Describe probability as a foundation of statistical modeling, including inference and maximum likelihood estimation.

Probability serves as a crucial foundation for statistical modeling, underpinning the methods used to understand and make inferences about data. Probability theory provides a mathematical framework for quantifying and managing uncertainty and randomness. It allows statisticians to describe the likelihood of various outcomes and model the inherent variability in data. Probability distributions, such as the normal, binomial, and Poisson distributions, describe how probabilities are distributed over possible outcomes. They are essential for modeling the behavior of random variables and understanding the data's structure. Normal distribution and binomial distribution usually help one to model uncertainty, allowing one to make predictions and understand the variability in data.
In simple linear regression, we model the relationship between a single independent variable x and a dependent variable y using the equation:

$$y=\beta0+\beta1x+\epsilon y=\beta0+\beta1x+\epsilon$$

where:

β0β0 is the intercept

β1β1 is the slope

When describing inference in Linear regression the primary goal is to estimate the coefficients, which is done using Ordinary Least Squares estimation. Inference can be done using hypothesis testing which can be done on the regression coefficients and the t-test is used to carry out the hypotheses if $\beta1=0\beta1=0$. Moreover, when including inference of we construct confidence interval for the regression coefficients.

Statistical Inference

Statistical inference involves drawing conclusions about a population based on a sample of data. Probability is fundamental to this process in several ways:

1. Sampling Distributions: When we take a sample from a population, the sample statistics (e.g., sample mean, sample variance) can be seen as random variables with their own probability distributions. These sampling distributions allow us to make probabilistic statements about population parameters.

2. Estimation:
Point Estimation: Using sample data to provide the best single guess (estimate) of a population parameter (e.g., the sample mean as an estimate of the population mean).
Interval Estimation: Using sample data to construct an interval (confidence interval) that, with a certain probability, contains the population parameter. This incorporates the idea of probability to provide a range of plausible values for the parameter.

3. Hypothesis Testing: Probability is used to determine the likelihood of observing the sample data under a null hypothesis. If this probability (p-value) is sufficiently low, the null hypothesis is rejected in favor of the alternative hypothesis.

Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation is a method for estimating the parameters of a statistical model. It is based on the principle of finding the parameter values that maximize the likelihood function, which measures how likely it is to observe the given sample data under various parameter values.

It can be seen in the likelihood function. In the context of linear regression, assuming the errors ϵ are normally distributed, the likelihood function is:

$$L(\beta 0,\beta 1,\ldots,\beta p,\sigma 2|y,X)=\prod i=1n12\pi\sigma 2\exp[f0](-(yi-\beta 0-\beta 1xi1-\ldots-\beta pxip)22\sigma 2)L(\beta 0,\beta 1$$

$$,\ldots,\beta p,\sigma 2|y,X)=\prod i=1n2\pi\sigma 21\exp(-2\sigma 2(yi-\beta 0-\beta 1xi1-\ldots-\beta pxip)2)$$

Application in Statistical Modeling

1. Model Selection: Probability models are chosen based on the nature of the data and the underlying phenomena. For instance, a Poisson distribution might be used for modeling the number of events occurring in a fixed interval of time or space.

2. Parameter Estimation: Once a model is selected, MLE is used to estimate the parameters, providing the best fit to the data according to the likelihood principle.

3. Model Checking and Validation: Probability-based methods, such as goodness-of-fit tests and likelihood ratio tests, are used to assess how well the model fits the data and whether assumptions (e.g., normality, independence) are reasonable.

4. Predictive Inference: The fitted model can be used to make probabilistic predictions about future or unobserved data points, allowing for decision-making under uncertainty.

Discussion and Results from the In-Class Activity on Simple and Multiple Linear Regression

Simple Linear Regression

Simple linear regression is a very straightforward approach for predicting a quantitative response Y based on a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y.
In our example, we explored the relationship between the personal freedom score, `pf_score`, and the political pressures and controls on media content index, `pf_expression_control`. Specifically, we used the political pressures and controls on media content index to predict a country's personal freedom score in 2016.

```
##$$
\hat{y} = b_0 + b_1 \times x_1
$$\hat{pf\\_score} = 3.361 + 0.428 \times x_1$$
```

$$\hat{pf}_{score} = 3.361 + 0.428 \times x_1$$

X is pf_expression_control and Y is the pf_score. Then we regressed pf_scores onto pf_expression_control by fitting the model pf_score ≈ β0 + β1 × x1. In the equation above, β0 and β1 are two unknown constants that represent the intercept and slope terms in the linear model. Together, 3.361 and 0.428 are intercept slope known as the model coefficients or parameters. Once we have used our coeffcient parameter training data to produce estimates βˆ0 and βˆ1 for the model coeffcients, we can predict future pf_score based on

pf_expression_control by computing yˆ = βˆ0 + βˆ1x, where yˆ indicates a prediction of Y on the basis of X = x.

For the X is pf_expression_control and Y is the pf_score., the least squares ft for the regression of pf_score onto pf_expression_control is shown. The fit is found by minimizing the residual sum of squares. Each grey line segment represents a residual. In this case a linear ft captures the essence of the relationship, although it overestimates the trend in the left of the plot. Let yˆi = βˆ0 + βˆ1xi be the prediction for Y based on the ith value of X.
Visualization

A scatter plot was used to display the relationship between the personal freedom score, `pf_score`, and the political pressures and controls on media content index,`pf_expression_control'  from the plot, there is a linear relationship between the personal freedom score, `pf_score`, and the political pressures and controls on media content index,`pf_expression_control`

Figure 1: Histogram displaying personal Freedom Score

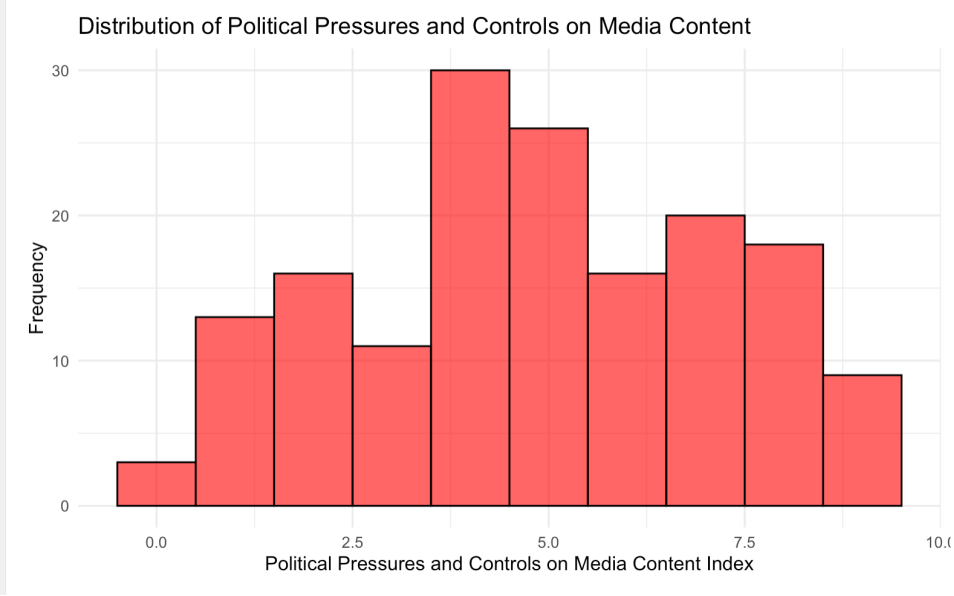

Figure 2: Histogram displaying the political pressures and control on Media Content Index

The distribution of pf scores is left skewed, and the center is at 5 while the distribution of pf_expression_control is normally distributed, the center is at 4 and there are no outliers.

The **scatter plot** below shows there is a linear relationship between the personal freedom score, `pf_score`, and the political pressures and controls on media content index,`pf_expression_control`. A scatter plot allows us to visualize the relationship between two continuous variables and can help identify patterns, correlations, and potential outliers.

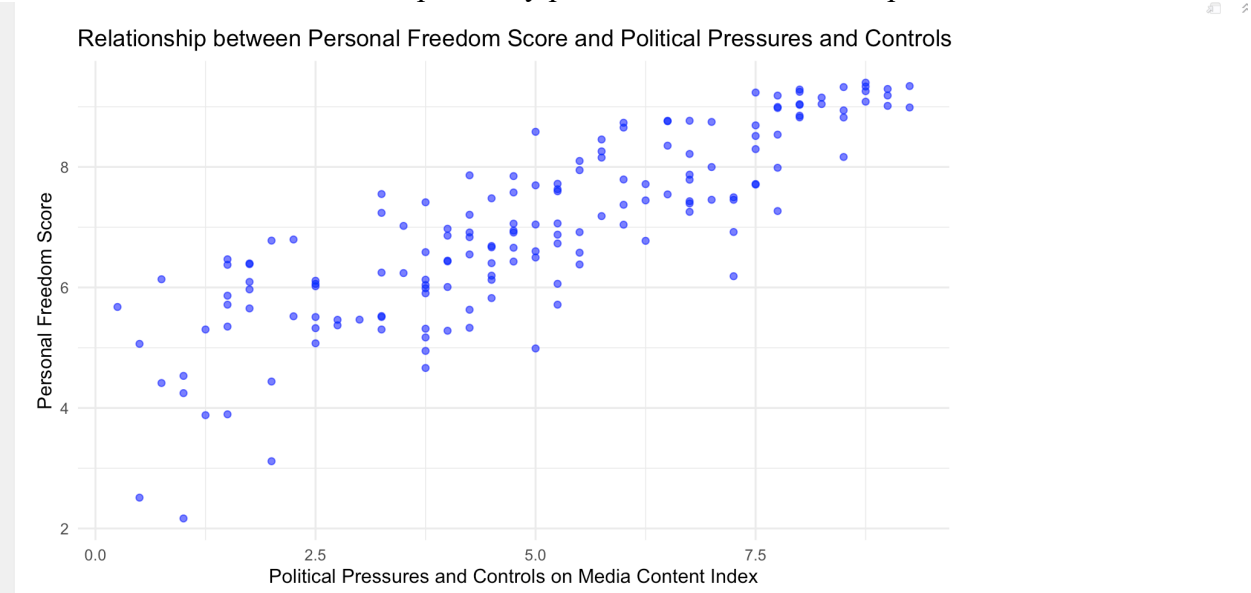

Figure 3: Personal freedom score vs Political Pressures and Control on Media Content Index

Correlation of Coefficients

The strong positive correlation coefficient of 0.8450646 indicates that personal freedom scores and political pressures and controls on media content index are strongly positively related. This suggests that in countries where personal freedoms are higher, political pressures and controls on media content are generally lower, demonstrating a significant linear relationship between these two variables. This relationship is visually supported by the scatter plot showing a clear upward trend.

Implications: The high correlation suggests that countries with higher personal freedom scores tend to have lower levels of political pressure and control on media content. This could imply that in environments where personal freedoms are respected, there is less political interference in media content.

Fit a Simple Linear Regression Model

pf_score = 4.283 + 0.541 × pf_expression_control this formula can now be used to predict the personal freedom score based on the value of the political pressures and controls on media content index.
In summary, for countries with a `pf_expression_control` of 0 we expect their mean personal freedom score to be 4.283. For every 1 unit increase in `pf_expression_control` we expect a country's mean personal freedom score to increase 0.541 units.

Assessing the Model

We got R2 of 0.7141342 or 71.41%. This means 71.41% of the variation in personal freedom scores across different countries can be accounted for by the variation in political pressures and controls on media content. This suggests that the model captures a significant portion that is 71.41% of the relationship between political pressures on media content and personal freedom scores.

Model Diagnostics

To assess whether the linear model is reliable, we checked for;
(1) linearity,
(2) nearly normal residuals, and
(3) constant variability.

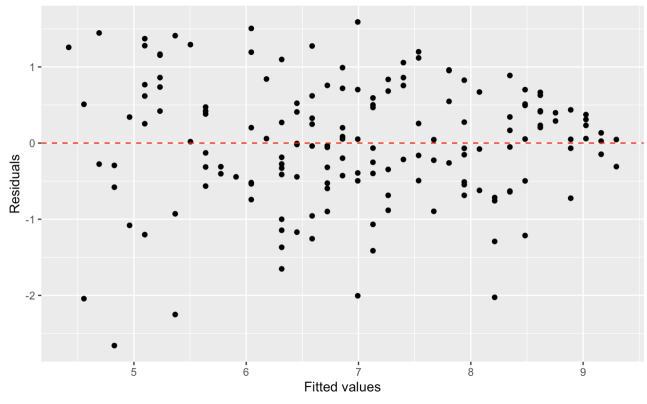

Figure 4: Residuals are randomly scattered around the horizontal line (y = 0) and don't form any specific pattern. What does this indicate about the linearity of the relationship between the two variables? The relationship between the predictors and the response variable are not linear.

Nearly Normal Residuals

Used a histogram of the residuals to check the Nearly normal residuals condition.

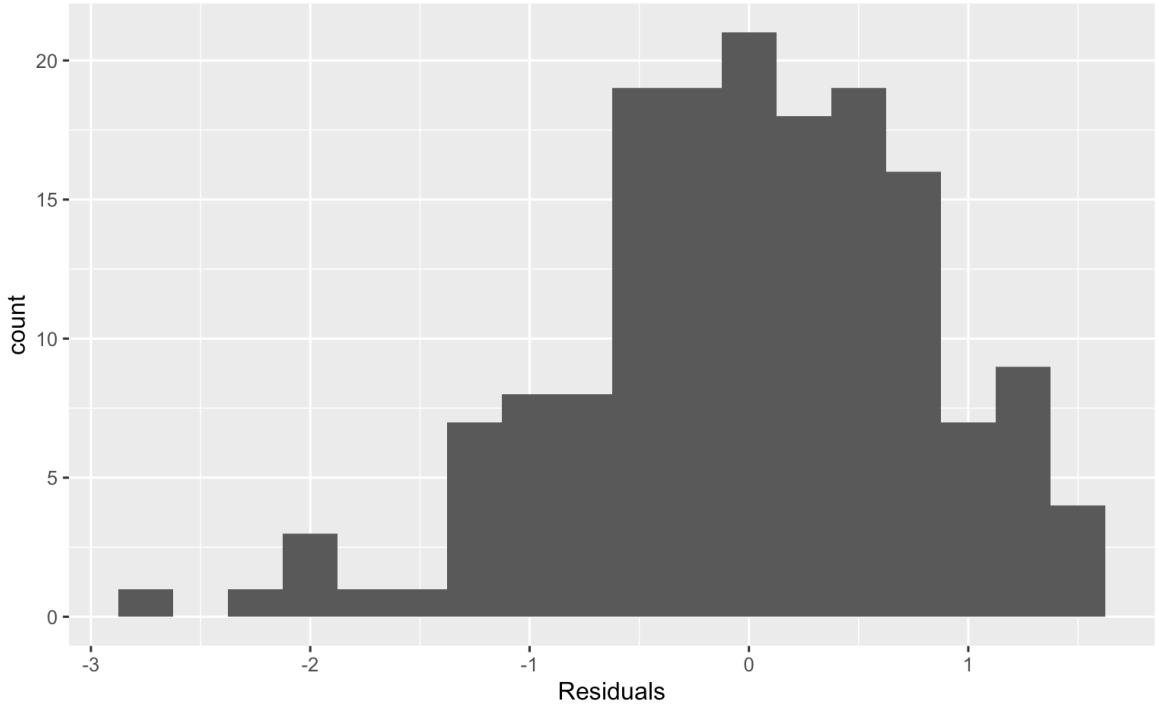

Figure 5: A Histogram of the residuals to check the Nearly normal residuals condition.

Based on the histogram, the residuals are approximately normally distributed, supporting the normality assumption.

Constant Variability

Based on the residuals vs. fitted plot, variability of the residuals is roughly constant across all levels of fitted values. The residuals fan out as fitted values increase, the variability of the residuals is not constant.

Multiple Linear Regression

Pairwise Relationships

There's a positive relationship between pf score and pf control. Additionally, there is a negative relationship between pf rank and pf score. As one's score increases, their rank decreases. Furthermore, the expression control has a negative relation with rank. As expression control increases, the rank decreases.

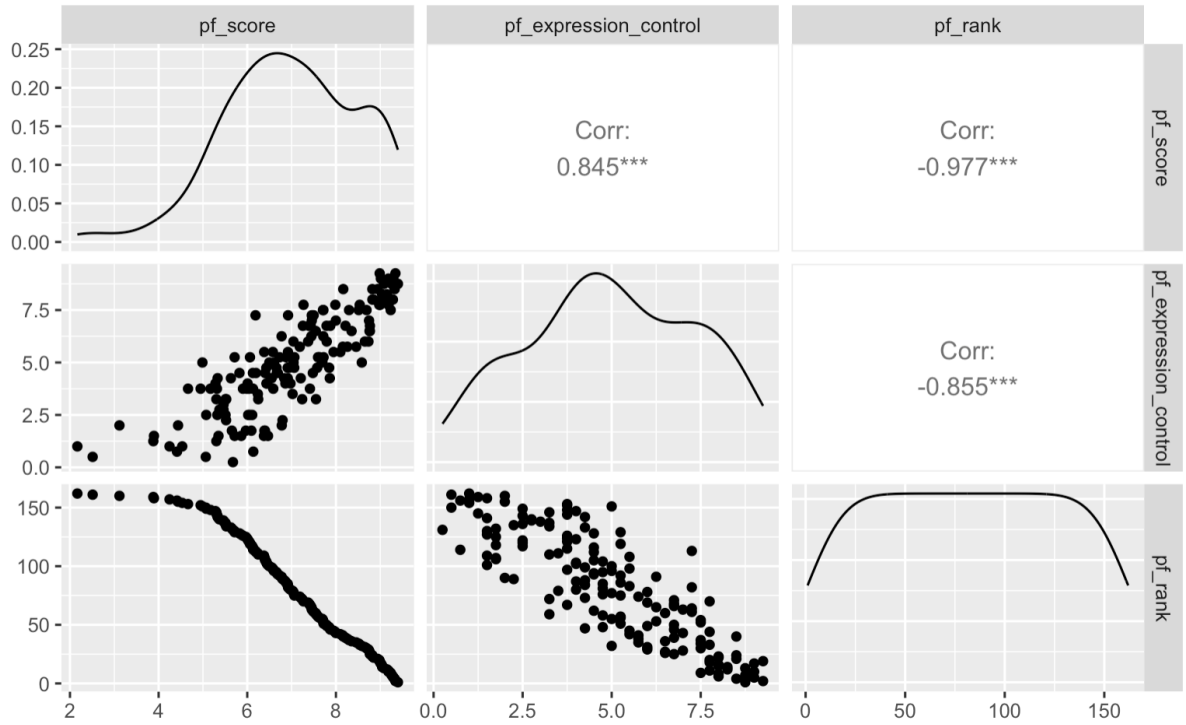

Figure 6: Pairwise relationships

Fit the Model

```
##$$
\hat{y} = b_0 + b_1 \times x_1
$$\hat{pf\\_score} = 3.361 + 0.428 \times x_1$$
```

$$\hat{pf}_{score} = 3.361 + 0.428 \times x_1$$

Intercept 3.361, the intercept represents the estimated personal freedom score when the political pressures and controls on media content index (\text{pf_expression_control}) is 0. If there were no political pressures and controls on media content (i.e., \text{pf_expression_control} = 0), the personal freedom score would be approximately 0.428. This provides a baseline level of personal freedom in the absence of political pressures on media content.

Slope 0.428, the slope represents the estimated change in the personal freedom score for each one-unit increase in the political pressures and controls on media content index (\texttt{pf_expression_control}). Specifically, for every one-unit increase in the \text{pf_expression_control} index, the personal freedom score is expected to increase by approximately 0.428 units.

This positive relationship suggests that higher political pressures and controls on media content are associated with higher personal freedom scores. This might seem counterintuitive at first, but it could imply that in environments where there are more political pressures and controls on media, there could be compensatory mechanisms or efforts in place to ensure personal freedoms are maintained or improved.

The intercept provides a starting point for personal freedom scores in the theoretical scenario where political pressures and controls on media content are completely absent.
The slope indicates the direction and strength of the relationship between political pressures on media and personal freedom scores. The positive slope suggests a direct, although not necessarily causal, relationship was increasing political pressures and controls on media content correlate with increasing personal freedom scores.

In summary, for countries with a `pf_expression_control` of 0 (those with the largest amount of political pressure on media content), we expect their mean personal freedom score to be 0.428. For every 1 unit increase in `pf_expression_control` (political pressure on media content index), we expect a country's mean personal freedom score to increase 0.428 units.

3-D Scatterplot

The 3D scatter plot visualizes the relationship between pf_expression_control (x-axis), pf_score (y-axis), and ef_legal (z-axis). Each point represents a country, with its position in the 3D space determined by these three variables. The 2D scatter plot shows the relationship between pf_expression_control and pf_score, with points colored by pf_rank.

From the 3D plot, we can observe clusters where countries with higher freedom of expression control tend to have higher overall freedom scores and varying levels of legal enforcement freedom. Outliers, represented by points deviating from the main clusters, indicate countries with unusual combinations of these freedoms. This visualization helps identify patterns and relationships that are not immediately apparent in 2D plots, such as how legal enforcement freedom interacts with the other two variables.

Using 3D scatter plots enhances the ability to visualize and interpret the relationships between multiple variables simultaneously, providing deeper insights into the data. The 3D scatter plot created using plotly in this example reveals complex relationships between freedom of expression control, overall freedom score, and legal enforcement freedom in the hfi_2016 dataset.

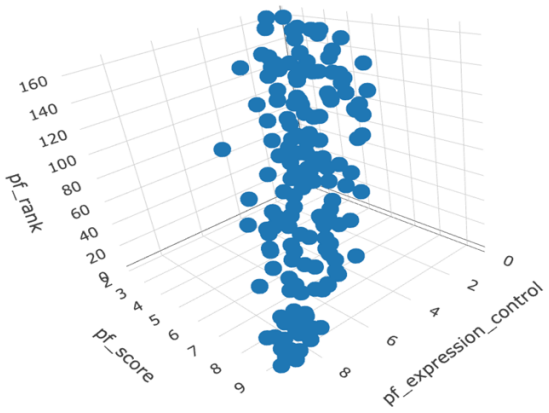

Figure 7: 3D Scatterplot

Fitting the overall model with $qualitative \times quantitative$ interaction

$$y = \beta_0 + \beta_1 \times \text{qualitative\_variable} + \beta_2 \times \text{quantitative\_variable} + \beta_3 \times (\text{qualitative\_variable} \times \text{quantitative\_variable}) + \varepsilon$$

pf_score=5.721−1.598×west_atlanticYes+0.296×pf_expression_control+0.275×(west_atlanticYes×pf_expression_control)

This equation indicates how pf_score is influenced by being in the Western Atlantic region, freedom of expression control, and their interaction. All coefficients are statistically significant, suggesting that these factors meaningfully contribute to predicting pf_score.

Using your background knowledge of F tests, what is the F test statistic and p-value for this test?

Based on an $\alpha = 0.05$ significant level, what should you conclude?

F test statistic F* = 3.284

p-value = 1.262236×10−3 (0.001262236)

For each slope, you are testing if that slope is zero (when including the other variables, the null) or if it is not zero (when including the other variables, the alternative).

Testing the Interaction Term

Based on an $\alpha = 0.05$ significant level, what should you conclude?

For the interaction term {west_atlanticYes:pf_expression_control}:

t-statistic: t*=3.283544

p-value: 0.001262236

Hypothesis Testing

Null hypothesis ($H0$): The coefficient of the interaction term is zero ($\beta3=0$)

Alternative hypothesis (HA): The coefficient of the interaction term is not zero ($\beta3\neq0$)

Calculating the t-Statistic and p-Value

t*=3.283544.

p-value

p=0.001262236.

Conclusion at α=0.05 Significance Level

To determine if the interaction term is statistically significant at the α=0.05, since p<0.05, we reject the null hypothesis.

Conclusion

Based on the *t*t-statistic of 3.283544 and a p-value of 0.001262236, we conclude that the interaction term {west_atlanticYes:pf_expression_control} is statistically significant at the α=0.05 significance level. This indicates that the interaction between west_atlanticYes and pf_expression_control significantly contributes to the prediction of pf_score.

If your interaction term was not significant, you could consider removing it.

Testing the two non-interaction terms

What are the t test statistic and p-value associated with these tests?

Based on an $\alpha = 0.05$ significant level, what should you conclude about these two predictors?

Testing the Predictors

For west_atlanticYes:

t-statistic: -3.304640

p-value: 0.001176686

For pf_expression_control:

t-statistic: 3.753059

p-value: 0.0002449722

Hypothesis Testing

Null hypothesis ($H0$): The coefficient is zero (β=0).

Alternative hypothesis (HA): The coefficient is not zero ($\beta \neq 0$)

Conclusion at α=0.05 Significance Level

For west_atlanticYes

t-statistic: t* = -3.304640

p-value: 0.001176686

Since p<0.05, we reject the null hypothesis. This means that west_atlanticYes is a statistically significant predictor of pf_score.

For pf_expression_control

t-statistic: t* = 3.753059

p-value: 0.0002449722

Since p<0.05, we reject the null hypothesis. This means that pf_expression_control is a statistically significant predictor of pf_score.

Conclusion

Based on an α=0.05 significance level, we conclude that both west_atlanticYes and pf_expression_control are statistically significant predictors of pf_score. This means that the coefficients for these predictors are significantly different from zero, indicating that they have a meaningful impact on the response variable pf_score.

Residual Assessment

By exploring the residuals from Activity 2 assess how well your model fits the data.

A tibble: 1 × 12

r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>	p.value <dbl>	df <dbl>	logLik <dbl>	AIC <dbl>	BIC <dbl>
0.7141342	0.7123476	0.7992708	399.7033	2.313214e-45	1	-192.5648	391.1297	400.3925

Interpretation of the Results

R^2 (R-squared) 0.7141342, indicates that approximately 71.41% of the variance in the response variable (pf_score) is explained by the predictor (pf_expression_control). This suggests a strong relationship between the predictor and the response variable.

Residual Standard Error (sigma) 0.7992708**,** this is the standard deviation of the residuals (prediction errors). A smaller sigma indicates that the model's predictions are close to the actual values.

F-statistic F*= 399.7033

p-value: 2.313214×10−45

The F-statistic tests whether the model explains a significant portion of the variance in the response variable. The very high F-statistic and extremely low p-value suggest that the model is highly significant.

Conclusion at α=0.05 Significance Level

1. **Significance of Predictors**:
    - Both the intercept and pf_expression_control are highly significant, given the very low p-values associated with their coefficients.
2. **Model Fit**:
    - **R^2**The high R^2 values indicate that the model explains a substantial portion of the variance in the response variable.
    - **Residual Standard Error (sigma)**: The relatively small value of sigma suggests that the model's predictions are fairly accurate.
    - **F-statistic**: The very high F-statistic with a p-value much smaller than 0.05 indicates that the overall model is statistically significant.

Based on these metrics, the model fits the data well, and both the predictor pf_expression_control and the interaction term (if considered) significantly contribute to explaining the variance in the response variable pf_score.

Proficiency for Simple and Multiple Linear Regression

Characterize whether you feel that you are proficient, aware, or unaware of each of the five course objectives (for simple and multiple linear regression only).

I have familiarized myself with the course objectives for simple and multiple linear regression, and I understand what is expected of me. As I continue to learn and develop my skills, I aim to become proficient in both simple linear regression and multiple linear regression. However, I still face challenges in interpreting multiple linear regression models.

Portfolio Plans

The portfolio will comprise of an activity most preferably the in-class activities demonstrating what I have learned, and the results will be a website. I plan to host this website on GitHub Page (github.io). My portfolio's primary focus will be to explore a dataset, analyze it, and present the findings. This project will serve as a steppingstone in my graduate studies and career, demonstrating my proficiency with R and GitHub.

I will independently develop the portfolio, using in-class activities as foundational examples. To ensure the quality of my work, I will seek feedback from Professor Bradford Dykes and my peers, incorporating their suggestions to refine my findings and improve the overall portfolio.

Through this project, I aim to derive meaningful conclusions from the chosen dataset and reflect on the insights gained. I will document my reflections in the portfolio, highlighting how the course has contributed to my personal and professional growth, as well as its impact on my community. The portfolio's direction may evolve over time as I delve deeper into the data and uncover new perspectives.

References

- Gareth, J., Daniela, W., Trevor, H., Robert T. (2023). An introduction to Statistical Learning with Applications in R (ISLR) (2nd ed.).
- Wasserman, L. (2004). All of Statistics: A Concise Course in Statistical Inference. Springer.
- DeGroot, M. H., & Schervish, M. J. (2012). Probability and Statistics (4th ed.). Pearson.