TU Dortmund

Introductory Case Studies

Project III: Regression Analysis

Lecturers:

Dr. Crystal Wiedner

Dr. Marlies Hafer

Dr. Rouven Michels

Author: Mohaiminul Islam

Group number: 5

Group members: Md Zihad Hossain, , Mosfiqun Nahid Hassan, Ilham Pambudi

December 15, 2024

Contents

1 Introduction

Bike rental services offer an efficient and eco-friendly transportation option, with demand often influenced by factors such as weather conditions, time of day, and seasonality. This project investigates the impact of various variables on bike rental demand. To explore the relationship between dependent and independent variables, linear and polynomial regression models are employed and evaluated comparatively using RSE and R^2. Additionally, backward selection based on the Akaike Information Criterion (AIC) on the polynomial regression is utilized to eliminate unnecessary variables and reduce model complexity. The resulting model coefficients are interpreted and assessed for statistical significance. Furthermore, the underlying assumptions of the regression models are evaluated, and the validity of the parameter estimates is critically examined.

The report is structured as follows: Section 2 provides an overview of the dataset and outlines the project's objectives. Section 3 details the statistical methods and their assumptions. Section 4 applies these techniques to the dataset and interprets the results, while Section 5 summarizes the findings, discusses implications, and offers suggestions for future research.

2 Problem statement

2.1 Data Description and Data Quality

The dataset analyzed in this report was provided by the instructors of the *Case Studies I* course at TU Dortmund University during the winter term 2023/24. It is a reduced version of the *Bike Sharing* dataset from the UC Irvine Machine Learning Repository, containing 731 observations across 9 variables. These include 8 independent variables (`mnth`, `weekday`, `workingday`, `weathersit`, `temp`, `atemp`, `hum`, and `windspeed`) and one dependent variable (`cnt`), which represents the daily number of bike rentals.

The dataset contains no missing values. Continuous features such as `temp`, `atemp`, `hum`, and `windspeed` are appropriately scaled or normalized, while categorical variables such as `mnth`, `weekday`, `workingday`, and `weathersit` are encoded as integers to facilitate compatibility with regression analysis. Most features are self-explanatory, except `atemp`, which represents the "feeling temperature" in Celsius. This clean and well-structured

dataset provided a reliable foundation for the modeling and analysis conducted in this report.

2.2 Project objective

The primary objective of this report is to analyze the influence of various predictors on daily bike rental demand through linear regression modeling techniques and to assess the goodness of fit of the resulting models. Furthermore, the analysis seeks to identify and exclude non-influential covariates to improve the model's predictive performance and interpretability. Finally, the report evaluates potential violations of linear regression assumptions and examines their implications for the validity of the model's results.

3 Statistical methods

3.1 Linear Regression

Linear regression is a statistical technique used to model the relationship between a dependent variable y and one or more independent variables X. In this study, a multiple linear regression model was employed, expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \tag{1}$$

where X_j represents the j-th predictor variable, β_j quantifies the association between X_j and the response variable Y, and ϵ is the error term. The coefficient β_j is interpreted as the average effect on Y of a one-unit increase in X_j, holding all other predictors constant (James et al., 2013, p. 71–72).

For $i = 1, 2, \ldots, n$, the model can be represented in matrix notation as:

$$Y = X\beta + \epsilon, \tag{2}$$

where Y is an $n \times 1$ vector of observed responses, X is an $n \times (p+1)$ matrix of predictors (including a column of ones for the intercept), β is a $(p+1) \times 1$ vector of regression coefficients, and ϵ is an $n \times 1$ vector of error terms.(Hastie et al., 2009, p. 45)

3.2 Polynomial Extension

The multiple linear regression model can be extended to account for nonlinear relationships between predictors and the response variable, resulting in a polynomial regression model. For instance, if the relationship between X_2 and Y is quadratic, Equation (1) can be modified as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2 + \cdots + \beta_{p+1} X_p + \epsilon. \tag{3}$$

This formulation includes transformed versions of predictors, such as X_2^2, allowing the model to capture non-linear patterns in the data. Thus, polynomial regression represents an extension of linear regression, enabling the explanation of non-linear relationships within the same modeling framework (James et al., 2013, p. 91–92).

3.3 Assumptions Regarding Linear Regression

The classical linear regression model is based on several key assumptions. First, there is a linear relationship between the dependent variable Y and the independent variables X. Second, the errors in the model are assumed to have a mean of zero, $E(\epsilon) = 0$. Third, the errors are homoscedastic, meaning their variance remains constant across all observations, $\text{Var}(\epsilon_i) = \sigma^2$. Additionally, the errors are assumed to be uncorrelated, so the covariance between errors at different observations is zero, $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. These two assumptions together imply that the error covariance matrix is $\text{Cov}(\epsilon) = E(\epsilon\epsilon^T) = \sigma^2 I$, where I is the identity matrix. Finally, the errors are assumed to follow a normal distribution with a mean of zero and constant variance, $\epsilon \sim N(0, \sigma^2 I)$. These assumptions are fundamental to the validity of inference and prediction in linear regression models (Fahrmeir et al., 2021, p. 90–93).

3.4 Parameter Estimation

In multiple linear regression, the true parameters $\beta_0, \beta_1, \ldots, \beta_p$ are unknown and must be estimated from the data. The estimation is performed using the Ordinary Least Squares (OLS) method, which minimizes the Residual Sum of Squares (RSS). The Residual Sum of Squares (RSS) is defined as the sum of the squared differences between the observed

values Y and the predicted values $\hat{Y}$, given by:

$$\text{RSS} = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = (Y - X\beta)^T (Y - X\beta),$$

where e_i represents the residual for the i-th observation, Y_i is the observed value of the dependent variable, and $\hat{Y}_i$ is the predicted value based on the estimated coefficients.

The OLS method estimates the regression coefficients β by minimizing the RSS:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum e_i^2 \right) = \arg\min_{\beta} \left(\|Y - X\beta\|^2 \right).$$

This optimization problem is solved by taking the derivative of the RSS with respect to β and setting it equal to zero:

$$\frac{\partial \text{RSS}}{\partial \beta} = -2X^T (Y - X\beta) = 0.$$

Solving for β, we obtain the OLS estimates of the regression coefficients:

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

These estimates are unbiased, consistent, and efficient under the assumptions of the linear regression model, including linearity, independence, homoscedasticity, and normality of the error terms.(Hastie et al., 2009, p. 44-45)

3.5 Confidence Interval of Parameter Estimation

Confidence intervals provide a range of values that are likely to contain the true parameter with a specified probability, such as 0.95. For simple linear regression, the 95 % confidence interval for the coefficients β_0 and β_1 is approximately:

$$\hat{\beta}_0 \pm 2 \cdot \text{SE}(\hat{\beta}_0), \quad \hat{\beta}_1 \pm 2 \cdot \text{SE}(\hat{\beta}_1),$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated coefficients, and $\text{SE}(\hat{\beta}_0)$ and $\text{SE}(\hat{\beta}_1)$ are their respective standard errors, calculated as:

$$\text{SE}(\hat{\beta}_0)^2 = \frac{\sigma^2}{n} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n \cdot \bar{x}^2}, \quad \text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

where $\sigma^2 = \text{Var}(\epsilon)$ is the error variance, n is the number of observations, $\bar{x}$ is the mean of the predictors, and $\sum_{i=1}^{n}(x_i - \bar{x})^2$ represents the variability in the predictor variable x.

These formulas assume homoscedastic and uncorrelated errors. The confidence interval implies that if repeated samples were taken and intervals were constructed for each sample, approximately 95% of those intervals would contain the true values of β_0 and β_1.(James et al., 2013, p. 65-66)

3.6 Significance of Parameter

In the multiple regression model, the goal is to determine whether a specific independent variable is a significant predictor, which implies assessing if the corresponding parameter estimate is significantly different from 0. This test evaluates whether there is a relationship between the predictor variable and the dependent variable. Based on the multiple regression equation, the hypothesis test can be formulated as follows:

$$H_0 : \beta_j = 0 \quad \text{(null hypothesis: no relationship between } x_j \text{ and } Y\text{)}$$

$$H_1 : \beta_j \neq 0 \quad \text{(alternative hypothesis: a significant relationship exists between } x_j \text{ and } Y\text{)}$$

The t-statistic is employed as the test statistic for this hypothesis test:

$$t = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)}$$

Here, t follows a t-distribution with $n - p$ degrees of freedom, where n is the number of observations, and p is the number of predictors in the model. The t-statistic measures how far $\hat{\beta}_j$ is from 0 in terms of its standard error. From this, the p-value can be calculated, representing the probability of observing such an extreme t-value under the assumption that the null hypothesis is true. A small p-value, typically less than 0.05, leads to rejecting the null hypothesis, indicating that the predictor x_j is a meaningful variable in explaining Y.(James et al., 2013, p. 67-68)

3.7 Goodness of Fit Statistics

3.7.1 Residual Standard Error (RSE)

The Residual Standard Error (RSE) is an estimate of the standard deviation of the error term ϵ. It measures the average deviation of the observed responses y_i from the true regression line, with smaller RSE values indicating a closer fit of the regression model to the data. The RSE for a multiple linear regression model with p predictors is calculated using the formula:

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where n represents the number of observations, p is the number of predictors, y_i is the actual observed value for the i-th observation, and $\hat{y}_i$ is the corresponding predicted value.

The Residual Standard Error (RSE) measures the lack of fit of a regression model, but its interpretation can be challenging as it is expressed in the units of the response variable Y.(James et al., 2013, p. 68-69)

3.7.2 R^2 Statistic

The R^2 statistic offers a scale-independent alternative, measuring the proportion of variability in Y explained by the predictors X. It ranges from 0 to 1, with values closer to 1 indicating a better fit. The formula for R^2 is:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}},$$

where

$$\text{TSS} = \sum(y_i - \bar{y})^2$$

is the Total Sum of Squares, representing the total variability in Y, and RSS is the Residual Sum of Squares, indicating the unexplained variability.

Thus, R^2 reflects the proportion of variability explained by the model. High R^2 values suggest a good fit, while low values indicate poor explanatory power, potentially due to model misspecification, high error variance, or both.(James et al., 2013, p. 68-69)

3.8 Model Selection Criterion

Although Residual Sum of Squares (RSS) and R^2 are commonly used to assess the goodness of fit of a regression model, these metrics are based on the training error. Consequently, models selected solely using RSS or R^2 often underestimate the test error, leading to overfitting. To address this limitation, alternative statistical criteria such as C_p, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Adjusted R^2 are employed. These metrics incorporate penalties for model complexity and provide a more robust basis for model selection by balancing fit and generalizability(James et al., 2013, p. 232-233).In this project, only AIC was employed.

3.8.1 Akaike Information Criterion (AIC)

In the context of a linear regression model with Gaussian errors, maximum likelihood estimation coincides with least squares. For such cases, the AIC is expressed as:

$$\text{AIC} = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right),$$

where RSS is the residual sum of squares, d is the number of predictors in the model, and $\hat{\sigma}^2$ is the estimated error variance.

Typically, $\hat{\sigma}^2$ is estimated using the full model that includes all predictors. The AIC statistic adjusts the training RSS by adding a penalty term $2d\hat{\sigma}^2$, which accounts for the tendency of training error to underestimate the test error. This penalty increases with the number of predictors d, reflecting the decrease in training RSS as model complexity grows. Consequently, AIC discourages overfitting by balancing the trade-off between model fit and complexity, helping to select a model that generalizes well to unseen data.(James et al., 2013, p. 233-234)

3.9 Variable Selection Method

Variable selection refers to the technique of selecting a subset of predictors that are genuinely associated with the response from a set of predictors. Among several variable selection techniques, the backward selection method with AIC as the criterion is employed in this project.

3.9.1 Backward Selection Method with AIC

Backward selection starts with all variables included in the model. The AIC value is calculated for this full model. The process then iteratively removes one predictor at a time, recalculating the AIC after each removal. At each step, the predictor whose removal results in the greatest reduction in AIC is eliminated. This process continues until removing any additional predictors no longer reduces the AIC value. The final model is the one that achieves the lowest AIC and balances goodness of fit with model simplicity.(James et al., 2013, p. 79)

3.10 Variance Inflation Factor (VIF)

The Variance Inflation Factor (VIF) is a statistical measure used to detect multicollinearity in a set of predictors in a regression model. Multicollinearity occurs when two or more predictors are highly correlated, making it difficult to determine the independent effect of each predictor on the response variable.

The VIF for a predictor X_j is calculated as:

$$\text{VIF}(X_j) = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination obtained by regressing X_j on all other predictors in the model. In the case of no multicollinearity, the VIF value for a predictor should be 1. A VIF greater than 1 indicates the presence of multicollinearity, with higher values suggesting more severe multicollinearity.(Kim, 2019)

3.11 Residual Plot

The residual plot serves as a diagnostic tool in regression analysis, assessing the model's adequacy and detect potential issues such as non-linearity, heteroscedasticity, and outliers. It is a scatter plot of residuals $e_i = Y_i - \hat{Y}_i$ on the Y-axis against the corresponding fitted values $\hat{Y}_i$ on the X-axis. Ideally, in a well-fitting model, the residuals should be randomly scattered around zero. A constant spread of residuals across predictor values suggests homoscedasticity, while a varying spread indicates heteroscedasticity. Any curvature or patterns in the residuals suggest non-linearity, and outliers are represented by points far from the zero line. Examining the residual plot provides critical insights

into the validity of the parameters estimated under the assumptions of linear regression (Fahrmeir et al., 2021, p. 90-96)

4 Statistical analysis

The statistical analysis was conducted using Python == 3.12.6(Van Rossum and Foundation, 2023) with VScode == 1.94.2 (Microsoft, 2023) as an IDE. The following Python packages are used: pandas == 2.2.3(McKinney et al., 2023), scikit-learn == 1.5.2(Pedregosa et al., 2011) for data manipulation, matplotlib == 3.9.2(Hunter et al., 2023) and seaborn == 0.13.2(Waskom et al., 2023) for visualization, and statsmodels == 0.14.4(Seabold et al., 2023) for statistical modeling. In this project, the significance level (alpha) for statistical tests is set to 0.05.

4.1 Relation Between Variables

The correlation heatmap provides an overview of the linear relationships between five variables: `cnt`, `temp`, `atemp`, `hum`, and `windspeed`. From figure 1, it is observed that the number of bike rentals (`cnt`) is moderately positively correlated with temperature (`temp`) and feeling temperature (`atemp`), both with a correlation coefficient of 0.63. This indicates that rentals tend to increase as temperatures rise. Additionally, `temp` and `atemp` exhibit a near-perfect correlation of 0.99, suggesting that these variables are almost identical.

On the other hand, `cnt` shows weak negative correlations with humidity (`hum`) at -0.10 and windspeed (`windspeed`) at -0.23, implying that higher humidity and windspeed may slightly reduce the number of rentals. Furthermore, `hum` and `windspeed` have a weak negative correlation of -0.25, indicating a slight inverse relationship between these variables.

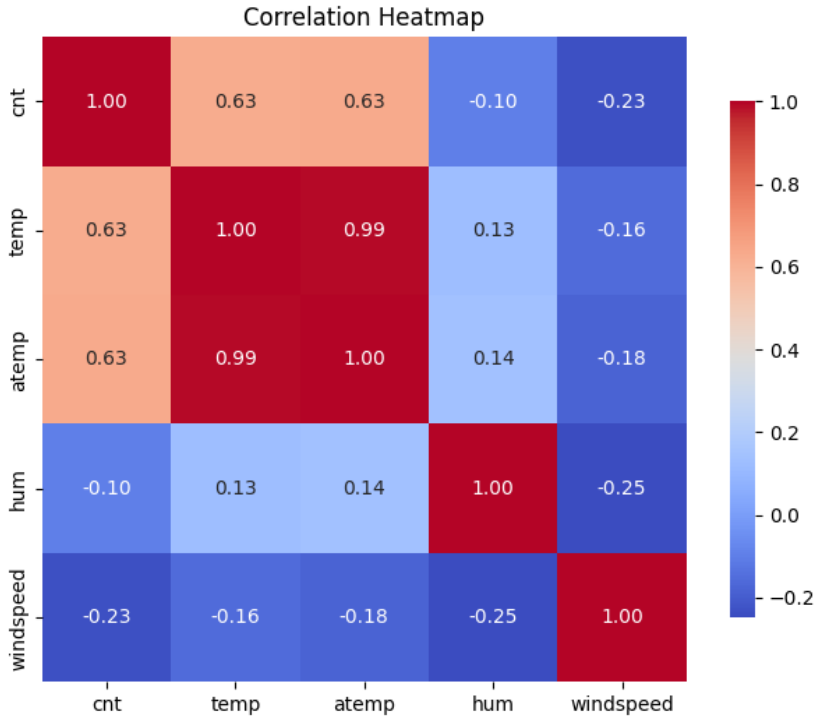

Figure 1: Correlation heatmap of numerical variables

4.2 Linear Regression

Two linear regression models, Model 1 and Model 2, are fitted to predict the number of rentals (`cnt`) using all the independent variables. Model 1 does not account for the polynomial effect of the covariates, while Model 2 incorporates quadratic and higher-order polynomial terms for `temp`, i.e., $\texttt{temp}^2$, $\texttt{temp}^3$, and $\texttt{temp}^4$. The mathematical notation for the two models is expressed as follows:

Model 1:

$$\begin{aligned}\texttt{cnt} = \beta_0 + \beta_1 \cdot \text{mnth} + \beta_2 \cdot \text{weekday} + \beta_3 \cdot \text{workingday} + \beta_4 \cdot \text{weathersit} \\ + \beta_5 \cdot \text{temp} + \beta_6 \cdot \text{atemp} + \beta_7 \cdot \text{hum} + \beta_8 \cdot \text{windspeed} + \epsilon\end{aligned}$$

Model 2:

$$\begin{aligned}\texttt{cnt} = \beta_0 + \beta_1 \cdot \text{mnth} + \beta_2 \cdot \text{weekday} + \beta_3 \cdot \text{workingday} + \beta_4 \cdot \text{weathersit} + \beta_5 \cdot \text{atemp} + \beta_6 \cdot \text{hum} \\ + \beta_7 \cdot \text{windspeed}\beta_8 \cdot \text{temp} + \beta_9 \cdot \text{temp}^2 + \beta_{10} \cdot \text{temp}^3 + \beta_{11} \cdot \text{temp}^4 + \epsilon\end{aligned}$$

From Table 1, it is observed that the Residual Standard Error (RSE) of Model 2 is 1223.89, which is lower than that of Model 1 at 1371.36. Furthermore, the R-squared (R^2) of Model 2 is 0.607, higher than Model 1's 0.504, indicating that Model 2 explains a greater proportion of the variance in the data. Additionally, the Akaike Information Criterion (AIC) also favors Model 2. Hence, incorporating the polynomial effects of the `temp` variable sufficiently improves the fit of the model.

Table 1: Comparison of Residual Standard Error (RSE), R-squared, and AIC for Linear and Polynomial Models

Model	**Residual Standard Error (RSE)**	**R-squared**	**AIC**
Linear	1371.36	0.504	1.264e+04
Polynomial	1223.89	0.607	1.248e+04

In Appendix Table 5, the summary of the coefficients for Model 2 is provided. It is observed that the coefficient estimates for the variables `workingday`, `atemp`, `temp`, `temp`^{2}, `temp`^{3}, and `temp`^{4} have p-values greater than 0.05, suggesting that these covariates may not exhibit a statistically significant relationship with the dependent variable `cnt`. Hence, their inclusion in the model may not substantially improve the model's predictive power. Consequently, further model refinement, such as the exclusion of these covariates, can be considered.

4.3 variable selection with Aic

In this project, the issue of redundant covariates was addressed using the backward selection technique, with the Akaike Information Criterion (AIC) used as the criterion to determine and eliminate noise covariates from the original model. Figure 2 illustrates the progression of AIC during backward elimination and exhibits that the minimum value of AIC is sustaited up to the inclusion of 7 covariates. Further removal of covariates increases AIC, indicating that the optimal model includes 7 of the initial 11 covariates from Model 2. This subset constitutes the most suitable predictors for the model.

Consequently, with backward selection, the four redundant variables i.e., `atemp`, `workingday`, `temp`, `temp`^{4} are dropped from the original Model 2. Hence, the resulting model, Model 3, is expressed as:

$$\text{cnt} = \beta_0 + \beta_1 \cdot \text{mnth} + \beta_2 \cdot \text{weekday} + \beta_3 \cdot \text{weathersit} + \beta_4 \cdot \text{hum} + \beta_5 \cdot \text{windspeed} + \beta_6 \cdot \text{temp}^2 + \beta_7 \cdot \text{temp}^3 + \epsilon$$

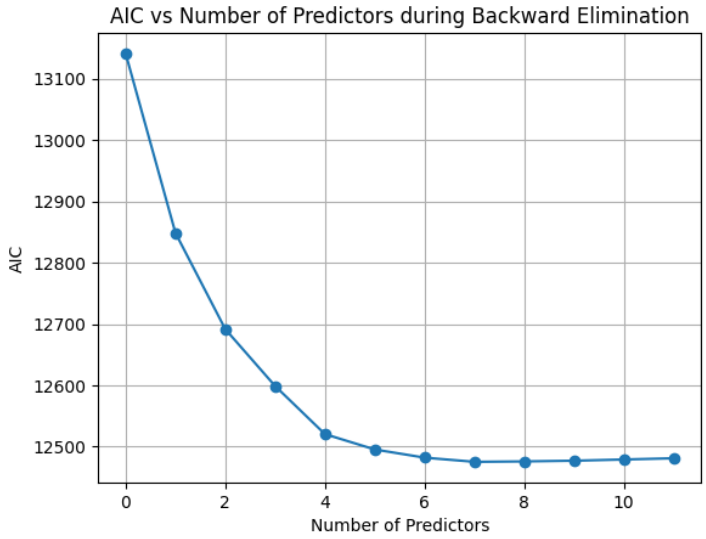

Figure 2: AIC vs Number of Predictors during Backward Elimination

The summary of the coefficients of Model 3 is provided in Table 2. From the table, it is observed that each of the coefficients has very low p-values, close to zero, suggesting that the corresponding covariates are significant and effectively explain the number of rentals (`cnt`). Additionally, none of the confidence intervals for these parameter estimates include zero, further reinforcing the notion of their statistical significance. Moreover, `mnth`, `weekday`, and $\texttt{temp}^2$ have a positive influence on `cnt`, while the other covariates, such as `weathersit`, `humidity`, `windspeed`, and $\texttt{temp}^3$ have negative effects on `cnt`.

The intercept of 3727.123 represents the expected number of bike rentals when all other variables are fixed to zero, indicating a baseline rental count. The coefficient for `weekday`, 67.1881, suggests that for each increase in the weekday value (moving from Sunday to Monday, Monday to Tuesday, and so on), the number of rentals is expected to increase by 67.1881 units, holding all other variables constant. In contrast, the coefficient for `humidity`, -3467.8760, indicates that with each one-unit increase in humidity, the number of rentals is expected to decrease by 3467.8760 units, holding all other variables constant.

Variable	Coef	Std Err	t	P>|t|	[0.025, 0.975]
const	3727.1222	297.047	12.547	0.000	[3143.946, 4310.299]
mnth	70.8182	13.962	5.072	0.000	[43.407, 98.230]
weekday	67.1881	22.683	2.962	0.003	[22.656, 111.720]
weathersit	-455.2309	109.178	-4.170	0.000	[-669.574, -240.888]
hum	-3467.8760	444.003	-7.810	0.000	[-4339.566, -2596.186]
windspeed	-4763.5940	633.352	-7.521	0.000	[-6007.022, -3520.166]
$temp^2$	3.809e+04	1904.210	20.002	0.000	[3.43e+04, 4.18e+04]
$temp^3$	-3.97e+04	2290.203	-17.334	0.000	[-4.42e+04, -3.52e+04]

Table 2: Summary of Coefficients for Model 3

Table 3 presents the goodness of fit statistics, including the RSE and R^2 values. The RSE and R^2 of Model 3 are 1222.33 and 0.607, respectively. In comparison, Model 2 has an RSE of 1223.04 and an R^2 of 0.607. Therefore, with the elimination of redundant features, Model 3 demonstrates either better or similar performance to Model 2.

Goodness of Fit Statistic	Value
Residual Standard Error (RSE)	1222.34
R-squared	0.606
Akaike Information Criterion (AIC)	1.248×10^4

Table 3: Goodness of Fit Statistics for Model 3

4.4 Assessment of Linear Regression Assumptions

Figure 3 shows the residual plot generated from Model 3. From the figure, it is evident that the residuals display a curvature or systematic pattern rather than being randomly distributed, suggesting a non-linear relationship between the predicted and observed values. Additionally, the residuals exhibit a funnel-shaped pattern with an increase in the target variable, suggesting heteroscedasticity. This implies that the variance of the residuals is not constant across all levels of the predicted values.

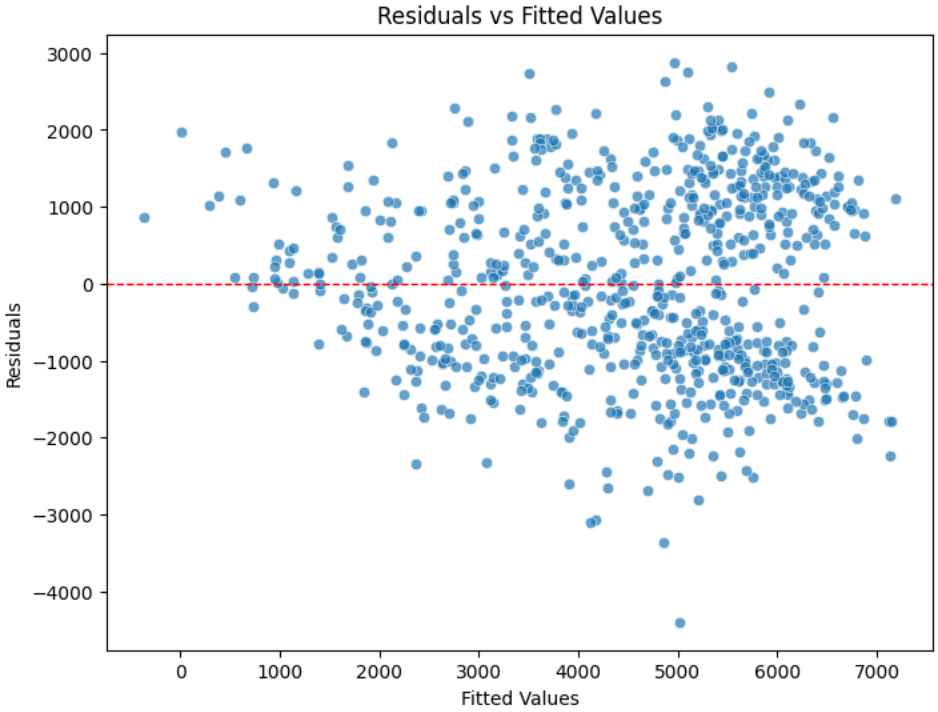

Figure 3: Residual Plot for Model 3

Figure 4 presents the QQ plot of the residuals for Model 3. The plot shows that the residuals do not perfectly align with the reference normality line. This deviation from the line suggests that the errors are not be normally distributed.

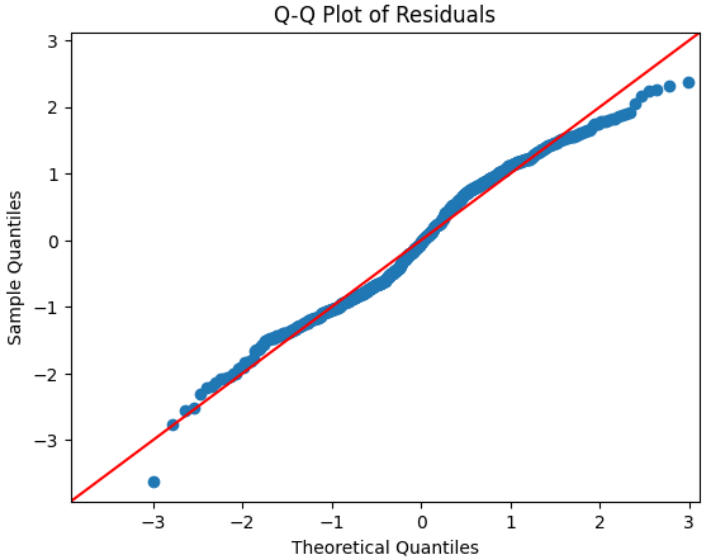

Figure 4: QQ Plot of Residuals for Model 3

Table 4 provides the Variance Inflation Factor (VIF) values for each covariate, which quantify the degree of multicollinearity between them. Most covariates have VIF values around 1, indicating minimal correlation with other predictors. However, $temp^2$, and $temp^3$, exhibit significantly higher VIF values, both around 58. Since VIF values greater than 10 typically indicate strong multicollinearity, it can be concluded that $temp^2$, and $temp^3$, demonstrate a substantial degree of correlation between them.

Feature	**VIF**
mnth	1.134938
weekday	1.010361
weathersit	1.729158
hum	1.953945
windspeed	1.177096
$temp^2$	58.850647
$temp^3$	58.184566

Table 4: Variance Inflation Factors (VIF) for the Model Predictors

4.5 Model Validation

The best fitted model,Model 3, is determined through backward selection with AIC and the corresponding residuals and multicollinearity behavior are assessed with the residual plot and VIF table. These assessments indicate violations of key assumptions in ordinary least squares (OLS) regression. Specifically, the assumption of linearity between predictors and the dependent variable is not satisfied and there is evidence of heteroskedasticity in the error terms. The presence of multicollinearity is also detected among covariates. These violations compromise the validity of the parameter estimates in the final model. For example, the coefficients β_j from Model 3, which appear to have lower p values, may in fact have higher p values, hence could cause us to erroneously conclude that a parameter is statistically significant. Additionally, the 95% confidence interval may in reality have a much lower probability than 0.95 of containing the true value of the parameter.

5 Summary

This report analyzed the relationship between predictors and the dependent variable using regression techniques, beginning with linear and polynomial regression models. A comparative evaluation with goodness of fit statistics i.e. R^2 and RSE, demonstrated that the polynomial regression model provided a better fit to the data, outperforming the linear model in capturing the underlying relationships. To further enhance the polynomial regression model, backward selection based on the Akaike Information Criterion (AIC) was employed. This process systematically removed redundant covariates i.e. `atemp`, `workingday`, `temp`, `temp`^{4}, resulting in an optimized and more interpretable final model with seven covariates retained from the initial set of ten. The underlying assumptions of linear regression were evaluated for the model through residual plots, QQ-plots of the residuals, and Variance Inflation Factor (VIF) values. These assessments revealed several key violations, including non-linearity between predictors and the dependent variable, heteroskedasticity in the error terms, non-normality of residuals, and multicollinearity among certain predictors. These violations undermine the validity of the parameter estimates, potentially leading to erroneous conclusions regarding statistical significance and confidence intervals.

However, the polynomial regression model refined through backward selection demonstrated strong predictive capabilities and the identified assumption violations underscore the need for careful interpretation of the results and consideration of alternative modeling approaches.

Bibliography

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian D. Marx. *Regression: Models, Methods and Applications*. Springer, Berlin, Germany, 2021.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, USA, 2nd edition, 2009. ISBN 978-0387848570.

John D. Hunter, Michael Droettboom, Thomas A. Caswell, Eric Firing, Benjamin Root, et al. Matplotlib: Visualization with python, 2023. URL `https://matplotlib.org/`. Version 3.7.2.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, New York, USA, 2nd edition, 2013. ISBN 978-1461471370.

Jong Hae Kim. Multicollinearity and misleading statistical results. *Korean journal of anesthesiology*, 72(6):558–569, 2019.

Wes McKinney, Jeff Reback, Joris Van den Bossche, Tom Augspurger, et al. pandas: Python data analysis library, 2023. URL `https://pandas.pydata.org/`. Version 2.1.3.

Microsoft. Visual studio code, 2023. URL `https://code.visualstudio.com/`. Version 1.83.1.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Skipper Seabold, Josef Perktold, Nathaniel Smith, Kevin Sheppard, et al. Statsmodels: Statistical modeling and econometrics in python, 2023. URL `https://www.statsmodels.org/`. Version 0.14.0.

Guido Van Rossum and Python Software Foundation. Python, 2023. URL `https://www.python.org/`. Version 3.12.6.

Michael Waskom, Martin L. Martin, Anna Lam, Daniel F. Larremore, et al. Seaborn: Statistical data visualization, 2023. URL https://seaborn.pydata.org/. Version 0.12.2.

Appendix

A Additional tables

Variable	Coefficient	Std. Error	t-Statistic	p-Value	[0.025]	[0.975]
const	4277.2195	1247.528	3.429	0.001	1827.987	6726.452
mnth	74.4079	14.454	5.148	0.000	46.031	102.785
weekday	67.1283	22.786	2.946	0.003	22.393	111.863
workingday	111.8113	97.961	1.141	0.254	-80.512	304.134
weathersit	-463.8966	110.004	-4.217	0.000	-679.864	-247.929
atemp	-247.6302	2414.376	-0.103	0.918	-4987.699	4492.438
hum	-3461.7602	447.334	-7.739	0.000	-4339.998	-2583.523
windspeed	-4732.9151	651.968	-7.259	0.000	-6012.904	-3452.926
temp	-4393.7596	12500.000	-0.351	0.726	-29000.000	20200.000
$temp^2$	48270.000	45700.000	1.057	0.291	-41400.000	138000.000
$temp^3$	-47110.000	68200.000	-0.691	0.490	-181000.000	86800.000

Table 5: Summary of Coefficients for Model 2