Linear Regression: quantitative description of the relationship between percentage_biomass_change_from_baseline and years_from_baseline

Mariano Viz

2024-10-14

Contents


```
# Read in Data
ma <- read_excel(here("data", "raw","Aggregated Data.xlsx"), sheet = "MA - Biomass by year")
reserve <- read_excel(here("data", "raw","Aggregated Data.xlsx"), sheet = "Reserve - Biomass by year")
```

Summary

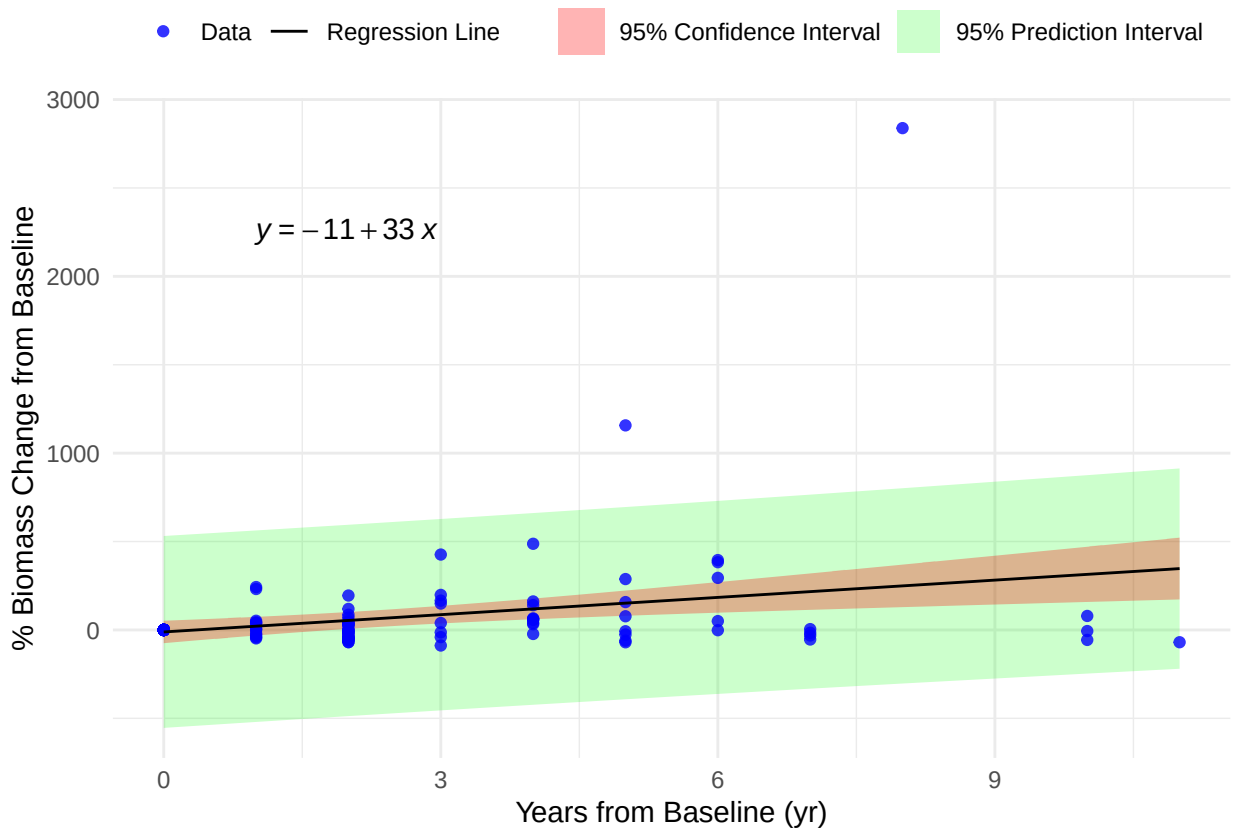
Managed Access Areas

1. MA (All Countries)

```
# Linear regression:
ma_all_lm <- lm(percentage_biomass_change_from_baseline ~ years_from_baseline, data = ma)
# Create a data frame with predictions and prediction intervals:
new_data <- ma %>%
  mutate(pred = predict(ma_all_lm, newdata = ma),
         pred_interval = predict(ma_all_lm, newdata = ma, interval = "prediction"))
# Separate out the lower and upper bounds of the prediction interval:
new_data <- new_data %>%
  mutate(lower_bound = pred_interval[, "lwr"],
         upper_bound = pred_interval[, "upr"])

# Plot + prediction interval:
ggplot(data = new_data, aes(x = years_from_baseline, y = percentage_biomass_change_from_baseline)) +
  geom_ribbon(aes(ymin = lower_bound, ymax = upper_bound, fill = "95% Prediction Interval"), alpha = 0.1
  geom_smooth(method = "lm", aes(fill = "95% Confidence Interval"), color = NA, size = 0.5, alpha = 0.3)
  geom_point(aes(color = "Data"), alpha = 0.8) +
  geom_smooth(method = "lm", aes(color = "Regression Line"), size = 0.5, se = FALSE) +
  theme_minimal() +
  ggpubr::stat_regline_equation(label.x = 1, label.y = 2250) +
  labs(x = "Years from Baseline (yr)",
       y = "% Biomass Change from Baseline") +
  scale_color_manual(name = NULL,
                     values = c("Data" = "blue", "Regression Line" = "black")) +
  scale_fill_manual(name = NULL,
                    values = c("95% Confidence Interval" = "red", "95% Prediction Interval" = "green"))
  guides(fill = guide_legend(override.aes = list(color = NA)),
         color = guide_legend(override.aes = list(fill = NA))) +
  theme(legend.position = "top")
```

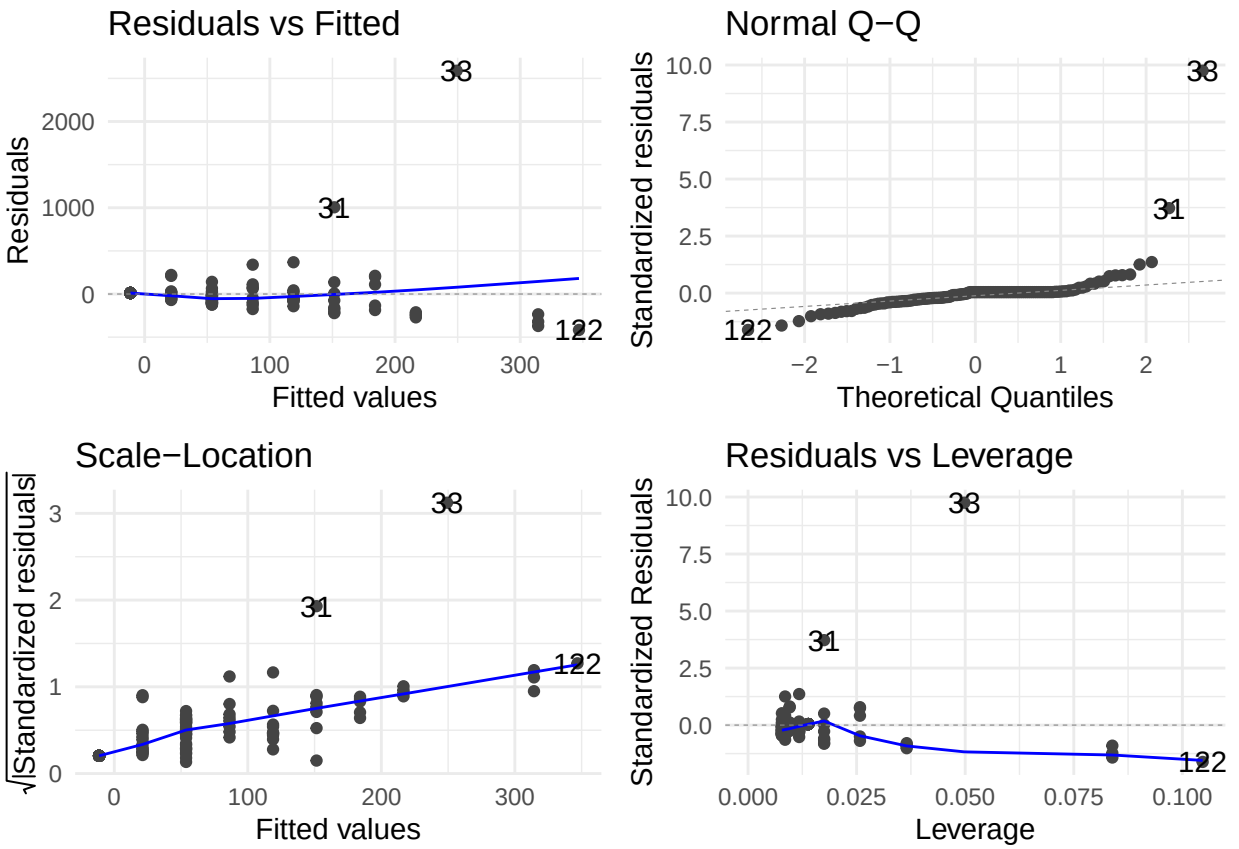
1.1. Linear Regression Analysis




```
# Linear regression:
ma_all_lm <- lm(percentage_biomass_change_from_baseline ~ years_from_baseline, data = ma)

#Model diagnostics:
# plot(ma_all_lm)
autoplot(ma_all_lm) +
  theme_minimal()
```

1.2. Diagnostic Plots



Residuals vs Fitted Plot (Top Left): The residuals should be randomly scattered around the horizontal line (residual = 0) if the linear model is appropriate. Here, residuals show some non-random patterns, and there seems to be a slight curvature, which might indicate non-linearity. Also, the spread of the residuals increases slightly as fitted values increase, suggesting heteroscedasticity (non-constant variance). *Normal*

Q-Q Plot (Top Right): The points should lie along the diagonal line if the residuals are normally distributed. In this plot, there is some deviation from the line at the tails, especially for points labeled 31 and 38, indicating that the residuals are not perfectly normally distributed. These points might be outliers.

Scale-Location Plot (Bottom Left): Ideally, the blue line should be horizontal, and the points should be randomly scattered around it. In this plot, the spread of points increases with the fitted values, as indicated by the upward-sloping line, suggesting heteroscedasticity. This confirms the impression from the Residuals vs Fitted plot.

Residuals vs Leverage Plot (Bottom Right): Points like 31, 38, and 122 seem to be influential outliers that are having a strong influence on the model. It might be worth investigating these points further to see if they are valid observations or if they need to be handled differently.

Summary:

- **Non-linearity** and **heteroscedasticity** are both potential issues, as indicated by the Residuals vs Fitted and Scale-Location plots.
- There may be some **outliers** or influential points (31, 38, and 122) that could be affecting the model.
- **Non-normality** in the residuals is visible, particularly at the tails, which could indicate the need for data transformation or a different regression model.

```r
# Tidy versions of the model output for in-line referencing:
ma_all_lm_tidy <- tidy(ma_all_lm)
ma_all_lm_glance <- glance(ma_all_lm)
#R^2 * 100 (for in-referencing as a percent):
ma_all_rsq_perc <- ma_all_lm_glance$r.squared*100
# Pearson's r correlation:
ma_all_cor <- cor.test(ma$percentage_biomass_change_from_baseline, ma$years_from_baseline)
# Tidy versions of correlation output for in-line referencing:
ma_all_cor_tidy <- tidy(ma_all_cor)


# Tables:
# Get the summary of the model
ma_all_lm_summary <- summary(ma_all_lm)
# Extract coefficients
coef_table <- as.data.frame(ma_all_lm_summary$coefficients)
# Extract R-squared and Adjusted R-squared
r_squared <- ma_all_lm_summary$r.squared
adj_r_squared <- ma_all_lm_summary$adj.r.squared
# Extract F-statistic and p-value
f_statistic <- ma_all_lm_summary$fstatistic[1]
p_value <- pf(f_statistic, ma_all_lm_summary$fstatistic[2], ma_all_lm_summary$fstatistic[3], lower.tail


# Display coefficients in a neat table
kable(coef_table, caption = "Coefficients from the Linear Model")
```

1.3. Model Summary

Table 1: Coefficients from the Linear Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-11.38445	32.087120	-0.354798	0.7233294
years_from_baseline	32.57084	9.647474	3.376101	0.0009759

```r
# Create a summary statistics table
model_summary <- data.frame(
  Statistic = c("R-squared", "Adjusted R-squared", "F-statistic", "p-value"),
  Value = c(r_squared, adj_r_squared, f_statistic, p_value)
)
# Display the model summary
kable(model_summary, caption = "Model Summary Statistics")
```

Table 2: Model Summary Statistics

Statistic	Value
R-squared	0.0823571
Adjusted R-squared	0.0751315
F-statistic	11.3980556
p-value	0.0009759

Coefficients:

- **Intercept**: The intercept is -11.384, but its p-value is quite high (p = 0.723329), which indicates that it is not statistically significant. This suggests that the intercept does not contribute meaningfully to the model.
- `years_from_baseline`: The coefficient for years_from_baseline is 32.571, and its p-value is highly significant (p = 0.000976). This means that for every additional year from the baseline, the percentage change in biomass is expected to increase by approximately 32.57%. This effect is statistically significant at the 0.001 level.

Residuals:

- The residuals show some variability, ranging from -416.51 to 2589.25, with a median value of 11.38. The relatively high maximum residual indicates potential outliers, as seen from the residual diagnostics plots.

Model Fit:

- **Residual standard error**: The residual standard error is 272.4 (deviating significantly from the fitted line), suggesting that the model is not fitting the data very well
- **Multiple R-squared**: The R-squared value is 0.08236, indicating that about 8.2% of the variance in the percentage biomass change can be explained by the years_from_baseline. This is relatively low, suggesting that the model does not explain a large portion of the variability in the biomass change.
- **Adjusted R-squared**: The adjusted R-squared is 0.07513, which is similar to the multiple R-squared and accounts for the number of predictors in the model. This confirms that the model's explanatory power is quite limited.

F-statistic:

- The F-statistic is 11.4 with a p-value of 0.0009759, which is highly significant. This indicates that the overall model is statistically significant, meaning that at least one of the coefficients (`years_from_baseline`) is significantly different from zero.

Summary:

- The `years_from_baseline` variable has a **statistically significant positive effect** on the percentage biomass change from baseline, with a 32.57% increase for every additional year.
- However, **the overall model fit is relatively weak**, as indicated by the low R-squared values. This suggests that while `years_from_baseline` is a significant predictor, there are other factors affecting biomass change that are not included in the model.
- The residuals show high variability, suggesting that there may be **outliers** or **heteroscedasticity**, as we observed in the diagnostic plots.

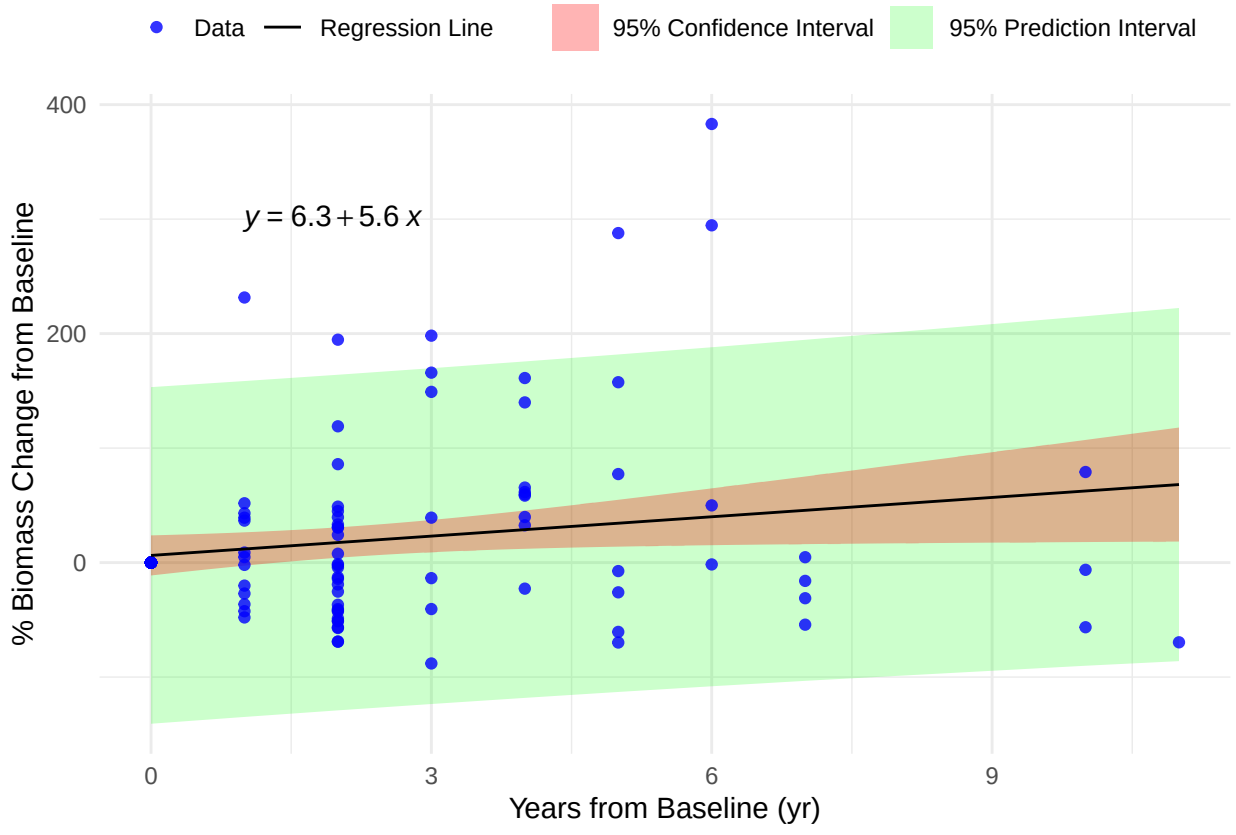
2. MA (All Countries - except Culasi)

```
# Linear regression:
ma_cul <- ma %>%
  filter(ma_name != "Culasi")
ma_cul_lm <- lm(percentage_biomass_change_from_baseline ~ years_from_baseline, data = ma_cul)
# Create a data frame with predictions and prediction intervals:
new_data_cul <- ma_cul %>%
  mutate(pred = predict(ma_cul_lm, newdata = ma_cul),
         pred_interval = predict(ma_cul_lm, newdata = ma_cul, interval = "prediction"))
# Separate out the lower and upper bounds of the prediction interval:
```

```r
new_data_cul <- new_data_cul %>%
  mutate(lower_bound = pred_interval[, "lwr"],
         upper_bound = pred_interval[, "upr"])

# Plot + prediction interval:
ggplot(data = new_data_cul, aes(x = years_from_baseline, y = percentage_biomass_change_from_baseline)) +
  geom_ribbon(aes(ymin = lower_bound, ymax = upper_bound, fill = "95% Prediction Interval"), alpha = 0.1
  geom_smooth(method = "lm", aes(fill = "95% Confidence Interval"), color = NA, size = 0.5, alpha = 0.3)
  geom_point(aes(color = "Data"), alpha = 0.8) +
  geom_smooth(method = "lm", aes(color = "Regression Line"), size = 0.5, se = FALSE) +
  theme_minimal() +
  ggpubr::stat_regline_equation(label.x = 1, label.y = 300) +
  labs(x = "Years from Baseline (yr)",
       y = "% Biomass Change from Baseline") +
  scale_color_manual(name = NULL,
                     values = c("Data" = "blue", "Regression Line" = "black")) +
  scale_fill_manual(name = NULL,
                    values = c("95% Confidence Interval" = "red", "95% Prediction Interval" = "green")) +
  guides(fill = guide_legend(override.aes = list(color = NA)),
         color = guide_legend(override.aes = list(fill = NA))) +
  theme(legend.position = "top")
```

2.1. Linear Regression Analysis

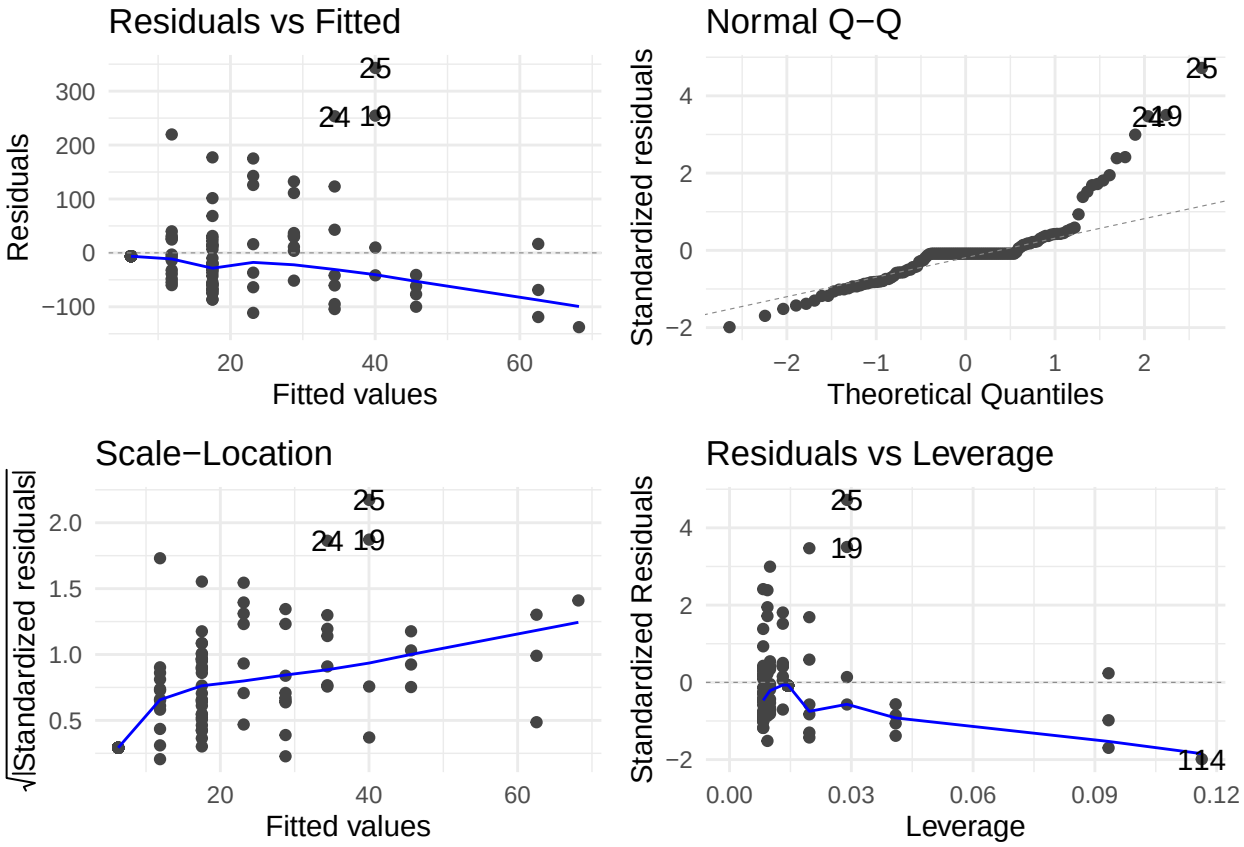



```r
# Linear regression:
```

```r
ma_cul_lm <- lm(percentage_biomass_change_from_baseline ~ years_from_baseline, data = ma_cul)

#Model diagnostics:
# plot(ma_all_lm)
autoplot(ma_cul_lm) +
  theme_minimal()
```

2.2. Diagnostic Plots



Residuals vs Fitted Plot (Top Left): The residuals show some non-random patterns, with slight curvature, indicating potential non-linearity. The spread of the residuals also increases slightly with fitted values, suggesting heteroscedasticity.

Normal Q-Q Plot (Top Right): There is noticeable deviation from the diagonal line at both the lower and upper tails, indicating that the residuals are not perfectly normally distributed, with potential outliers at points 25, 24, and 19.

Scale-Location Plot (Bottom Left): The upward-sloping blue line, along with the increasing spread of points, suggests the presence of heteroscedasticity, confirming what was observed in the Residuals vs Fitted plot.

Residuals vs Leverage Plot (Bottom Right): Points 19, 25, and 114 have relatively high leverage, particularly 114, indicating these points may be influential in the model and warrant further investigation.

Summary:

- **Non-linearity** and **heteroscedasticity** appear to be issues in this model, as indicated by the Residuals vs Fitted and Scale-Location plots.

- There are potential influential **outliers**, particularly points 19, 25, and 114, that could be affecting the model.
- **Non-normality** in the residuals is visible, particularly at the tails, which could indicate the need for data transformation or a different regression model.

```
# Tables:
# Get the summary of the model
ma_cul_lm_summary <- summary(ma_cul_lm)
# Extract coefficients
coef_table_cul <- as.data.frame(ma_cul_lm_summary$coefficients)
# Extract R-squared and Adjusted R-squared
r_squared_cul <- ma_cul_lm_summary$r.squared
adj_r_squared_cul <- ma_cul_lm_summary$adj.r.squared
# Extract F-statistic and p-value
f_statistic_cul <- ma_cul_lm_summary$fstatistic[1]
p_value_cul <- pf(f_statistic_cul, ma_cul_lm_summary$fstatistic[2], ma_cul_lm_summary$fstatistic[3], lo
# Display coefficients in a neat table
kable(coef_table_cul, caption = "Coefficients from the Linear Model")
```

2.3. Model Summary

Table 3: Coefficients from the Linear Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.258064	8.844060	0.7076008	0.4805767
years_from_baseline	5.628450	2.727726	2.0634217	0.0412474

```
# Create a summary statistics table
model_summary_cul <- data.frame(
  Statistic = c("R-squared", "Adjusted R-squared", "F-statistic", "p-value"),
  Value = c(r_squared_cul, adj_r_squared_cul, f_statistic_cul, p_value_cul)
)
# Display the model summary
kable(model_summary_cul, caption = "Model Summary Statistics")
```

Table 4: Model Summary Statistics

Statistic	Value
R-squared	0.0345431
Adjusted R-squared	0.0264301
F-statistic	4.2577092
p-value	0.0412474

```
#summary(ma_cul_lm)
```

Coefficients:

- **Intercept**: The intercept is 6.258, with a p-value of 0.4806, indicating that it is not statistically significant (i.e., the intercept does not contribute meaningfully to the model).

- `years_from_baseline`: The coefficient for `years_from_baseline` is 5.628, and the p-value is significant at the 0.05 level ($p = 0.0412$). This means that for every additional year from the baseline, the percentage change in biomass is expected to increase by approximately 5.63%, a smaller but still statistically significant effect.

Residuals:

- The residuals show some variability, ranging from -137.79 to 343.07, with a median value of -6.26. This indicates some outliers, though the variability is smaller compared to the previous model.

Model Fit:

- **Residual standard error**: The residual standard error is 73.72, which is relatively lower, indicating a better fit compared to the previous model.
- **Multiple R-squared**: The R-squared value is 0.0345, meaning only 3.45% of the variance in biomass change is explained by years_from_baseline, suggesting a weak fit.
- **Adjusted R-squared**: The adjusted R-squared is 0.0264, which is similar to the multiple R-squared, confirming limited explanatory power.

F-statistic:

- The F-statistic is 4.258 with a p-value of 0.04125, meaning the overall model is statistically significant at the 0.05 level, though only marginally.

Summary:

- The `years_from_baseline` variable has a **statistically significant positive effect** on biomass change, with a 5.63% increase per year.
- The **overall model fit is weak**.
- The residuals suggest less variability than the previous model but still indicate potential **outliers**.

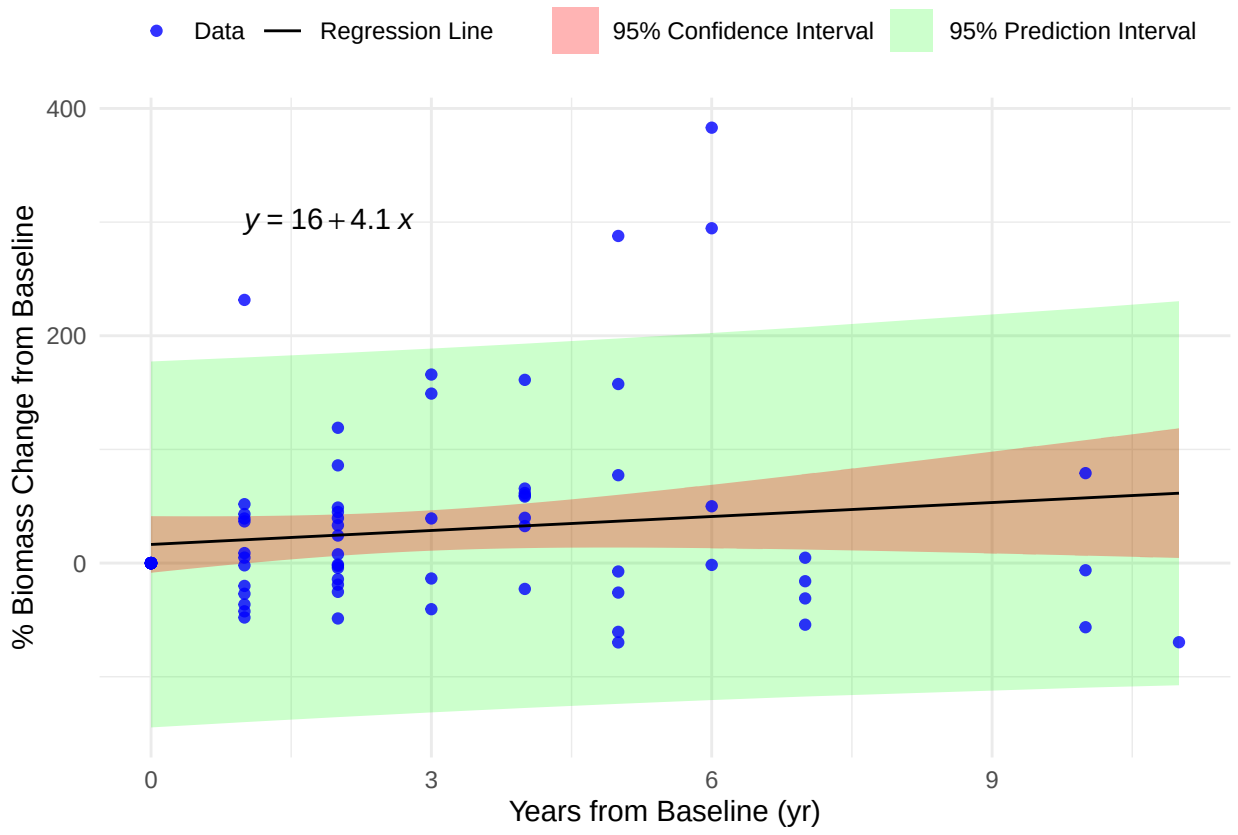
3. MA (Philipinnes - except Culasi)

```
# Linear regression:
ma_cul_phi <- ma_cul %>%
  filter(country == "Philippines")
ma_cul_phi_lm <- lm(percentage_biomass_change_from_baseline ~ years_from_baseline, data = ma_cul_phi)
# Create a data frame with predictions and prediction intervals:
new_data_cul_phi <- ma_cul_phi %>%
  mutate(pred = predict(ma_cul_phi_lm, newdata = ma_cul_phi),
         pred_interval = predict(ma_cul_phi_lm, newdata = ma_cul_phi, interval = "prediction"))
# Separate out the lower and upper bounds of the prediction interval:
new_data_cul_phi <- new_data_cul_phi %>%
  mutate(lower_bound = pred_interval[, "lwr"],
         upper_bound = pred_interval[, "upr"])

# Plot + prediction interval:
ggplot(data = new_data_cul_phi, aes(x = years_from_baseline, y = percentage_biomass_change_from_baselin
  geom_ribbon(aes(ymin = lower_bound, ymax = upper_bound, fill = "95% Prediction Interval"), alpha = 0.
  geom_smooth(method = "lm", aes(fill = "95% Confidence Interval"), color = NA, size = 0.5, alpha = 0.3
  geom_point(aes(color = "Data"), alpha = 0.8) +
  geom_smooth(method = "lm", aes(color = "Regression Line"), size = 0.5, se = FALSE) +
  theme_minimal() +
  ggpubr::stat_regline_equation(label.x = 1, label.y = 300) +
```

```
  labs(x = "Years from Baseline (yr)",
       y = "% Biomass Change from Baseline") +
  scale_color_manual(name = NULL,
                     values = c("Data" = "blue", "Regression Line" = "black")) +
  scale_fill_manual(name = NULL,
                    values = c("95% Confidence Interval" = "red", "95% Prediction Interval" = "green")) +
  guides(fill = guide_legend(override.aes = list(color = NA)),
         color = guide_legend(override.aes = list(fill = NA))) +
  theme(legend.position = "top")
```

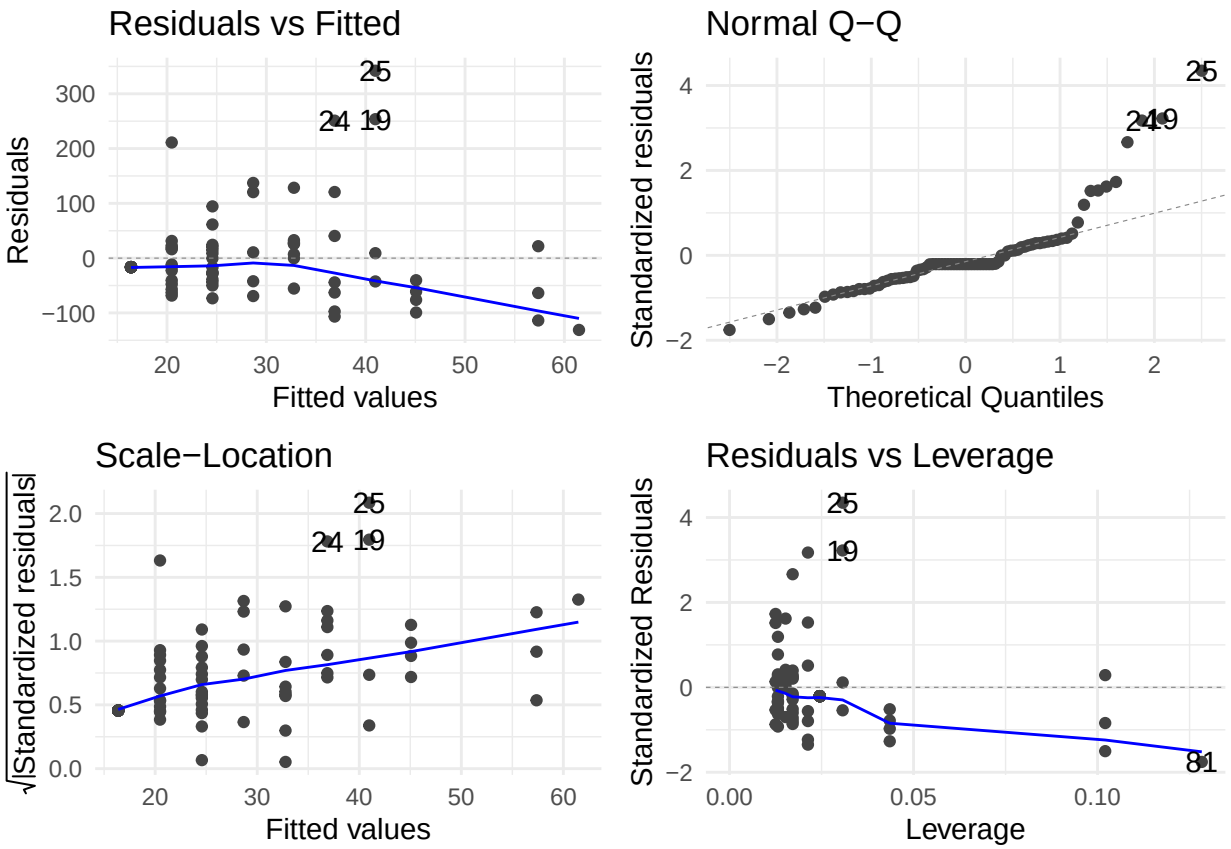
3.1. Linear Regression Analysis




```
# Linear regression:
ma_cul_phi_lm <- lm(percentage_biomass_change_from_baseline ~ years_from_baseline, data = ma_cul_phi)

#Model diagnostics:
# plot(ma_all_lm)
autoplot(ma_cul_phi_lm) +
  theme_minimal()
```

3.2. Diagnostic Plots



Residuals vs Fitted Plot (Top Left): The residuals show some non-random patterns, with a slight downward curvature, suggesting potential non-linearity. There is also a mild indication of heteroscedasticity, as the spread of residuals increases slightly at higher fitted values.

Normal Q-Q Plot (Top Right): There is noticeable deviation from the diagonal line at the upper tail, particularly with points 25, 24, and 19, indicating that the residuals are not perfectly normally distributed and suggesting potential outliers.

Scale-Location Plot (Bottom Left): The blue line slopes upwards, and the points fan out slightly, confirming the presence of heteroscedasticity, where the variability of residuals increases with fitted values.

Residuals vs Leverage Plot (Bottom Right): Points 19, 25, and 81 have relatively high leverage, particularly point 81, indicating these points may be influential in the model and should be investigated further.

Summary:

- The diagnostic plots suggest issues with **non-linearity** and **heteroscedasticity**, as observed in the Residuals vs Fitted and Scale-Location plots.
- Potential **outliers** and influential points, especially points 19, 25, and 81, could be affecting the model's accuracy.
- **Non-normality** in the residuals is visible, which could indicate the need for data transformation or a different regression model.

```
# Tables:
# Get the summary of the model
```

```r
ma_cul_phi_lm_summary <- summary(ma_cul_phi_lm)
# Extract coefficients
coef_table_cul_phi <- as.data.frame(ma_cul_phi_lm_summary$coefficients)
# Extract R-squared and Adjusted R-squared
r_squared_cul_phi <- ma_cul_phi_lm_summary$r.squared
adj_r_squared_cul_phi <- ma_cul_phi_lm_summary$adj.r.squared
# Extract F-statistic and p-value
f_statistic_cul_phi <- ma_cul_phi_lm_summary$fstatistic[1]
p_value_cul_phi <- pf(f_statistic_cul_phi, ma_cul_phi_lm_summary$fstatistic[2], ma_cul_phi_lm_summary$fs
# Display coefficients in a neat table
kable(coef_table_cul_phi, caption = "Coefficients from the Linear Model")
```

3.3. Model Summary

Table 5: Coefficients from the Linear Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.377467	12.512572	1.308881	0.1943694
years_from_baseline	4.099382	3.275521	1.251521	0.2144378

```r
# Create a summary statistics table
model_summary_cul_phi <- data.frame(
  Statistic = c("R-squared", "Adjusted R-squared", "F-statistic", "p-value"),
  Value = c(r_squared_cul_phi, adj_r_squared_cul_phi, f_statistic_cul_phi, p_value_cul_phi)
)
# Display the model summary
kable(model_summary_cul_phi, caption = "Model Summary Statistics")
```

Table 6: Model Summary Statistics

Statistic	Value
R-squared	0.0194412
Adjusted R-squared	0.0070290
F-statistic	1.5663043
p-value	0.2144378

```r
#summary(ma_cul_phi_lm)
```

Coefficients:

- **Intercept**: The intercept is 16.377, with a p-value of 0.194, indicating that it is not statistically significant. This means the intercept does not contribute meaningfully to the model.
- `years_from_baseline`: The coefficient for `years_from_baseline` is 4.099, with a p-value of 0.214, which is also not statistically significant. This suggests that `years_from_baseline` does not have a significant effect on the percentage biomass change in this model.

Residuals:

- The residuals range from -131.09 to 342.13, with a median of -16.38. The spread of residuals suggests possible outliers but less extreme variability than some previous models.

Model Fit:

- **Residual standard error**: The residual standard error is 79.92, suggesting moderate deviations from the fitted line, but less than other models analyzed.
- **Multiple R-squared**: The R-squared value is 0.01944, indicating that only about 1.94% of the variance in the percentage biomass change is explained by years_from_baseline, suggesting a very weak fit.
- **Adjusted R-squared**: The adjusted R-squared is 0.007029, further confirming the model's weak explanatory power.

F-statistic:

- The F-statistic is 1.566 with a p-value of 0.2144, indicating that the overall model is not statistically significant. This means that neither the intercept nor years_from_baseline provides a good explanation of the biomass change.
-

Summary:

- In this model, the `years_from_baseline` variable **does not have a statistically significant effect** on the percentage biomass change.
- The **model fit is very weak**, explaining only a minimal portion of the variability in biomass change.
- Both coefficients and the overall model lack significance, and the residuals suggest some potential **outliers**.

Reserve Areas

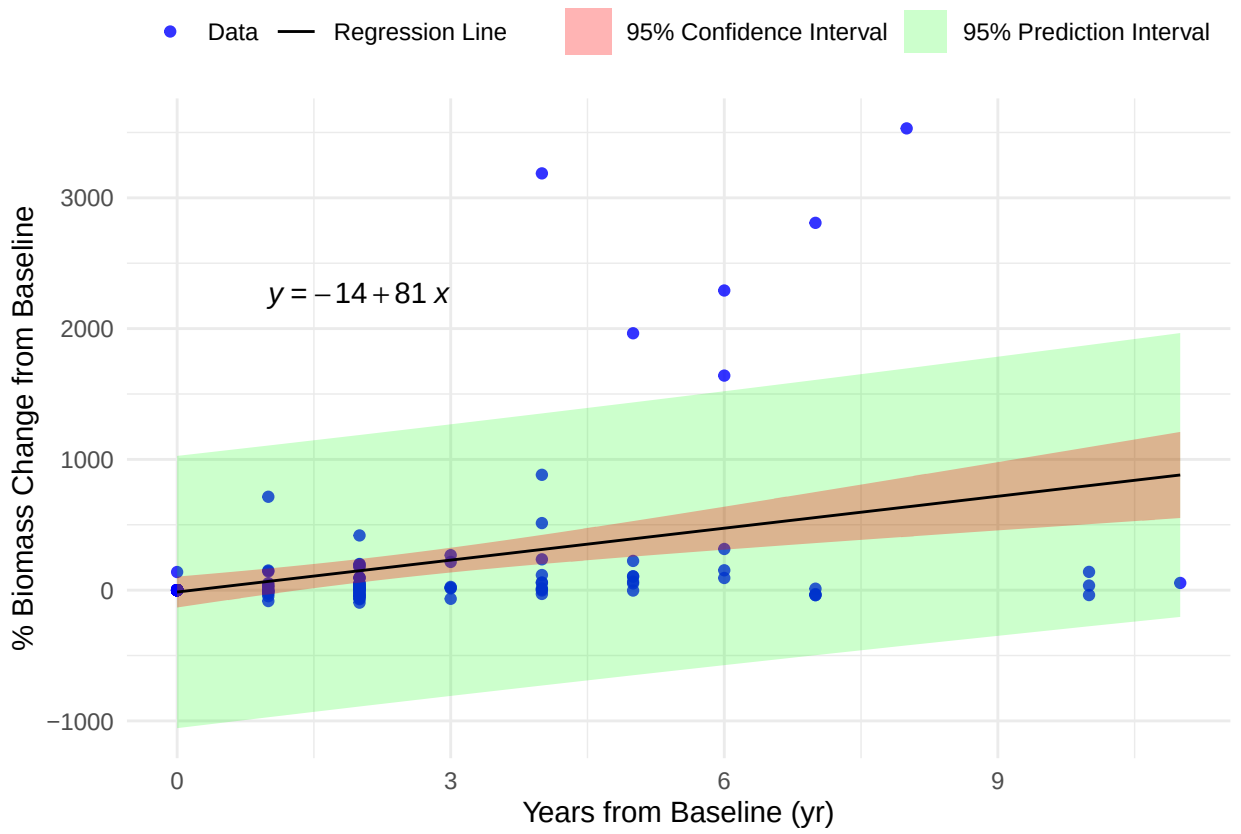
1. Reserve Areas (All Countries)

```
# Linear regression:
reserve_lm <- lm(percentage_biomass_change_from_baseline ~ years_from_baseline, data = reserve)
# Create a data frame with predictions and prediction intervals:
new_data_reserve <- reserve %>%
  mutate(pred = predict(reserve_lm, newdata = reserve),
         pred_interval = predict(reserve_lm, newdata = reserve, interval = "prediction"))
# Separate out the lower and upper bounds of the prediction interval:
new_data_reserve <- new_data_reserve %>%
  mutate(lower_bound = pred_interval[, "lwr"],
         upper_bound = pred_interval[, "upr"])

# Plot + prediction interval:
ggplot(data = new_data_reserve, aes(x = years_from_baseline, y = percentage_biomass_change_from_baselin
  geom_point(aes(color = "Data"), alpha = 0.8) +
  geom_ribbon(aes(ymin = lower_bound, ymax = upper_bound, fill = "95% Prediction Interval"), alpha = 0.
  geom_smooth(method = "lm", aes(fill = "95% Confidence Interval"), color = NA, size = 0.5, alpha = 0.3
  geom_smooth(method = "lm", aes(color = "Regression Line"), size = 0.5, se = FALSE) +
  theme_minimal() +
  ggpubr::stat_regline_equation(label.x = 1, label.y = 2250) +
  labs(x = "Years from Baseline (yr)",
       y = "% Biomass Change from Baseline") +
  scale_color_manual(name = NULL,
                     values = c("Data" = "blue", "Regression Line" = "black")) +
  scale_fill_manual(name = NULL,
                    values = c("95% Confidence Interval" = "red", "95% Prediction Interval" = "green"))
```

```
  guides(fill = guide_legend(override.aes = list(color = NA)),
         color = guide_legend(override.aes = list(fill = NA))) +
  theme(legend.position = "top")
```

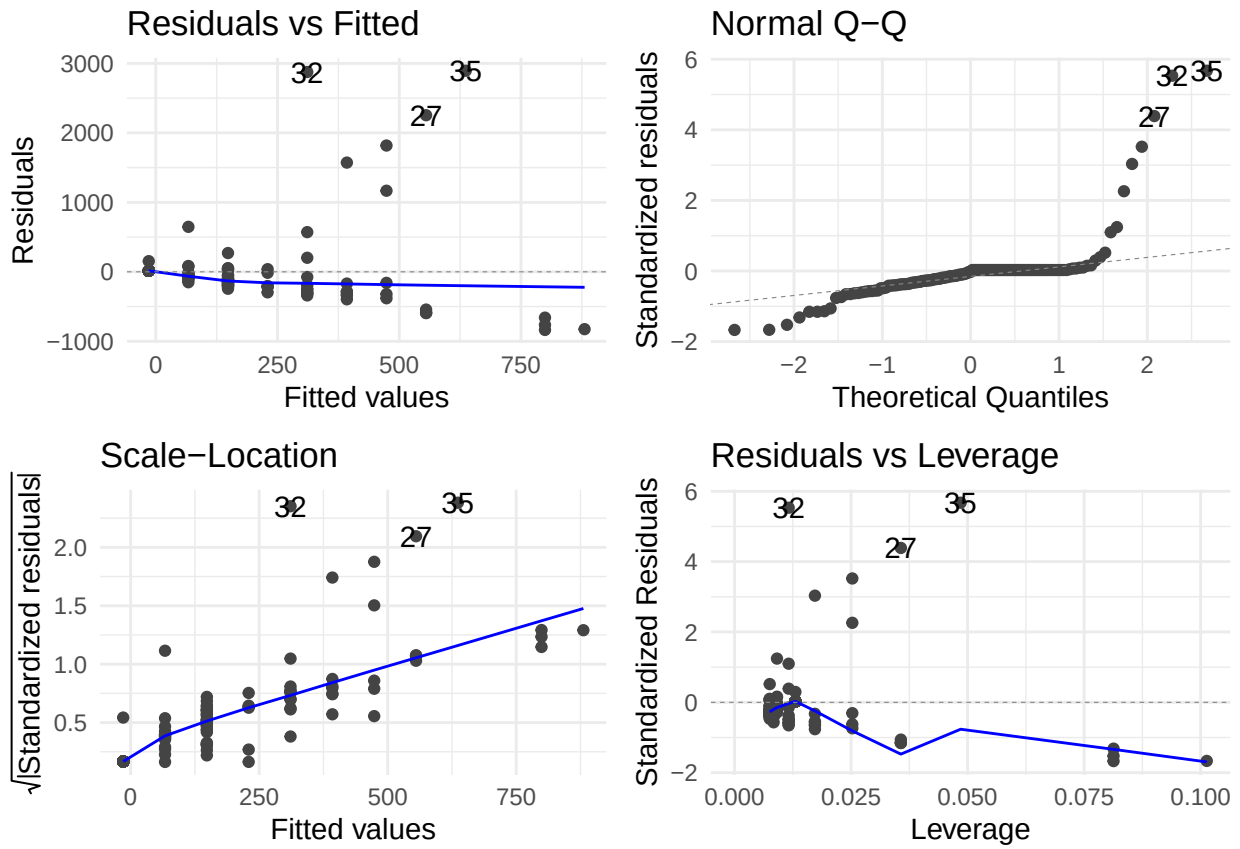
1.1. Linear Regression Analysis




```
# Linear regression:
reserve_lm <- lm(percentage_biomass_change_from_baseline ~ years_from_baseline, data = reserve)

#Model diagnostics:
# plot(ma_all_lm)
autoplot(reserve_lm) +
  theme_minimal()
```

1.2. Diagnostic Plots



Residuals vs Fitted Plot (Top Left): There is a slight curvature in the residuals, indicating potential non-linearity. Additionally, the spread of residuals increases at higher fitted values, suggesting heteroscedasticity.

Normal Q-Q Plot (Top Right): There is significant deviation from the diagonal line at the upper tail, especially for points 32, 35, and 27, indicating that the residuals are not perfectly normally distributed and suggesting the presence of outliers.

Scale-Location Plot (Bottom Left): The upward-sloping blue line indicates increasing variability in residuals as fitted values increase, confirming heteroscedasticity.

Residuals vs Leverage Plot (Bottom Right): Points 32, 35, and 27 show relatively high leverage, particularly point 32, which suggests these points may be influential in the model and should be further examined.

Summary:

- This analysis suggests **non-linearity** and **heteroscedasticity**, as observed in the Residuals vs Fitted and Scale-Location plots.
- Potential **outliers** and influential points, particularly 32, 35, and 27, could be affecting the model.
- **Non-normality** in the residuals is visible, which could indicate the need for data transformation or a different regression model.

```
# Tables:
# Get the summary of the model
reserve_lm_summary <- summary(reserve_lm)
# Extract coefficients
```

```r
coef_table_res <- as.data.frame(reserve_lm_summary$coefficients)
# Extract R-squared and Adjusted R-squared
r_squared_res <- reserve_lm_summary$r.squared
adj_r_squared_res <- reserve_lm_summary$adj.r.squared
# Extract F-statistic and p-value
f_statistic_res <- reserve_lm_summary$fstatistic[1]
p_value_res <- pf(f_statistic_res, reserve_lm_summary$fstatistic[2], reserve_lm_summary$fstatistic[3], l
# Display coefficients in a neat table
kable(coef_table_res, caption = "Coefficients from the Linear Model")
```

1.3. Model Summary

Table 7: Coefficients from the Linear Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-14.47112	59.74770	-0.2422037	0.8090011
years_from_baseline	81.37722	18.09518	4.4971756	0.0000150

```r
# Create a summary statistics table
model_summary_res <- data.frame(
  Statistic = c("R-squared", "Adjusted R-squared", "F-statistic", "p-value"),
  Value = c(r_squared_res, adj_r_squared_res, f_statistic_res, p_value_res)
)
# Display the model summary
kable(model_summary_res, caption = "Model Summary Statistics")
```

Table 8: Model Summary Statistics

Statistic	Value
R-squared	0.1337388
Adjusted R-squared	0.1271261
F-statistic	20.2245884
p-value	0.0000150

```r
#summary(reserve_lm)
```

Coefficients:

- **Intercept**: The intercept is -14.47, with a p-value of 0.809, indicating that it is not statistically significant. This suggests that the intercept does not meaningfully contribute to the model.
- `years_from_baseline`: The coefficient for `years_from_baseline` is 81.38, and it is highly significant, with a p-value of 1.5e-05. This suggests that for each additional year from baseline, the percentage biomass change is expected to increase by 81.38%, and this effect is statistically significant.

Residuals:

- The residuals range from -836.84 to 2894.72, with a median of -13.85. The large spread of residuals, especially the high maximum value, suggests potential outliers, which might affect the model fit.

Model Fit:

- **Residual standard error**: The residual standard error is 522.9, indicating that the actual values deviate substantially from the predicted values, suggesting a less precise model fit.

- **Multiple R-squared**: The R-squared value is 0.1337, indicating that about 13.4% of the variance in percentage biomass change can be explained by years_from_baseline. This is a modest fit, meaning that other factors likely affect biomass change.
- Adjusted R-squared: The adjusted R-squared is 0.1271, confirming that the model explains some variability, but the explanatory power remains relatively low.

F-statistic:

- The F-statistic is 20.22, with a highly significant p-value of 1.501e-05. This indicates that the overall model is statistically significant, meaning the years_from_baseline variable is a meaningful predictor of biomass change.

Summary:

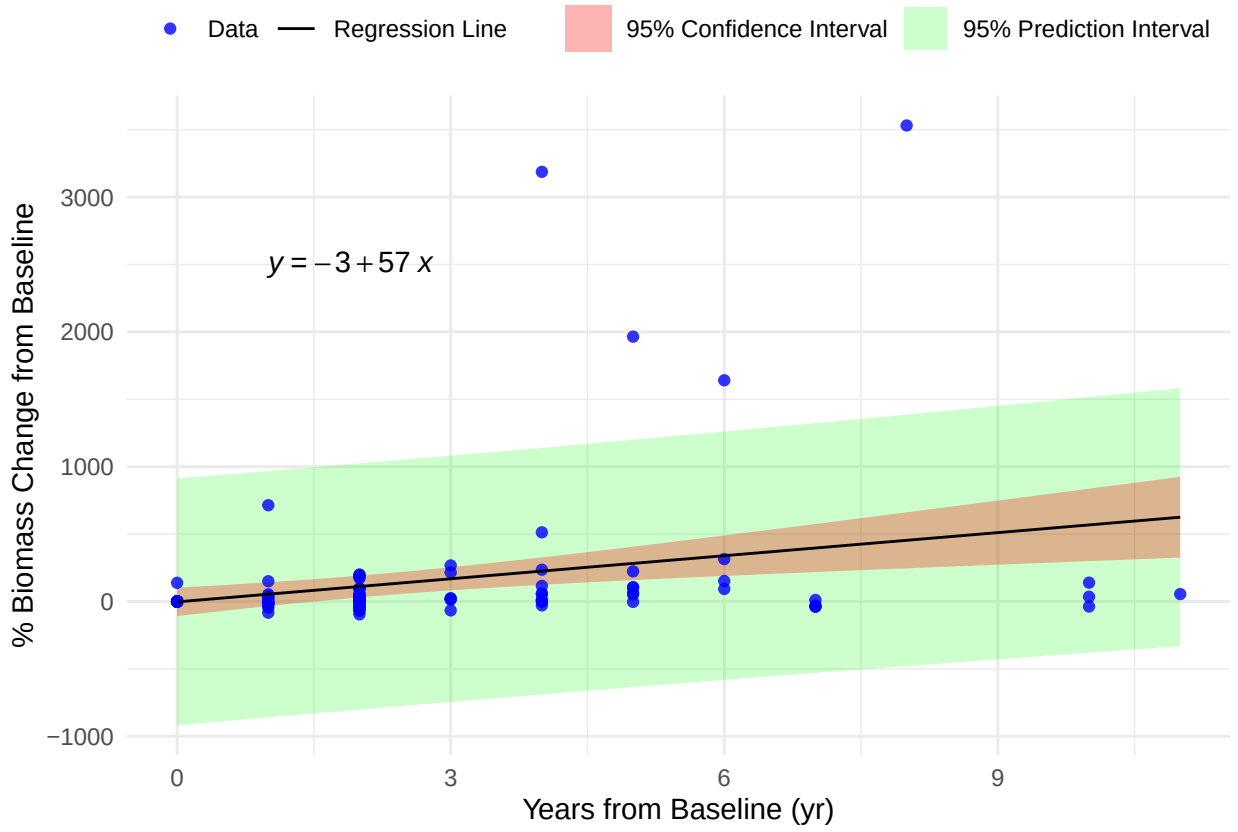
- The `years_from_baseline` variable has a **statistically significant positive effect** on the percentage biomass change, with an 81.38% increase for each additional year.
- The **model fit is still moderate**, as indicated by the relatively low R-squared values, suggesting other factors not included in the model may be influencing biomass change.
- The residuals indicate potential **outliers**, contributing to the high residual standard error, which reduces the model's precision.

2. Reserve Areas (All Countries - except Cortes)

```
# Linear regression:
reserve_cor <- reserve %>%
  filter(ma_name != "Cortes")
reserve_cor_lm <- lm(percentage_biomass_change_from_baseline ~ years_from_baseline, data = reserve_cor)
# Create a data frame with predictions and prediction intervals:
new_data_reserve_cor <- reserve_cor %>%
  mutate(pred = predict(reserve_cor_lm, newdata = reserve_cor),
         pred_interval = predict(reserve_cor_lm, newdata = reserve_cor, interval = "prediction"))
# Separate out the lower and upper bounds of the prediction interval:
new_data_reserve_cor <- new_data_reserve_cor %>%
  mutate(lower_bound = pred_interval[, "lwr"],
         upper_bound = pred_interval[, "upr"])

# Plot + prediction interval:
ggplot(data = new_data_reserve_cor, aes(x = years_from_baseline, y = percentage_biomass_change_from_base
  geom_ribbon(aes(ymin = lower_bound, ymax = upper_bound, fill = "95% Prediction Interval"), alpha = 0.1
  geom_smooth(method = "lm", aes(fill = "95% Confidence Interval"), color = NA, size = 0.5, alpha = 0.3)
  geom_point(aes(color = "Data"), alpha = 0.8) +
  geom_smooth(method = "lm", aes(color = "Regression Line"), size = 0.5, se = FALSE) +
  theme_minimal() +
  ggpubr::stat_regline_equation(label.x = 1, label.y = 2500) +
  labs(x = "Years from Baseline (yr)",
       y = "% Biomass Change from Baseline") +
  scale_color_manual(name = NULL,
                     values = c("Data" = "blue", "Regression Line" = "black")) +
  scale_fill_manual(name = NULL,
                    values = c("95% Confidence Interval" = "red", "95% Prediction Interval" = "green"))
  guides(fill = guide_legend(override.aes = list(color = NA)),
         color = guide_legend(override.aes = list(fill = NA))) +
  theme(legend.position = "top")
```

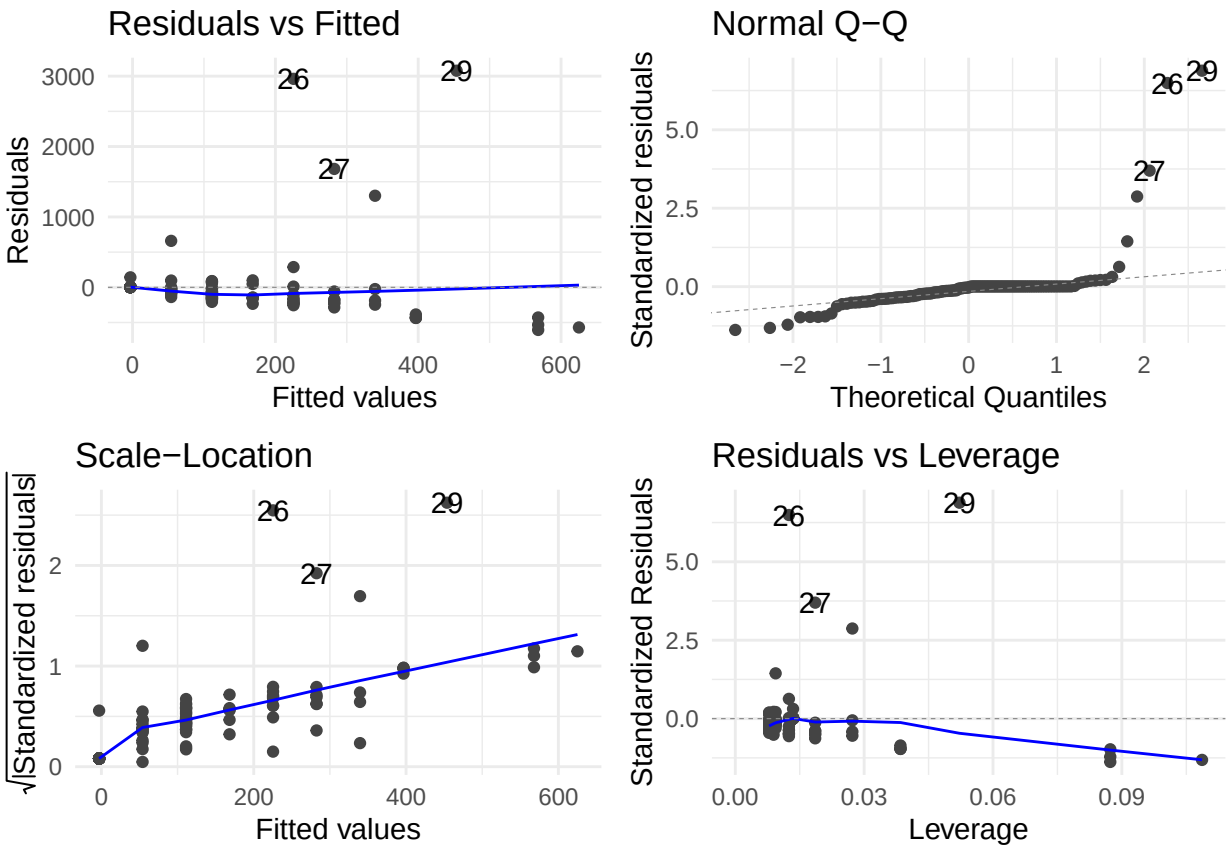
2.1. Linear Regression Analysis




```
# Linear regression:
reserve_cor_lm <- lm(percentage_biomass_change_from_baseline ~ years_from_baseline, data = reserve_cor)

#Model diagnostics:
# plot(ma_all_lm)
autoplot(reserve_cor_lm) +
  theme_minimal()
```

2.2. Diagnostic Plots



Residuals vs Fitted Plot (Top Left): There is some slight curvature, suggesting potential non-linearity. The spread of residuals increases slightly at higher fitted values, indicating possible heteroscedasticity.

Normal Q-Q Plot (Top Right): Significant deviation from the diagonal line at the upper tail is noticeable, especially for points 26, 27, and 29, indicating non-normality in the residuals and the presence of potential outliers.

Scale-Location Plot (Bottom Left): The blue line slopes upwards, and the spread of points increases, confirming heteroscedasticity with increasing fitted values.

Residuals vs Leverage Plot (Bottom Right): Points 26, 27, and 29 have relatively high leverage, suggesting they may be influential in the model and should be examined for their impact.

Summary:

- The diagnostic plots suggest **non-linearity** and **heteroscedasticity**, as seen in the Residuals vs Fitted and Scale-Location plots.
- **Outliers** and influential points, particularly 26, 27, and 29, could be affecting the model.
- **Non-normality** in the residuals is visible, which could indicate the need for data transformation or a different regression model.

```
# Tables:
# Get the summary of the model
reserve_cor_lm_summary <- summary(reserve_cor_lm)
# Extract coefficients
coef_table_res_cor <- as.data.frame(reserve_cor_lm_summary$coefficients)
```

```r
# Extract R-squared and Adjusted R-squared
r_squared_res_cor <- reserve_cor_lm_summary$r.squared
adj_r_squared_res_cor <- reserve_cor_lm_summary$adj.r.squared
# Extract F-statistic and p-value
f_statistic_res_cor <- reserve_cor_lm_summary$fstatistic[1]
p_value_res_cor <- pf(f_statistic_res_cor, reserve_cor_lm_summary$fstatistic[2], reserve_cor_lm_summary$
# Display coefficients in a neat table
kable(coef_table_res_cor, caption = "Coefficients from the Linear Model")
```

2.3. Model Summary

Table 9: Coefficients from the Linear Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.018334	53.26231	-0.0566692	0.9548992
years_from_baseline	57.099370	16.37277	3.4874588	0.0006737

```r
# Create a summary statistics table
model_summary_res_cor <- data.frame(
  Statistic = c("R-squared", "Adjusted R-squared", "F-statistic", "p-value"),
  Value = c(r_squared_res_cor, adj_r_squared_res_cor, f_statistic_res_cor, p_value_res_cor)
)
# Display the model summary
kable(model_summary_res_cor, caption = "Model Summary Statistics")
```

Table 10: Model Summary Statistics

Statistic	Value
R-squared	0.0886713
Adjusted R-squared	0.0813807
F-statistic	12.1623687
p-value	0.0006737

```r
#summary(reserve_cor_lm)
```

Coefficients:

- **Intercept**: The intercept is -3.018, with a p-value of 0.954899, indicating that it is not statistically significant. This suggests that the intercept does not meaningfully contribute to the model.
- `years_from_baseline`: The coefficient for `years_from_baseline` is 57.099, and it is highly significant, with a p-value of 0.000674. This suggests that for each additional year from baseline, the percentage biomass change is expected to increase by 57.1%, and this effect is statistically significant.

Residuals:

- The residuals range from -605.52 to 3077.49, with a median of -13.64. The large spread of residuals, especially the high maximum value, suggests potential outliers, which might affect the model fit.

Model Fit:

- **Residual standard error**: The residual standard error is 459.3, indicating that the actual values deviate substantially from the predicted values, suggesting a moderate model fit.

- **Multiple R-squared**: The R-squared value is 0.08867, indicating that about 8.9% of the variance in percentage biomass change can be explained by years_from_baseline. This is a relatively low value, suggesting that the model only explains a small portion of the variability in biomass change.
- **Adjusted R-squared:** The adjusted R-squared is 0.08138, confirming that the model's explanatory power remains low, even when accounting for the number of predictors.

F-statistic:

- The F-statistic is 12.16, with a significant p-value of 0.0006737. This indicates that the overall model is statistically significant, meaning the `years_from_baseline` variable is a meaningful predictor of biomass change.

Summary:

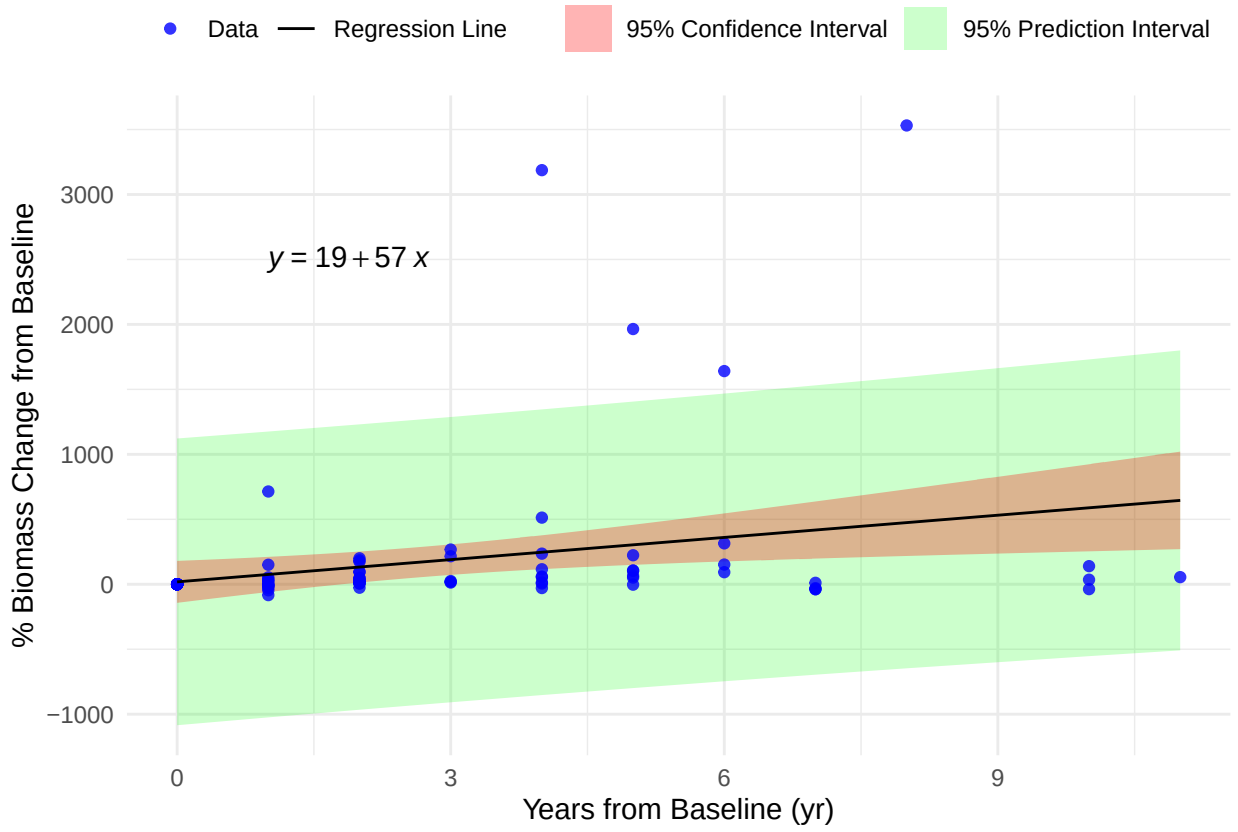
- The `years_from_baseline` variable has a **statistically significant positive effect** on the percentage biomass change, with a 57.1% increase for each additional year.
- **The model fit is weak**, as indicated by the low R-squared values, suggesting that other factors not included in the model may be influencing biomass change.
- The residuals indicate potential **outliers**, contributing to the high residual standard error, which reduces the model's precision.

3. Reserve Areas (Philippines - except Cortes)

```
# Linear regression:
reserve_cor_phi <- reserve_cor %>%
  filter(country == "Philippines")
reserve_cor_phi_lm <- lm(percentage_biomass_change_from_baseline ~ years_from_baseline, data = reserve_
# Create a data frame with predictions and prediction intervals:
new_data_reserve_cor_phi <- reserve_cor_phi %>%
  mutate(pred = predict(reserve_cor_phi_lm, newdata = reserve_cor_phi),
         pred_interval = predict(reserve_cor_phi_lm, newdata = reserve_cor_phi, interval = "prediction"
# Separate out the lower and upper bounds of the prediction interval:
new_data_reserve_cor_phi <- new_data_reserve_cor_phi %>%
  mutate(lower_bound = pred_interval[, "lwr"],
         upper_bound = pred_interval[, "upr"])

# Plot + prediction interval:
ggplot(data = new_data_reserve_cor_phi, aes(x = years_from_baseline, y = percentage_biomass_change_from_
  geom_ribbon(aes(ymin = lower_bound, ymax = upper_bound, fill = "95% Prediction Interval"), alpha = 0.1
  geom_smooth(method = "lm", aes(fill = "95% Confidence Interval"), color = NA, size = 0.5, alpha = 0.3
  geom_point(aes(color = "Data"), alpha = 0.8) +
  geom_smooth(method = "lm", aes(color = "Regression Line"), size = 0.5, se = FALSE) +
  theme_minimal() +
  ggpubr::stat_regline_equation(label.x = 1, label.y = 2500) +
  labs(x = "Years from Baseline (yr)",
       y = "% Biomass Change from Baseline") +
  scale_color_manual(name = NULL,
                     values = c("Data" = "blue", "Regression Line" = "black")) +
  scale_fill_manual(name = NULL,  # Correct the fill for the legend
                    values = c("95% Confidence Interval" = "red", "95% Prediction Interval" = "green"))
  guides(fill = guide_legend(override.aes = list(color = NA)),
         color = guide_legend(override.aes = list(fill = NA))) +
  theme(legend.position = "top")
```

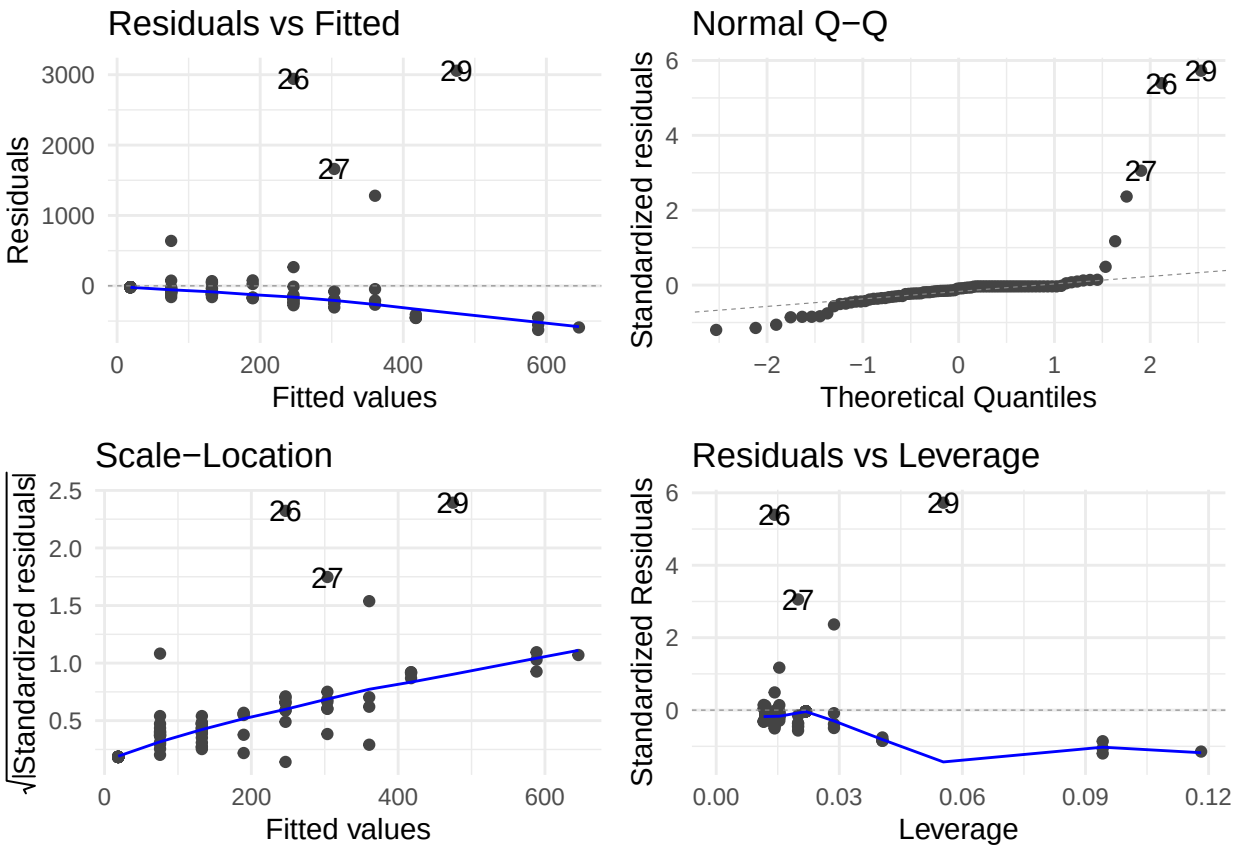
3.1. Linear Regression Analysis




```
# Linear regression:
reserve_cor_phi_lm <- lm(percentage_biomass_change_from_baseline ~ years_from_baseline, data = reserve_

#Model diagnostics:
# plot(ma_all_lm)
autoplot(reserve_cor_phi_lm) +
  theme_minimal()
```

3.2. Diagnostic Plots



Residuals vs Fitted Plot (Top Left): There is slight downward curvature in the residuals, suggesting potential non-linearity. The spread of the residuals increases slightly at higher fitted values, indicating possible heteroscedasticity.

Normal Q-Q Plot (Top Right): There is significant deviation from the diagonal line at the upper tail, particularly for points 26, 27, and 29, indicating non-normality in the residuals and the presence of potential outliers.

Scale-Location Plot (Bottom Left): The upward-sloping blue line and increasing spread of points confirm heteroscedasticity, where the variability of residuals increases as fitted values rise.

Residuals vs Leverage Plot (Bottom Right): Points 26, 27, and 29 have relatively high leverage, indicating these points may be influential in the model and should be examined for their impact on the analysis.

Summary:

- This analysis suggests issues with **non-linearity** and **heteroscedasticity**, as shown by the Residuals vs Fitted and Scale-Location plots.
- **Outliers** and influential points, particularly 26, 27, and 29, could be affecting the model's accuracy.
- **Non-normality** in the residuals is visible, which could indicate the need for data transformation or a different regression model.

```
# Tables:
# Get the summary of the model
reserve_cor_phi_lm_summary <- summary(reserve_cor_phi_lm)
```

```r
# Extract coefficients
coef_table_res_cor_phi <- as.data.frame(reserve_cor_phi_lm_summary$coefficients)
# Extract R-squared and Adjusted R-squared
r_squared_res_cor_phi <- reserve_cor_phi_lm_summary$r.squared
adj_r_squared_res_cor_phi <- reserve_cor_phi_lm_summary$adj.r.squared
# Extract F-statistic and p-value
f_statistic_res_cor_phi <- reserve_cor_phi_lm_summary$fstatistic[1]
p_value_res_cor_phi <- pf(f_statistic_res_cor_phi, reserve_cor_phi_lm_summary$fstatistic[2], reserve_co
# Display coefficients in a neat table
kable(coef_table_res_cor_phi, caption = "Coefficients from the Linear Model")
```

3.3. Model Summary

Table 11: Coefficients from the Linear Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18.57082	81.18687	0.2287417	0.8196127
years_from_baseline	56.99891	21.43016	2.6597516	0.0093266

```r
# Create a summary statistics table
model_summary_res_cor_phi <- data.frame(
  Statistic = c("R-squared", "Adjusted R-squared", "F-statistic", "p-value"),
  Value = c(r_squared_res_cor_phi, adj_r_squared_res_cor_phi, f_statistic_res_cor_phi, p_value_res_cor_
)
# Display the model summary
kable(model_summary_res_cor_phi, caption = "Model Summary Statistics")
```

Table 12: Model Summary Statistics

Statistic	Value
R-squared	0.0760068
Adjusted R-squared	0.0652627
F-statistic	7.0742784
p-value	0.0093266

```r
#summary(reserve_cor_phi_lm)
```

Coefficients:

- **Intercept**: The intercept is 18.57, but its p-value is 0.81961, indicating it is not statistically significant. This suggests the intercept does not meaningfully contribute to the model.
- `years_from_baseline`: The coefficient for `years_from_baseline` is 57.00, with a p-value of 0.00933, indicating that the effect of years from baseline is statistically significant. This means that for each additional year from baseline, the percentage biomass change is expected to increase by 57.0%.
- **Residuals**:
- The residuals range from -626.10 to 3056.71, with a median of -50.21. The large range, particularly the high maximum, suggests the presence of potential outliers affecting the model fit.

Model Fit:

- **Residual standard error**: The residual standard error is 549.1, suggesting considerable deviation of actual data points from the predicted values, indicating moderate model fit.
- **Multiple R-squared**: The R-squared value is 0.07601, meaning that around 7.6% of the variance in percentage biomass change is explained by the years_from_baseline variable. This is quite low, indicating that the model does not explain much of the variability in biomass change.
- **Adjusted R-squared**: The adjusted R-squared is 0.06526, which further confirms the low explanatory power of the model.

F-statistic:

- The F-statistic is 7.074 with a p-value of 0.009327, indicating that the model as a whole is statistically significant, meaning that the `years_from_baseline` variable is a meaningful predictor in this model.

Summary:

- The `years_from_baseline` variable has a **statistically significant positive effect** on the percentage biomass change, with an expected 57.0% increase for each additional year.
- The **model fit remains weak**, as indicated by the low R-squared values, meaning other factors influencing biomass change are not captured by this model.
- There are potential **outliers**, reflected by the high residual range, which might be affecting the overall fit of the model.