Lecture 11 - Class Activity: Multiple Regression

Your Name

# Analysis of Net Primary Production in Forests: A Modern Tidyverse Approach

*Based on Michaletz et al. (2014) data*

## Introduction

This analysis examines the relationships between Net Primary Production (npp) and various climate and forest characteristics across global forest sites. We’ll explore multicollinearity, model selection, and variable transformations.

**Key Learning Objectives:**

* Understand multicollinearity in multiple regression
* Learn model diagnostics and assumption checking
* Practice variable selection techniques
* Apply data transformations appropriately

## Load Required Packages

# Load required packages  
library(tidyverse) # For data manipulation and visualization  
library(car) # For regression diagnostics (VIF, etc.)  
library(corrplot) # For correlation plots  
library(GGally) # For pairs plots  
library(broom) # For tidy model outputs  
library(performance) # For model performance metrics  
library(see) # For better diagnostic plots

## Load and Explore the Data

# Load the forest npp data  
forest\_data <- read\_csv("data/michaletz\_etal\_2014.csv")  
  
# Display top lines  
head(forest\_data)

# A tibble: 6 × 15  
 Index Lat Long NPPTotal\_gm2y Age\_yr BiomassTotal\_gm2 Lgs\_mo MAT\_C MAP\_mm  
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 1243 -13.1 -52.4 2084 104. 18198. 11 25.3 1888   
2 1241 -1.72 -51.4 2234 333. 54523. 12 26.9 2348.  
3 1242 -1.72 -51.4 2714 213. 41358. 12 26.9 2348.  
4 1235 -3.95 -73.4 2828 114. 31557. 12 26.7 2784.  
5 1236 -3.95 -73.4 2882 113. 21417 12 26.7 2784.  
6 672 51.7 113. 774 79 11188 3 -3.53 408.  
# ℹ 6 more variables: LeafType <chr>, TEB\_DD <dbl>, Source <chr>,  
# SourceIndex <chr>, SourcePrimary <chr>,  
# `Methodology for biomass and production` <chr>

## Data Preparation

Following the original analysis, we’ll focus on the key variables and create our working dataframe:

# Create working dataset with key variables  
# Following the original script's variable selection  
forest\_clean <- forest\_data %>%  
 select(  
 npp = NPPTotal\_gm2y, # Net Primary Production (response variable)  
 age = Age\_yr, # Stand age in years  
 biomass = BiomassTotal\_gm2, # Total biomass  
 season = Lgs\_mo, # Length of growing season in months  
 temp = MAT\_C, # Mean annual temperature  
 precip = MAP\_mm, # Mean annual precipitation  
 teb = TEB\_DD, # Thermal environment  
 leaf = LeafType # Leaf type (categorical)  
 ) %>%  
 # Remove any rows with missing values in key variables  
 drop\_na()  
  
forest\_clean <- forest\_clean %>% mutate(leaf = as.factor(leaf))  
  
  
# Display the cleaned dataset  
glimpse(forest\_clean)

Rows: 1,220  
Columns: 8  
$ npp <dbl> 2084, 2234, 2714, 2828, 2882, 774, 1052, 3100, 2254, 785, 1673…  
$ age <dbl> 104.429, 333.268, 213.382, 114.231, 113.477, 79.000, 60.000, 6…  
$ biomass <dbl> 18198.4, 54522.6, 41357.8, 31556.8, 21417.0, 11188.0, 24520.0,…  
$ season <dbl> 11, 12, 12, 12, 12, 3, 6, 12, 12, 6, 6, 5, 5, 5, 5, 5, 5, 5, 5…  
$ temp <dbl> 25.283333, 26.916667, 26.916667, 26.725000, 26.725000, -3.5333…  
$ precip <dbl> 1888.0, 2348.5, 2348.5, 2784.4, 2784.4, 407.8, 1128.7, 1079.3,…  
$ teb <dbl> 0.55, 0.60, 0.60, 1.25, 1.25, 1.80, 2.05, 2.05, 2.05, 2.05, 2.…  
$ leaf <fct> broadleaf, broadleaf, broadleaf, broadleaf, broadleaf, broadle…

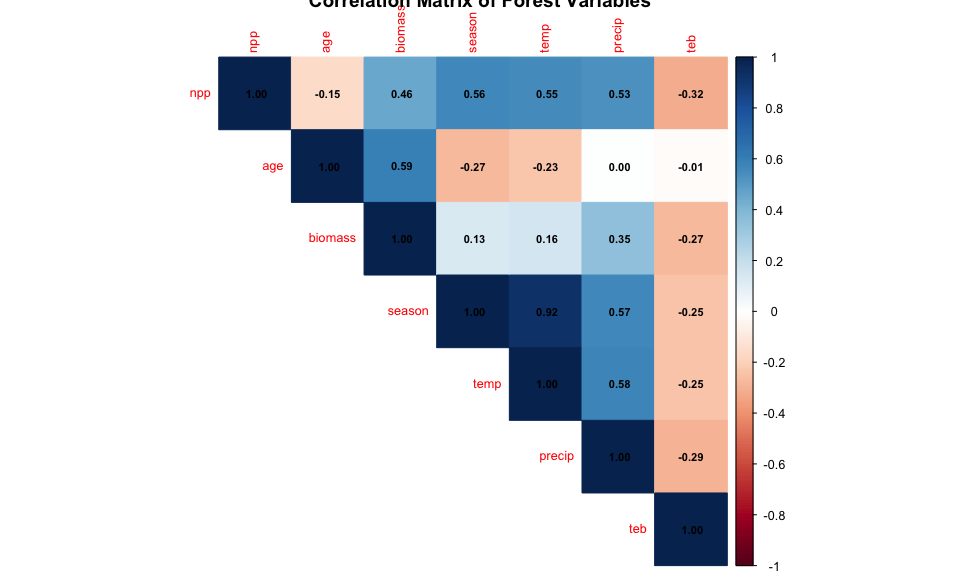
## 1. Initial Exploration: Variable Relationships and Multicollinearity

### Correlation Matrix and Visualization

# Create correlation matrix for numeric variables only  
numeric\_vars <- forest\_clean %>%  
 select\_if(is.numeric)  
  
# Calculate correlation matrix  
cor\_matrix <- cor(numeric\_vars, use = "complete.obs")  
  
# Display correlation matrix  
print(round(cor\_matrix, 3))

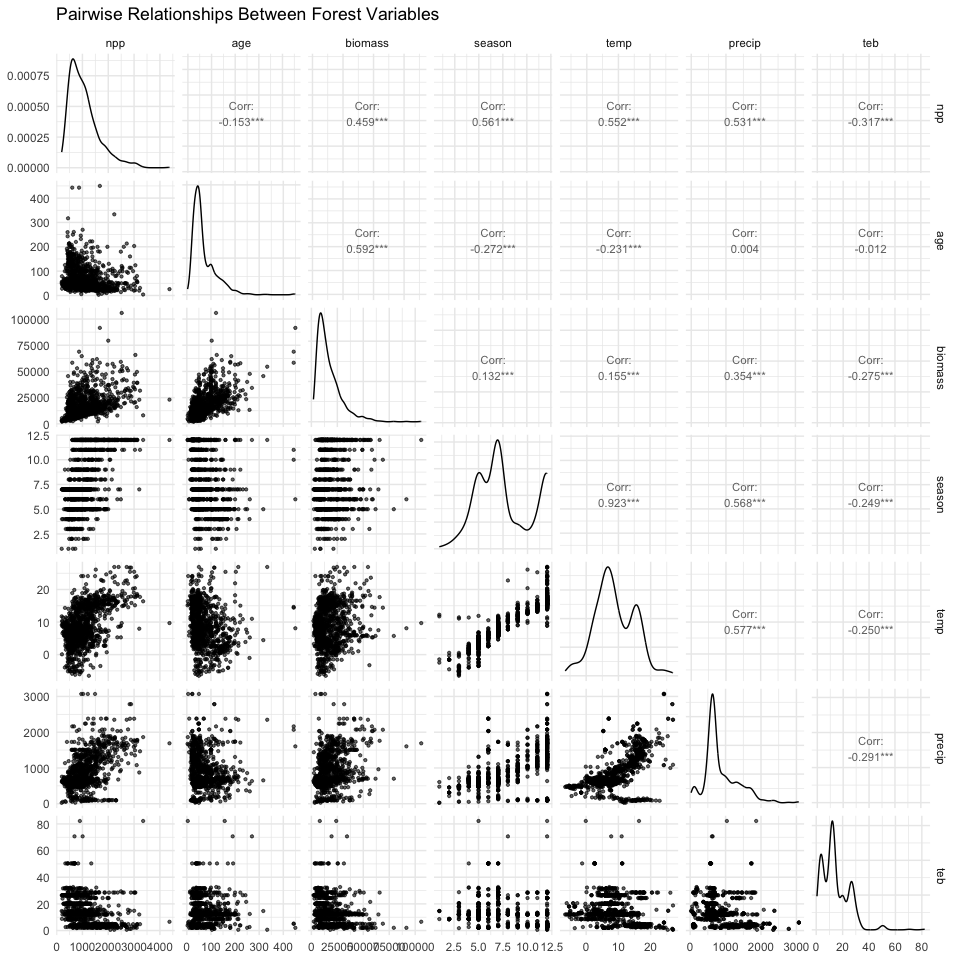
npp age biomass season temp precip teb  
npp 1.000 -0.153 0.459 0.561 0.552 0.531 -0.317  
age -0.153 1.000 0.592 -0.272 -0.231 0.004 -0.012  
biomass 0.459 0.592 1.000 0.132 0.155 0.354 -0.275  
season 0.561 -0.272 0.132 1.000 0.923 0.568 -0.249  
temp 0.552 -0.231 0.155 0.923 1.000 0.577 -0.250  
precip 0.531 0.004 0.354 0.568 0.577 1.000 -0.291  
teb -0.317 -0.012 -0.275 -0.249 -0.250 -0.291 1.000

# Create a visual correlation plot  
corrplot(cor\_matrix, method = "color", type = "upper",   
 addCoef.col = "black", tl.cex = 0.8, number.cex = 0.7,  
 title = "Correlation Matrix of Forest Variables")



### Pairs Plot for Visual Inspection

# Create pairs plot to visualize relationships  
# This replaces the original pairs() function with ggplot2  
forest\_clean %>%  
 select(-leaf) %>% # Exclude categorical variable for pairs plot  
 ggpairs(  
 title = "Pairwise Relationships Between Forest Variables",  
 upper = list(continuous = wrap("cor", size = 3)),  
 lower = list(continuous = wrap("points", alpha = 0.6, size = 0.8))  
 ) +  
 theme\_minimal()



## 2. Initial Multiple Regression Model

Let’s start with a full model including all predictors:

# Fit initial model with all predictors (Model 1)  
model1 <- lm(npp ~ age + biomass + season + temp +   
 precip + teb + leaf, data = forest\_clean)  
  
# Get model summary  
summary(model1)

Call:  
lm(formula = npp ~ age + biomass + season + temp + precip + teb +   
 leaf, data = forest\_clean)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-1331.10 -206.27 -34.09 166.94 2760.41   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 583.92959 53.81169 10.851 < 2e-16 \*\*\*  
age -4.78822 0.28518 -16.790 < 2e-16 \*\*\*  
biomass 0.03154 0.00122 25.848 < 2e-16 \*\*\*  
season 41.18220 9.49073 4.339 1.55e-05 \*\*\*  
temp 4.61281 4.37372 1.055 0.291788   
precip 0.09674 0.02852 3.392 0.000716 \*\*\*  
teb -2.12880 1.08408 -1.964 0.049794 \*   
leafneedle -267.00569 22.43078 -11.904 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 351.7 on 1212 degrees of freedom  
Multiple R-squared: 0.6415, Adjusted R-squared: 0.6394   
F-statistic: 309.8 on 7 and 1212 DF, p-value: < 2.2e-16

### Check for Multicollinearity

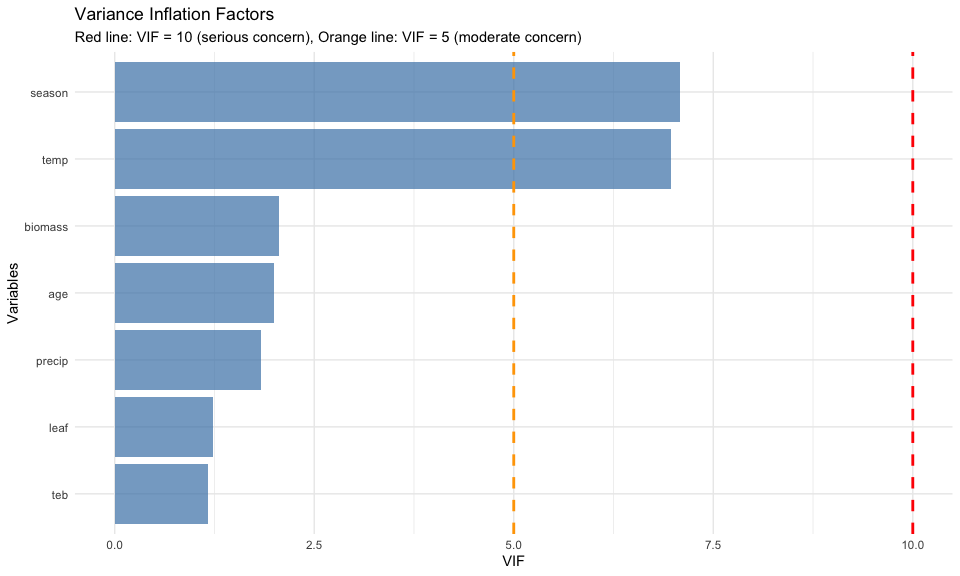
# Calculate Variance Inflation Factors (VIF)  
vif\_values <- vif(model1)  
print("Variance Inflation Factors:")

[1] "Variance Inflation Factors:"

print(vif\_values)

age biomass season temp precip teb leaf   
1.993348 2.061766 7.079004 6.972147 1.831127 1.167007 1.236588

# Create a data frame for better visualization  
vif\_df <- data.frame(  
 Variable = names(vif\_values),  
 VIF = as.numeric(vif\_values)  
) %>%  
 arrange(desc(VIF))  
  
# Visualize VIF values  
ggplot(vif\_df, aes(x = reorder(Variable, VIF), y = VIF)) +  
 geom\_col(fill = "steelblue", alpha = 0.7) +  
 geom\_hline(yintercept = 10, color = "red", linetype = "dashed",   
 linewidth = 1) +  
 geom\_hline(yintercept = 5, color = "orange", linetype = "dashed",   
 linewidth = 1) +  
 coord\_flip() +  
 labs(  
 title = "Variance Inflation Factors",  
 subtitle = "Red line: VIF = 10 (serious concern), Orange line: VIF = 5 (moderate concern)",  
 x = "Variables",  
 y = "VIF"  
 ) +  
 theme\_minimal()



### Address Multicollinearity by Removing Growing Season

Based on the original analysis, season and Temperature are highly correlated. Let’s remove season:

# Model 2: Remove season due to multicollinearity  
model2 <- lm(npp ~ age + biomass + temp + precip + teb + leaf,   
 data = forest\_clean)  
  
# Model summary  
summary(model2)

Call:  
lm(formula = npp ~ age + biomass + temp + precip + teb + leaf,   
 data = forest\_clean)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-1309.94 -209.40 -42.65 167.71 2898.47   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 7.511e+02 3.785e+01 19.845 < 2e-16 \*\*\*  
age -5.004e+00 2.829e-01 -17.690 < 2e-16 \*\*\*  
biomass 3.180e-02 1.228e-03 25.905 < 2e-16 \*\*\*  
temp 2.112e+01 2.173e+00 9.718 < 2e-16 \*\*\*  
precip 1.120e-01 2.851e-02 3.929 9.02e-05 \*\*\*  
teb -2.255e+00 1.092e+00 -2.066 0.039 \*   
leafneedle -2.634e+02 2.258e+01 -11.666 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 354.3 on 1213 degrees of freedom  
Multiple R-squared: 0.6359, Adjusted R-squared: 0.6341   
F-statistic: 353.1 on 6 and 1213 DF, p-value: < 2.2e-16

# Check VIF again  
vif\_values2 <- vif(model2)  
print("VIF values after removing season:")

[1] "VIF values after removing season:"

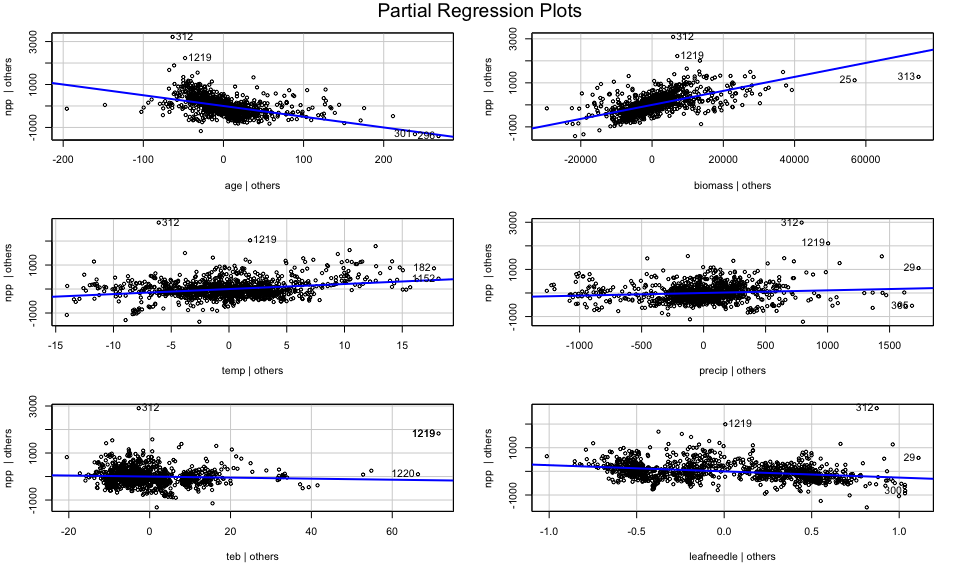
print(vif\_values2)

age biomass temp precip teb leaf   
1.932806 2.056724 1.696787 1.803266 1.166162 1.234906

## 3. Exploring Variable Transformations

### Check the Shape of Relationships with Partial Regression Plots

# Create partial regression plots (Added Variable Plots)  
# This helps us see the relationship between each predictor and response  
# after accounting for other variables  
  
par(mfrow = c(2, 3))  
avPlots(model2, main = "Partial Regression Plots")



par(mfrow = c(1, 1))

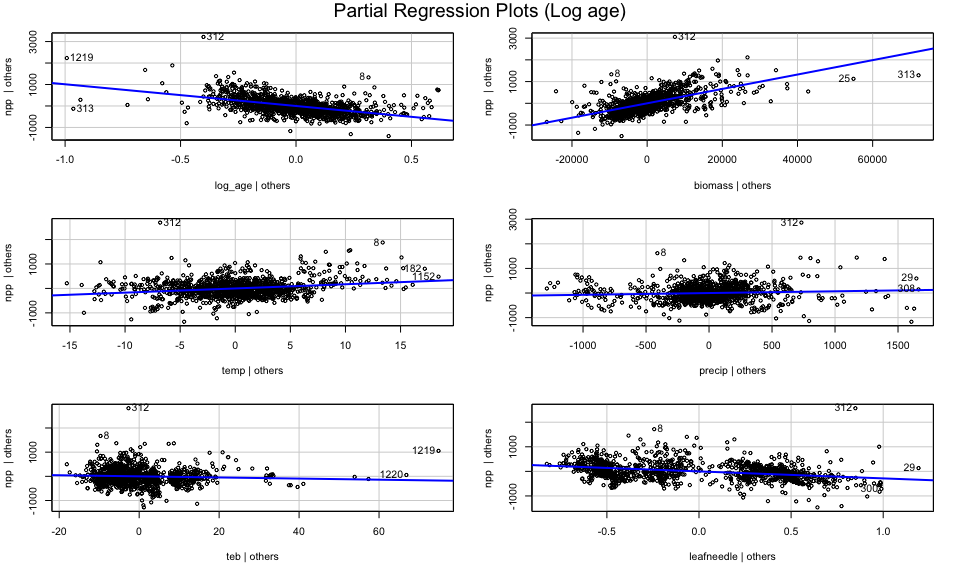
### Apply Log Transformation to age

The original analysis found that age showed a curvy relationship. Let’s try log transformation:

# Create dataset with log-transformed age  
forest\_transformed <- forest\_clean %>%  
 mutate(log\_age = log10(age))  
  
# Model 3: With log-transformed age  
model3 <- lm(npp ~ log\_age + biomass + temp + precip + teb + leaf,   
 data = forest\_transformed)  
  
  
# Model summary  
summary(model3)

Call:  
lm(formula = npp ~ log\_age + biomass + temp + precip + teb +   
 leaf, data = forest\_transformed)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-1282.97 -203.31 -23.13 163.75 2810.76   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 2.237e+03 8.961e+01 24.962 < 2e-16 \*\*\*  
log\_age -1.012e+03 5.064e+01 -19.973 < 2e-16 \*\*\*  
biomass 3.313e-02 1.194e-03 27.755 < 2e-16 \*\*\*  
temp 1.723e+01 2.159e+00 7.982 3.33e-15 \*\*\*  
precip 7.194e-02 2.781e-02 2.587 0.00981 \*\*   
teb -2.315e+00 1.061e+00 -2.183 0.02925 \*   
leafneedle -2.852e+02 2.177e+01 -13.105 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 344.7 on 1213 degrees of freedom  
Multiple R-squared: 0.6553, Adjusted R-squared: 0.6536   
F-statistic: 384.4 on 6 and 1213 DF, p-value: < 2.2e-16

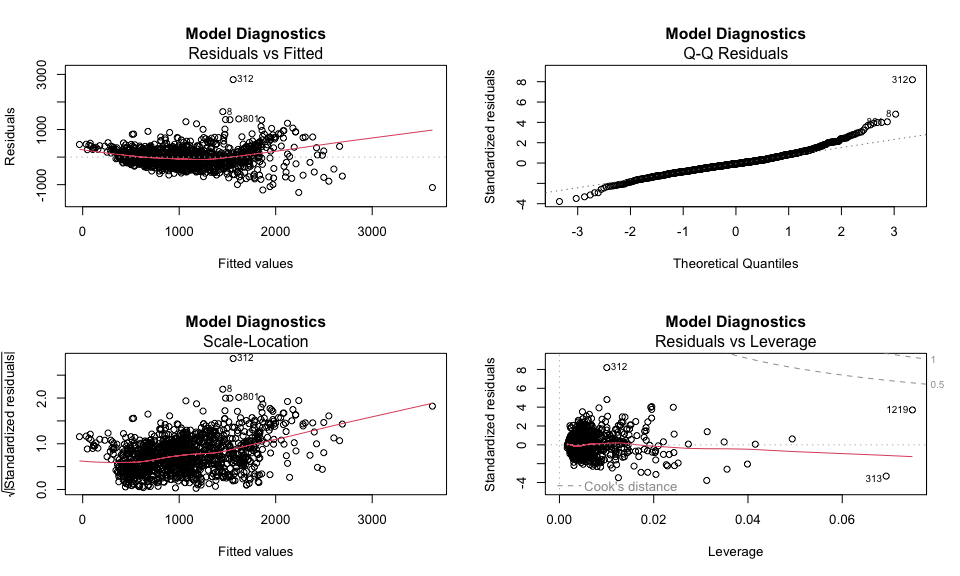
# Compare partial regression plots  
par(mfrow = c(2, 3))  
avPlots(model3, main = "Partial Regression Plots (Log age)")



par(mfrow = c(1, 1))

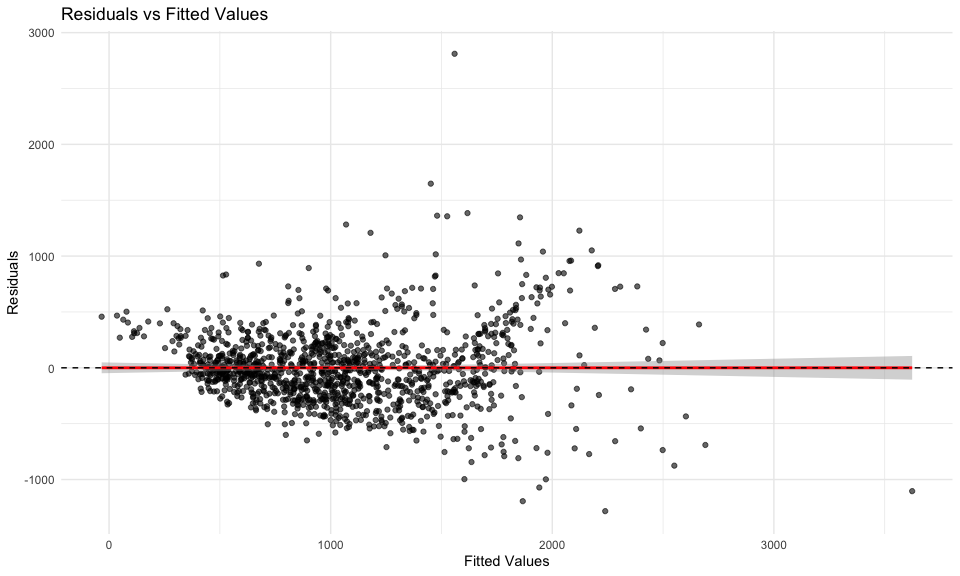
## 4. Model Diagnostics and Assumption Checking

# Create diagnostic plots  
par(mfrow = c(2, 2))  
plot(model3, main = "Model Diagnostics")



par(mfrow = c(1, 1))

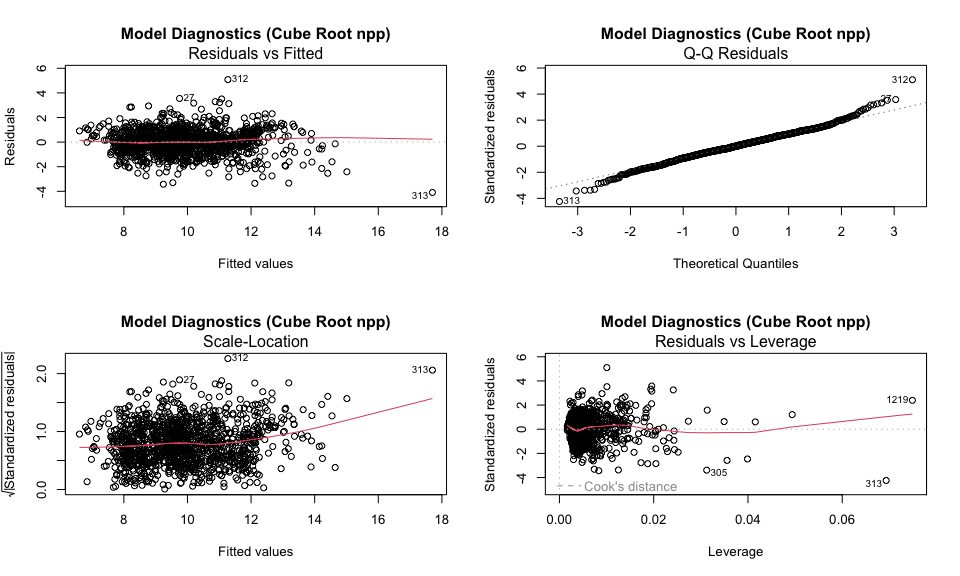
# Check for normality of residuals  
residuals\_data <- data.frame(  
 Fitted = fitted(model3),  
 Residuals = residuals(model3),  
 Standardized\_Residuals = rstandard(model3)  
)  
  
# Residuals vs Fitted plot using ggplot  
ggplot(residuals\_data, aes(x = Fitted, y = Residuals)) +  
 geom\_point(alpha = 0.6) +  
 geom\_smooth(method = "lm", color = "red") +  
 geom\_hline(yintercept = 0, linetype = "dashed") +  
 labs(  
 title = "Residuals vs Fitted Values",  
 x = "Fitted Values",  
 y = "Residuals"  
 ) +  
 theme\_minimal()



### Try Response Variable Transformation

Following the original analysis, let’s try a cube root transformation of npp:

# Model 4: Cube root transformation of npp  
forest\_transformed <- forest\_transformed %>%  
 mutate(npp\_cuberoot = npp^(1/3))  
  
model4 <- lm(npp\_cuberoot ~ log\_age + biomass + temp + precip +   
 teb + leaf, data = forest\_transformed)  
  
# Check diagnostics  
par(mfrow = c(2, 2))  
plot(model4, main = "Model Diagnostics (Cube Root npp)")



par(mfrow = c(1, 1))  
  
# Model summary  
tidy(model4)

# A tibble: 7 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 13.9 0.260 53.5 2.06e-321  
2 log\_age -3.15 0.147 -21.4 1.07e- 86  
3 biomass 0.000103 0.00000346 29.8 8.32e-147  
4 temp 0.0451 0.00627 7.19 1.10e- 12  
5 precip 0.0000636 0.0000807 0.788 4.31e- 1  
6 teb -0.0127 0.00308 -4.13 3.79e- 5  
7 leafneedle -1.03 0.0632 -16.3 4.57e- 54

glance(model4)

# A tibble: 1 × 12  
 r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC  
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 0.679 0.678 1.00 428. 2.43e-295 6 -1728. 3472. 3513.  
# ℹ 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

## 5. Model Simplification and Comparison

### Remove Non-significant Variables

# Model 5: Remove non-significant Precipitation  
model5 <- lm(npp\_cuberoot ~ log\_age + biomass + temp + teb + leaf,   
 data = forest\_transformed)  
  
# Compare models using AIC  
model\_comparison <- data.frame(  
 Model = c("Model 4 (Full)", "Model 5 (No precip)"),  
 AIC = c(AIC(model4), AIC(model5)),  
 R\_squared = c(summary(model4)$r.squared, summary(model5)$r.squared),  
 Adj\_R\_squared = c(summary(model4)$adj.r.squared, summary(model5)$adj.r.squared)  
)  
  
print(model\_comparison)

Model AIC R\_squared Adj\_R\_squared  
1 Model 4 (Full) 3472.092 0.6792946 0.6777082  
2 Model 5 (No precip) 3470.717 0.6791304 0.6778089

# Final model summary  
cat("\n=== FINAL MODEL SUMMARY ===\n")

=== FINAL MODEL SUMMARY ===

summary(model5)

Call:  
lm(formula = npp\_cuberoot ~ log\_age + biomass + temp + teb +   
 leaf, data = forest\_transformed)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-4.1085 -0.5919 0.0019 0.6459 5.1261   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.396e+01 2.532e-01 55.120 < 2e-16 \*\*\*  
log\_age -3.159e+00 1.465e-01 -21.559 < 2e-16 \*\*\*  
biomass 1.040e-04 3.306e-06 31.465 < 2e-16 \*\*\*  
temp 4.719e-02 5.658e-03 8.342 < 2e-16 \*\*\*  
teb -1.296e-02 3.064e-03 -4.231 2.5e-05 \*\*\*  
leafneedle -1.040e+00 6.150e-02 -16.910 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 1 on 1214 degrees of freedom  
Multiple R-squared: 0.6791, Adjusted R-squared: 0.6778   
F-statistic: 513.9 on 5 and 1214 DF, p-value: < 2.2e-16

### Model Performance and Interpretation

# Get tidy summary of final model  
final\_summary <- tidy(model5, conf.int = TRUE)  
  
# Get tidy summary of final model  
summary(model5, conf.int = TRUE)

Call:  
lm(formula = npp\_cuberoot ~ log\_age + biomass + temp + teb +   
 leaf, data = forest\_transformed)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-4.1085 -0.5919 0.0019 0.6459 5.1261   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.396e+01 2.532e-01 55.120 < 2e-16 \*\*\*  
log\_age -3.159e+00 1.465e-01 -21.559 < 2e-16 \*\*\*  
biomass 1.040e-04 3.306e-06 31.465 < 2e-16 \*\*\*  
temp 4.719e-02 5.658e-03 8.342 < 2e-16 \*\*\*  
teb -1.296e-02 3.064e-03 -4.231 2.5e-05 \*\*\*  
leafneedle -1.040e+00 6.150e-02 -16.910 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 1 on 1214 degrees of freedom  
Multiple R-squared: 0.6791, Adjusted R-squared: 0.6778   
F-statistic: 513.9 on 5 and 1214 DF, p-value: < 2.2e-16

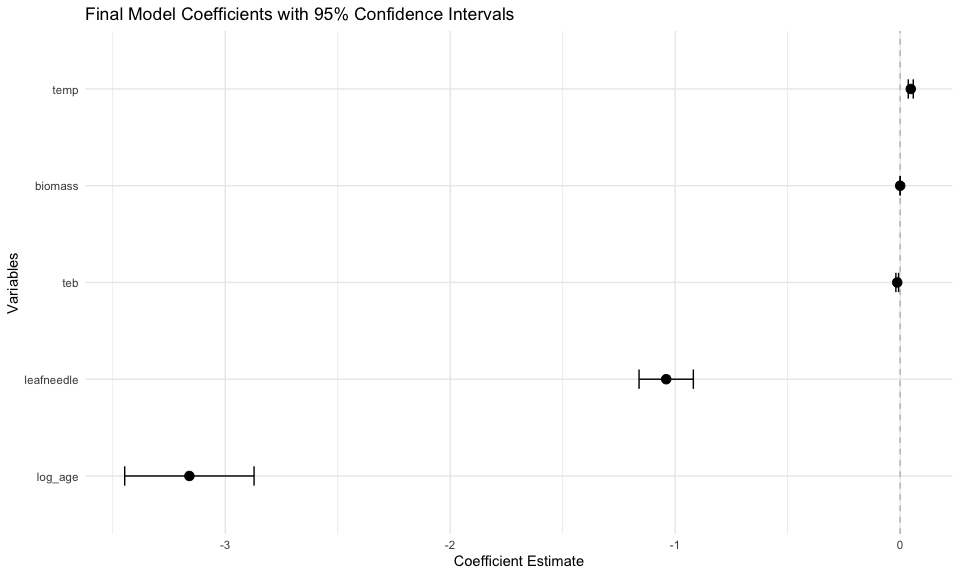
# Calculate partial R-squared for each variable  
# Using the sensemakr package equivalent or manual calculation  
anova\_table <- anova(model5)  
print("ANOVA Table:")

[1] "ANOVA Table:"

print(anova\_table)

Analysis of Variance Table  
  
Response: npp\_cuberoot  
 Df Sum Sq Mean Sq F value Pr(>F)   
log\_age 1 193.32 193.32 193.239 < 2.2e-16 \*\*\*  
biomass 1 1886.95 1886.95 1886.142 < 2.2e-16 \*\*\*  
temp 1 179.34 179.34 179.264 < 2.2e-16 \*\*\*  
teb 1 24.88 24.88 24.866 7.038e-07 \*\*\*  
leaf 1 286.08 286.08 285.957 < 2.2e-16 \*\*\*  
Residuals 1214 1214.52 1.00   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Create coefficient plot  
final\_summary %>%  
 filter(term != "(Intercept)") %>%  
 mutate(term = fct\_reorder(term, estimate)) %>%  
 ggplot(aes(x = estimate, y = term)) +  
 geom\_vline(xintercept = 0, linetype = "dashed", color = "gray") +  
 geom\_point(size = 3) +  
 geom\_errorbarh(aes(xmin = conf.low, xmax = conf.high), height = 0.2) +  
 labs(  
 title = "Final Model Coefficients with 95% Confidence Intervals",  
 x = "Coefficient Estimate",  
 y = "Variables"  
 ) +  
 theme\_minimal()



## 6. Alternative Approach: Standardized Variables

Following the original analysis, let’s also try the standardized approach:

# Create standardized variables  
forest\_standardized <- forest\_clean %>%  
 mutate(  
 npp\_sqrt\_scaled = scale(sqrt(npp))[,1],  
 log\_age\_scaled = scale(log10(age))[,1],  
 biomass\_scaled = scale(biomass)[,1],  
 temp\_scaled = scale(temp)[,1],  
 precip\_scaled = scale(precip)[,1],  
 teb\_scaled = scale(teb)[,1]  
 )  
  
# Standardized model  
model\_std <- lm(npp\_sqrt\_scaled ~ log\_age\_scaled + biomass\_scaled +   
 temp\_scaled \* precip\_scaled + teb\_scaled,   
 data = forest\_standardized)  
  
# Summary  
summary(model\_std)

Call:  
lm(formula = npp\_sqrt\_scaled ~ log\_age\_scaled + biomass\_scaled +   
 temp\_scaled \* precip\_scaled + teb\_scaled, data = forest\_standardized)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-3.14438 -0.38836 -0.03068 0.39178 3.01216   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.064952 0.020075 -3.235 0.001247 \*\*   
log\_age\_scaled -0.543699 0.024500 -22.192 < 2e-16 \*\*\*  
biomass\_scaled 0.691415 0.025313 27.315 < 2e-16 \*\*\*  
temp\_scaled 0.215693 0.023304 9.256 < 2e-16 \*\*\*  
precip\_scaled -0.008188 0.028714 -0.285 0.775563   
teb\_scaled -0.071775 0.018879 -3.802 0.000151 \*\*\*  
temp\_scaled:precip\_scaled 0.112727 0.017068 6.605 5.94e-11 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.6113 on 1213 degrees of freedom  
Multiple R-squared: 0.6281, Adjusted R-squared: 0.6263   
F-statistic: 341.5 on 6 and 1213 DF, p-value: < 2.2e-16

## 7. Key Findings and Conclusions

# Create a summary of key findings  
cat("=== KEY FINDINGS ===\n\n")

=== KEY FINDINGS ===

cat("1. MULTICOLLINEARITY:\n")

1. MULTICOLLINEARITY:

cat(" - Growing season length and temperature were highly correlated\n")

- Growing season length and temperature were highly correlated

cat(" - Removed growing season to address multicollinearity\n\n")

- Removed growing season to address multicollinearity

cat("2. VARIABLE TRANSFORMATIONS:\n")

2. VARIABLE TRANSFORMATIONS:

cat(" - Log transformation of age improved model fit\n")

- Log transformation of age improved model fit

cat(" - Cube root transformation of npp addressed assumption violations\n\n")

- Cube root transformation of npp addressed assumption violations

cat("3. FINAL MODEL RESULTS:\n")

3. FINAL MODEL RESULTS:

final\_r2 <- summary(model5)$r.squared  
cat(" - R-squared:", round(final\_r2, 3), "\n")

- R-squared: 0.679

cat(" - Significant predictors: age (negative), biomass (positive), temp (positive)\n")

- Significant predictors: age (negative), biomass (positive), temp (positive)

cat(" - teb had negative effect, Leaf type differences were significant\n\n")

- teb had negative effect, Leaf type differences were significant

cat("4. BIOLOGICAL INTERPRETATION:\n")

4. BIOLOGICAL INTERPRETATION:

cat(" - Younger stands had higher npp (for given biomass)\n")

- Younger stands had higher npp (for given biomass)

cat(" - Higher biomass associated with higher npp\n")

- Higher biomass associated with higher npp

cat(" - temp positively related to npp\n")

- temp positively related to npp

cat(" - Coniferous forests had lower npp than broadleaf forests\n")

- Coniferous forests had lower npp than broadleaf forests

## References and Additional Notes

This analysis is based on:

* **Michaletz, S.T., Cheng, D., Kerkhoff, A.J. & Enquist, B.J.** (2014). Convergence of terrestrial plant production across global climate gradients. *Nature*, 512, 39-43.

**Key Learning Points:** 1. **Multicollinearity Detection**: Use VIF values and correlation matrices 2. **Variable Transformations**: Log and power transformations can improve model fit 3. **Model Diagnostics**: Always check residual plots and assumption violations 4. **Model Comparison**: Use AIC and other criteria for model selection 5. **Interpretation**: Focus on biologically meaningful relationships

**Modern R Practices Used:** - tidyverse for data manipulation and visualization - broom for tidy model outputs - car for regression diagnostics - ggplot2 for modern visualizations - Pipe operators (%>%) for readable code - Tidy data principles throughout

# Session information for reproducibility  
sessionInfo()

R version 4.5.0 (2025-04-11)  
Platform: aarch64-apple-darwin20  
Running under: macOS Sequoia 15.5  
  
Matrix products: default  
BLAS: /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRblas.0.dylib   
LAPACK: /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRlapack.dylib; LAPACK version 3.12.1  
  
locale:  
[1] en\_US.UTF-8/en\_US.UTF-8/en\_US.UTF-8/C/en\_US.UTF-8/en\_US.UTF-8  
  
time zone: America/Chicago  
tzcode source: internal  
  
attached base packages:  
[1] stats graphics grDevices utils datasets methods base   
  
other attached packages:  
 [1] see\_0.11.0 performance\_0.14.0 broom\_1.0.8 GGally\_2.2.1   
 [5] corrplot\_0.95 car\_3.1-3 carData\_3.0-5 lubridate\_1.9.4   
 [9] forcats\_1.0.0 stringr\_1.5.1 dplyr\_1.1.4 purrr\_1.0.4   
[13] readr\_2.1.5 tidyr\_1.3.1 tibble\_3.2.1 ggplot2\_3.5.2   
[17] tidyverse\_2.0.0   
  
loaded via a namespace (and not attached):  
 [1] gtable\_0.3.6 xfun\_0.52 insight\_1.3.0 lattice\_0.22-7   
 [5] tzdb\_0.5.0 vctrs\_0.6.5 tools\_4.5.0 generics\_0.1.4   
 [9] parallel\_4.5.0 pkgconfig\_2.0.3 Matrix\_1.7-3 RColorBrewer\_1.1-3  
[13] lifecycle\_1.0.4 compiler\_4.5.0 farver\_2.1.2 codetools\_0.2-20   
[17] htmltools\_0.5.8.1 yaml\_2.3.10 Formula\_1.2-5 pillar\_1.10.2   
[21] crayon\_1.5.3 abind\_1.4-8 nlme\_3.1-168 ggstats\_0.9.0   
[25] tidyselect\_1.2.1 digest\_0.6.37 stringi\_1.8.7 labeling\_0.4.3   
[29] splines\_4.5.0 fastmap\_1.2.0 grid\_4.5.0 cli\_3.6.5   
[33] magrittr\_2.0.3 utf8\_1.2.5 withr\_3.0.2 scales\_1.4.0   
[37] backports\_1.5.0 bit64\_4.6.0-1 timechange\_0.3.0 rmarkdown\_2.29   
[41] bit\_4.6.0 hms\_1.1.3 evaluate\_1.0.3 knitr\_1.50   
[45] mgcv\_1.9-3 rlang\_1.1.6 Rcpp\_1.0.14 glue\_1.8.0   
[49] rstudioapi\_0.17.1 vroom\_1.6.5 jsonlite\_2.0.0 R6\_2.6.1   
[53] plyr\_1.8.9