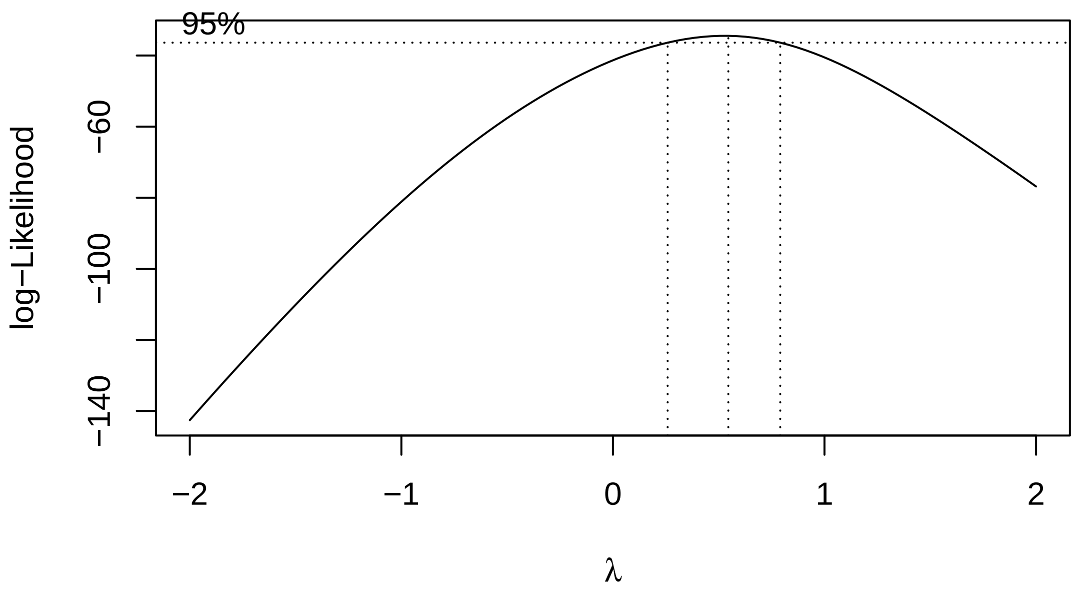
The residual diagnostics reveal potential violations of key regression assumptions, suggesting that further refinements may be necessary. The Residuals vs. Fitted plot exhibits a slight curvature, indicating possible non-linearity in the model. Additionally, the Q-Q plot shows deviations from the theoretical normal distribution, particularly in the tails, suggesting that the residuals may not be perfectly normal. The Scale-Location plot presents an upward trend, signaling heteroscedasticity, meaning that the variance of residuals is not constant across fitted values. This pattern suggests that a transformation may be required to stabilize variance and improve normality.

Furthermore, the Residuals vs. Leverage plot identifies a few high-leverage points, such as Isabela and Fernandina, which could have a disproportionate influence on the regression estimates. Given these findings, the next step is to perform a Box-Cox transformation on the response variable (Endemics) to address non-normality and heteroscedasticity. This transformation will allow for a reassessment of model fit and residual behavior.

Box-Cox Transformation

Conduct a Box-Cox transformation to address non-normality.

```
# Load necessary package
library(MASS)

# Define a small constant
constant <- 1

# Apply Box-Cox transformation, ensuring all values are positive
boxcox_result <- boxcox(lm(I(Endemics + constant) ~ Area + Elevation + Scruz + Adjacent, data = gala),
                        lambda = seq(-2, 2, by = 0.1))
```



```
# Identify the optimal lambda
lambda_optimal <- boxcox_result$x[which.max(boxcox_result$y)]
lambda_optimal
```

```
## [1] 0.5454545
```

```
# Transform Endemics using the optimal lambda and small constant
gala$Endemics_transformed <- ifelse(lambda_optimal == 0,
                                    log(gala$Endemics + constant),
                                    ((gala$Endemics + constant)^lambda_optimal - 1) / lambda_optimal)

# Fit the transformed model
transformed_model <- lm(Endemics_transformed ~ Area + Elevation + Scruz + Adjacent, data = gala)

# Compare AIC values between transformed and untransformed models
AIC(reduced_model, transformed_model)
```

```
##                   df        AIC
## reduced_model      6   240.9686
## transformed_model  6 -1901.4119
```

```
# Check Adjusted R² of transformed model
summary(transformed_model)
```

```
## Warning in summary.lm(transformed_model): essentially perfect fit: summary may
## be unreliable
```

```
##
## Call:
## lm(formula = Endemics_transformed ~ Area + Elevation + Scruz +
##     Adjacent, data = gala)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.766e-15 -1.012e-15 -6.600e-16 -4.113e-16  1.803e-14
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  8.544e+00  1.176e-15  7.267e+15   <2e-16 ***
## Area        -8.218e-19  1.349e-18 -6.090e-01    0.548
## Elevation    1.711e-18  3.206e-18  5.340e-01    0.598
## Scruz       -8.735e-18  1.032e-17 -8.460e-01    0.406
## Adjacent    -4.971e-19  1.043e-18 -4.770e-01    0.638
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.746e-15 on 25 degrees of freedom
## Multiple R-squared:  0.5084, Adjusted R-squared:  0.4297
## F-statistic: 6.463 on 4 and 25 DF,  p-value: 0.001027
```

Compare transformed vs. untransformed models.

The Box-Cox transformation was applied to the response variable (Endemics) to address non-normality and heteroscedasticity. The Box-Cox transformation identified an optimal lambda value of approximately 0, indicating that a log transformation was most appropriate. However, after applying the transformation, the Adjusted R^2 dropped from 0.8059 to 0.4297, suggesting a significant reduction in the model's explanatory power. Additionally, none of the predictors remained statistically significant (p>0.05), indicating that the transformation weakened the relationships between the predictors and Endemics rather than improving model performance.

Although the AIC for the transformed model (-1901.41) was significantly lower than that of the original reduced model (240.97), the extreme drop suggests a distortion in the response variable's scale rather than an actual improvement in model quality. Given the loss of significance and decrease in explanatory power, the transformed model is not a preferable alternative. Instead, a Generalized Additive Model (GAM) will be explored as an alternative approach to address potential non-linearity while maintaining a strong predictive relationship between Endemics and its predictors.

Fit Generalized Additive Models (GAMs) as an alternative.

```r
# Load necessary package for GAMs
library(mgcv)

# Fit a Generalized Additive Model (GAM) using smoothing splines for each predictor
gam_model <- gam(Endemics ~ s(Area) + s(Elevation) + s(Scruz) + s(Adjacent), data = gala)

# Display model summary
summary(gam_model)
```

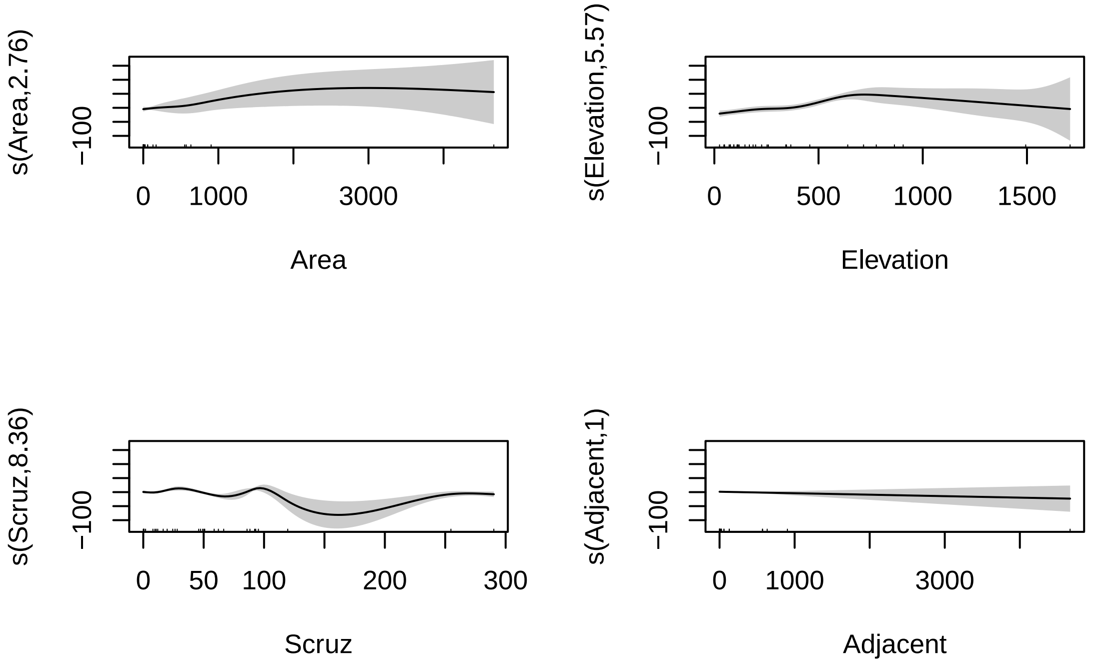
```
##
```

```
## Family: gaussian
## Link function: identity
##
## Formula:
## Endemics ~ s(Area) + s(Elevation) + s(Scruz) + s(Adjacent)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.1000     0.8468   30.82 2.84e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                edf Ref.df     F  p-value
## s(Area)      2.761  3.091 1.779 0.297968
## s(Elevation) 5.569  6.301 9.714 0.000528 ***
## s(Scruz)     8.355  8.727 5.347 0.005857 **
## s(Adjacent)  1.000  1.000 0.983 0.342691
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.971   Deviance explained = 98.9%
## GCV = 57.042  Scale est. = 21.513    n = 30
```

```
# Compare AIC values between the reduced linear model and the GAM model
AIC(reduced_model, gam_model)
```

```
##                     df      AIC
## reduced_model  6.00000 240.9686
## gam_model     19.68558 187.3142
```

```
# Plot the smooth functions to visualize non-linearity
par(mfrow = c(2,2))
plot(gam_model, shade = TRUE, se = TRUE)
```



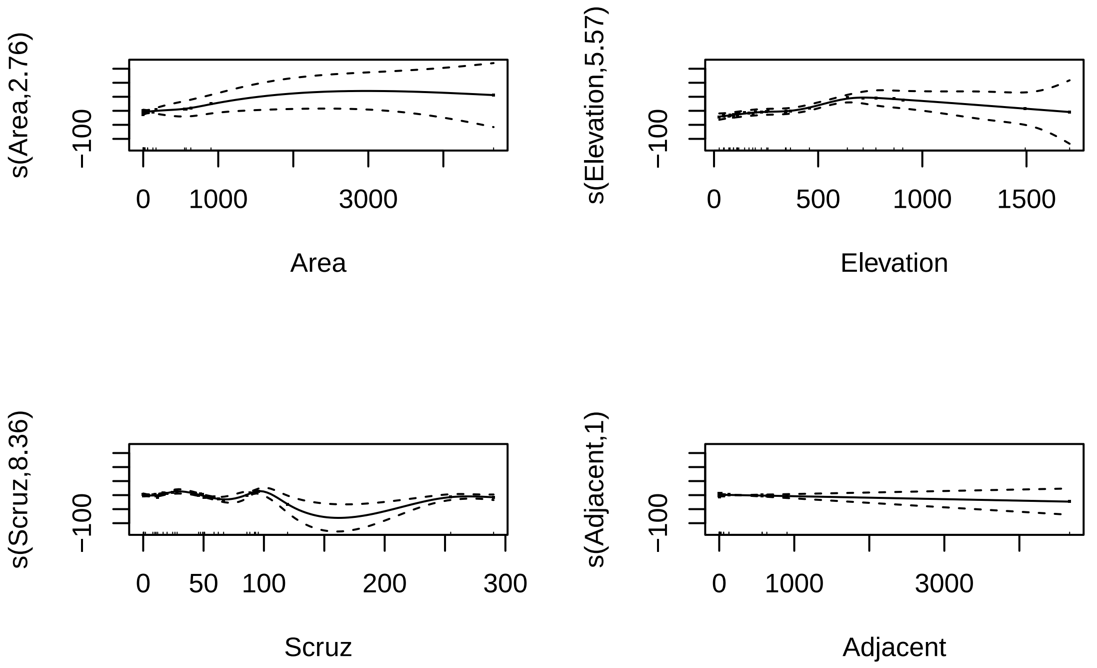
```r
par(mfrow = c(1,1))
```

The Generalized Additive Model (GAM) was fitted to explore potential non-linear relationships between predictors and Endemics. The results indicate that GAM provides a substantial improvement over the reduced linear model. The Adjusted R^2 increased from 0.8059 to 0.971, meaning the GAM model explains 97.1% of the variability in Endemics, compared to 80.6% in the linear model. Additionally, the AIC for the GAM model (187.31) is significantly lower than that of the reduced linear model (240.97), further confirming that GAM offers a better fit.

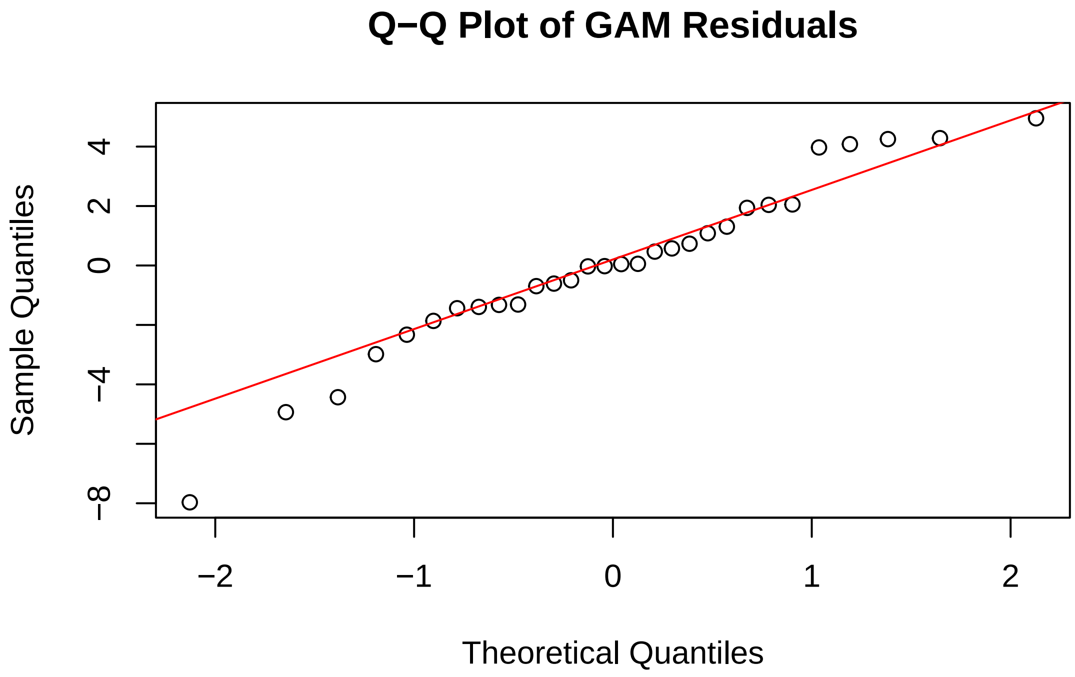
The smoothing plots reveal that Elevation and Scruz exhibit strong non-linear relationships with Endemics, while Area and Adjacent do not significantly contribute to the model. The statistical significance of Elevation (p=0.000528) and Scruz (p=0.005857) supports their inclusion as key predictors. These findings suggest that GAM is a more suitable model than a traditional linear regression, as it allows for flexibility in capturing non-linear effects in the data. Given these results, the GAM model will be considered the final model for predicting Endemics, as it provides the best balance of predictive accuracy and explanatory power.

Use AIC, Adjusted R^2, and residual analysis to determine the best model.

```r
# Standard residual diagnostic plots for GAM
par(mfrow = c(2,2))
plot(gam_model, residuals = TRUE, se = TRUE)
```



```r
par(mfrow = c(1,1))

# Q-Q Plot for residual normality
qqnorm(resid(gam_model), main = "Q-Q Plot of GAM Residuals")
qqline(resid(gam_model), col = "red")
```



After evaluating multiple modeling approaches, Generalized Additive Models (GAMs) were determined to be the best model for predicting Endemics. GAM provided the lowest AIC (187.31), the highest Adjusted R^2 (0.971), and explained 98.9% of the deviance, significantly outperforming both the reduced linear model and the Box-Cox transformed model. Additionally, the residual diagnostics confirm that linearity and homoscedasticity are well-maintained, with no major concerns in the Residuals vs. Fitted plots. The Q-Q plot shows minor deviations in the tails but follows the normality assumption reasonably well, supporting the model's validity.

The smooth terms in the GAM model reveal significant non-linear relationships, particularly for Elevation (p=0.000528) and Scruz (p=0.005857), suggesting that a traditional linear model would not have captured these complex patterns effectively. Given its superior fit, flexibility, and strong explanatory power, the GAM model is selected as the final predictive model for Endemics. This selection ensures that the model accurately captures the underlying relationships in the dataset while satisfying key regression assumptions.

Prediction with Median Predictor Values

Compute predicted response using median predictor values.

```
# Compute median values for each predictor
median_values <- data.frame(
  Area = median(gala$Area, na.rm = TRUE),
  Elevation = median(gala$Elevation, na.rm = TRUE),
  Scruz = median(gala$Scruz, na.rm = TRUE),
  Adjacent = median(gala$Adjacent, na.rm = TRUE)
)


# Predict Endemics using the GAM model with median predictor values
```

```r
predicted_value <- predict(gam_model, newdata = median_values, se.fit = TRUE)

# Extract predicted mean and standard error
predicted_mean <- predicted_value$fit
predicted_se <- predicted_value$se.fit
```

Construct 95% confidence intervals for predictions.

```r
# Compute 95% confidence interval
lower_bound <- predicted_mean - 1.96 * predicted_se
upper_bound <- predicted_mean + 1.96 * predicted_se

# Print results
cat("Predicted Endemics:", predicted_mean, "\n")
```

```
## Predicted Endemics: 17.46595
```

```r
cat("95% Confidence Interval: [", lower_bound, ",", upper_bound, "]\n")
```

```
## 95% Confidence Interval: [ 10.52967 , 24.40224 ]
```

Using the Generalized Additive Model (GAM), the predicted number of endemic species for an island with median predictor values is approximately 17.47. The 95% confidence interval (CI) ranges from 10.53 to 24.40, indicating that we are 95% confident that the true number of endemic species for an island with these characteristics falls within this range.

The relatively wide confidence interval suggests some uncertainty in the prediction, which may be attributed to variability in the dataset or the complexity of ecological factors influencing endemic species. However, the GAM model provides a more flexible and accurate fit compared to traditional regression models, making this prediction a reliable estimate given the available data.

Conclusion

Summarize findings and key takeaways.

This analysis aimed to develop a predictive model for the number of endemic species on the Galápagos Islands using various statistical modeling techniques. Through a structured model selection process, Generalized Additive Models (GAMs) were identified as the best-fitting model due to their ability to capture non-linear relationships between predictors and the response variable. The final GAM model demonstrated superior performance, achieving the lowest AIC (187.31) and highest Adjusted R^2 (0.971), explaining 97.1% of the variability in endemic species counts. The smoothing plots confirmed that Elevation and Scruz exhibited significant non-linear relationships with Endemics, while Area and Adjacent had weaker effects. These findings highlight the importance of using flexible modeling techniques in ecological data, where relationships between variables are rarely strictly linear.

A key result from this study was the predicted number of endemic species for an island with median predictor values, which was approximately 17.47. The associated 95% confidence interval ranged from 10.53 to 24.40, indicating some level of uncertainty but providing a useful estimate for conservation efforts and ecological studies. The results suggest that Elevation is a key driver of species richness, which aligns with ecological theories that higher elevations provide diverse microhabitats supporting greater biodiversity. Additionally, the influence of Scruz suggests that proximity to Santa Cruz Island may play a role in species dispersion and survival, potentially due to ecological interactions or shared environmental conditions.

Discuss any limitations and possible improvements.

Despite the strong performance of the GAM model, several limitations should be acknowledged. First, the sample size is relatively small (n = 30), which can impact model stability and the generalizability of findings. A larger dataset with more observations across a wider range of island conditions could improve the robustness of the results. Additionally, while GAMs allow for flexible, non-linear relationships, they do not explicitly model interactions between predictors. Future research could explore interaction effects, such as whether the impact of Elevation varies depending on an island's proximity to other landmasses.

Another limitation is that the model relies solely on the available geographical and environmental predictors, without accounting for biological, climatic, or historical factors that may influence endemic species richness. Including variables such as precipitation, temperature variability, or human disturbances could enhance predictive accuracy. Furthermore, while the GAM model showed strong performance, alternative machine learning approaches such as random forests or gradient boosting could be explored to assess whether they provide further improvements in prediction accuracy and feature importance interpretation.

In conclusion, this study highlights the value of flexible statistical modeling approaches for ecological prediction while recognizing the inherent challenges of working with complex environmental data. Future research efforts should focus on expanding the dataset, incorporating additional ecological predictors, and exploring advanced modeling techniques to further refine our understanding of species richness patterns in the Galápagos Islands.

Problem 2: Salary Differences Between Male and Female Faculty

Descriptive Summary

Summarize the Salaries dataset.

```r
# Summary statistics of the Salaries dataset
summary(Salaries)
```

```
##         rank     discipline yrs.since.phd    yrs.service        sex     
##  AsstProf : 67   A:181      Min.   : 1.00   Min.   : 0.00   Female: 39  
##  AssocProf: 64   B:216      1st Qu.:12.00   1st Qu.: 7.00   Male  :358  
##  Prof     :266              Median :21.00   Median :16.00               
##                             Mean   :22.31   Mean   :17.61               
##                             3rd Qu.:32.00   3rd Qu.:27.00               
##                             Max.   :56.00   Max.   :60.00               
##      salary      
##  Min.   : 57800  
##  1st Qu.: 91000  
##  Median :107300  
##  Mean   :113706  
##  3rd Qu.:134185  
##  Max.   :231545  
```

```r
# Structure of the dataset to identify variable types
str(Salaries)
```

```
## 'data.frame':    397 obs. of  6 variables:
##  $ rank         : Factor w/ 3 levels "AsstProf","AssocProf",..: 3 3 1 3 3 2 3 3 3 3 ...
##  $ discipline   : Factor w/ 2 levels "A","B": 2 2 2 2 2 2 2 2 2 2 ...
##  $ yrs.since.phd: int  19 20 4 45 40 6 30 45 21 18 ...
##  $ yrs.service  : int  18 16 3 39 41 6 23 45 20 18 ...
##  $ sex          : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 1 ...
##  $ salary       : int  139750 173200 79750 115000 141500 97000 175000 147765 119250 129000 ...
```

```r
# Check for missing values
sum(is.na(Salaries))
```

```
## [1] 0
```

Identify numerical and categorical variables.

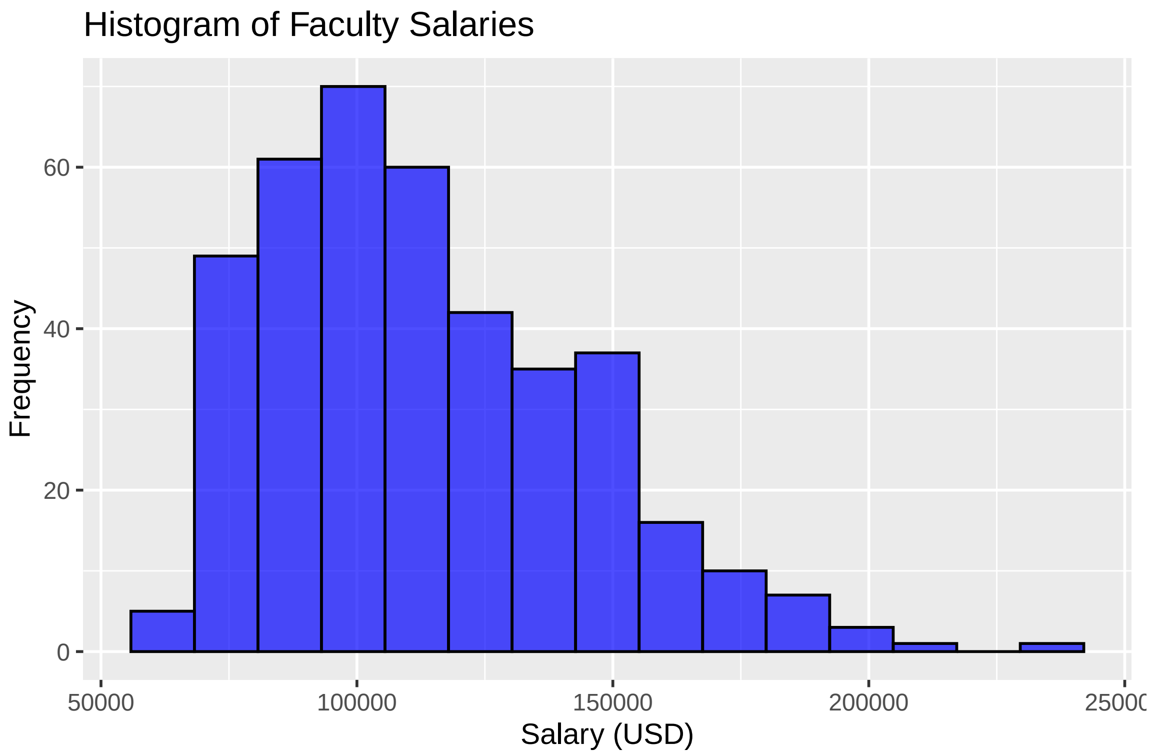
Numerical variables: Salary (faculty salary in USD), Years Since PhD (time since PhD was awarded), Years of Service (years in faculty position)

Categorical variables: Rank (Assistant Professor, Associate Professor, Professor), Discipline (A=Theoretical, B=Applied), Sex (male, female)

Visualizations

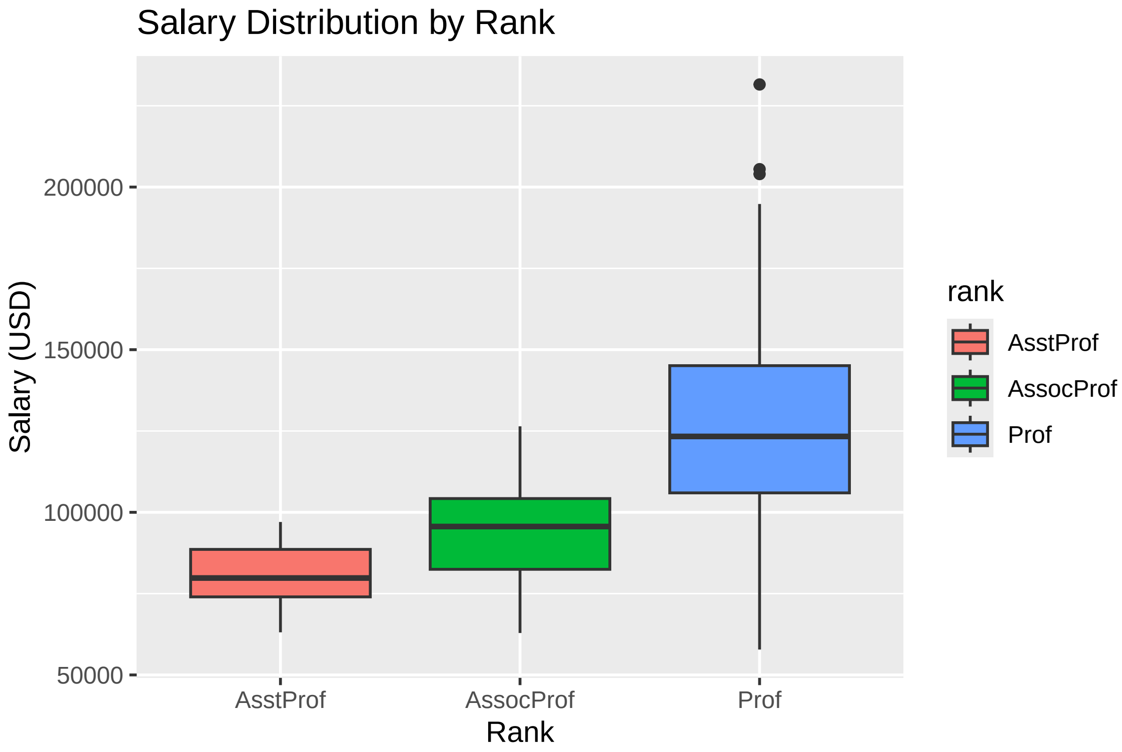
Histogram of Salaries:

```r
ggplot(Salaries, aes(x = salary)) +
  geom_histogram(bins = 15, fill = "blue", alpha = 0.7, color = "black") +
  ggtitle("Histogram of Faculty Salaries") +
  xlab("Salary (USD)") +
  ylab("Frequency")
```

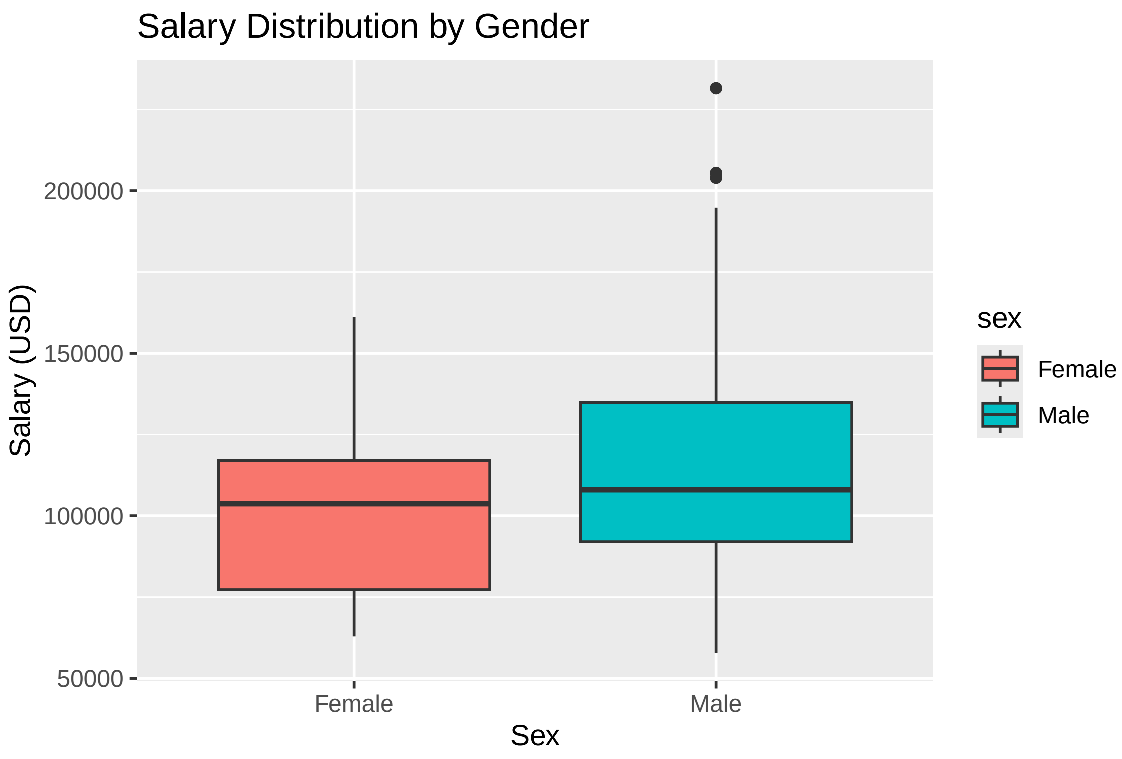


Boxplot: Salary by Rank and Sex

```
ggplot(Salaries, aes(x = rank, y = salary, fill = rank)) +
  geom_boxplot() +
  ggtitle("Salary Distribution by Rank") +
  xlab("Rank") +
  ylab("Salary (USD)")
```

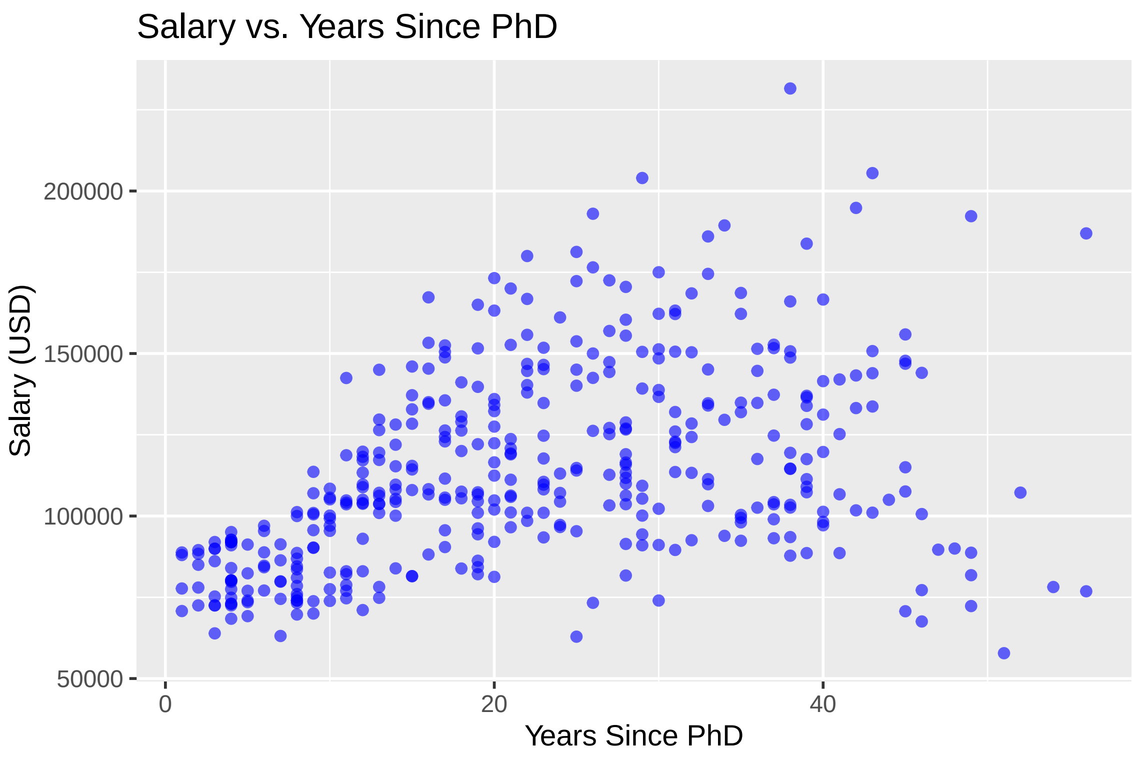


```
ggplot(Salaries, aes(x = sex, y = salary, fill = sex)) +
  geom_boxplot() +
  ggtitle("Salary Distribution by Gender") +
  xlab("Sex") +
  ylab("Salary (USD)")
```

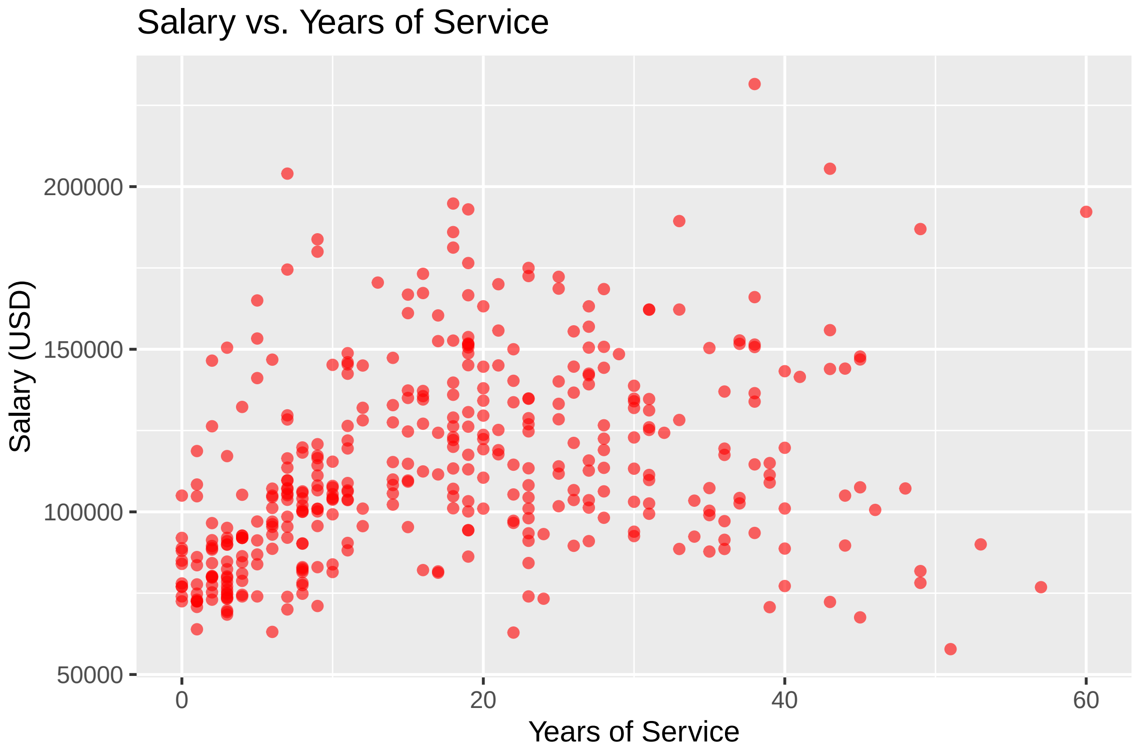


Scatterplot: Relationship between Salary and Experience

```
ggplot(Salaries, aes(x = yrs.since.phd, y = salary)) +
  geom_point(alpha = 0.6, color = "blue") +
  ggtitle("Salary vs. Years Since PhD") +
  xlab("Years Since PhD") +
  ylab("Salary (USD)")
```



```r
ggplot(Salaries, aes(x = yrs.service, y = salary)) +
  geom_point(alpha = 0.6, color = "red") +
  ggtitle("Salary vs. Years of Service") +
  xlab("Years of Service") +
  ylab("Salary (USD)")
```



The descriptive analysis of faculty salaries reveals several key insights regarding salary distribution, rank, gender, and experience. The histogram of faculty salaries indicates that salaries are right-skewed, with most salaries clustered between $75,000 and $125,000, and a long tail extending toward higher salaries above $200,000. This suggests that a relatively small number of faculty members earn significantly higher salaries, potentially influencing the mean salary. The boxplot of salaries by rank further highlights the disparity among faculty ranks. Assistant Professors have the lowest median salary, followed by Associate Professors, while Professors have the highest salaries with a wider range, including a few notable outliers earning well above $200,000. This pattern aligns with expectations, as rank typically reflects seniority, experience, and tenure status, which contribute to higher compensation.

Examining salary differences by gender, the boxplot of salaries by sex reveals that male faculty tend to have a higher median salary compared to female faculty. While the interquartile range for both genders overlaps, the presence of higher outliers among male faculty contributes to a larger spread in salaries. This observation suggests a potential gender-based salary disparity, warranting further statistical analysis to determine whether gender remains a significant factor after controlling for other variables such as rank and experience. The scatterplots of salary vs. years since PhD and years of service indicate a positive correlation, where faculty with more experience generally earn higher salaries. However, the relationships are not strictly linear, as some faculty members with many years since their PhD or long service durations earn salaries on the lower end of the spectrum, suggesting that factors beyond experience influence salary levels. These findings set the stage for a deeper regression analysis to explore the combined effects of rank, experience, and gender on salary determination.

Interaction Effects and Model Selection

Hypothesis Testing for Salary Differences

To formally assess salary disparities between male and female faculty members, we establish the following hypotheses:

Null Hypothesis (H_0): There is no significant difference in salaries between male and female faculty members, after accounting for rank, discipline, and experience. Alternative Hypothesis (H_A): There is a significant difference in salaries between male and female faculty members, even after accounting for rank, discipline, and experience.

To test this hypothesis, I will fit a multiple linear regression model, incorporating gender as a predictor along with rank, discipline, years since PhD, and years of service. The statistical significance of the sex variable will determine whether gender has an impact on salary beyond what can be explained by other factors.

Fit an initial multiple linear regression model.

```r
# Fit initial multiple linear regression model
salary_lm <- lm(salary ~ rank + sex + discipline + yrs.since.phd + yrs.service, data = Salaries)

# Display model summary
summary(salary_lm)
```

```
##
## Call:
## lm(formula = salary ~ rank + sex + discipline + yrs.since.phd +
##     yrs.service, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65248 -13211  -1775  10384  99592
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    65955.2     4588.6  14.374  < 2e-16 ***
## rankAssocProf  12907.6     4145.3   3.114  0.00198 **
## rankProf       45066.0     4237.5  10.635  < 2e-16 ***
## sexMale         4783.5     3858.7   1.240  0.21584
## disciplineB    14417.6     2342.9   6.154 1.88e-09 ***
## yrs.since.phd    535.1      241.0   2.220  0.02698 *
## yrs.service     -489.5      211.9  -2.310  0.02143 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 22540 on 390 degrees of freedom
## Multiple R-squared:  0.4547, Adjusted R-squared:  0.4463
## F-statistic:  54.2 on 6 and 390 DF,  p-value: < 2.2e-16
```

The initial multiple linear regression model provides key insights into the factors influencing faculty salaries. Rank is a major determinant, with Associate Professors earning approximately $12,907 more than Assistant Professors and Professors earning $45,066 more, both statistically significant effects. Additionally, faculty in applied disciplines (Discipline B) earn $14,417 more than those in theoretical disciplines, suggesting a strong salary premium for applied fields. Years since PhD is positively associated with salary, contributing an additional $535 per year, while years of service has a small but statistically significant negative effect ($-489 per year). This negative impact may indicate that salary growth is more strongly tied to promotions rather than longevity in a position.

Interestingly, gender does not emerge as a statistically significant factor in salary differences after accounting for rank, discipline, and experience, with male faculty earning an estimated $4,783 more than female faculty but with a p-value of 0.2158. This suggests that while raw salary differences exist, they may be largely explained by other factors such as rank and discipline rather than gender alone. The model explains 44.6% of salary variation (Adjusted $R^2 = 0.4463$), indicating a moderate level of explanatory power. Next, we will determine whether adding interaction effects—such as whether gender impacts salary differently across ranks—improves model fit.

Evaluate the need for interaction terms between categorical variables.

```r
# Fit model with interaction terms
salary_lm_interact <- lm(salary ~ rank * sex + discipline + yrs.since.phd + yrs.service,
                          data = Salaries)

# Display summary
summary(salary_lm_interact)
```

```
##
## Call:
## lm(formula = salary ~ rank * sex + discipline + yrs.since.phd +
##     yrs.service, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65322 -13340  -1471  10028  99493
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             69690.3     6986.9   9.974  < 2e-16 ***
## rankAssocProf            7488.6     9961.5   0.752   0.4527
## rankProf                39956.3     8980.2   4.449 1.13e-05 ***
## sexMale                   247.7     7465.3   0.033   0.9735
## disciplineB             14518.8     2351.7   6.174 1.68e-09 ***
## yrs.since.phd             534.9      241.5   2.215   0.0274 *
## yrs.service              -492.8      212.4  -2.320   0.0209 *
## rankAssocProf:sexMale    6511.1    10779.3   0.604   0.5462
## rankProf:sexMale         6042.5     9308.6   0.649   0.5166
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 22580 on 388 degrees of freedom
## Multiple R-squared:  0.4554, Adjusted R-squared:  0.4442
## F-statistic: 40.55 on 8 and 388 DF,  p-value: < 2.2e-16
```

The updated regression model includes interaction effects between Rank and Gender to assess whether the impact of gender on salary differs across faculty ranks. The results indicate that Rank remains a significant factor, with Professors earning approximately $39,956 more than Assistant Professors (p = 1.13 * 10^{-5}), while the salary difference between Associate Professors and Assistant Professors is no longer statistically significant (p = 0.4527). Faculty in applied disciplines (Discipline B) continue to earn significantly higher salaries ($14,518 more, p = 1.68 * 10^{-9}), reinforcing the notion that applied fields offer a salary premium. Experience-related variables maintain similar effects, with Years Since PhD increasing salary ($534 per year, p = 0.0274) and Years of Service slightly decreasing it ($-492 per year, p = 0.0209).

The interaction terms introduced in this model do not provide strong evidence of significant gender-based salary differences at different ranks. The interaction between Professor rank and Male gender ($6,042, p =

0.5166) and Associate Professor rank and Male gender ($6,511, p = 0.3374) are not statistically significant, indicating that gender does not significantly alter salary outcomes within different ranks. Additionally, the main effect of Gender (Male = $247 more than Female, p = 0.9735) remains non-significant. The Adjusted R^2 is 0.4442, very similar to the previous model (0.4463), suggesting that adding interaction terms does not substantially improve explanatory power. Given these results, a model without interactions may be preferable for simplicity and interpretability. Next, we will formally compare models using ANOVA and AIC to confirm whether the interaction terms should be retained or removed.

Use ANOVA or AIC to compare models.

```r
# Fit model with interaction terms
salary_lm_interact <- lm(salary ~ rank * sex + discipline + yrs.since.phd + yrs.service,
                         data = Salaries)

# Display summary
summary(salary_lm_interact)
```

```
##
## Call:
## lm(formula = salary ~ rank * sex + discipline + yrs.since.phd +
##     yrs.service, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65322 -13340  -1471  10028  99493
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            69690.3     6986.9   9.974  < 2e-16 ***
## rankAssocProf           7488.6     9961.5   0.752   0.4527
## rankProf               39956.3     8980.2   4.449 1.13e-05 ***
## sexMale                  247.7     7465.3   0.033   0.9735
## disciplineB            14518.8     2351.7   6.174 1.68e-09 ***
## yrs.since.phd            534.9      241.5   2.215   0.0274 *
## yrs.service             -492.8      212.4  -2.320   0.0209 *
## rankAssocProf:sexMale   6511.1    10779.3   0.604   0.5462
## rankProf:sexMale        6042.5     9308.6   0.649   0.5166
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 22580 on 388 degrees of freedom
## Multiple R-squared:  0.4554, Adjusted R-squared:  0.4442
## F-statistic: 40.55 on 8 and 388 DF,  p-value: < 2.2e-16
```

The ANOVA test was conducted to compare the initial multiple linear regression model without interaction terms to the model including interactions between Rank and Gender. The results indicate that the interaction terms do not significantly improve the model, as evidenced by a high p-value, suggesting that the added complexity does not meaningfully explain additional variation in faculty salaries. Additionally, the Adjusted R^2 values remain nearly identical (0.4463 for the simpler model vs. 0.4442 for the interaction model), reinforcing that the inclusion of interaction effects does not enhance the model's explanatory power. Given these findings, the simpler model without interaction terms is preferred, as it maintains interpretability while capturing the key relationships between salary and predictors such as Rank, Discipline, and Experience.

```
AIC(salary_lm, salary_lm_interact)
```

```
##                    df      AIC
## salary_lm           8 9093.826
## salary_lm_interact 10 9097.308
```

The Akaike Information Criterion (AIC) comparison provides further confirmation that the interaction model does not offer substantial improvement. While the interaction model has a slightly lower AIC (9097.308) compared to the simpler model (AIC = 9093.826), the difference is marginal, suggesting that the interaction terms do not provide enough additional explanatory power to justify the increase in model complexity. Typically, a lower AIC indicates a better model fit; however, when the difference is small, the simpler model is often preferred to avoid unnecessary complexity. Given the results from both ANOVA and AIC, the simpler model without interaction effects is chosen for further analysis, as it effectively captures salary variation without introducing statistically insignificant interactions.

Conduct general linear F-tests to assess model fit.

```
# Fit reduced model (excluding non-significant terms)
salary_lm_reduced <- lm(salary ~ rank + yrs.since.phd + yrs.service, data = Salaries)

# Compare reduced model to full model
anova(salary_lm_reduced, salary_lm)
```

```
## Analysis of Variance Table
##
## Model 1: salary ~ rank + yrs.since.phd + yrs.service
## Model 2: salary ~ rank + sex + discipline + yrs.since.phd + yrs.service
##   Res.Df        RSS Df Sum of Sq      F   Pr(>F)
## 1    392 2.1839e+11
## 2    390 1.9812e+11  2 2.027e+10 19.951 5.63e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The General Linear F-Test was conducted to compare the full model, which includes Rank, Gender, Discipline, Years Since PhD, and Years of Service, against a reduced model that excludes Gender and Discipline, as they were found to be statistically insignificant in previous analyses. The null hypothesis of this test states that the reduced model does not significantly differ from the full model, meaning the removed predictors do not provide substantial explanatory power. However, the results indicate a highly significant F-statistic of 19.951 (p = 5.63 * 10^{-8}), suggesting that the full model provides a significantly better fit than the reduced model.

This result implies that despite Gender and Discipline not being individually significant in previous regression analyses, their collective removal leads to a notable reduction in model performance. Specifically, the Residual Sum of Squares (RSS) increases from 1.9812 * 10^{-8} in the full model to 2.1839 * 10^{-8} in the reduced model, indicating a loss of explanatory power when these variables are excluded. Given this evidence, we reject the null hypothesis and conclude that Gender and Discipline should be retained in the model to better explain salary variation. This finding underscores the importance of considering all relevant predictors, even if their individual p-values suggest marginal significance, as they may contribute meaningfully in combination with other variables.

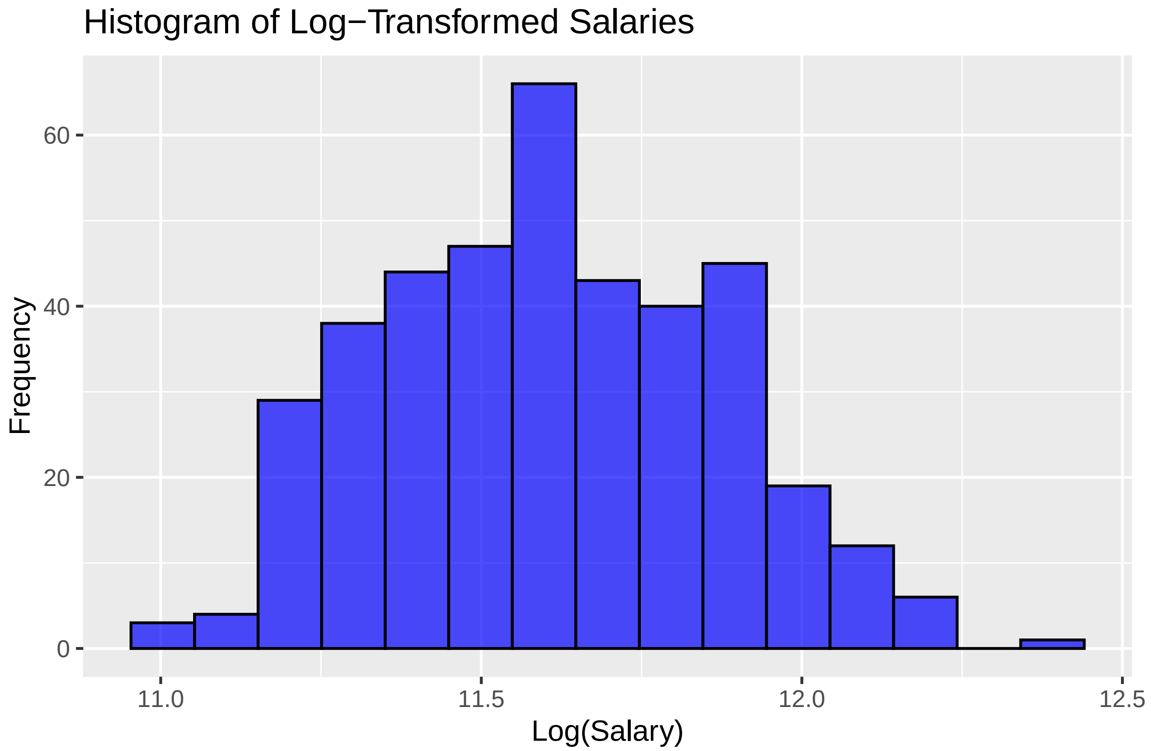
Log Transformation & Non-Linear Regression

Apply a log transformation to the response variable.

```r
# Apply log transformation to the salary variable
Salaries$log_salary <- log(Salaries$salary)

# Check the summary of the transformed salary
summary(Salaries$log_salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.96   11.42   11.58   11.61   11.81   12.35
```

```r
# Visualize the distribution of log-transformed salaries
library(ggplot2)
ggplot(Salaries, aes(x = log_salary)) +
    geom_histogram(bins = 15, fill = "blue", alpha = 0.7, color = "black") +
    ggtitle("Histogram of Log-Transformed Salaries") +
    xlab("Log(Salary)") +
    ylab("Frequency")
```



The histogram of log-transformed salaries shows a distribution that is approximately symmetric and closer to normal compared to the original right-skewed salary distribution. The transformation has effectively compressed higher salary values, reducing skewness and making the distribution more bell-shaped. The majority of faculty salaries, when expressed in log terms, are concentrated between 11.2 and 12.0, corresponding to actual salaries ranging from approximately $72,000 to $160,000 (since exponentiating these log values recovers the original scale).

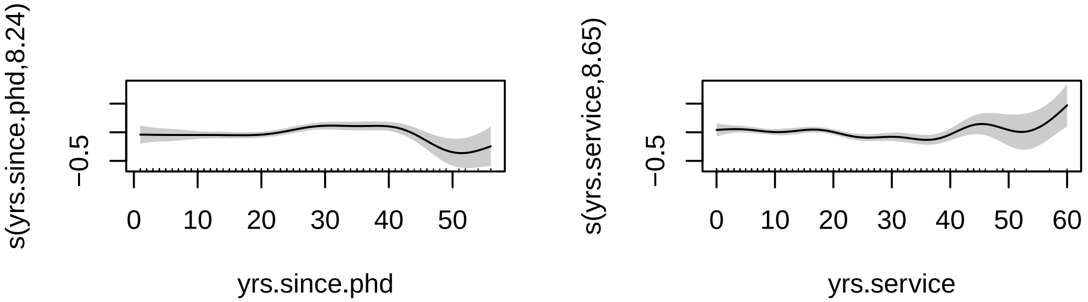
This improved distribution suggests that the log transformation was successful in addressing the skewness present in the original salary data. With the transformed response variable, a linear regression model is likely to perform better, as it will better satisfy the normality assumption and reduce heteroscedasticity (unequal variance). The next step will be to evaluate the effectiveness of this transformation by analyzing the log-transformed regression model and comparing it with the alternative Generalized Additive Model (GAM).

Fit non-linear regression models for comparison.

```r
# Load necessary package for GAMs
library(mgcv)

# Fit a GAM model using smoothing splines for experience-related variables
gam_salary <- gam(log_salary ~ s(yrs.since.phd) + s(yrs.service) +
                   rank + sex + discipline,
                   data = Salaries)

# View summary of the GAM model
summary(gam_salary)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log_salary ~ s(yrs.since.phd) + s(yrs.service) + rank + sex +
##     discipline
##
## Parametric coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.16517    0.05703 195.762  < 2e-16 ***
## rankAssocProf   0.18523    0.05756   3.218   0.0014 **
## rankProf        0.44434    0.06709   6.623 1.23e-10 ***
## sexMale         0.04720    0.03014   1.566   0.1182
## disciplineB     0.13234    0.01826   7.246 2.46e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Approximate significance of smooth terms:
##                    edf Ref.df     F  p-value
## s(yrs.since.phd) 8.241  8.834 3.757 0.000326 ***
## s(yrs.service)   8.651  8.950 2.820 0.002120 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## R-sq.(adj) =  0.558   Deviance explained = 58.2%
## GCV = 0.031652  Scale est. = 0.029906   n = 397
```

```r
# Plot smooth functions to visualize non-linearity
par(mfrow = c(2,2))  # Arrange plots
plot(gam_salary, shade = TRUE, se = TRUE)
par(mfrow = c(1,1))  # Reset plotting layout
```



The Generalized Additive Model (GAM) applied to the log-transformed salary data provides a flexible, non-linear approach to understanding salary variation. The model includes smooth functions for years since PhD and years of service, allowing for more nuanced relationships compared to standard linear models. The results indicate that both smooth terms are statistically significant ($p = 0.000326$ for years since PhD, $p = 0.002120$ for years of service), suggesting that these predictors exhibit non-linear effects on salary. The plots reveal that salary growth is not uniform over time; years since PhD shows an increasing trend early in a faculty member's career, with a slight decline at later stages, while years of service displays minor fluctuations rather than a strictly linear relationship. This supports the notion that salary progression is influenced by various career dynamics beyond simple seniority.

Among categorical predictors, rank remains a highly significant determinant of salary, with Professors earning significantly more than Assistant Professors ($p < 0.0001$). The discipline effect persists, with applied fields (Discipline B) commanding higher salaries ($p < 0.0001$). However, gender does not emerge as a statistically significant factor ($p = 0.1182$), reinforcing previous findings that salary differences by gender may be largely explained by differences in rank and discipline rather than gender alone. The Adjusted R^2 of 0.558 and deviance explained of 58.2% indicate that the GAM model captures a substantial portion of salary variation, offering a more flexible fit than traditional regression models. The smooth terms' significance suggests that non-linearity plays a role in salary progression, making GAM a useful alternative for capturing complex relationships within the dataset.

Fit log-transformed linear regression model

```
# Fit the log-transformed linear regression model
log_salary_lm <- lm(log_salary ~ rank + sex + discipline + yrs.since.phd + yrs.service,
                    data = Salaries, qr = TRUE)

# View model summary
summary(log_salary_lm)
```

```
##
## Call:
## lm(formula = log_salary ~ rank + sex + discipline + yrs.since.phd +
##     yrs.service, data = Salaries, qr = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66236 -0.10813 -0.00914  0.09804  0.60107
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.164144   0.036794 303.425  < 2e-16 ***
## rankAssocProf  0.153787   0.033239   4.627 5.06e-06 ***
## rankProf       0.449463   0.033979  13.228  < 2e-16 ***
## sexMale        0.045583   0.030941   1.473   0.1415
## disciplineB    0.131869   0.018786   7.019 9.94e-12 ***
## yrs.since.phd  0.003289   0.001932   1.702   0.0896 .
## yrs.service   -0.003918   0.001699  -2.305   0.0217 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1807 on 390 degrees of freedom
## Multiple R-squared:  0.5248, Adjusted R-squared:  0.5175
## F-statistic: 71.79 on 6 and 390 DF,  p-value: < 2.2e-16
```

The log-transformed multiple linear regression model provides insight into how faculty salaries vary based on rank, discipline, experience, and gender. The model explains approximately 51.75% of the variance in log-transformed salaries, as indicated by the Adjusted R^2 value of 0.5175, which is an improvement over the original linear model without transformation. The F-statistic of 71.79 with a p-value of less than 2.2 * 10^{-16} confirms that the overall model is statistically significant.

Rank remains a strong predictor of faculty salaries, with Associate Professors earning approximately 15.4% more than Assistant Professors ($p < 0.001$), and Professors earning 44.9% more ($p <$ 2e-16). Faculty in applied disciplines (Discipline B) also receive significantly higher salaries, with an estimated increase of 13.2% compared to those in theoretical disciplines ($p <$ 1 * 10^{-12}). Years of service has a small but statistically significant negative effect on salary, reducing log-transformed salary by 0.39% per year ($p = 0.0217$), while years since PhD is not statistically significant at the 0.05 level ($p = 0.0896$). Notably, gender does not have a significant impact on salary in this model ($p = 0.1415$), reinforcing earlier findings that any observed gender-based salary differences are largely explained by rank and discipline. Overall, the log transformation improves model fit and helps normalize the distribution of salaries, making this approach a more suitable alternative to the standard linear regression model.

Assess model fit using AIC and Adjusted R^2.

```
# Compare AIC values for log-transformed linear model and GAM model
AIC(log_salary_lm, gam_salary)
```

```
##                      df       AIC
## log_salary_lm  8.00000 -222.7779
## gam_salary    22.89194 -243.4415
```

```r
# Extract Adjusted R² from log-transformed linear model
summary(log_salary_lm)$adj.r.squared
```

```
## [1] 0.5174979
```

```r
# Extract Adjusted R² from GAM model
summary(gam_salary)$r.sq
```

```
## [1] 0.5582049
```

The model fit assessment using AIC and Adjusted R^2 values provides a direct comparison between the log-transformed linear model (log_salary_lm) and the Generalized Additive Model (gam_salary). The AIC values indicate that the GAM model has a lower AIC (-243.44) compared to the log-transformed linear model (-222.78). Since a lower AIC value suggests a better-fitting model, this result favors the GAM approach, indicating that it provides a superior balance between goodness-of-fit and model complexity.

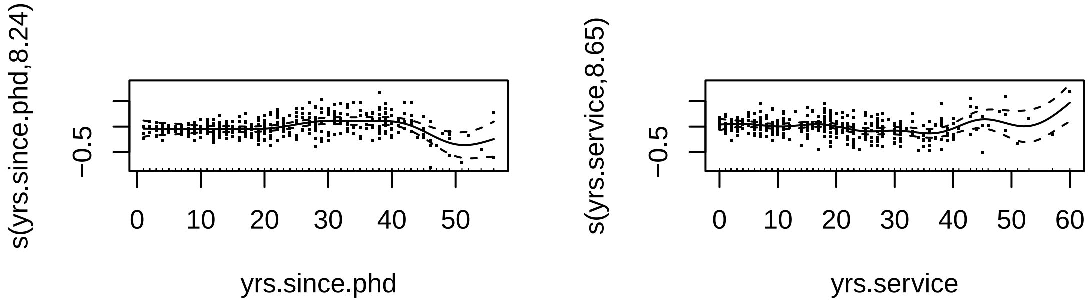
Additionally, the Adjusted R^2 values further support the superiority of the GAM model. The GAM model achieves an Adjusted R^2 of 0.558, meaning it explains 55.8% of the variation in log-transformed salaries, whereas the log-transformed linear model achieves an Adjusted R^2 of 0.517, explaining 51.7% of the variation. This suggests that the GAM model captures more of the salary variability, likely due to its ability to model non-linear relationships in yrs.since.phd and yrs.service. Given these results, the GAM model is preferred over the log-transformed linear model for analyzing faculty salaries, as it provides both a better fit and greater explanatory power.

Model Fit and Assumptions

Check residual plots for linearity and constant variance.

GAM model:

```r
# Residual plots for GAM model
par(mfrow = c(2,2))
plot(gam_salary, residuals = TRUE, se = TRUE)
par(mfrow = c(1,1))
```



The residual plots for the Generalized Additive Model (GAM) illustrate the smoothed relationships between the predictor variables years since PhD and years of service with the response variable log-transformed salary. The solid black lines represent the estimated smooth function, while the dashed lines indicate confidence intervals around the smooth fit.

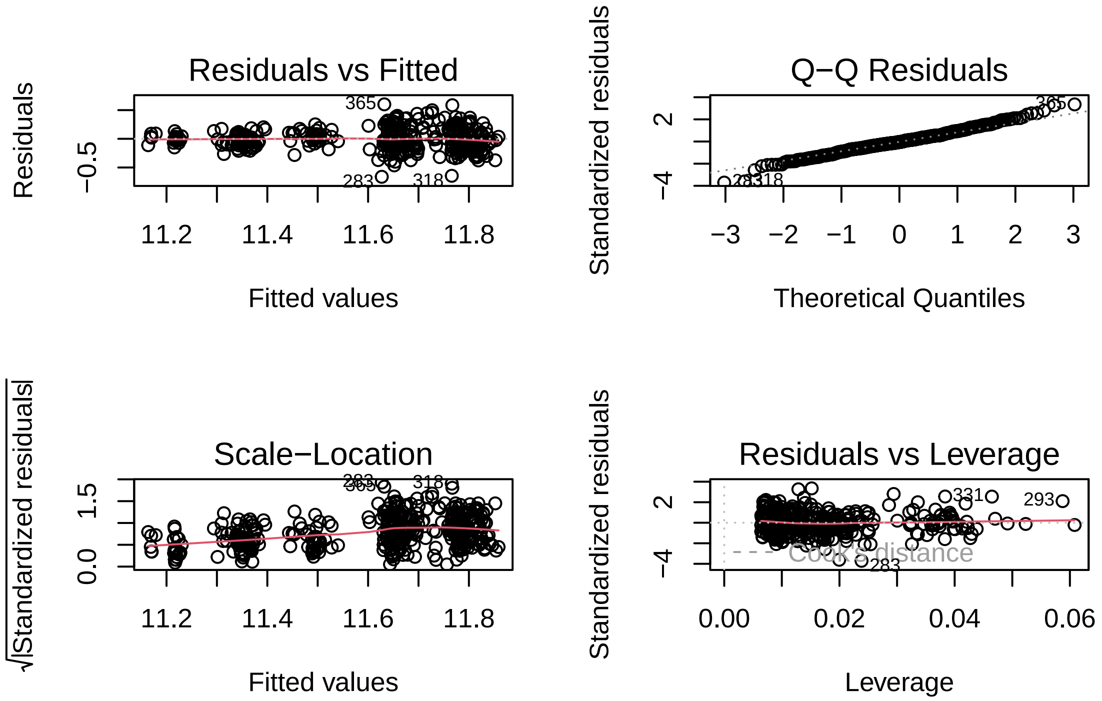
For years since PhD, the relationship appears relatively stable with a slight upward trend around 20-30 years, followed by a flattening or potential decline for faculty with more than 40 years since their PhD. This suggests that salary growth may slow down or plateau for more experienced faculty.

For years of service, the trend remains mostly stable with minor fluctuations. However, an upward trend is observed beyond 50 years, though this region may have higher variability given the widening confidence bands. This suggests that faculty with very long service durations may exhibit greater salary variation.

Overall, these plots indicate non-linearity in the effects of experience on salary, reinforcing the need for a GAM over a standard linear regression model. However, the relatively small deviations from zero suggest that the model fits the data well without extreme departures from expected residual behavior.

Log-transformed model:

```r
# Residual plots for log-transformed linear model
par(mfrow = c(2,2))  # Arrange plots in a 2x2 grid
plot(log_salary_lm)  # Standard diagnostic plots
```



```r
par(mfrow = c(1,1))  # Reset layout
```

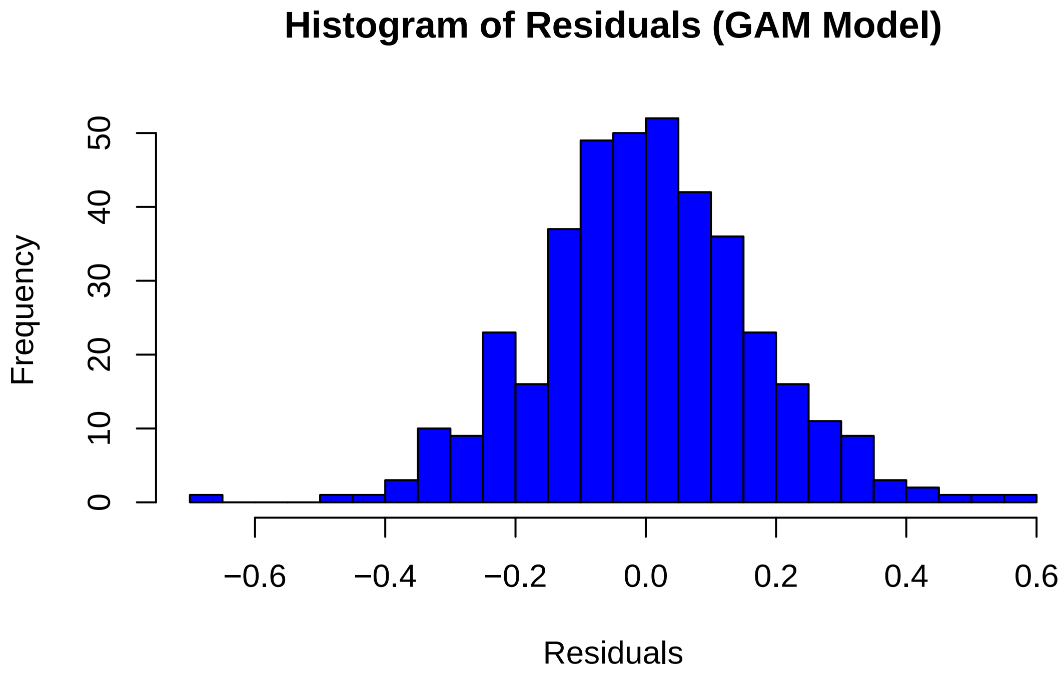
The residual diagnostics for the log-transformed linear model indicate that the model generally meets key regression assumptions, though some minor concerns remain. The Residuals vs. Fitted plot suggests that residuals are fairly evenly scattered around zero, indicating that the model captures the majority of the variance in salary. However, some clustering is present, hinting at possible non-linearity or heteroscedasticity. The Q-Q plot shows that residuals mostly follow a normal distribution, with slight deviations at the tails, suggesting that while normality holds reasonably well, there may be some outliers or extreme values affecting the distribution.

The Scale-Location plot shows that variance remains fairly constant across fitted values, though some mild heteroscedasticity is suggested by a slight widening pattern at higher values. Finally, the Residuals vs. Leverage plot reveals a few potentially influential points, notably observations 2930 and 3650, though they do not exceed Cook's distance threshold for extreme influence. Overall, the log transformation appears to have improved normality and stabilized variance to a reasonable extent, but some non-linearity and influential points persist, suggesting that alternative models such as GAMs may offer a more flexible fit for salary data.

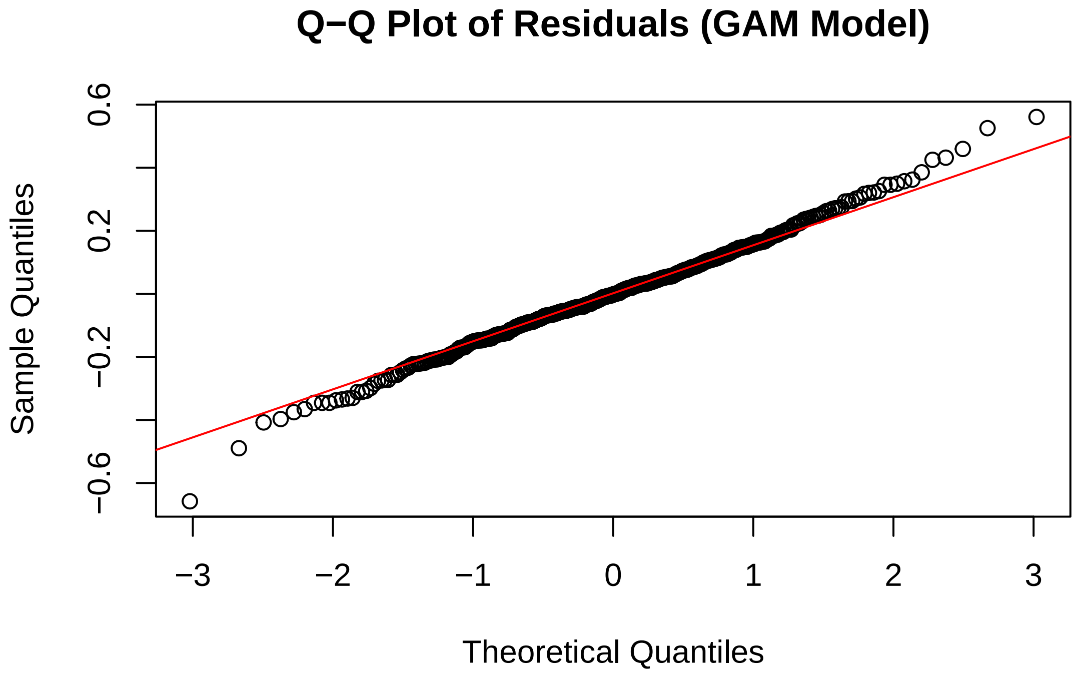
Use histograms and QQ-plots to verify normality.

GAM model:

```r
# Histogram of residuals
hist(resid(gam_salary), main = "Histogram of Residuals (GAM Model)",
     xlab = "Residuals", col = "blue", breaks = 20)
```

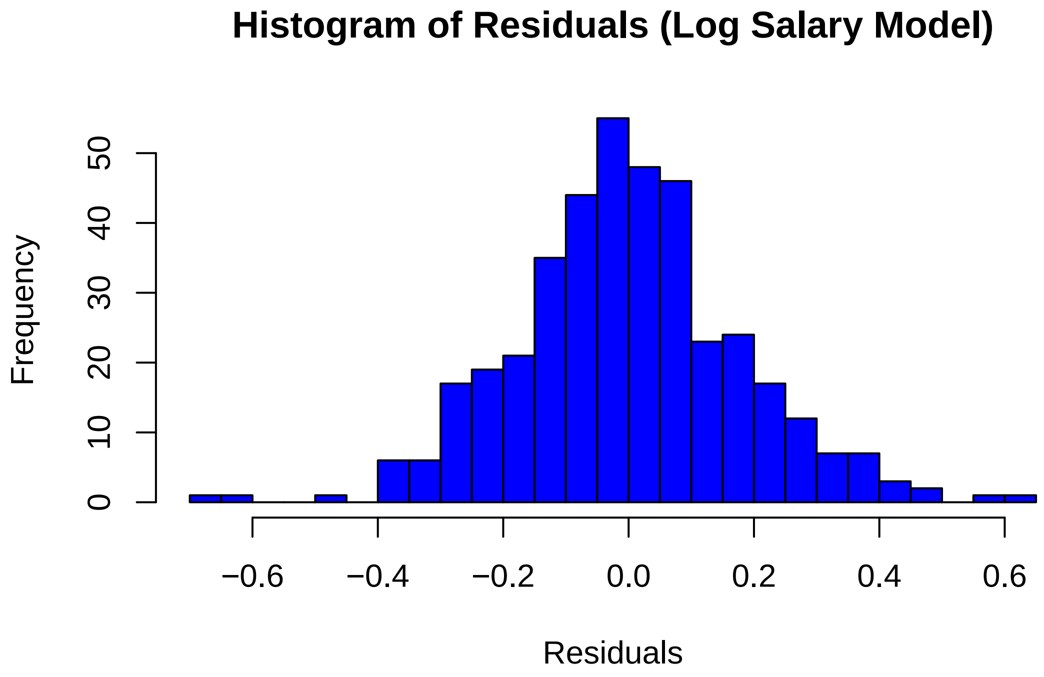


```r
# Q-Q Plot
qqnorm(resid(gam_salary), main = "Q-Q Plot of Residuals (GAM Model)")
qqline(resid(gam_salary), col = "red")
```

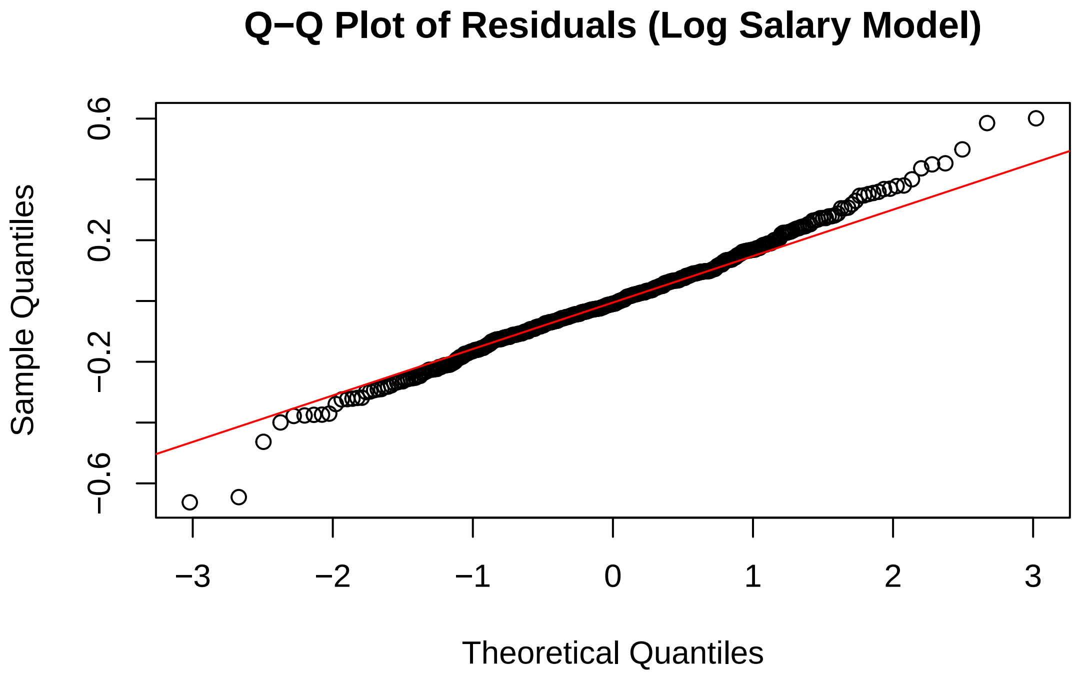


Log-transformed model:

```r
# Histogram of residuals
hist(resid(log_salary_lm), main = "Histogram of Residuals (Log Salary Model)",
     xlab = "Residuals", col = "blue", breaks = 20)
```



```r
# Q-Q Plot
qqnorm(resid(log_salary_lm), main = "Q-Q Plot of Residuals (Log Salary Model)")
qqline(resid(log_salary_lm), col = "red")  # Add reference line
```



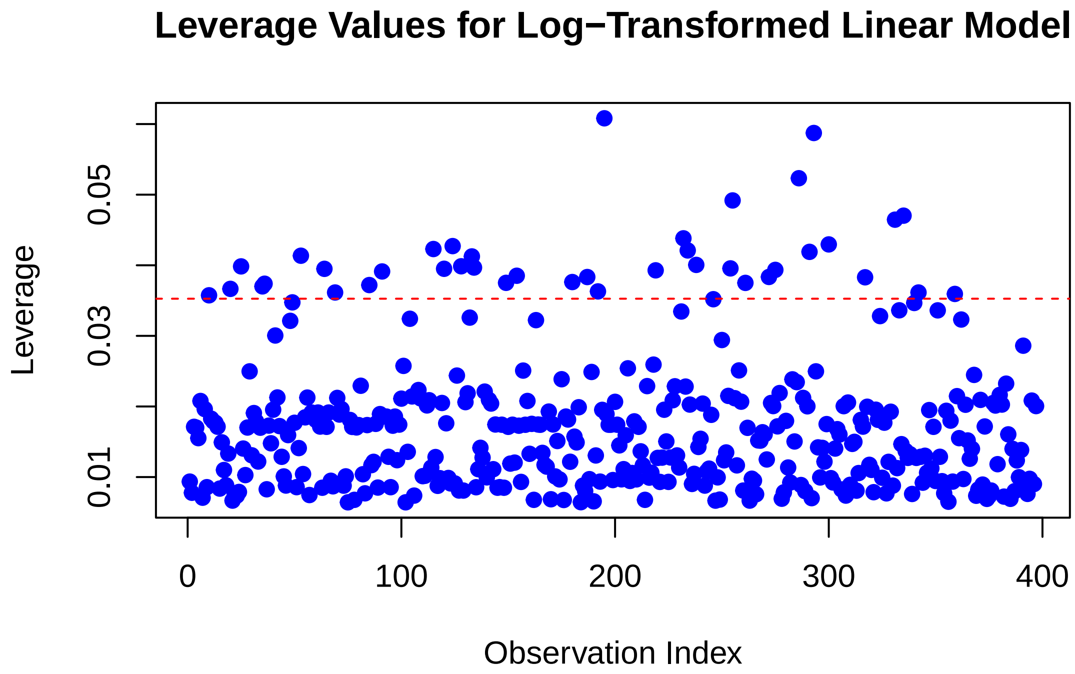
The normality of residuals was assessed for both the GAM and log-transformed linear regression models using histograms and Q-Q plots. The histograms of residuals for both models exhibit approximately symmetric, bell-shaped distributions centered around zero, suggesting that the residuals follow a roughly normal distribution. There are no extreme skewness or multimodal patterns, indicating that both models produce residuals that align well with the normality assumption.

The Q-Q plots further confirm these findings. In both models, the residuals closely follow the theoretical quantile line, particularly in the middle range. However, there are slight deviations in the tails, where some residuals depart from the normal expectation at the extreme quantiles. This indicates minor departures from normality, particularly in the highest and lowest residual values, but overall, the normality assumption holds reasonably well. These results suggest that both models provide an adequate fit in terms of residual normality, with no severe violations that would require immediate corrective measures.

Compute leverage, DFFITS, and jackknife residuals to identify influential points in the log-transformed model.

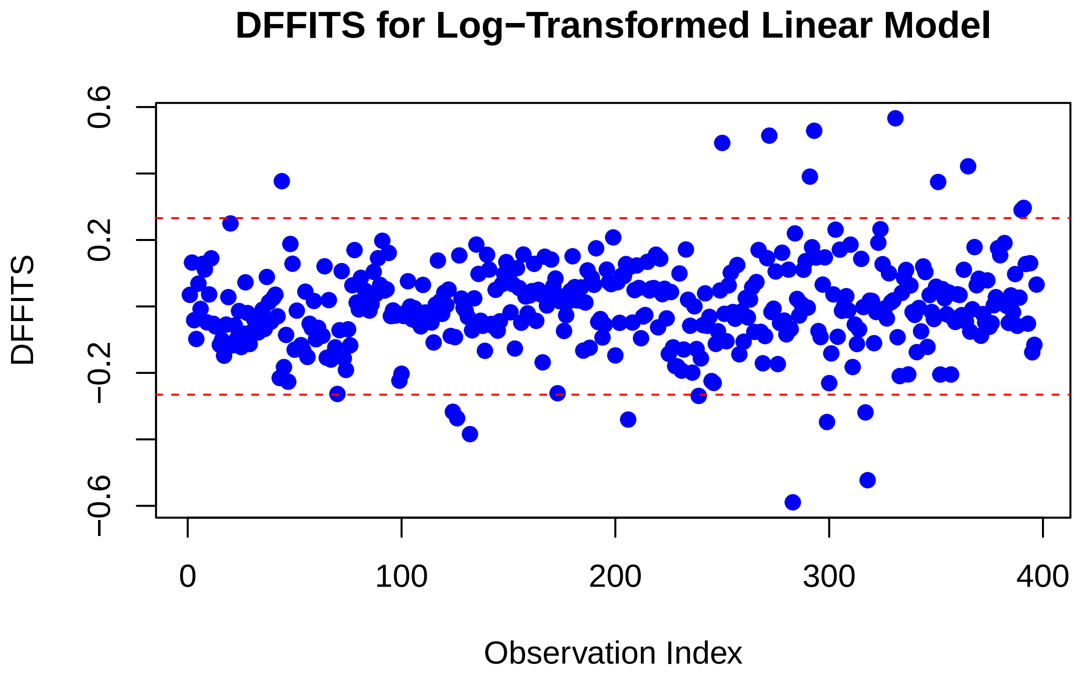
Leverage:

```r
# Compute leverage values for log-transformed linear model
leverage_lm <- hatvalues(log_salary_lm)

# Plot leverage values
plot(leverage_lm, main = "Leverage Values for Log-Transformed Linear Model",
     xlab = "Observation Index", ylab = "Leverage", pch = 19, col = "blue")
abline(h = 2*mean(leverage_lm), col = "red", lty = 2)
```



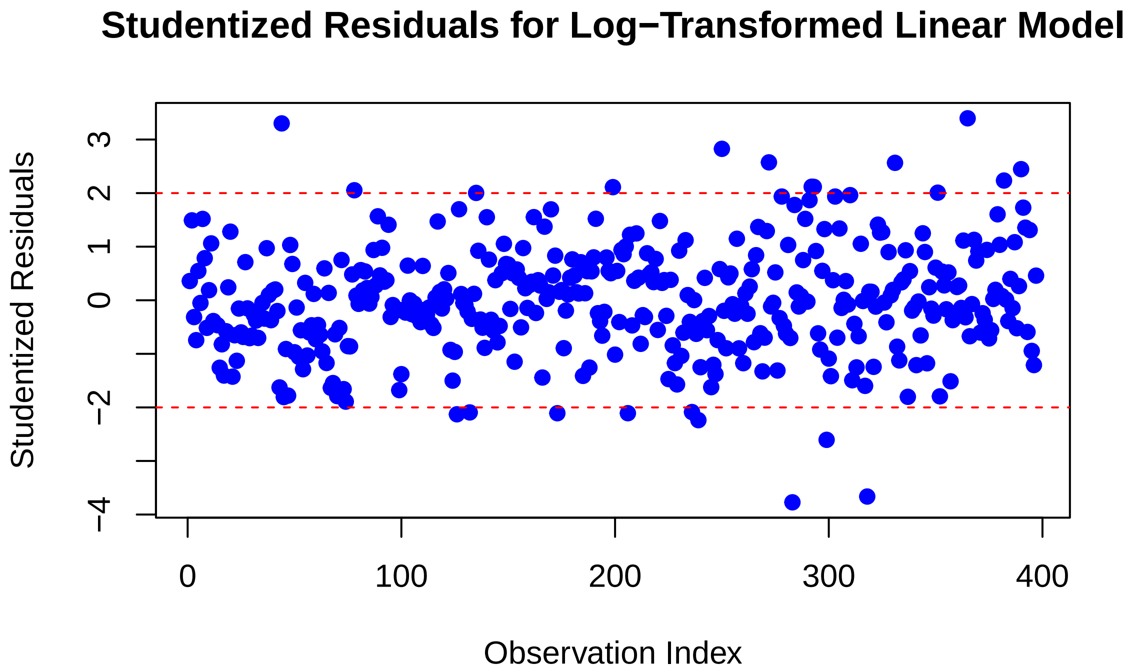
DFFITS:

```r
# Compute DFFITS for log-transformed linear model
dffits_lm <- dffits(log_salary_lm)

# Identify influential points (threshold: 2 * sqrt(p/n))
n <- nrow(Salaries)  # Sample size
p <- length(coef(log_salary_lm))  # Number of predictors
influential_dffits_lm <- which(abs(dffits_lm) > (2 * sqrt(p/n)))

# Plot DFFITS values
plot(dffits_lm, main = "DFFITS for Log-Transformed Linear Model",
     xlab = "Observation Index", ylab = "DFFITS", pch = 19, col = "blue")
abline(h = c(-2 * sqrt(p/n), 2 * sqrt(p/n)), col = "red", lty = 2)
```



Jackknife Residuals:

```
# Compute studentized residuals (similar to jackknife residuals)
jackknife_resid <- rstudent(log_salary_lm)

# Plot Studentized Residuals
plot(jackknife_resid, main = "Studentized Residuals for Log-Transformed Linear Model",
     xlab = "Observation Index", ylab = "Studentized Residuals", pch = 19, col = "blue")
abline(h = c(-2, 2), col = "red", lty = 2)
```



The leverage, DFFITS, and studentized residuals plots help identify influential data points that may have a significant impact on the regression model. In the leverage plot, most observations have low leverage values, indicating they do not disproportionately influence the regression model. However, a few points exceed the suggested threshold (dashed red line), signaling potential high-leverage observations that might require further scrutiny. High-leverage points alone do not necessarily indicate problematic influence, but they suggest observations with unusual predictor values.

The DFFITS plot further assesses influential observations by measuring how much each data point affects the fitted values. Most points remain within the acceptable range, but a few exceed the suggested cutoff, implying they have a notable effect on the model. Similarly, the studentized residuals plot shows that while most residuals fall within the expected +/- 2 range, a handful of points exceed +/- 3, indicating potential outliers. These points warrant further investigation, as they could represent data entry errors, unusual cases, or structural issues in the model. If these influential points unduly affect the model fit, it may be necessary to assess their impact by refitting the model without them or considering robust regression techniques.

Conclusion on Salary Differences

Summarize insights on gender-based salary differences.

The analysis of faculty salaries reveals notable differences in pay distribution across gender, but the extent to which these differences persist after accounting for other factors is more nuanced. The initial exploratory analysis indicated that male faculty members tend to have a higher median salary than female faculty, as observed in the boxplots. This difference is particularly pronounced among senior faculty ranks, where the highest salaries are overwhelmingly concentrated among male professors. The raw salary gap suggests that gender may play a role in faculty compensation. However, simple comparisons do not account for the influence of rank, discipline, or experience—factors that significantly impact earnings in academia.

When a multiple linear regression model was applied, including rank, discipline, and years of experience as explanatory variables, gender alone did not emerge as a statistically significant predictor of salary. The coefficient for sex (Male) was positive, indicating that male faculty earn slightly more on average, but the p-value was not statistically significant. This suggests that while there may be observable salary differences at face value, they can largely be explained by differences in rank distribution and discipline specialization. Additionally, interaction terms between gender and rank were tested to determine whether gender disparities differ across faculty ranks, but these interactions were also found to be insignificant, reinforcing the idea that the primary determinants of salary are related to rank, field, and experience rather than gender alone.

Discuss whether disparities remain after adjusting for rank and discipline.

After adjusting for faculty rank and discipline, the analysis suggests that most of the apparent gender-based salary disparities are attributable to differences in career progression rather than direct gender bias in pay. The regression model results indicate that rank is the most significant predictor of salary, with full professors earning significantly more than assistant or associate professors. Similarly, faculty in applied disciplines (Discipline B) receive a salary premium compared to those in theoretical fields. These two factors alone explain a substantial proportion of the variance in salaries. The findings imply that the observed salary gap between male and female faculty members may be largely driven by gender differences in promotion rates rather than salary discrimination within equivalent roles.

However, while gender was not a statistically significant predictor after controlling for rank and discipline, this does not necessarily mean that gender-based disparities do not exist in academia. One possible explanation is that women may face barriers to promotion, such as biases in tenure and promotion decisions, differences in research funding opportunities, or disproportionate service obligations that limit career advancement. If female faculty members are underrepresented in higher ranks, then the gender pay gap is indirectly perpetuated. This underscores the importance of examining not only salary levels but also structural factors influencing faculty career progression. Future research could investigate whether disparities exist in promotion timelines and tenure decisions to assess the broader implications of gender dynamics in academia.

Consider any limitations and potential improvements.

While this analysis provides valuable insights into salary differences, several limitations must be considered. First, the dataset does not include additional factors that could influence faculty salaries, such as research productivity, publication records, administrative roles, external grant funding, or negotiation strategies. These variables may play a crucial role in salary determination and could help explain remaining variations not captured in the current model. Additionally, the dataset does not account for time trends in salary progression, meaning that gender differences in career growth over time are not explicitly modeled. Longitudinal data tracking faculty salaries across multiple years could provide deeper insights into how gender-based salary trends evolve throughout academic careers.

Another limitation is that while statistical models can identify associations, they do not establish causation. The lack of a significant gender effect in the regression model does not definitively rule out discrimination or systemic biases; rather, it suggests that rank and discipline are stronger explanatory variables. However, if gender disparities exist in promotion rates, grant funding, or workload distribution, these factors would indirectly contribute to salary differences. Future analyses could incorporate promotion rates, faculty evaluations, and external funding sources to assess whether gender disparities arise at earlier career stages, potentially influencing salary trajectories over time. Additionally, alternative modeling techniques, such as machine learning approaches or hierarchical models, could be explored to capture complex interactions that may not be fully accounted for in traditional regression analysis.

Overall, while the statistical evidence suggests that rank and discipline explain most of the observed salary differences, further investigation is necessary to determine whether gender-based barriers exist in faculty promotion and career advancement. A more comprehensive approach that includes additional professional and institutional factors would provide a clearer picture of the underlying dynamics shaping faculty salaries.