STAT0028 ICA

Group 1 - 24185360 & 20082011

19 Nov 2024

Question 1

Data Description

For this question, we are investigating a dataset which has 2022 measurements of **NOx** (nitrogen oxide) pollution content in the ambient air. There are three response variables are collected:

noxem : Sum of NOx emission of cars on this motorway.
ws : Wind speed in m/s.
humidity : Absolute humidity in the air in g/kg air.

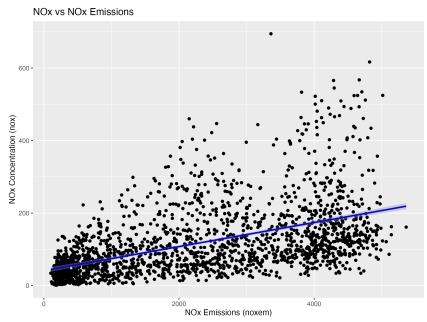

Figure 1: NOx vs noxem

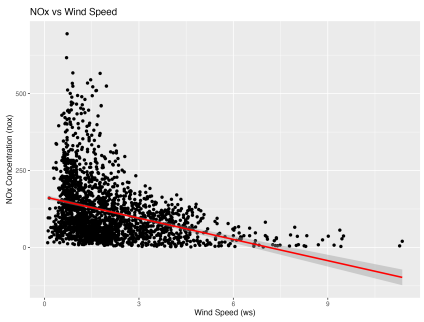

Figure 2: NOx vs windspeed

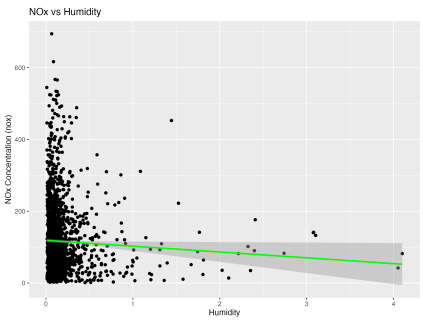

Figure 3: NOx vs humidity

Figure 4: Relationship between target variables and response variables. (Code: Appendix 1)

Figure 4 illustrates the relationship between the target variable (NOx) and the three predictor variables. While the first plot in Figure 5 (NOx vs noxem) suggests a positive linear relationship, the other two plots in Figure 6 and Figure 7 (NOx vs ws and NOx vs humidity) indicate significant non-linear associations as well as some potential outliers that might affect the model's performance. As a result, appropriate transformations will be applied to the data to improve model fitting and capture these non-linear effects more effectively.

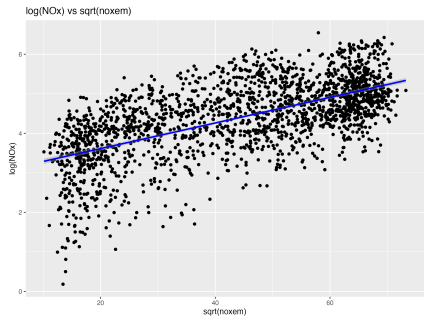

Figure 5: log(NOx) vs sqrt(noxem)

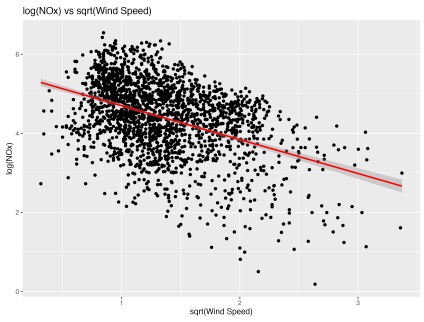

Figure 6: log(NOx) vs sqrt(ws)

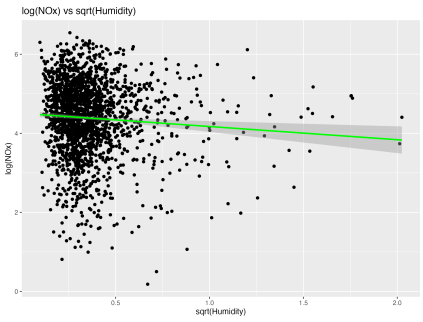

Figure 7: log(NOx) vs sqrt(hum)

Figure 8: Relationship between target variables and response variables after transformation. (Code: Appendix 2)

Statistical Models

Linear Model (Baseline Model)

The **linear model**, used as a baseline, fits the response variable nox_i as a linear combination of the original predictors humidity_i, noxem_i, and ws_i without any transformations. It assumes linear relationships, normally distributed residuals, and constant variance, but may struggle with non-linear patterns or skewed data. The mathematical equation of linear model performs as followed.

$$\text{nox}_i = \beta_0 + \beta_1 \cdot \text{humidity}_i + \beta_2 \cdot \text{noxem}_i + \beta_3 \cdot \text{ws}_i + \epsilon_i$$

Where:

nox_i : The target variable (NOx concentration for observation i).
β_0 : The intercept term.
β_1 : Effect of humidity_i (absolute humidity).
β_2 : Effect of noxem_i (NOx emissions).
β_3 : Effect of ws_i (wind speed).
ϵ_i : The residual error term for observation i.

Generalised Linear Models

The **generalized linear model (GLM)** applies transformations to address these limitations. The response variable nox_i is log-transformed, and the predictors are square-root transformed to capture non-linear relationships. Two distributions will be considered: the **Gamma distribution** and the **Gaussian distribution**. The mathematical equation of GLM is presended below.

$$\log(\text{nox}_i) = \beta_0 + \beta_1 \cdot \sqrt{\text{noxem}_i} + \beta_2 \cdot \sqrt{\text{ws}_i} + \beta_3 \cdot \sqrt{\text{humidity}_i} + \epsilon_i$$

Where:

nox_i : The target variable (NOx concentration for observation i).
$\sqrt{\text{noxem}_i}$: The square root transformation of NOx emissions for observation i.
$\sqrt{\text{ws}_i}$: The square root transformation of wind speed for observation i.
$\sqrt{\text{humidity}_i}$: The square root transformation of absolute humidity for observation i.
β_0 : The intercept term.
$\beta_1, \beta_2, \beta_3$: The coefficients of the respective predictors.
ϵ_i : The residual error term for observation i.

Results of Fitted Models

Before fitting the models, the original dataset was split into an 80% training set and a 20% testing set to enable robust model evaluation. Model comparison was conducted using MSE, MAE,R-squared and AIC as evaluation metrics. Since data transformations were applied to the GLMs, the predicted values are on a different scale compared to the baseline linear model. To ensure a fair comparison, the MSE and MAE values were recalculated on the original scale of the target variable. Table 1 below summarizes the performance of the three models on both the training and testing datasets. The code for this result in the table is given in Appendix 3, 4, 5)

Model	**Dataset**	**MSE**	**MAE**	**R-squared**	**AIC**
Linear Model (Baseline)	Training	5765.019	52.53818	0.439158	–
	Testing	4215.312	47.67469	0.410928	–
GLM with Gamma Distribution	Training	7038.046	53.15608	0.616754	–
	Testing	5440.196	47.15623	0.548968	3442.646
GLM with Gaussian Distribution	Training	5499.205	46.83815	0.6578908	–
	Testing	3803.385	40.61301	0.5974634	2735.202

Table 1: Models Performance Results including MSE, MAE, R-squared, and AIC.

Non-recommended Models

Based on the model performance results, we do not recommend the linear model fitted with the original variables (baseline model) or the GLM with a Gamma distribution. The linear model is not recommended due to its relatively poor performance across all evaluation metrics compared to the GLM with a Gaussian distribution. Specifically, its R-square values of 0.439 for the training set and 0.411 for the testing set indicate that it explains only a modest proportion of the variance in the target variable.

Similarly, while the GLM with a Gamma distribution achieves higher R-square values for both training and testing sets compared to the linear model, its MSE and MAE values are significantly higher than those of the baseline model. Therefore, this model is also not recommended.

From the data description section, it is evident that the relationships between the target variable and the predictors do not strictly follow a linear pattern. This observation is further supported by the low R-square values of the linear model, which shows the necessity for data transformations such as the logarithmic and square root transformations applied in the GLMs. Moreover, the high MSE and MAE values of the GLM with a Gamma distribution emphasizes the importance of selecting an appropriate distribution for the target variable. This result ultimately led to the adoption of the GLM with a Gaussian distribution, which is discussed further in the next section.

Recommended Models

The GLM with a Gaussian distribution is the recommended model based on the best performance metrics compared with the previous two models. As shown in Table 1, this model achieved the lowest MSE and MAE values across both the training and testing datasets, which indicates higher predictive accuracy. On the other hand, the R-square values of 0.658 for training and 0.597 for testing show that this model explains a substantial proportion of the variance in the target variable, which is better than the alternative models. The low AIC value of 2735.202 compared to GLM with Gamma distribution further supports the model's efficiency in balancing goodness-of-fit and complexity.

The Figure 9 demonstrates the diagnostic plot for Gaussian model. The **Residuals vs Fitted** plot shows no clear patterns and suggests that the model reasonably satisfies the assumptions of linearity and homoscedasticity. The **Normal Q-Q** plot indicates that residuals follow a normal distribution, supporting the appropriateness of the Gaussian assumption. The **Scale-Location** plot demonstrates consistent variance across fitted values, with minor heteroscedasticity that does not significantly affect model reliability. The **Residuals vs Leverage** plot shows that most residuals are centered around zero with no clear pattern, which demonstrates unbiased predictions and a good model fit.

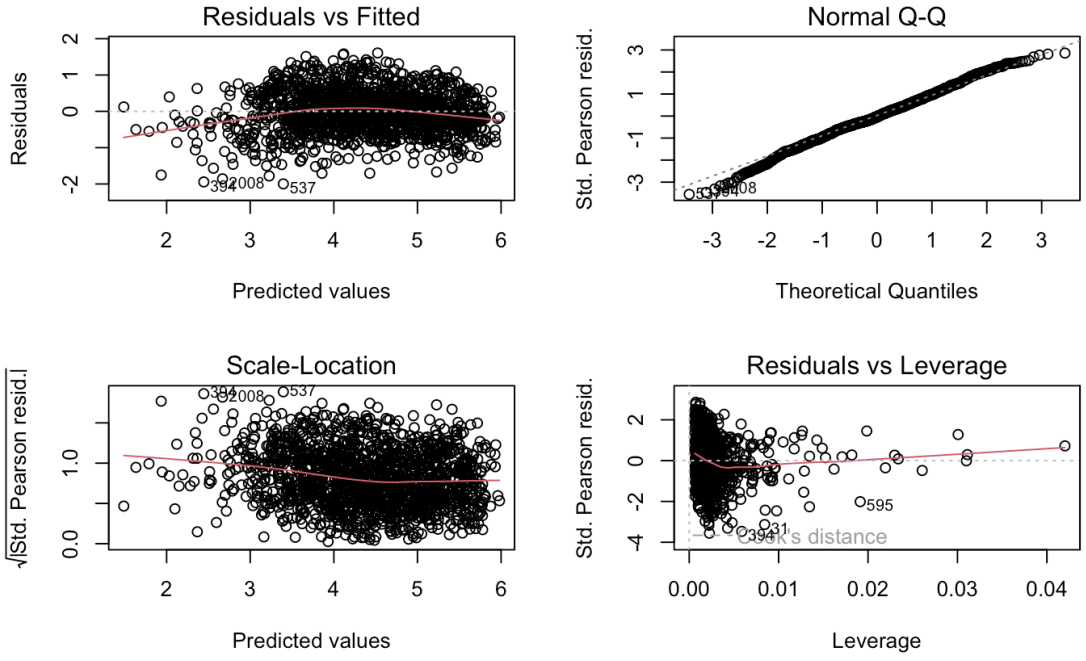

Figure 9: Diagnostic Plot for GLM Gaussian Model. (Code: Appendix 6)

Hypothesis Testing (Code: Appendix 7 and 8)

In the GLM with a Gaussian distribution, we aim to test the null hypothesis that the effects of $\sqrt{\text{noxem}}$ and $\sqrt{\text{ws}}$ as well as $\sqrt{\text{humidity}}$ and $\sqrt{\text{ws}}$ are identical. First, we conduct the test for the first pair. From the model summary, we observe that the estimated coefficient for $\sqrt{\text{noxem}}$ is 0.036, while for $\sqrt{\text{ws}}$, it is -1.041. To determine whether the difference in these coefficients is statistically significant, we formulate the hypotheses as follows:

$$H_0 : \beta_{\text{noxem}} = \beta_{\text{ws}} \quad \text{(The effects of sqrt(noxem) and sqrt(ws) are the same)}$$

$$H_a : \beta_{\text{noxem}} \neq \beta_{\text{ws}} \quad \text{(The effects of sqrt(noxem) and sqrt(ws) are different)}$$

For the hypothesis test, we reformulate the null hypothesis to assess the difference between the coefficients:

$$H_0 : \beta_{\text{noxem}} - \beta_{\text{ws}} = 0$$

$$H_a : \beta_{\text{noxem}} - \beta_{\text{ws}} \neq 0$$

Given the estimated coefficients for $\sqrt{\text{noxem}}$ and $\sqrt{\text{ws}}$ from the GLM with Gaussian distribution, we aim to calculate the 95% confidence interval for the difference between their effects. The difference in the estimated coefficients is denoted as:

$$\Delta = \hat{\beta}_{\sqrt{\text{noxem}}} - \hat{\beta}_{\sqrt{\text{ws}}}$$

The standard error of this difference is given as:

$$\text{SE}_\Delta = 0.0299854$$

The 95% confidence interval is then computed as:

$$CI = \Delta \pm t_{\alpha/2} \cdot \text{SE}_\Delta$$

Plugging in the values, we get:

$$CI = 1.076882 \pm 1.96 \cdot 0.0299854$$

$$CI = 1.076882 \pm 0.058811$$

$$CI = [1.018281, 1.135483]$$

This interval provides the range of plausible values for the true difference between the effects of $\sqrt{\text{noxem}}$ and $\sqrt{\text{ws}}$. Since the interval does not include 0, we reject the null hypothesis and conclude that there is a statistically significant difference between the effects of $\sqrt{\text{noxem}}$ and $\sqrt{\text{ws}}$ at the 95% confidence level.

Using the same method, the null hypothesis for $\sqrt{\text{humidity}}$ and $\sqrt{\text{ws}}$ is:

$$H_0 : \beta_{\text{humidity}} - \beta_{\text{ws}} = 0$$

After computation, the 95% confidence intervalis given as:

$$CI = [0.7629484, 1.063325]$$

The result also does not include 0 in the confidence interval, therefore, we can conclude that the effect of $\sqrt{\text{humidity}}$ and $\sqrt{\text{ws}}$ is distinct.

NOx Pollution Report

Introduction

The purpose of this report is to analyze the influence of the three variables **noxem** NOx emissions (amount of NOx produced by cars), wind speed, and absolute humidity on the response variable **NOx** concentrations (a key measure of air pollution) using three different typed of models. The models evaluated include a simple linear model and two generalized linear models (GLMs) using different approaches. These models aim to find the best fit to the data with the results explained in accessible terms.

Overview of the Models

The linear model directly relates NOx concentrations to the regressors without transforming the data. This straightforward approach assumes that relationships between the variables are linear and provides a baseline model for furthre comparison.

To improve model's reliability, the GLMs apply transformations to the data. NOx concentrations are log-transformed to address variability and skewness, while the regressors, such as NOx emissions, wind speed, and humidity, are square-root transformed to better capture non-linear relationships. Two GLMs with different distribution assumptions were evaluated.

Analysis of Model Results

For the results, the linear model struggled to capture the complexity of the relationships between NOx concentrations and the regressors. It did not account for non-linear patterns in the data and showed limited accuracy and reliability.

The GLM with Gamma distribution showed improvement over the linear model but still had limitations. Its assumptions about the data distribution were not fully consistent with the observed characteristics, resulting in higher errors.

The GLM with Gaussian distribution performed the best overall. By applying log and square-root transformations, it effectively captured the non-linear relationships in the data and aligned well with the distribution of NOx concentrations. This made it the most reliable model in this study.

Interpreting the model's results, the sum of NOx emissions from cars on this motorway shows the most significant positive correlation with NOx pollution in the air. On the other hand, wind speed has a negative correlation with pollution, indicating that stronger wind speeds result in lower levels of NOx pollution in the air. While humidity does not have a significant impact on pollution, it still demonstrates a negative relationship with pollution levels.

To further investigate the effects of the regression coefficients, we conducted hypothesis tests to compare the effects between **noxem** and **ws**, as well as between **humidity** and **ws**. After statistical testing, we found that the effects of both pairs are significantly distinct.

Limitations and Improvements

Considering the potential limitations, the log transformation of NOx concentrations and square-root transformations of predictors improved model performance but may complicate direct interpretation of results. Moreover, the current model assumes that the patterns observed in the data remain consistent. Any significant changes in data characteristics may require adjustments to the model or additional transformations. Lastly, though the GLM with Gaussian distribution achieves high performance, it still does not capture a substantial proportion of the variance. More complex models, such as Generalized Additive Models (GAMs), could be applied to this study to better capture intricate non-linear relationships and improve model performance.

Conclusion

This study analyzed NOx concentrations using a baseline linear model and two GLMs with Gaussian and Gamma distributions. The baseline model performed poorly due to its inability to capture non-linear relationships. The GLM with Gaussian distribution, using log-transformed responses and square-root-transformed predictors, proved most effective. Furthermore, the analysis highlights the dominant role of NOx emissions in contributing to pollution, the mitigating effect of wind speed, and the minor influence of humidity, with hypothesis tests confirming the distinct impacts of these factors.

Question 2

We dispose of the following equation for the expression of D_i:

$$D_i = \frac{1}{p\hat{\sigma}^2}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)^T \mathbf{X}^T\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right) \tag{1}$$

We are given the following:

$$\mathbf{C}^{-1} = (\mathbf{X}^T\mathbf{X})^{-1} \tag{2}$$

$$e_i = y_i - \hat{y}_i \tag{3}$$

$$\hat{y}_i = \mathbf{x}_i^T\hat{\boldsymbol{\beta}} \tag{4}$$

And:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X} \tag{5}$$

where $\mathbf{H}$ satisfies $\mathbf{HY} = \hat{\mathbf{y}}$ and $h_{i,i}$ is the i^{th} diagonal term of the matrix, which can be written as:

$$h_{i,i} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i = \mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i \tag{6}$$

We start by writing the equation of $\hat{\boldsymbol{\beta}}_{(i)}$, representing the least squares estimator of $\boldsymbol{\beta}$ without the i^{th} observation. We introduce $\mathbf{d}_{(i)} = \mathbf{d} - \boldsymbol{x}_i y_i = \mathbf{X}^T\mathbf{Y} - \boldsymbol{x}_i y_i$, where $\mathbf{d} = \mathbf{X}^T\mathbf{Y}$. Then, the least square estimate $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}} = \mathbf{C}^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{C}^{-1}\mathbf{d}$$

Then, removing the i^{th} observation:

$$\hat{\boldsymbol{\beta}}_{(i)} = \mathbf{C}_{(i)}^{-1}\mathbf{d}_{(i)}$$

By definition:

$$\mathbf{C}_{(i)}^{-1} = \mathbf{C}^{-1} - \mathbf{x}_i\mathbf{x}_i^T$$

Which can be rewritten as:

$$\mathbf{C}_{(i)}^{-1} = \mathbf{C}^{-1} - \frac{\mathbf{C}^{-1}\mathbf{x}_i\mathbf{x}_i^T\mathbf{C}^{-1}}{1 - \mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i}$$

Replacing in the difference expression:

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \mathbf{C}^{-1}\mathbf{d} - \mathbf{C}_{(i)}^{-1}\mathbf{d}_{(i)} \tag{7}$$

$$= \mathbf{C}^{-1}\mathbf{d} - \left(\mathbf{C}^{-1} - \frac{\mathbf{C}^{-1}\mathbf{x}_i\mathbf{x}_i^T\mathbf{C}^{-1}}{1 - \mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i}\right)\mathbf{d}_{(i)}$$

$$= \mathbf{C}^{-1}\mathbf{d} - \left(\mathbf{C}^{-1} - \frac{\mathbf{C}^{-1}\mathbf{x}_i\mathbf{x}_i^T\mathbf{C}^{-1}}{1 - \mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i}\right)(\mathbf{d} - \mathbf{x}_i y_i)$$

$$= \mathbf{C}^{-1}\mathbf{d} - \mathbf{C}^{-1}\mathbf{d} + \mathbf{C}^{-1}\mathbf{x}_i y_i - \frac{\mathbf{C}^{-1}\mathbf{x}_i\mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{d}}{1 - \mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i} + \frac{\mathbf{C}^{-1}\mathbf{x}_i\mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i y_i}{1 - \mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i}$$

$$= \mathbf{C}^{-1}\mathbf{x}_i y_i - \frac{\mathbf{C}^{-1}\mathbf{x}_i\mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{d}}{1 - \mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i} + \frac{\mathbf{C}^{-1}\mathbf{x}_i\mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i y_i}{1 - \mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i} \tag{8}$$

Identifying the expression of $h_{i,i}$ and of $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \mathbf{C}^{-1}\mathbf{x}_i y_i + \frac{\mathbf{C}^{-1}\mathbf{x}_i(\mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i)y_i}{1 - \mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i} - \frac{\mathbf{C}^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{C}^{-1}\mathbf{X}^T\mathbf{Y})}{1 - \mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i}$$

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \mathbf{C}^{-1}\mathbf{x}_i\mathbf{y}_i + \frac{\mathbf{C}^{-1}\mathbf{x}_i h_{i,i}\mathbf{y}_i}{1-h_{i,i}} - \frac{\mathbf{C}^{-1}\mathbf{x}_i\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}{1-h_{i,i}} \tag{9}$$

And therefore, remembering $\hat{y}_i = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$, we deduce:

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \mathbf{C}^{-1}\mathbf{x}_i y_i + \frac{\mathbf{C}^{-1}\mathbf{x}_i h_{i,i} y_i}{1-h_{i,i}} - \frac{\mathbf{C}^{-1}\mathbf{x}_i(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}{1-h_{i,i}} \tag{10}$$

$$= \mathbf{C}^{-1}\mathbf{x}_i\left(y_i + \frac{h_{i,i}y_i}{1-h_{i,i}}\right) - \frac{\mathbf{C}^{-1}\mathbf{x}_i\hat{y}_i}{1-h_{i,i}}$$

$$= \mathbf{C}^{-1}\mathbf{x}_i\left(1 + \frac{h_{i,i}}{1-h_{i,i}}\right)y_i - \frac{\mathbf{C}^{-1}\mathbf{x}_i\hat{y}_i}{1-h_{i,i}}$$

$$= \mathbf{C}^{-1}\mathbf{x}_i\left(\frac{1-h_{i,i}}{1-h_{i,i}} + \frac{h_{i,i}}{1-h_{i,i}}\right)y_i - \frac{\mathbf{C}^{-1}\mathbf{x}_i\hat{y}_i}{1-h_{i,i}}$$

$$= \mathbf{C}^{-1}\mathbf{x}_i\left(\frac{1-h_{i,i}}{1-h_{i,i}} + \frac{h_{i,i}}{1-h_{i,i}}\right)y_i - \frac{\mathbf{C}^{-1}\mathbf{x}_i\hat{y}_i}{1-h_{i,i}}$$

$$= \mathbf{C}^{-1}\mathbf{x}_i\left(\frac{1-h_{i,i}+h_{i,i}}{1-h_{i,i}}\right)y_i - \frac{\mathbf{C}^{-1}\mathbf{x}_i\hat{y}_i}{1-h_{i,i}}$$

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \mathbf{C}^{-1}\mathbf{x}_i\left(\frac{1}{1-h_{i,i}}\right)y_i - \frac{\mathbf{C}^{-1}\mathbf{x}_i\hat{y}_i}{1-h_{i,i}} \tag{11}$$

Then, factorizing and identifying the expression of the error e_i, we get:

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \mathbf{C}^{-1}\mathbf{x}_i\left(\frac{y_i}{1-h_{i,i}} - \frac{\hat{y}_i}{1-h_{i,i}}\right)$$

$$= \mathbf{C}^{-1}\mathbf{x}_i\left(\frac{y_i-\hat{y}_i}{1-h_{i,i}}\right)$$

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \mathbf{C}^{-1}\mathbf{x}_i\frac{e_i}{1-h_{i,i}} \tag{12}$$

which is the result we were expecting to find.
We can rewrite the expression of D_i as:

$$D_i = \frac{1}{p\hat{\sigma}^2}\left(\mathbf{C}^{-1}\mathbf{x}_i\frac{e_i}{1-h_{i,i}}\right)\mathbf{X}^T\mathbf{X}\left(\mathbf{C}^{-1}\mathbf{x}_i\frac{e_i}{1-h_{i,i}}\right) \tag{13}$$

Appendix: Q1 R code

Listing 1: Correlations between variables

```r
library(ggplot2)

ggplot(emissionssw, aes(x = noxem, y = nox)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "NOx vs NOx Emissions",
       x = "NOx Emissions (noxem)",
       y = "NOx Concentration (nox)")
ggsave("NOx_vs_noxem.png", plot = plot1, width = 8, height = 6, dpi = 300)

ggplot(emissionssw, aes(x = ws, y = nox)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  labs(title = "NOx vs Wind Speed",
       x = "Wind Speed (ws)",
       y = "NOx Concentration (nox)")
ggsave("NOx_vs_WindSpeed.png", plot = plot2, width = 8, height = 6, dpi = 300)


ggplot(emissionssw, aes(x = humidity, y = nox)) +
  geom_point() +
  geom_smooth(method = "lm", col = "green") +
  labs(title = "NOx vs Humidity",
       x = "Humidity",
       y = "NOx Concentration (nox)")
```

Listing 2: Correlations between variables after transformation

```
ggplot(emissionssw, aes(x = sqrt(noxem), y = log(nox))) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "log(NOx) vs sqrt(noxem)",
       x = "sqrt(noxem)",
       y = "log(NOx)")
ggsave("NOx_vs_sqrtnoxem.png", plot = plot4, width = 8, height = 6, dpi = 300)

ggplot(emissionssw, aes(x = sqrt(ws), y = log(nox))) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  labs(title = "log(NOx) vs sqrt(Wind Speed)",
       x = "sqrt(Wind Speed)",
       y = "log(NOx)")
ggsave("NOx_vs_sqrtws.png", plot = plot5, width = 8, height = 6, dpi = 300)


ggplot(emissionssw, aes(x = sqrt(humidity), y = log(nox))) +
  geom_point() +
  geom_smooth(method = "lm", col = "green") +
  labs(title = "log(NOx) vs sqrt(Humidity)",
       x = "sqrt(Humidity)",
       y = "log(NOx)")
ggsave("NOx_vs_sqrthum.png", plot = plot6, width = 8, height = 6, dpi = 300)
```

Listing 3: Linear Model (Baseline Model)

```r
# Model fitting using lm function
model_linear <- lm(nox ~ humidity + noxem + ws, data = train_data)
summary(model_linear)

# Calculate and print the model performance metrics
train_predictions <- predict(model_linear, newdata = train_data)
train_actual <- train_data$nox

ss_total_train <- sum((train_actual - mean(train_actual))^2)
ss_residual_train <- sum((train_actual - train_predictions)^2)

r_squared_train <- 1 - (ss_residual_train / ss_total_train)
mse_train <- mean((train_actual - train_predictions)^2)

mae_train <- mean(abs(train_actual - train_predictions))

test_predictions <- predict(model_linear, newdata = test_data)
test_actual <- test_data$nox

mse_test <- mean((test_actual - test_predictions)^2)
ss_total_test <- sum((test_actual - mean(test_actual))^2)
ss_residual_test <- sum((test_actual - test_predictions)^2)
r_squared_test <- 1 - (ss_residual_test / ss_total_test)

mae_test <- mean(abs(test_actual - test_predictions))

cat("Linear Model (Baseline Model) Performance:\n")
cat("Training Data:\n")
cat("Mean Squared Error (Training):", mse_train, "\n")
cat("Mean Absolute Error (Training):", mae_train, "\n")
cat("R-squared (Training):", r_squared_train, "\n")
cat("\nTesting Data:\n")
cat("Mean Squared Error (Testing):", mse_test, "\n")
cat("Mean Absolute Error (Testing):", mae_test, "\n")
cat("R-squared (Testing):", r_squared_test, "\n")
cat("AIC:", AIC(model_linear), "\n")
```

Listing 4: GLM with Gamma Distribution

```r
# Model fitting using glm function
model_gamma <- glm(log(nox) ~ sqrt(noxem) + sqrt(ws) + sqrt(humidity),
    family = Gamma(link = "log"), data = train_data)

summary(model_gamma)

# Calculate and print the model performance metrics
# MSE and MAE for original scale
train_predictions <- predict(model_gamma, newdata = train_data, type = "response")
train_predictions_original <- exp(train_predictions)
train_actual_original <- train_data$nox

mse_train_original <- mean((train_actual_original - train_predictions_original)^2)
mae_train <- mean(abs(train_actual_original - train_predictions_original))

test_predictions <- predict(model_gamma, newdata = test_data, type = "response")
test_predictions_original <- exp(test_predictions)
test_actual_original <- test_data$nox

mse_test_original <- mean((test_actual_original - test_predictions_original)^2)
mae_test <- mean(abs(test_actual_original - test_predictions_original))

cat("GLM with Gamma Distribution Model Performance:\n")
cat("Mean Squared Error (Original Scale):\n")
cat("Training Data:", mse_train_original, "\n")
cat("Testing Data:", mse_test_original, "\n")
cat("Mean Absolute Error (Training):", mae_train, "\n")
cat("Mean Absolute Error (Testing):", mae_test, "\n")

# Model performance for transformed scale
train_actual <- log(train_data$nox)

ss_total_train <- sum((train_actual - mean(train_actual))^2)
ss_residual_train <- sum((train_actual - train_predictions)^2)

r_squared_train <- 1 - (ss_residual_train / ss_total_train)

mse_train <- mean((train_actual - train_predictions)^2)
mae_train <- mean(abs(train_actual - train_predictions))

test_actual <- log(test_data$nox)

mse_test <- mean((test_actual - test_predictions)^2)
mae_test <- mean(abs(test_actual - test_predictions))

ss_total_test <- sum((test_actual - mean(test_actual))^2)
ss_residual_test <- sum((test_actual - test_predictions)^2)
r_squared_test <- 1 - (ss_residual_test / ss_total_test)

cat("GLM with Gamma Distribution Model Performance:\n")
cat("Training Data:\n")
cat("Mean Squared Error (Training):", mse_train, "\n")
cat("Mean Absolute Error (Training):", mae_train, "\n")
cat("R-squared (Training):", r_squared_train, "\n")
cat("\nTesting Data:\n")
cat("Mean Squared Error (Testing):", mse_test, "\n")
cat("Mean Absolute Error (Testing):", mae_test, "\n")
cat("R-squared (Testing):", r_squared_test, "\n")
cat("AIC:", AIC(model_gamma), "\n")
```

Listing 5: GLM with Gaussian Distribution

```r
# Fit the model
model_gaussian <- glm(
  log(nox) ~ sqrt(noxem) + sqrt(ws) + sqrt(humidity),
  family = gaussian(link = "identity"),
  data = train_data
)

summary(model_gaussian)

# Calculate and print the model performance metrics
# MSE and MAE for original scale

train_predictions <- predict(model_gaussian, newdata = train_data, type = "response")
train_predictions_original <- exp(train_predictions)
train_actual_original <- train_data$nox

mse_train_original <- mean((train_actual_original - train_predictions_original)^2)
mae_train <- mean(abs(train_actual_original - train_predictions_original))

test_predictions <- predict(model_gaussian, newdata = test_data, type = "response")
test_predictions_original <- exp(test_predictions)
test_actual_original <- test_data$nox

mse_test_original <- mean((test_actual_original - test_predictions_original)^2)
mae_test <- mean(abs(test_actual_original - test_predictions_original))

cat("GLM␣with␣Gaussian␣Distribution␣Model␣Performance:\n")
cat("Mean␣Squared␣Error␣(Original␣Scale):\n")
cat("Training␣Data:", mse_train_original, "\n")
cat("Testing␣Data:", mse_test_original, "\n")
cat("Mean␣Absolute␣Error␣(Training):", mae_train, "\n")
cat("Mean␣Absolute␣Error␣(Testing):", mae_test, "\n")

# For transformed scale
train_actual <- log(train_data$nox)

ss_total_train <- sum((train_actual - mean(train_actual))^2)
ss_residual_train <- sum((train_actual - train_predictions)^2)

r_squared_train <- 1 - (ss_residual_train / ss_total_train)

mse_train <- mean((train_actual - train_predictions)^2)

test_actual <- log(test_data$nox)

mse_test <- mean((test_actual - test_predictions)^2)

ss_total_test <- sum((test_actual - mean(test_actual))^2)
ss_residual_test <- sum((test_actual - test_predictions)^2)
r_squared_test <- 1 - (ss_residual_test / ss_total_test)

cat("GLM␣with␣Gaussian␣Distribution␣Model␣Performance:\n")
cat("Training␣Data:\n")
cat("Mean␣Squared␣Error␣(Training):", mse_train, "\n")
cat("R-squared␣(Training):", r_squared_train, "\n")
cat("\nTesting␣Data:\n")
cat("Mean␣Squared␣Error␣(MSE):", mse_test, "\n")
cat("R-squared␣(Testing):", r_squared_test, "\n")
cat("AIC:", AIC(model_gaussian), "\n")
```

Listing 6: Diagnostic plot for GLM with Gaussian Distribution

```
par(mfrow = c(2, 2))
plot(model_gaussian)
```

Listing 7: Hypothesis Testing 1

```
coefs <- coef(model_gaussian)
cov_matrix <- vcov(model_gaussian)

coef_diff <- coefs["sqrt(noxem)"] - coefs["sqrt(ws)"]

se_diff <- sqrt(
  cov_matrix["sqrt(noxem)", "sqrt(noxem)"] +
  cov_matrix["sqrt(ws)", "sqrt(ws)"] -
  2 * cov_matrix["sqrt(noxem)", "sqrt(ws)"]
)

ci_lower <- coef_diff - 1.96 * se_diff
ci_upper <- coef_diff + 1.96 * se_diff

cat("Difference in coefficients (sqrt(noxem) - sqrt(ws)): ", coef_diff, "\n")
cat("95% Confidence Interval: [", ci_lower, ", ", ci_upper, "]\n")
cat("se:", se_diff, "\n")

if (ci_lower <= 0 && ci_upper >= 0) {
  cat("Fail to reject the null hypothesis:
  The effects of noxem and ws are not significantly different.\n")
} else {
  cat("Reject the null hypothesis:
  The effects of noxem and ws are significantly different.\n")
}
```

Listing 8: Hypothesis Testing 2

```
coefs <- coef(model_gaussian)
cov_matrix <- vcov(model_gaussian)

coef_diff <- coefs["sqrt(humidity)"] - coefs["sqrt(ws)"]

se_diff <- sqrt(
  cov_matrix["sqrt(humidity)", "sqrt(humidity)"] +
  cov_matrix["sqrt(ws)", "sqrt(ws)"] -
  2 * cov_matrix["sqrt(humidity)", "sqrt(ws)"]
)

ci_lower <- coef_diff - 1.96 * se_diff
ci_upper <- coef_diff + 1.96 * se_diff

cat("Difference in coefficients (sqrt(humidity) - sqrt(ws)): ", coef_diff, "\n")
cat("95% Confidence Interval: [", ci_lower, ", ", ci_upper, "]\n")
cat("se:", se_diff, "\n")

if (ci_lower <= 0 && ci_upper >= 0) {
  cat("Fail to reject the null hypothesis:
  The effects of noxem and ws are not significantly different.\n")
} else {
  cat("Reject the null hypothesis: The effects of noxem and ws are significantly different.\n")
}
```