DATA210P HW2 - Bike Sharing (hour.csv): Linear Modeling, Selection, Validation, and Ridge & Lasso

Joe Nguyen, Haesung Becker, Jared Lyon, Tao Chen

Table of contents

1 Project Overview

1.1 Libraries & packages

```
# Standard library imports
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Third-party imports
from dataclasses import dataclass
from typing import Callable, Dict, Optional, Tuple, List, Any
from textwrap import dedent
import patsy

# Statsmodels imports
import statsmodels.api as sm
import statsmodels.formula.api as smf
from statsmodels.stats.outliers_influence import variance_inflation_factor
from statsmodels.graphics.gofplots import qqplot
from statsmodels.nonparametric.smoothers_lowess import lowess
from statsmodels.stats.diagnostic import het_breuschpagan

# Sklearn imports
from sklearn.model_selection import KFold
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LinearRegression, RidgeCV, LassoCV, Ridge, Lasso, lasso_path

# Set global configurations
np.random.seed(42)
pd.set_option("display.max_columns", 200)
pd.set_option("display.width", 120)
```

1.2 Data Source & Import:

Import UCI ML Repo and load dataset (hour.csv). We decided to use the hour.csv dataset for this homework assignment because it contains a larger sample size and time-of-day effects for a more robust analysis.

Table 1: Shape and Columns of Data Frame.

```
Shape:  (17379, 14)
Columns:
dteday, season, yr, mnth, hr, holiday, weekday, workingday, weathersit, temp,
atemp, hum, windspeed, cnt
```

1.3 Data Dictionary:

From data source:

1) **Outcome (Response) Variable**: cnt (integer) - count of total rental bikes including both casual and registered

2) **Predictor (Feature) Variables**:

- instant (integer) - record index
- dteday (date) - date
- season (categorical) - season (1:winter, 2:spring, 3:summer, 4:fall)
- yr (categorical) - year (0: 2011, 1:2012)
- mnth (categorical) - month (1 to 12)
- hr (categorical) - hour (0 to 23)
- holiday (binary) - whether the day is a holiday or not
- weekday (categorical) - day of the week
- workingday (binary) - if day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit (categorical) -
  - 1: Clear, Few clouds, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp (continuous) - Normalized temperature in Celsius. The values are dervied via (t - t_min)/(t_max - t_min), where t_min=-8, t_max=39 (only in hourly scale)
- atemp (continuous) - Normalized feeling temperature in Celsius. The values are derived via (t - t_min)/(t_max - t_min), where t_min=-16, t_max=50 (only in hourly scale)
- hum (continuous) - Normalized humidity. The values are divided to 100 (max)
- windspeed (continuous) - Normalized wind speed. The values are divided to 67 (max)
- casual (integer) - count of casual users
- registered (integer) - count of registered users

Initialize dataframe and perform initial data exploration:

1.4 Initial Exploratory Data Analysis (EDA) with Visualization

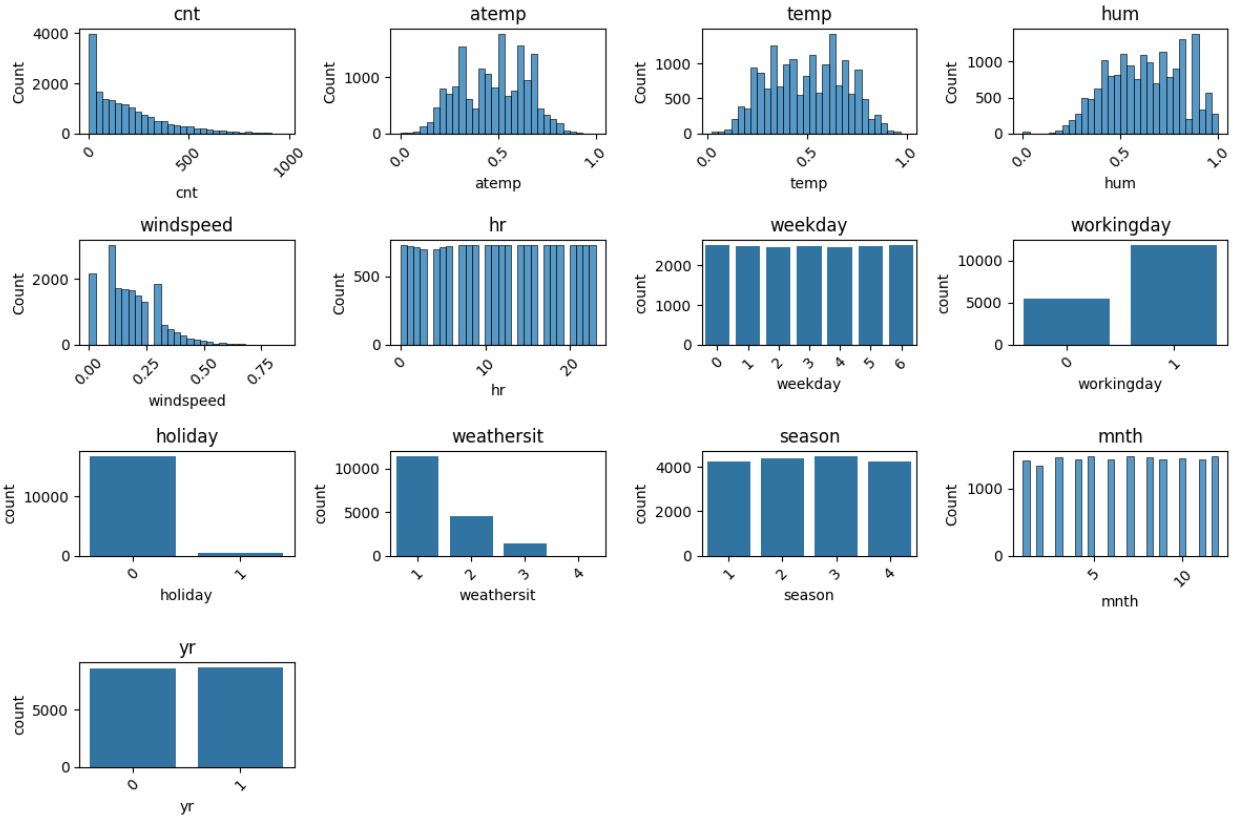

Figure 1: Distributions of continuous and categorical predictors.

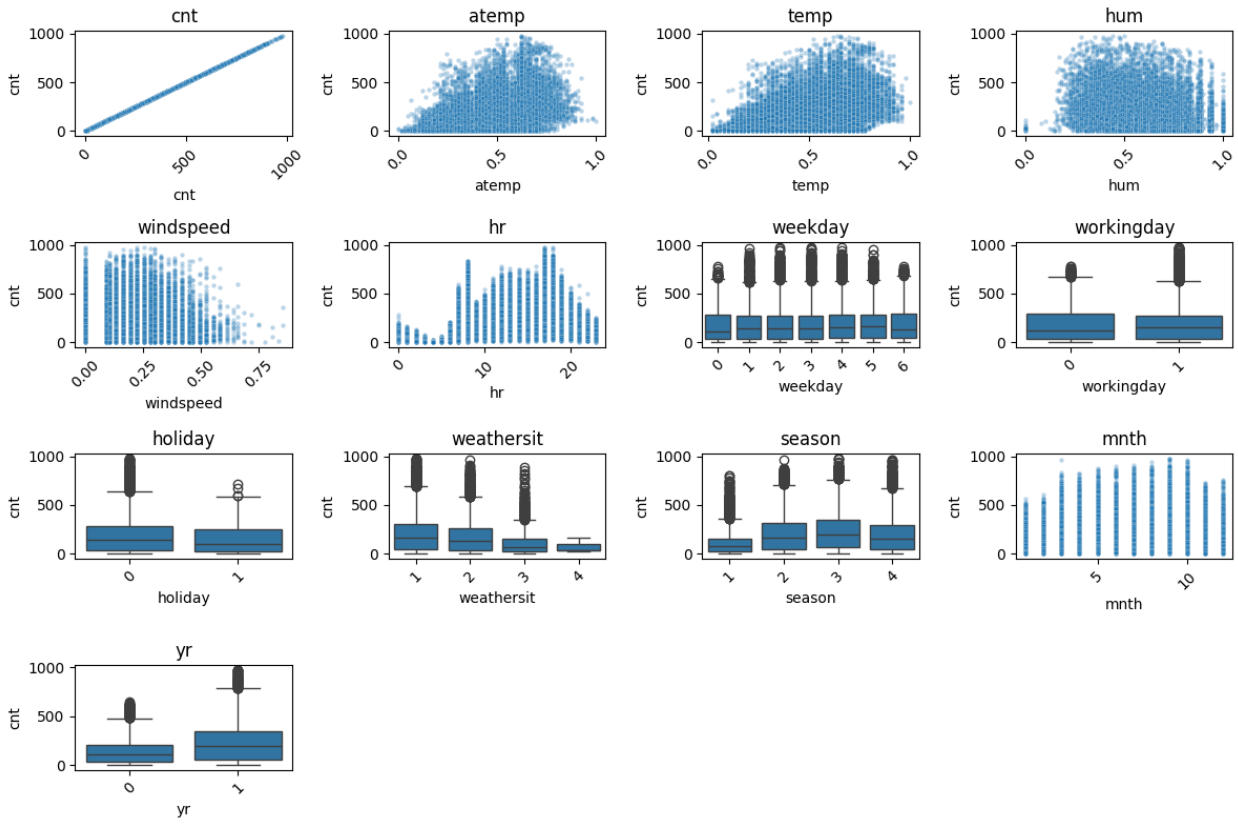

Figure 2: Relationships between predictors and hourly bike rentals (cnt).

2 Linear Model and Interpretation

2.1 Predictor selection and justification

After careful consideration from our initial EDA in Figure 1 and Figure 2 , we've decided to break down the predictor selection into several categories to fit into the baseline OLS model while excluding the variables that may lead to redundancy or multicollinearity (#4):

1) **Calendar Variables**: hr, weekday, workingday, holiday
   - These variables capture the time-related patterns in bike rentals.
2) **Seasonal Variables**: season, yr
   - Seasonal trends and yearly changes can significantly impact bike rental behavior.
3) **Weather Variables**: weathersit, temp, hum, windspeed
   - Weather conditions significantly influence bike rental behavior and demand.
4) **Exclusion of leakage Variables**: dteday, atemp, mnth
   - The variable atemp is highly collinear with temp and doesn't much predictive power beyond temp. Similar, mnth is highly collinear to season and possibly temp.

Additionally, we've also broken them down into specific data types for clarity:

1) **Categorical Variables**: season, yr, mnth, hr, holiday, weekday, workingday, weathersit
    - These variables represent distinct categories or groups.
2) **Continuous Variables**: temp, hum, windspeed
    - These variables represent measurable quantities that can take on a wide range of values.

2.2 Fit and show results of OLS model

Table 2 summarizes the baseline model fit statistics whereas Table 3 provides a coefficient table that's sorted for lowest p-values.

2.3 Interpretation of coefficients

Before we begin, all interpretations for the individual variables below assume that we are **holding all other variables constant** and that **cnt is bike rentals per hour** since we're using the hour.csv dataset. Out of the 10 variables fitted into the model, we've selected 7 variables to interpret. Among them are 3 continuous weather variables with strongly supported effects, 1 encoded categorical variable to discuss the referencing, 2 calendar variables with unstable and unsupported effects with potential multicollinearity issues, and the hour variable enabled by hour.csv to demonstrate its analytical value. The exclusion of the other 3 variables (weekday, season, and yr) does not simply that they're not important to our analysis and will be accounted for in our analysis in later sections.

Our interpretations of the selected coefficients from the baseline OLS model are as follows:

1) Continuous Weather Variables:
    - **Temperature (temp)**: From Table 3, a one-unit increase in normalized temperature (temp) is associated with an increase of approximately 233.3 additional bike rentals per hour on average. The data strongly rejects the null hypothesis of no effect ($p < 0.001$) and the 95% confidence interval (CI): [215, 252]. This suggests that warmer temperatures may potentially encourage more bike rentals.
    - **Humidity (hum)**: From Table 3, a one-unit increase in normalized humidity (hum) is associated with a decrease of approximately 81.3 fewer bike rentals per hour on average. The data strongly rejects the null hypothesis of no effect ($p < 0.001$) and the 95% CI: [-70, -92]. This indicates that higher humidity levels may deter people from renting bikes.
    - **Windspeed (windspeed)**: From Table 3, a one-unit increase in normalized windspeed (windspeed) is associated with a decrease of approximately 36.3 fewer bike rentals per hour on average. The data strongly rejects the null hypothesis of no effect ($p < 0.001$) and the 95% CI:[-23, -50]. This suggests that windier conditions may discourage bike rentals.
2) Categorical Weather Variable:

Table 2: Baseline OLS model fit statistics.

```
--- Statsmodels Summary ---

                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.682
Model:                            OLS   Adj. R-squared:                  0.681
Method:                 Least Squares   F-statistic:                     929.0
Date:                Fri, 30 Jan 2026   Prob (F-statistic):               0.00
Time:                        01:01:48   Log-Likelihood:            -1.0509e+05
No. Observations:               17379   AIC:                         2.103e+05
Df Residuals:                   17338   BIC:                         2.106e+05
Df Model:                          40
Covariance Type:            nonrobust
================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept        -87.3998      6.258    -13.966      0.000     -99.666     -
75.134
C(weathersit)[T.2]   -10.8506      1.928     -5.629      0.000     -14.629      -
7.072
C(weathersit)[T.3]   -66.4743      3.248    -20.468      0.000     -72.840     -
60.108
C(weathersit)[T.4]   -70.5084     59.255     -1.190      0.234    -
186.654      45.637
C(hr)[T.1]       -17.5302      5.382     -3.257      0.001     -28.080      -
6.981
C(hr)[T.2]       -26.6938      5.400     -4.943      0.000     -37.278     -
16.110
C(hr)[T.3]       -37.3660      5.439     -6.870      0.000     -48.027     -
26.705
C(hr)[T.4]       -40.6120      5.442     -7.462      0.000     -51.279     -
29.945
C(hr)[T.5]       -23.8207      5.405     -4.407      0.000     -34.415     -
13.226
C(hr)[T.6]        35.0407      5.392      6.498      0.000      24.471      45.610
C(hr)[T.7]       170.2513      5.382     31.631      0.000     159.701     180.802
C(hr)[T.8]       310.9038      5.380     57.793      0.000     300.359     321.448
C(hr)[T.9]       163.4304      5.384     30.354      0.000     152.877     173.984
C(hr)[T.10]      108.8632      5.404     20.146      0.000      98.271     119.455
C(hr)[T.11]      134.5404      5.437     24.746      0.000     123.884     145.197
C(hr)[T.12]      174.0695      5.475     31.795      0.000     163.338     184.800
C(hr)[T.13]      169.1885      5.502     30.751      0.000     158.404     179.973
C(hr)[T.14]      153.3648      5.527     27.750      0.000     142.532     164.198
C(hr)[T.15]      162.7674      5.534     29.412      0.000     151.920     173.615
C(hr)[T.16]      224.7619      5.526     40.677      0.000     213.931     235.592
C(hr)[T.17]      378.3610      5.499     68.805      0.000     367.582     389.140
C(hr)[T.18]      346.4628      5.470     63.334      0.000     335.740     357.185
C(hr)[T.19]      237.6637      5.428     43.784      0.000     227.024     248.303
C(hr)[T.20]      157.9849      5.405     29.229      0.000     147.390     168.579
C(hr)[T.21]      108.3892      5.386     20.125      0.000      97.832     118.946
```

Table 3: Baseline OLS model fit coefficient table.

--- Coefficient Table ---

	term	estimate	std_err	p_value	conf_low	conf_high
0	C(hr)[T.8]	3.109038e+02	5.379574e+00	0.000000e+00	3.003593e+02	3.214483e+02
1	C(hr)[T.18]	3.464628e+02	5.470435e+00	0.000000e+00	3.357402e+02	3.571854e+02
2	C(hr)[T.19]	2.376637e+02	5.428111e+00	0.000000e+00	2.270241e+02	2.483034e+02
3	C(hr)[T.16]	2.247619e+02	5.525528e+00	0.000000e+00	2.139313e+02	2.355925e+02
4	C(hr)[T.17]	3.783610e+02	5.499041e+00	0.000000e+00	3.675824e+02	3.891397e+02
5	C(yr)[T.1]	8.548581e+01	1.568613e+00	0.000000e+00	8.241117e+01	8.856045e+01
6	temp	2.449964e+02	7.110919e+00	1.136117e-		
251	2.310583e+02	2.589345e+02				
7	C(hr)[T.12]	1.740695e+02	5.474718e+00	1.127114e-		
215	1.633385e+02	1.848005e+02				
8	C(hr)[T.7]	1.702513e+02	5.382477e+00	1.567761e-		
213	1.597011e+02	1.808015e+02				
9	C(hr)[T.13]	1.691885e+02	5.501973e+00	3.130365e-		
202	1.584041e+02	1.799730e+02				
10	C(hr)[T.9]	1.634304e+02	5.384154e+00	3.118441e-		
197	1.528769e+02	1.739839e+02				
11	C(hr)[T.15]	1.627674e+02	5.534014e+00	1.350689e-		
185	1.519202e+02	1.736147e+02				
12	C(hr)[T.20]	1.579849e+02	5.405104e+00	2.277420e-		
183	1.473904e+02	1.685795e+02				
13	C(hr)[T.14]	1.533648e+02	5.526624e+00	7.170303e-		
166	1.425321e+02	1.641976e+02				
14	C(season)[T.4]	6.578348e+01	2.439501e+00	6.302680e-		
157	6.100181e+01	7.056514e+01				
15	C(hr)[T.11]	1.345404e+02	5.436805e+00	6.824192e-		
133	1.238837e+02	1.451971e+02				
16	C(weathersit)[T.3]	-6.647434e+01	3.247787e+00	5.115518e-92	-7.284033e+01	-
6.010835e+01						
17	C(hr)[T.10]	1.088632e+02	5.403667e+00	3.049689e-		
89	9.827149e+01	1.194550e+02				
18	C(hr)[T.21]	1.083892e+02	5.385804e+00	4.632920e-		
89	9.783245e+01	1.189459e+02				
19	C(season)[T.2]	4.287260e+01	2.822412e+00	8.866599e-		
52	3.734039e+01	4.840481e+01				
20	Intercept	-8.739983e+01	6.257868e+00	4.333326e-44	-9.966588e+01	-
7.513378e+01						
21	C(hr)[T.22]	7.123408e+01	5.378033e+00	7.494517e-		
40	6.069260e+01	8.177557e+01				
22	hum	-6.878424e+01	5.439703e+00	1.730957e-36	-7.944660e+01	-
5.812187e+01						
23	C(season)[T.3]	2.734460e+01	3.648229e+00	6.932776e-		
14	2.019370e+01	3.449549e+01				
24	C(hr)[T.4]	-4.061204e+01	5.442219e+00	8.897576e-14	-5.127934e+01	-
2.994474e+01						
25	C(hr)[T.3]	-3.736604e+01	5.438879e+00	6.630284e-12	-4.802679e+01	-
2.670529e+01						
26	C(hr)[T.6]	3.504071e+01	5.392208e+00	8.339101e-		

- **Weather Situation (weathersit)**: From Table 3, the top 2 weather situations are weathersit 3 (T.3) and 2 (T.2) in terms of the strength of the statistical evidence supporting an association with hourly bire rentals. However, note that the categorical coefficients of T.3 and T.2 are relative to the reference weather situation (T.1 or clear/few clouds/partly cloudly). T.2 or misty/cloudy weather is associated with approximately 10.7 fewer bike rentals per hour on average compared to clear weather (T.1). Similarly, T.3 or light rain/snowy weather is associated with approximately 65.9 fewer bike rentals per hour on average compared to clear weather with a 95% CI:[-60, -73]. Both effects are supported by strong statistical evidence in the data ($p < 0.001$). This potentially highlights the negative impact of adverse weather conditions on bike rental demand.

3) Calendar & Human Behavior Variables:

- **Holiday (holiday)**: From Table 3, the model estimates that holidays are associated with a decrease of approximately 2.47e+13 bike rentals compared to to the baseline reference of non-holidays which is huge and unstable. In this case, the data does not support rejecting the null hypothesis of no effect ($p = 0.998$) with an extremely wide 95% CI:[-1.68e+14, 2.18e+14] spanning large negative and positive values - indicating that the model cannot reliably determine the effect and its direction.
- **Working Day (workingday)**: From Table 3, the model also estimates that working day is associated with a decrease of approximately bike rentals per hour compared to non-working days. Similarly to holidays, this estimate is not statistically supported ($p = 0.978$) with an implausibly large estimated coefficient of 2.47e+13 and the 95% CI (-1.68e+14, 2.18e+14) is extremely wide. This suggests that the model cannot reliably determine the effect of working days on bike rentals.

NOTE: Not only are the holiday and workingday variables not statistically supported in this model, their estimated coefficients are exactly the same indicating potential multicollinearity between these two variables. This is likely due to the fact that non-working days are holidays and weekends. We should expect an inverse relationship between workingday and holiday.

- **Hour of the Day (hr)**: From Table 3, we observe that certain hours (e.g., T.8 (8am) and T.16-19 (4-7pm)) dominates the significance levels (p-values) among the top 5. However, we need to be careful in interpreting these coefficients as they are relative to the reference hour (midnight-1am). For example, T.8 has a coefficient of approximately +310.9 with a 95% CI: [300, 321], indicating that bike rentals (cnt) increase significantly during this hour compared to midnight, likely due to morning commute patterns. In Figure 3, we can see the peaks during typical commuting hours (8am and 5pm), suggesting that people are more likely to rent bikes during these times for commuting purposes.

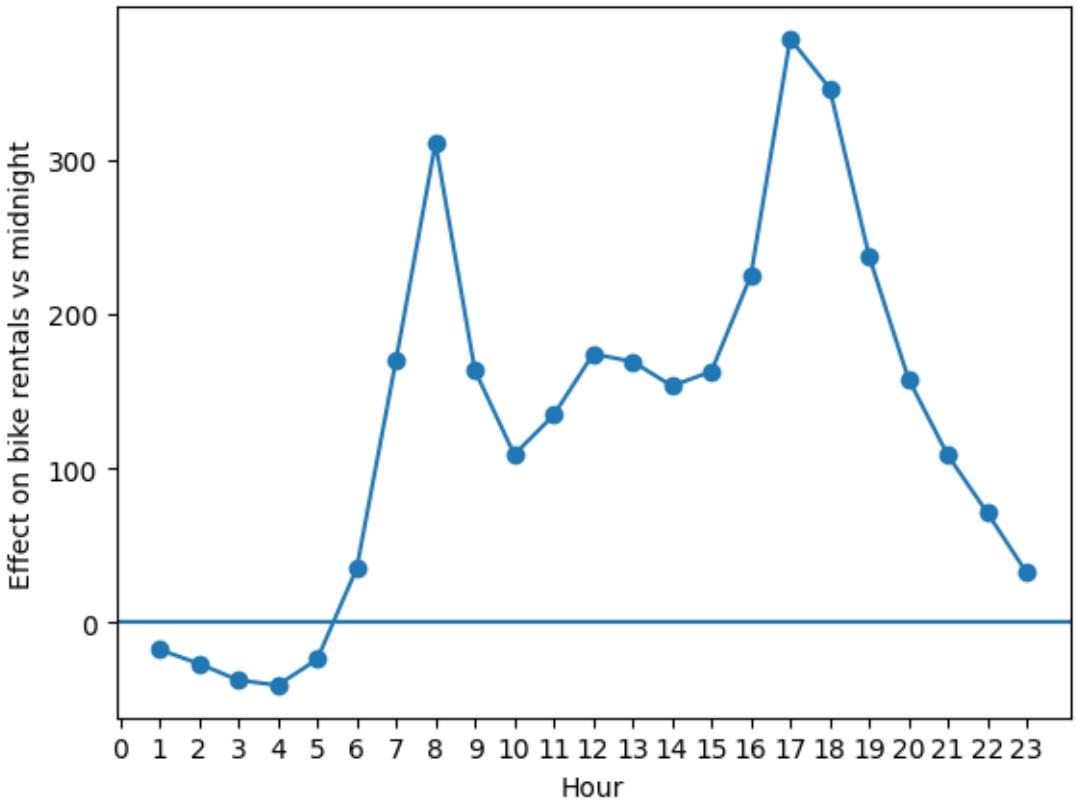

Figure 3: Estimated hour-of-day effects relative to the reference hour (midnight/ hr=0).

2.4 Significance vs. practical importance

In Table 3 , we observe many predictors in the baseline OLS model exhibit strong statistical evidence of association with hourly bike rentals (cnt). However, this does not imply practical importance. Given the large sample size of the hourly dataset (n = 17,379), even **relatively small effects can achieve extremely low p-values**. As a result, p-values in this context primarily reflect the precision and consistency of the estimates rather than the magnitude of the effects.

To assess practical importance, we should consider the actual effect sizes (coefficients) alongside their **95% confidence intervals and their real-world implications**. For example, hour of the day (hr) and temperature (temp) both show overwhelming statistical support and large, meaninful effects. During peak commute hours as shown in Figure 3, bike rentals increase by several hundred bikes per hour compared to the baseline hour (midnight). Similarly, temperature also show substantial effects on hourly bike rentals, with warmer temperatures leading to significant increases in rentals. Other weather variables like humidity, windspeed, and weather situation also show meaningful effects, although smaller in magnitude.

In contrast, a select few calendar-based variables carry **unstable estimated coefficients and wide confidence intervals** e.g., workingday with a 95% CI: [-1.69e+14, 2.18e+14] suggest that, under the current model, do not contribute reliably to explaining variation in hourly bike rentals. Although not discussed in our interpretation, the categorical variable C(weathersit)[T.4] or severe weather show a p-value of 0.234 and a 95% CI: [-186.65, 45.64] indicating uncertain direction.

Conceptually, severe weather should reduce demand, however, from Figure 1, we can observe that weathersit = 4 is very rare in the dataset. Despite being conceptually practical, this is another example of a variable being impractical for inference in our model likely due to the data sparsity.

In summary, this analysis highlights the **importance of interpreting statistical significance alongside effect magnitude, direction, and conceptual relevance**, rather than relying on p-values or any one metric alone. To improve our analysis and the model's predictive and inferential performance, further model diagnostics, transformations, and validation techniques will be considered below.

3 Transformations and Model Diagnostics

3.1 Baseline OLS Model diagnostics

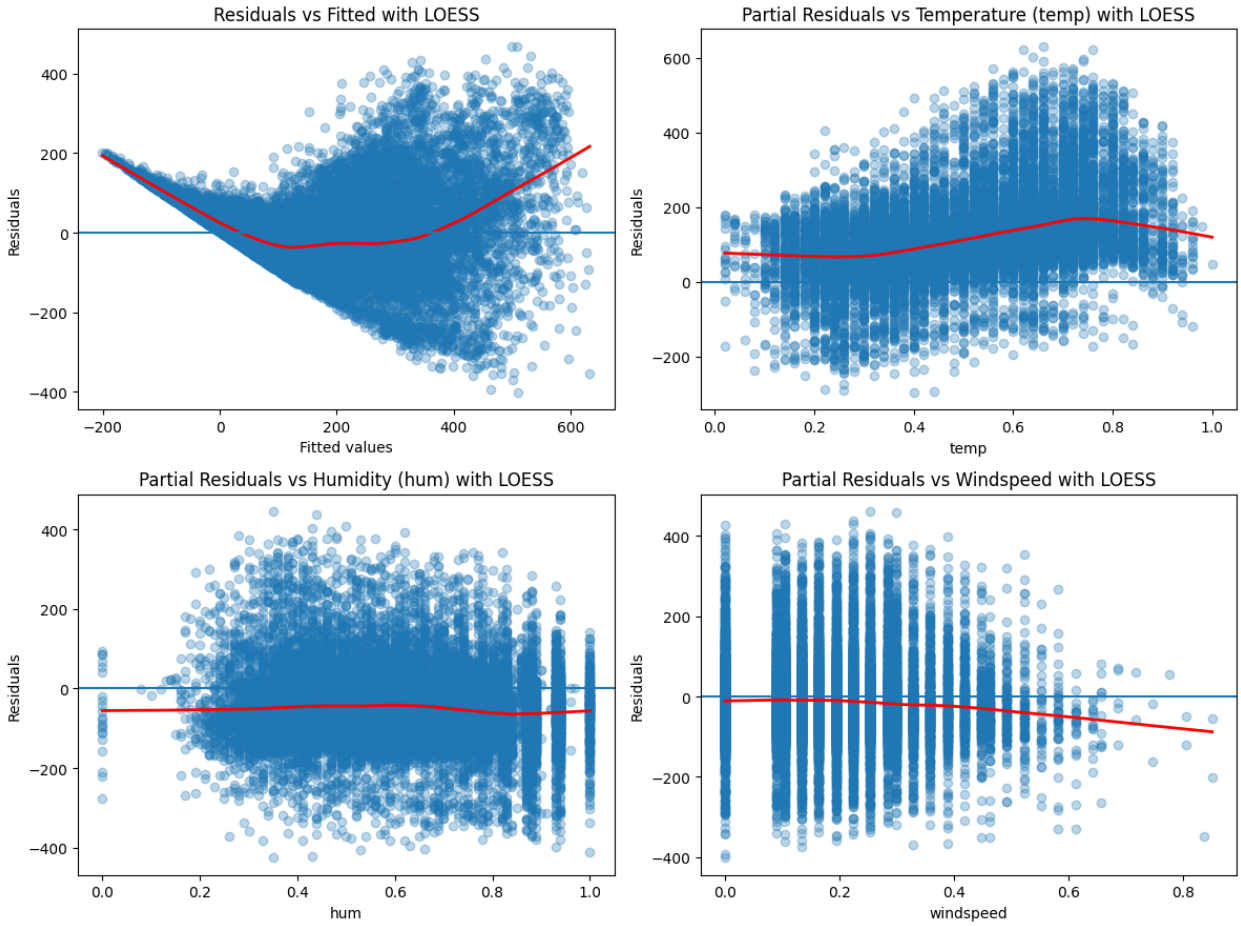

Figure 4: Residual diagnostics for the baseline OLS model.

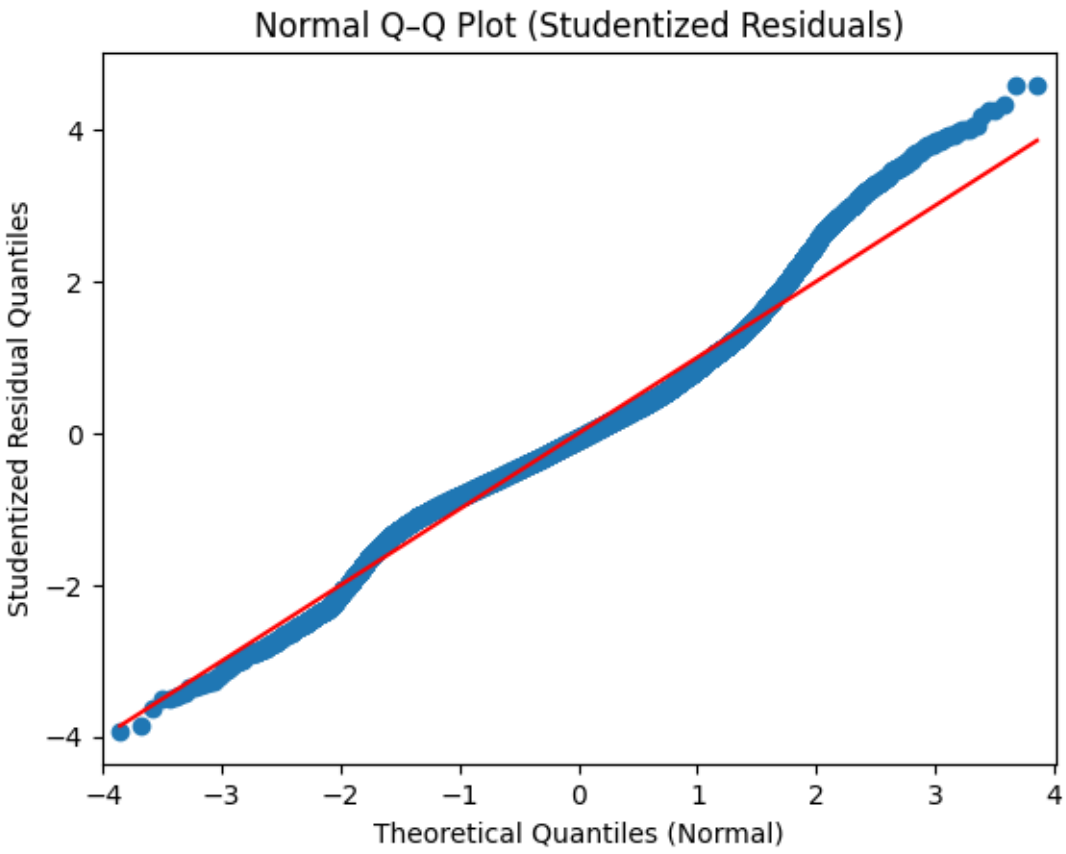

Figure 5: Normal Q–Q plot of studentized residuals from the baseline OLS model.

Table 4: Breusch–Pagan test for heteroskedasticity in the baseline OLS model.

LM Statistic	5504.267929
LM p-value	0.000000
F Statistic	200.916107
F p-value	0.000000
dtype:	float64

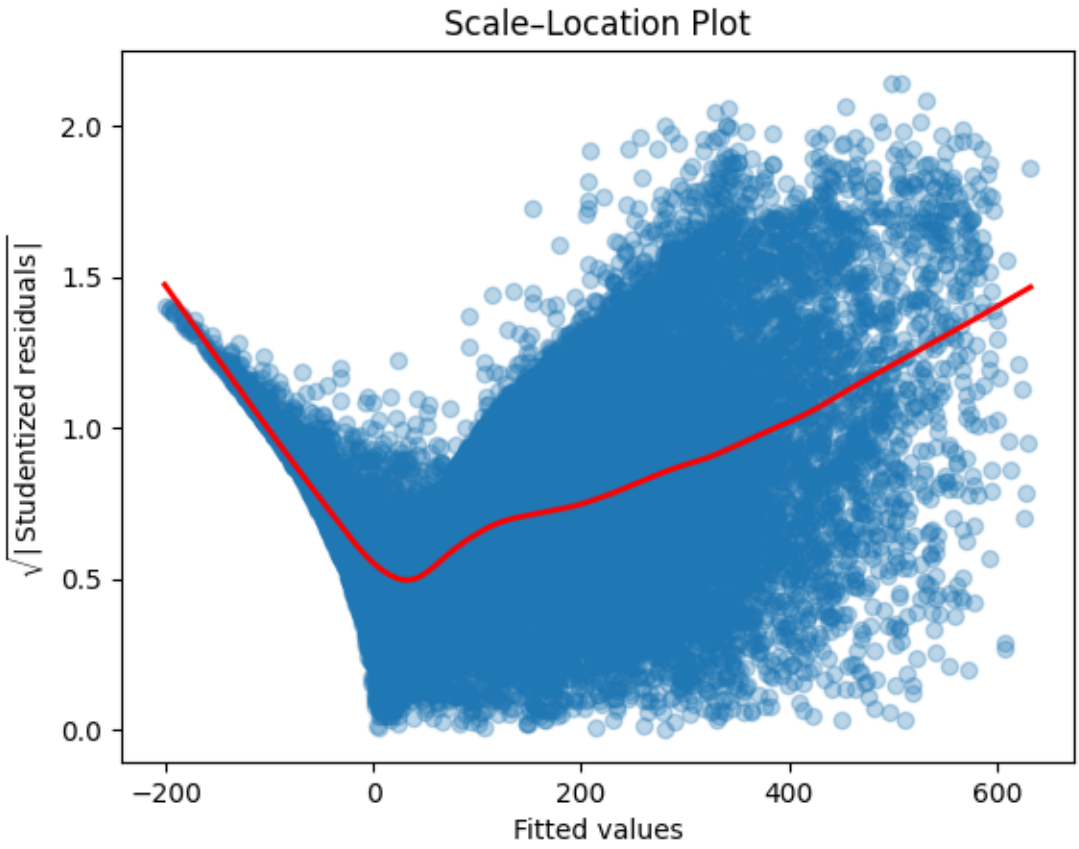

Figure 6: Scale–location plot of studentized residuals for the baseline OLS model.

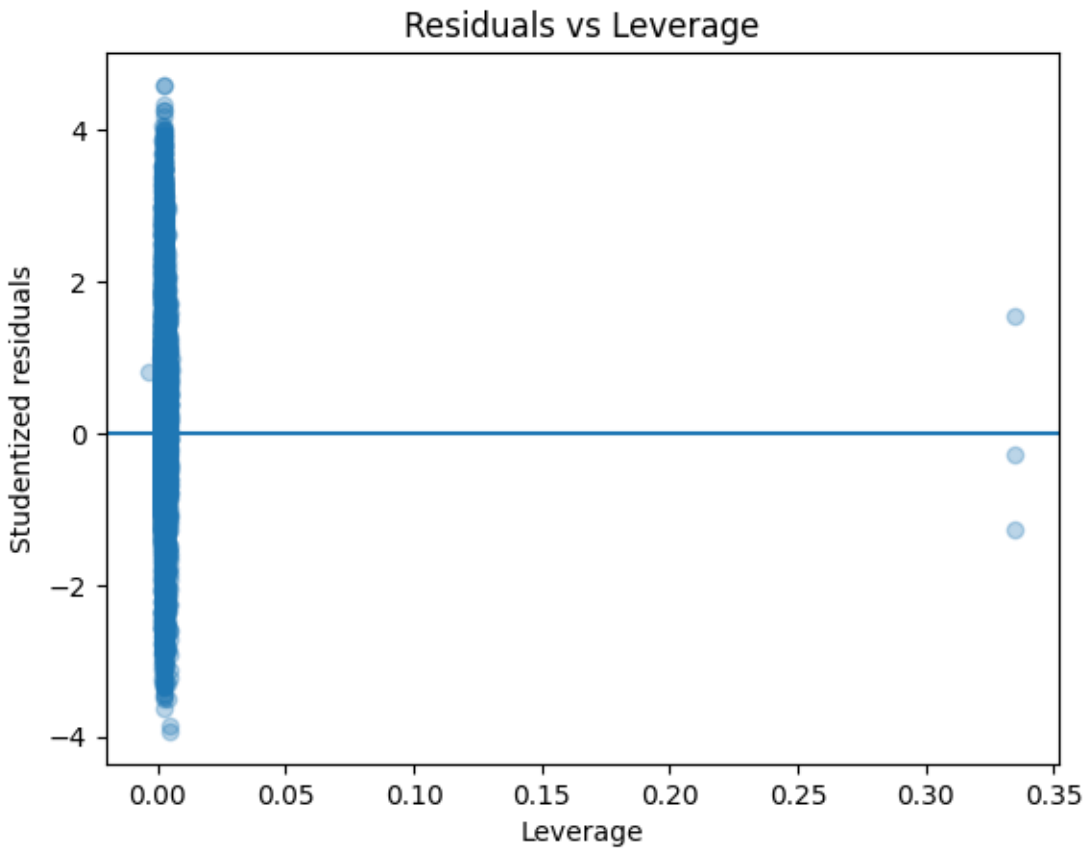

Figure 7: Studentized residuals versus leverage for the baseline OLS model.

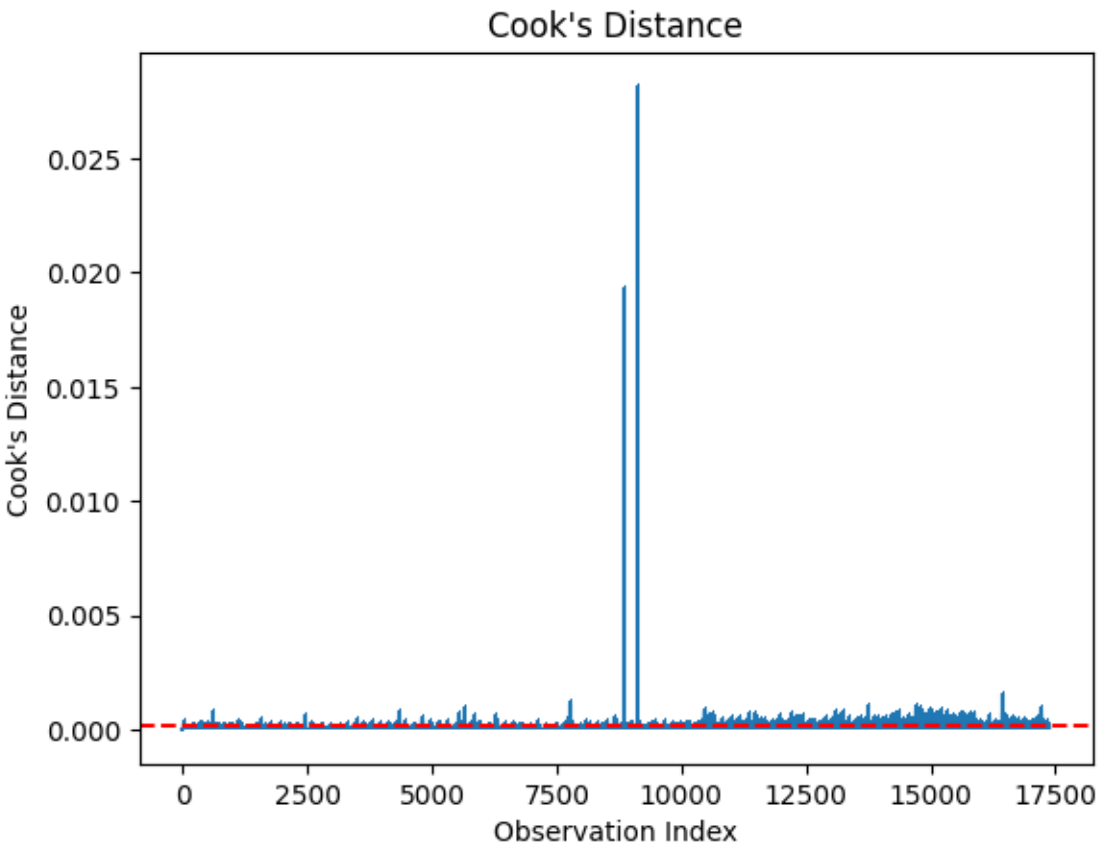

Figure 8: Cook's distance for the baseline OLS model.

From our diagnostic plots above, we can visually and numerically evaluate our baseline OLS model's assumptions as follows:

With respect to linearity in Figure 4, there is a U-shaped curvature in the LOESS line in the Residuals vs. Fitted plot indicating **systematic departures from linearity** in the overall mean function indicating global nonlinearity at low and high fitted values. This could lead to prediction bias, misleading coefficients, and unreliable statistical inference. As a result, these issues motivate the consideration of transformations or alternative modeling approaches that better capture nonlinear relationships.

In further analysis, we plotted the 3 continuous predictors (temp, hum, and windspeed) against their respective partial residuals in Figure 4. When plotted using the partial residuals vs temperature (temp), the LOESS line has mild slope changes at low and high temperatures, suggesting the **linearity assumption is not violated**. Similarly for humidity (hum) and windspeed, we observe no significant curvatures that would hint at nonlinearity other than mild curvatures at the tails.

In Figure 5 for checking normality, using the studentized residuals, the normal Q-Q plot shows that while most of the distribution is approximately normal, there is clear departures occuring at the upper tail. This suggests **heavy-tailed errors and a violation of the normality assumption** and that our p-values and CIs may not be reliable for inference. Once again, another motivation for transformations or alternative modeling methods.

In Figure 6 for validating homoskedasticity, we observe a funnel shape in which studentized residual spread as fitted values increase. The lower tail is sharply bounded whereas the upper tail grows. This observation hints at **heteroskedasticity and our assumption of constant residual variance is violated**. This motivates further variance-stabilizing transformations. **Supported by the Breusch-Pagan test** in Table 4 with a p-value < 0.05 i.e., there is strong support in rejecting the null hypothesis of homoskedasticity - there's evidence of heteroskedasticity.

In Figure 7 for confirming leverage and influence in observations, there are a small number of observations with high leverage and moderate studentized residuals whereas the rest of the dataset remain small and well below thresholds associated with undue influence. Consistent with Figure 8, only two observations with high leverage and moderate residual show up as high influence on Cook's D near 0.02 and 0.025 which are very minimal in comparison to the 0.5 threshold that would be of concern. As a result, there's **no evidence that the fitted model is overly sensitive to any small number of influential observations**.

3.2 Transformations

To address the assumption violations above, we chose to apply a **log transformation to the response variable (cnt)**. In our early EDA, we observed that cnt has a heavily skewed right-tail as observed in Figure 1. Additionally, this transformation is also motivated by the presence of heteroskedasticity, heavy-tailed residuals, and global nonlinearity in the baseline OLS model. The log transformation should stabilize the variance and reduce the influence of extreme outliers. A constant of one is added to accomodate for zero counts.

Important note: the log transformation changes our interpretation of estimated coefficients to **approximate percent change in hourly bike rentals**.

3.3 Refit model and compare

Table 5: In-sample model comparison metrics for the baseline and log-transformed OLS models.

	Model	Adj_R^2	AIC	BIC
0	Baseline OLS	0.681128	210260.540315	210578.824047
1	Log-Transformed OLS	0.825787	31126.973333	31445.257065

Our model comparison includes in-sample (data seen) and out-of-sample (data unseen) metrics. In-sample statistics such as adjusted R^2, AIC, and BIC provide descriptive measures of fit for their respective response scales (baseline vs. log-transformed OLS). In Table 5, the log-transformed model has a substantially higher adjusted R^2 indicating **improved explanatory power** i.e., closer fit to the observed data, but not a definitive improvement in fit from the original scale.

Whereas the large decrease in AIC and BIC values, this tells us that the log-transformed model improved in-sample likelihood fit i.e., the **model's assumptions about noise and stability match the data better** - likely due to improved normality and stabilizing residual variance.

In contrast, the out-of-sample evaluation using k-fold cross-validated RMSE is **deferred to part IV** and will serve as the primary criterion for predictive performance and model selection.

Table 6: Correlation matrix for numeric predictors.

	temp	hum	windspeed
temp	1.000000	-0.069881	-0.023125
hum	-0.069881	1.000000	-0.290105
windspeed	-0.023125	-0.290105	1.000000

4 Collinearity Assessment

4.1 Correlation matrix with numeric predictors

4.2 VIF analysis

Table 7: Top 15 Variance Inflation Factors (VIF) for predictors in the log-transformed OLS model.

	variable	VIF
30	C(weekday)[T.4]	inf
31	C(weekday)[T.5]	inf
27	C(weekday)[T.1]	inf
29	C(weekday)[T.3]	inf
28	C(weekday)[T.2]	inf
41	holiday	inf
40	workingday	inf
34	C(season)[T.3]	4.228093
37	temp	3.105377
33	C(season)[T.2]	2.498219
18	C(hr)[T.15]	2.038059
19	C(hr)[T.16]	2.033286
17	C(hr)[T.14]	2.031779
20	C(hr)[T.17]	2.014788
16	C(hr)[T.13]	2.014241

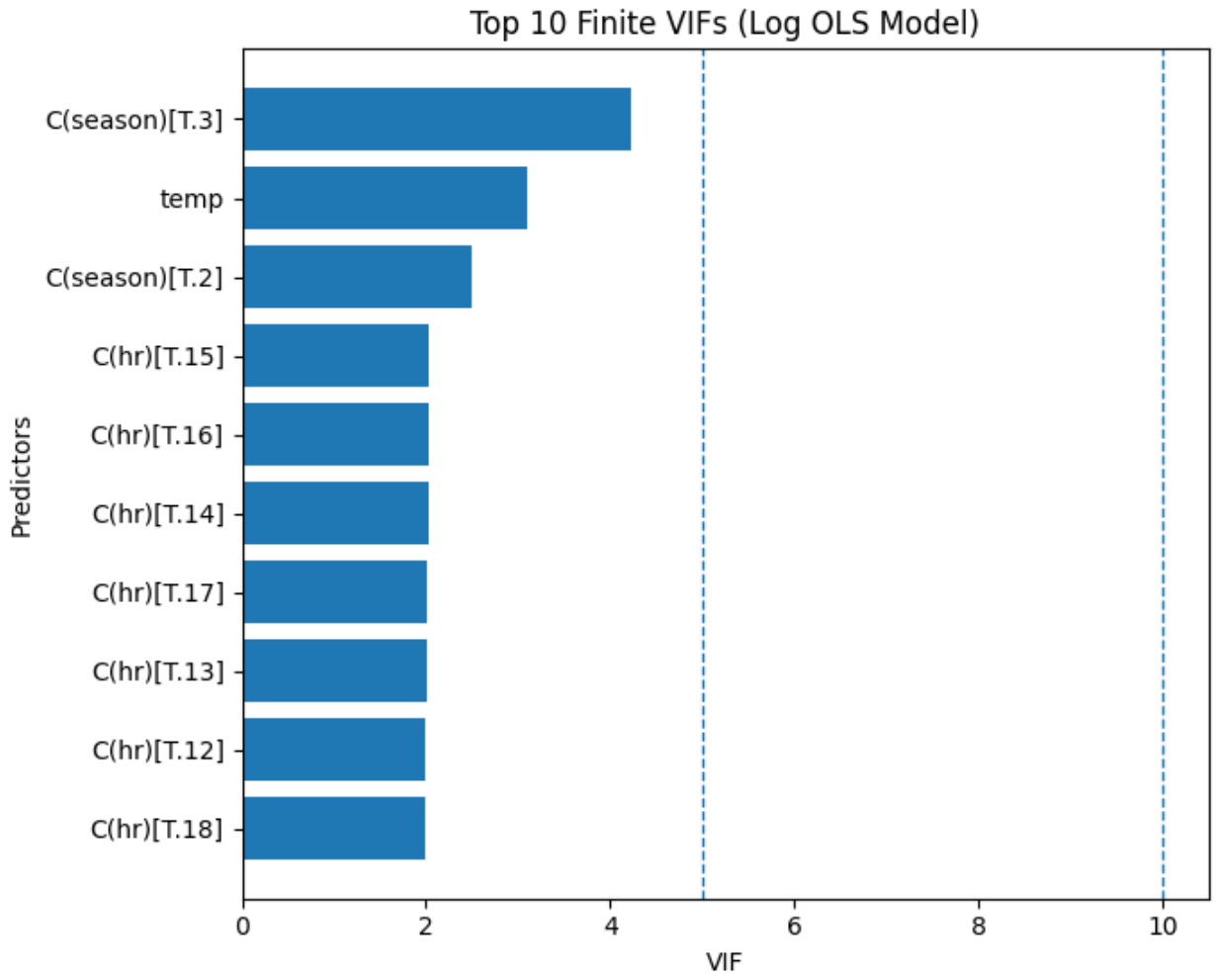

Figure 9: Top finite variance inflation factors (VIF) for predictors in the log-transformed OLS model. Predictors with infinite VIF arise from exact multicollinearity among dummy-encoded categorical variables and are omitted from this plot.

4.3 Discussion of collinearity effects

In Table 6, the correlation matrix among numerically continous predictors (temp, hum, and windspeed) indicates no strong pairwise linear relationships. While humidity and windspeed show a modest negative association (~ -0.29), none of the absolute correlation values approach problematic levels for collinearity (e.g., $|p| > 0.7$).

In Table 7, VIF values are reported in descending order. Several categorical and binary variables with **inf VIF values indicate perfect multicollinearity or exact linear combinations** of one another once encoding is applied e.g., workingday is defined by weekday and holiday with overlapping calendar information. The remaining variables as shown in Figure 9 are considered in moderate VIF values (~ 2-5) or low values (<= 1) without raising concerns for multicollinearity. However, it's important to highlight the potential collinearity between season and temp at the top

Table 8: RMSE mean and standard deviation for baseline log cnt OLS model.

```
Baseline log cnt OLS model RMSE: 0.5925789002917841 SD: 0.006101317168347299
```

of the list. Since temperature changes with the season, these two variables are likely to provide overlapping information. Something we will explore more in section IV.3 for model selection.

As a result, calendar-related categorical and binary variables with perfect multicollinearity, due to overlapping definitions, can lead to **variance inflation of standard errors and complicated coefficient interpretations** as discusses in section I.4. Although collinearity complicates interpretability and statistical inference, it **does not necessarily harm predictive performance** because predictions depend on combined linear predictors with redundant variables that may jointly capture meaningful structure in the data.

Several strategies can be used to address collinearity, depending on **whether the primary goal is interpretability or predictive performance**. Common strategies include: 1) Remove redundant predictors: remove workingday and keep weekday & holiday to improve interpretation 2) Redesign variable encodings: represent calendar effects solely with weekday indicators 3) Combine correlated variables: temp_index = (temp + atemp) / 2 4) Apply regularization techniques: Ridge or Lasso to preserve predictive performance

5 Model Selection and Validation

5.1 Cross-validation & Utilities Setup

Here, we define 3 functions to leverage in model selection and validation.

1) cv_metric_formula: to calculate MSE or RMSE given the parameters (nested inside stepwise_selection_cv)
2) stepwise_selection_cv: to conduct stepwise selection cross-validation based on RMSE or MSE as primary criterion
3) stepwise_selection_aic_bic: to conduct stepwise selection cross-validation based on AIC or BIC as primary criterion

Table 9: Stepwise five-fold cross-validation with Root Mean Squared Error (RMSE) for candidate log-transformed OLS models.

```
Start: RMSE=1.417834 (sd=0.014040) | log_cnt ~ 1
ADD  C(hr)                -> RMSE=0.774129 (sd=0.003304)
ADD  temp                 -> RMSE=0.683957 (sd=0.005311)
ADD  C(yr)                -> RMSE=0.643233 (sd=0.005164)
ADD  C(weathersit)        -> RMSE=0.621278 (sd=0.005742)
ADD  C(season)            -> RMSE=0.597358 (sd=0.005879)
ADD  C(weekday)           -> RMSE=0.594297 (sd=0.006039)
ADD  hum                  -> RMSE=0.593732 (sd=0.005565)
ADD  windspeed            -> RMSE=0.593172 (sd=0.005992)
ADD  workingday           -> RMSE=0.592592 (sd=0.006158)

Best formula: log_cnt ~ C(hr) + temp + C(yr) + C(weathersit) + C(season) + C(weekday) + hum + w
RMSE: 0.5925916101069012 SD: 0.006158486580407104
```

5.2 Baseline model CV RMSE results

5.3 Stepwise selection

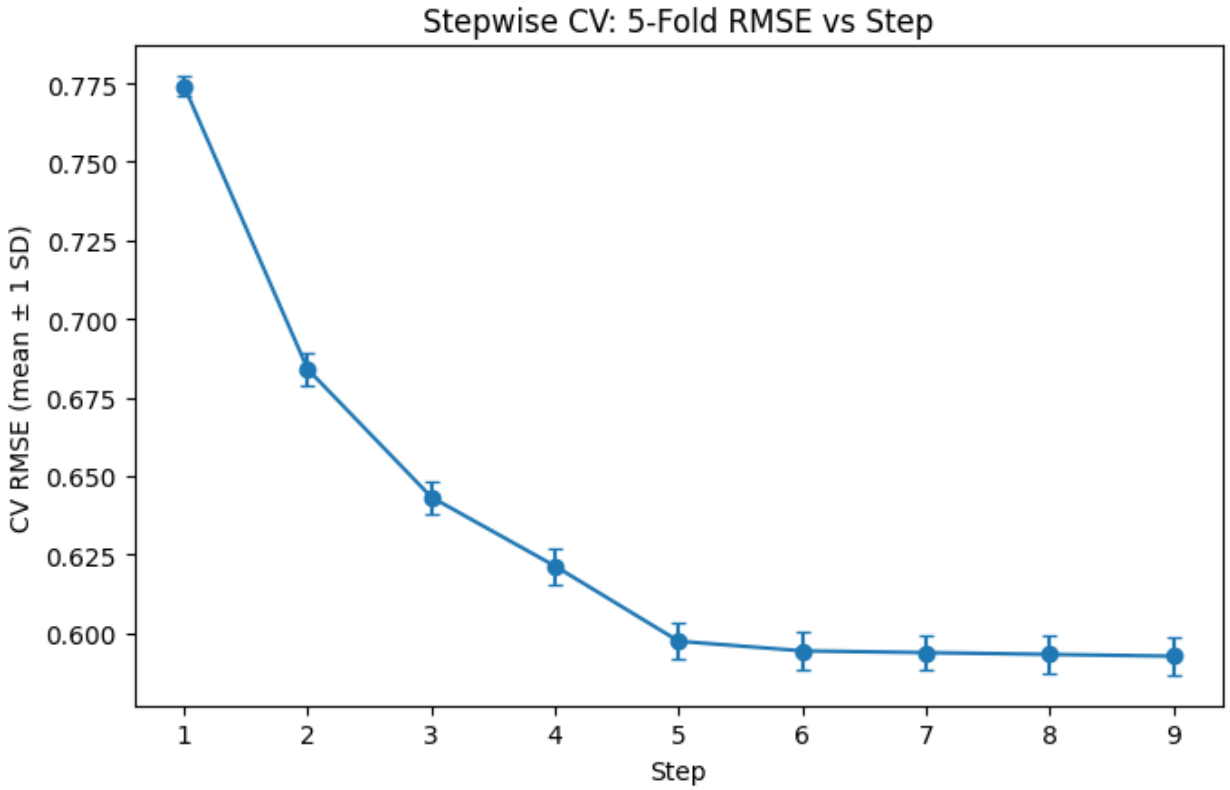

Figure 10: Stepwise five-fold cross-validation path showing mean RMSE (±1 SD) versus step index for log-transformed OLS models.

Table 10: Stepwise five-fold cross-validation with Mean Squared Error (MSE) for candidate log-transformed OLS models.

```
Start: MSE=2.010411 (sd=0.039871) | log_cnt ~ 1
ADD  C(hr)                -> MSE=0.599285 (sd=0.005114)
ADD  temp                 -> MSE=0.467819 (sd=0.007279)
ADD  C(yr)                -> MSE=0.413770 (sd=0.006639)
ADD  C(weathersit)        -> MSE=0.386012 (sd=0.007119)
ADD  C(season)            -> MSE=0.356864 (sd=0.007008)
ADD  C(weekday)           -> MSE=0.353218 (sd=0.007165)
ADD  hum                  -> MSE=0.352542 (sd=0.006600)
ADD  windspeed            -> MSE=0.351882 (sd=0.007098)
ADD  workingday           -> MSE=0.351195 (sd=0.007287)

Best formula: log_cnt ~ C(hr) + temp + C(yr) + C(weathersit) + C(season) + C(weekday) + hum + w
MSE: 0.3511951579346585 SD: 0.007286777545519475
```

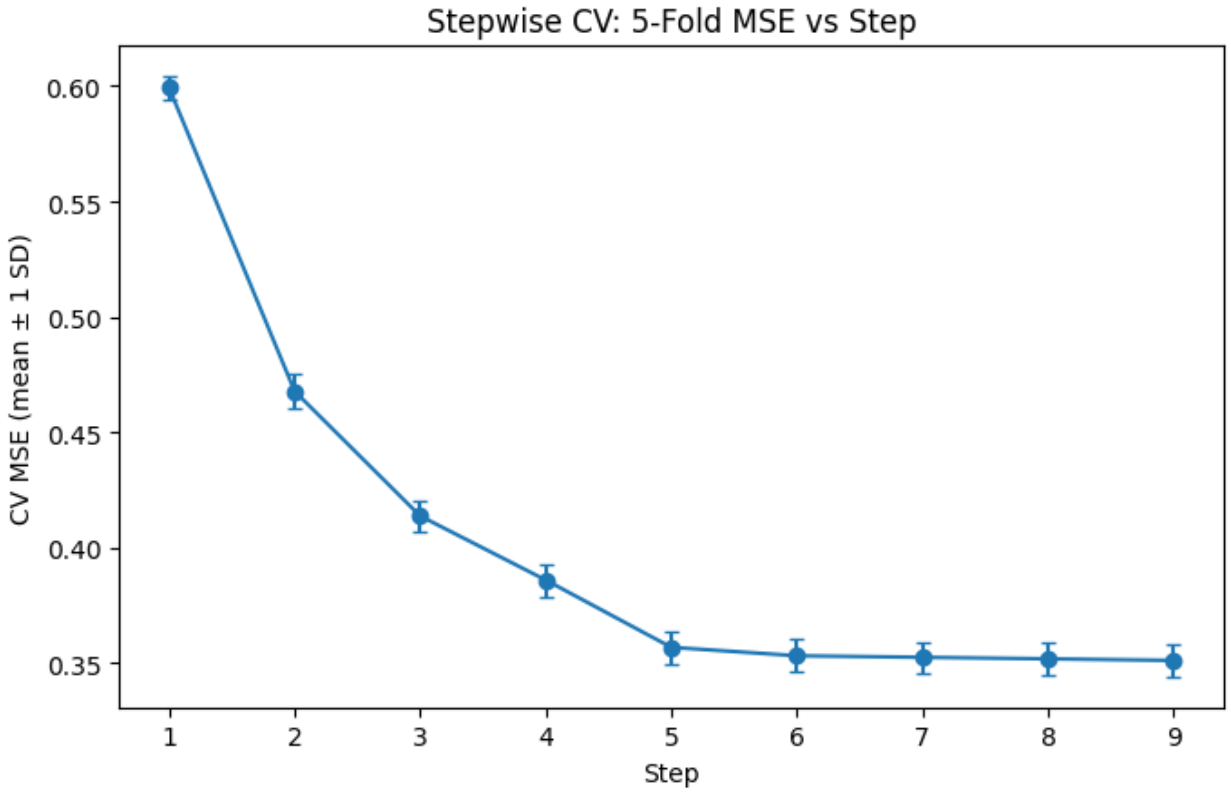

Figure 11: Stepwise five-fold cross-validation path showing mean MSE (±1 SD) versus step index for log-transformed OLS models.

Table 11: Stepwise five-fold cross-validation with BIC for candidate log-transformed OLS models.

```
Start: BIC=61464.164802 | log_cnt ~ 1
ADD  C(hr)                -> BIC=40593.280281
ADD  temp                 -> BIC=36299.400639
ADD  C(yr)                -> BIC=34178.461775
ADD  C(weathersit)        -> BIC=32991.065329
ADD  C(season)            -> BIC=31652.933488
ADD  C(weekday)           -> BIC=31523.283631
ADD  hum                  -> BIC=31495.422792
ADD  windspeed            -> BIC=31469.720397
ADD  workingday           -> BIC=31444.832698

Best formula: log_cnt ~ C(hr) + temp + C(yr) + C(weathersit) + C(season) + C(weekday) + hum + w

Score: 31444.83269754891
```

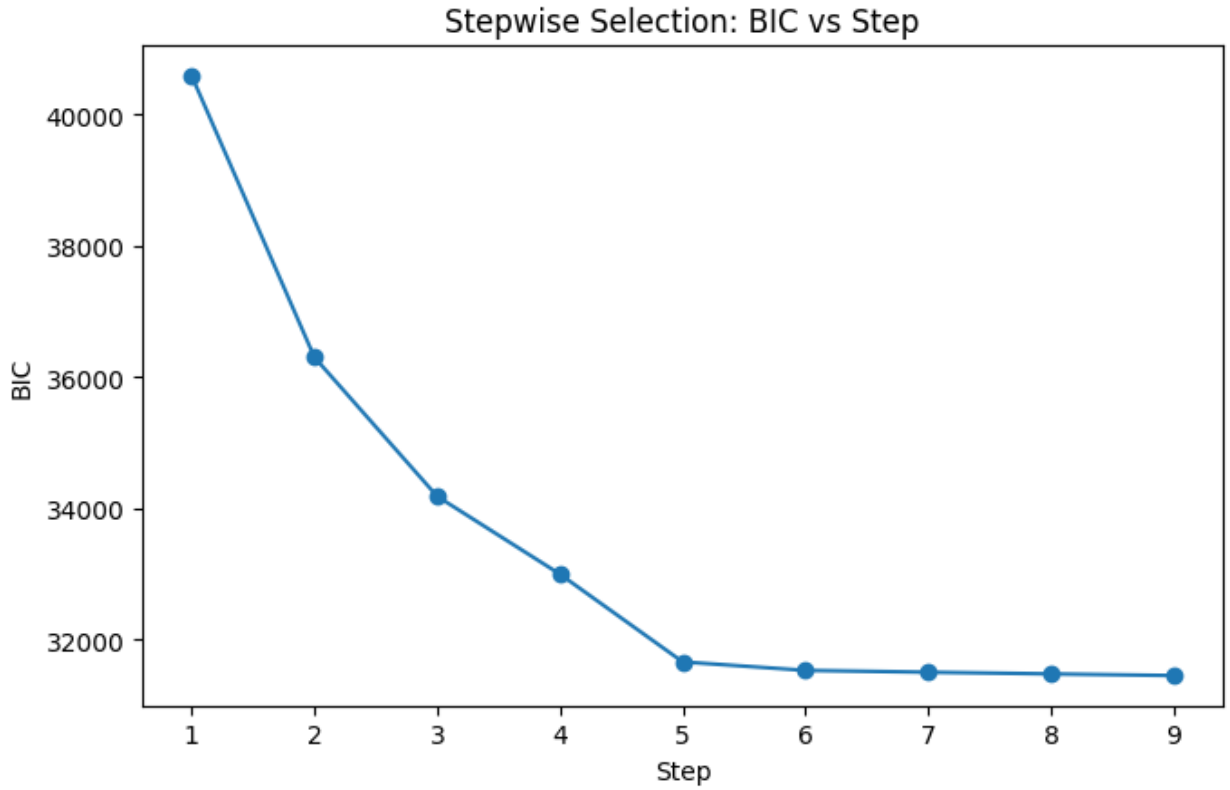

Figure 12: Stepwise selection path showing Bayesian Information Criterion (BIC) versus step index for log-transformed OLS models.

Our model selection was performed using **stepwise 5-fold cross-validation (CV) with RMSE** as the selection criterion as mentioned in section III.3 and shown in Table 9 and Figure 10. Given

Table 13: CV RMSE evaluation for selected OLS model for log cnt.

```
Start: RMSE=1.417834 (sd=0.014040) | log_cnt ~ 1
ADD  C(hr)                -> RMSE=0.774129 (sd=0.003304)
ADD  temp                 -> RMSE=0.683957 (sd=0.005311)
ADD  C(yr)                -> RMSE=0.643233 (sd=0.005164)
ADD  C(weathersit)        -> RMSE=0.621278 (sd=0.005742)
ADD  C(season)            -> RMSE=0.597358 (sd=0.005879)
RMSE: 0.5973581846445841 SD: 0.0058789993280516616
```

the size of the dataset, 5-fold CV was adequate for our analysis within considerations for time and resources. The motivation for using RMSE is that CV RMSE directly **evaluates out-of-sample predictive performance** compared to AIC/BIC and better for interpretation than MSE with the same penalization for outliers. Unlike likelihood-based criteria (AIC/BIC), CV RMSE makes minimal assumptions and provides a robust assessment of **generalization error** especially in the presence of correlated predictors and large sample sizes.

In the stepwise selection logs for RMSE in Table 9 and depicted in Figure 10, significant reductions in error were observed when adding core time and weather-related predictors (C(hr), temp, C(yr), C(weathersit), and C(season)). However, beyond these core predictors, additional predictors produced only **marginal improvements in RMSE**, with reductions within one standard deviation of the CV estimates. This flattening or convergence of the RMSE curve (~0.592-0.597) shows a **diminishing predictive returns** and led us to a selection of a reduced model that balanced **predictive accuracy with minimal complexity for interpretation**.

Among these, one key observation was the **complete exclusion of the holiday predictor** while retaining both weekday and workingday. These specifications could highlight the value of preserving some relevant calendar and weather structures while eliminating weak or redundant predictors like hum, windspeed, and workingday, without sacrificing out-of-sample predictive performance.

Lastly, we also experimented with stepwise 5-fold cross-validation with MSE and BIC as shown in Table 10 and Table 11. Overall, both experiments (MSE and BIC) were **consistent with the same observations** as mentioned above. Criterion scores begin to converge (MSE: ~0.351-0.353 / BIC: ~31,444-31,523) after adding C(hr), temp, C(yr), C(weathersit), and C(season) with minimal additional benefit from including remaining variables.

5.4 Compare selected models vs. baseline

Table 12: In-sample evaluation for Baseline vs. Selected OLS model for log cnt using Adjusted R^2, AIC, and BIC.

	Model	Adj_R^2	AIC	BIC
0	Baseline log OLS	0.825786	31127.080796	31445.364528
1	Selected model	0.822891	31404.516917	31652.933488

Table 14: Ridge regression: cross-validated selection of the regularization parameter (alpha) and predictive performance.

```
Ridge Regression with 5-fold CV selection
Best ridge alpha: 54.62277217684348
```

The reduced/selected model, where the variables with negligible out-of-sample CV RMSE improvements were excluded, was compared to the baseline log-transformed OLS model using both in-sample metrics (adjusted R^2, AIC, and BIC) and out-of-sample CV RMSE. The in-sample comparison is shown in Table 12, whereas the CV RMSE comparison is presented in Table 9 and Table 13, as discussed in section IV.3.

The adjusted R^2 of the selected model slightly lower than the baseline model, reflecting the intentional removal of predictors. However, the difference is small, showing that the **reduced model retains most of the explanatory power with fewer predictors**. This is expected since adjusted R^2 penalizes complexity minimally and, as a result, not designed as a primary model selection criterion.

Both the AIC and BIC increased for the reduced model. This indicates a preference for the larger baseline model under likelihood-based criteria. While this was initially surprising since we expected BIC to penalize complexity heavily, it is likely driven by the large sample size because **small improvements in log-likelihood over many observations can outweigh the complexity penalty** especially for predictors with subtle but detectable effects.

Most importantly, this does not contradict our CV RMSE findings. AIC and BIC evaluate goodness-of-fit through the lens of likelihood i.e., goodness-of-fit to observed data, whereas CV RMSE evaluates predictive stability on unseen data. Our findings highlight a **modeling trade-off** between likelihood criteria that favor richer models in large samples, while CV emphasizes gains in predictive performance. In our case, our decision for the reduced model prioritizes generalization performance over marginal in-sample improvements and avoids over-interpreting weak effects or collinearity among time and calendar-related predictors.

In the next section, we'll explore regularization techniques that may allow all predictors to remain in the baseline model while shrinking correlated coefficients toward each other or possibly shrinking the same weaker variables that we removed to zero as part of feature selection.

Table 15: Lasso regression: cross-validated selection of the regularization parameter (alpha) and predictive performance.

```
Lasso Regression with 5-fold CV selection
alpha_min: 0.0009326033468832199
alpha_1se: 0.010476157527896652
min mean MSE: 0.3698843993068294
1-SE threshold: 0.3898582268015863
```

6 Ridge and Lasso Regression

6.1 Ridge Regression

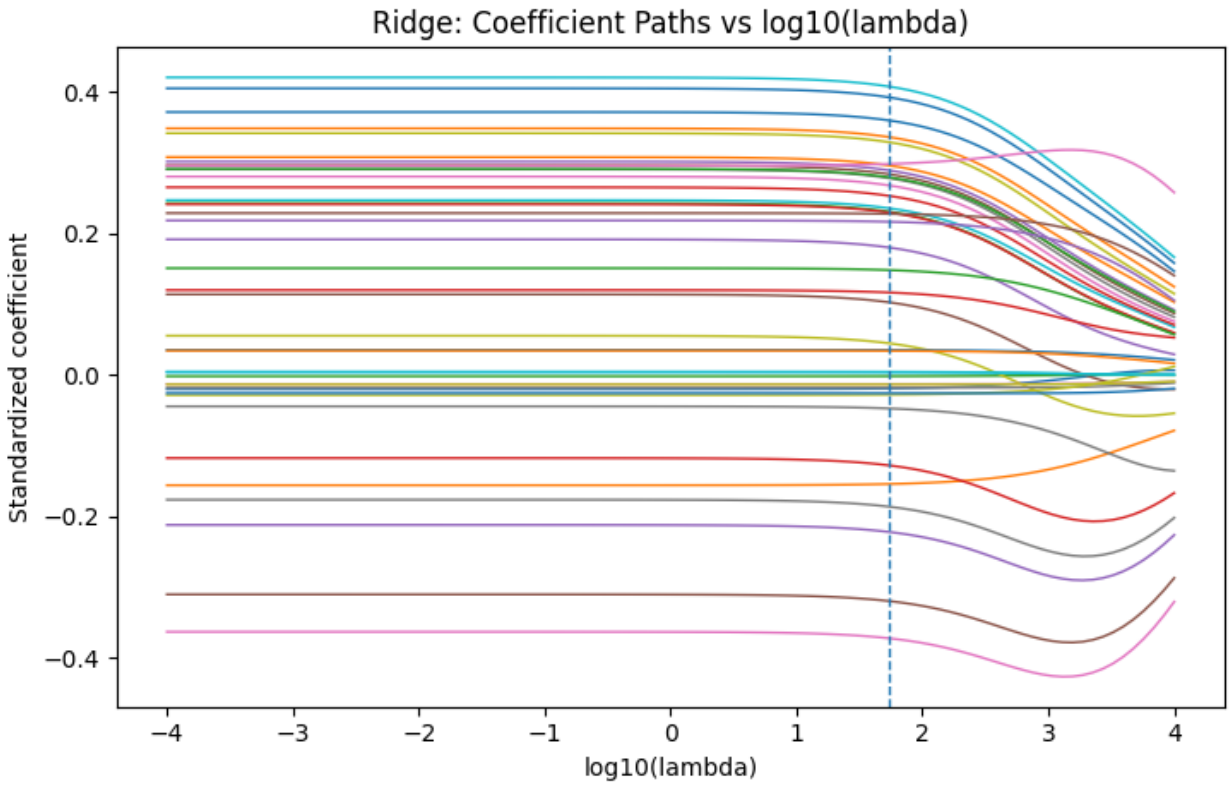

Figure 13: Ridge regression coefficient paths showing standardized coefficients as a function of log10( ), with the dashed line indicating the cross-validated optimal regularization parameter.

6.2 Lasso regression

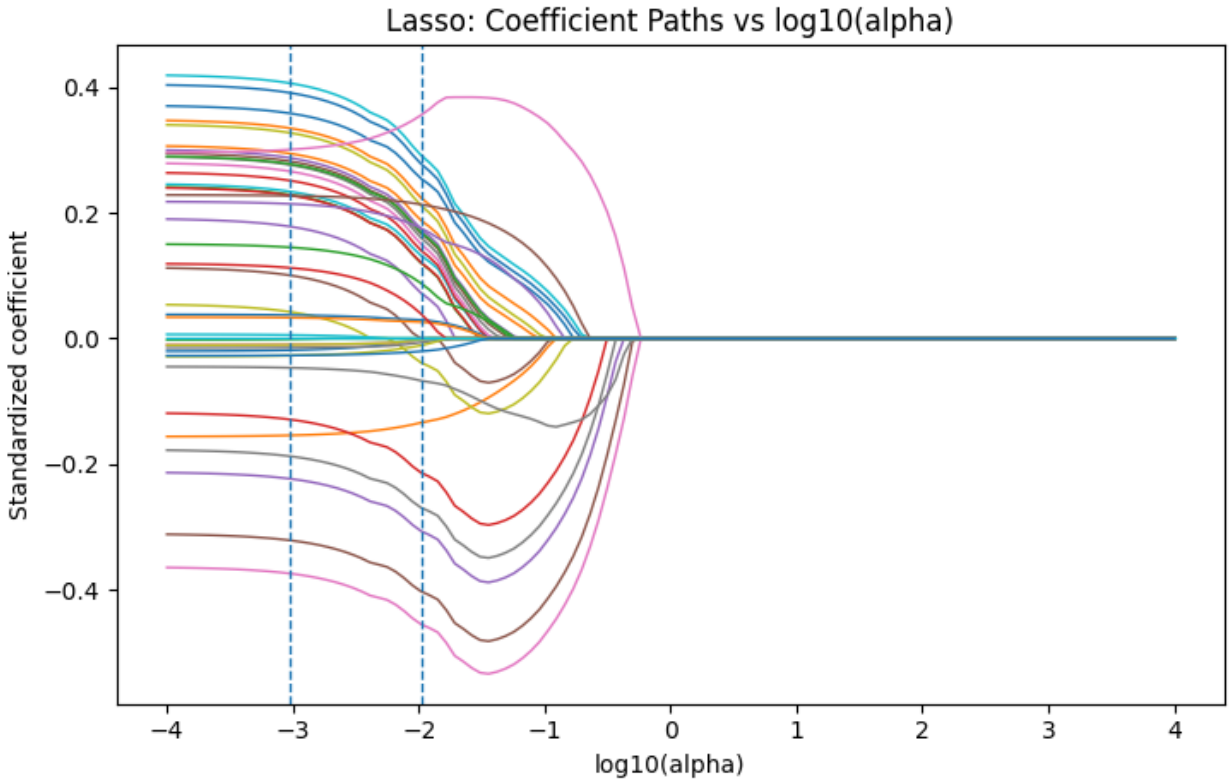

Figure 14: Lasso coefficient paths showing standardized coefficients as a function of log10(alpha), with dashed lines indicating alpha_min (minimum CV error) and alpha_1se (one-standard-error rule).

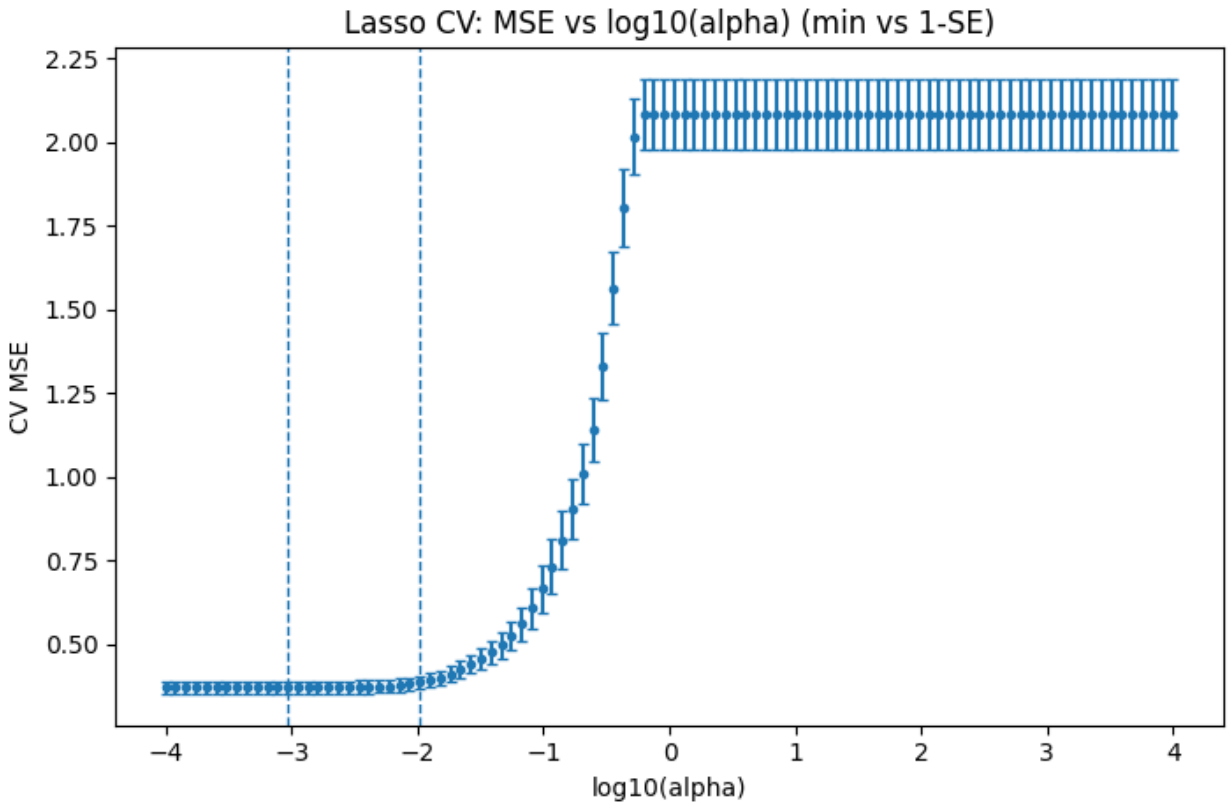

Figure 15: Lasso cross-validation curve showing mean CV MSE (±1 SE) as a function of log10(alpha), with dashed lines indicating alpha_min and alpha_1se.

Table 16: Comparison of predictors selected by lasso regression under the minimum cross-validated error rule (alpha_min) and the one-standard-error rule (alpha_1se). For each model, retained predictors and their corresponding coefficients are shown; blank cells indicate predictors not selected under the given regularization level.

	Predictor (alpha_min)	Coefficient (alpha_min)	Predictor (alpha_1se)	Coefficient (alpha_1se)
0	C(hr)[T.17]	0.406811	C(hr)[T.4]	-0.455284
1	C(hr)[T.18]	0.391467	C(hr)[T.3]	-0.402966
2	C(hr)[T.4]	-0.374228	temp	0.356028
3	C(hr)[T.8]	0.358817	C(hr)[T.2]	-0.306849
4	C(hr)[T.19]	0.335117	C(hr)[T.17]	0.291569
5	C(hr)[T.16]	0.327749	C(hr)[T.18]	0.277849
6	C(hr)[T.3]	-0.321401	C(hr)[T.5]	-0.269552
7	temp	0.300879	C(hr)[T.8]	0.254554
8	C(hr)[T.9]	0.294801	C(hr)[T.19]	0.223850
9	C(hr)[T.12]	0.287836	C(hr)[T.1]	-0.213825
10	C(hr)[T.13]	0.282424	C(yr)[T.1]	0.213509
11	C(hr)[T.20]	0.278079	C(hr)[T.16]	0.211511
12	C(hr)[T.15]	0.276665	C(hr)[T.9]	0.187935
13	C(hr)[T.14]	0.266417	C(season)[T.4]	0.175481

Table 16: Comparison of predictors selected by lasso regression under the minimum cross-validated error rule (alpha_min) and the one-standard-error rule (alpha_1se). For each model, retained predictors and their corresponding coefficients are shown; blank cells indicate predictors not selected under the given regularization level.

	Predictor (alpha_min)	Coefficient (alpha_min)	Predictor (alpha_1se)	Coefficient (alpha_1se)
14	C(hr)[T.11]	0.251702	C(hr)[T.12]	0.174027
15	C(hr)[T.7]	0.234265	C(hr)[T.20]	0.168766
16	C(hr)[T.10]	0.229266	C(hr)[T.13]	0.167215
17	C(hr)[T.21]	0.228082	C(hr)[T.15]	0.160057
18	C(yr)[T.1]	0.227399	C(hr)[T.14]	0.150109
19	C(hr)[T.2]	-0.223349	C(hr)[T.11]	0.140110
20	C(season)[T.4]	0.214505	C(weathersit)[T.3]	-0.133944
21	C(hr)[T.5]	-0.187458	C(hr)[T.7]	0.131777
22	C(hr)[T.22]	0.178496	C(hr)[T.21]	0.120582
23	C(weathersit)[T.3]	-0.154328	C(hr)[T.10]	0.119946
24	C(season)[T.2]	0.145269	C(season)[T.2]	0.087729
25	C(hr)[T.1]	-0.128913	C(hr)[T.22]	0.072219
26	C(season)[T.3]	0.112694	hum	-0.067306
27	C(hr)[T.23]	0.100751	C(season)[T.3]	0.038873
28	hum	-0.046635	C(hr)[T.6]	-0.038587
29	C(hr)[T.6]	0.042595	C(weekday)[T.5]	0.029429
30	C(weekday)[T.5]	0.037252	C(weekday)[T.6]	0.026240
31	C(weekday)[T.6]	0.033168	holiday	-0.020356
32	windspeed	-0.027126	windspeed	-0.010901
33	holiday	-0.026923	C(weekday)[T.2]	-0.006324
34	C(weathersit)[T.2]	-0.018609	C(weekday)[T.1]	-0.002233
35	C(weekday)[T.2]	-0.014622	C(weekday)[T.3]	-0.001446
36	C(weekday)[T.3]	-0.009896	C(weathersit)[T.2]	-0.000602
37	C(weekday)[T.1]	-0.009535	NaN	NaN
38	C(weekday)[T.4]	0.005620	NaN	NaN
39	C(weathersit)[T.4]	-0.001120	NaN	NaN

Table 17: Predictors shrunk to zero by lasso under the minimum cross-validated error rule (alpha_min) and the one-standard-error rule (alpha_1se).

	Zeroed out (alpha_min)	Zeroed out (alpha_1se)
0	workingday	C(hr)[T.23]
1		C(weathersit)[T.4]
2		C(weekday)[T.4]
3		workingday

6.3 Selected OLS vs. Ridge vs. Lasso comparison

Table 18: Comparison of selected OLS, ridge, and lasso models based on cross-validated predictive performance and regularization characteristics, including the one-standard-error (1-SE) lasso solution.

	Model	Selected alpha	CV RMSE (mean)	CV RMSE (SD)
0	Selected OLS	—	0.597358	0.005879
1	Ridge	54.622772	0.592777	0.006014
2	Lasso (alpha_min)	0.000933	0.592737	0.005986
3	Lasso (alpha_1se)	0.010476	0.607567	0.005406

In this section (V), we explore Ridge and Lasso regularization techniques intended to **improve generalization in the presence of predictors with strong collinearity**. By penalizing large coefficients and reducing variance, regularization stabilizes estimates of correlated predictors (e.g., many encoded hour (hr) or categorical predictors). Additionally, it provides a bias-variance tradeoff that could help ensure a balance in-sample fit to reliable out-of-sample predictive performance. Both regularization techniques were **performed on the baseline log transformed OLS model** before reduction in predictors.

In Table 14, we leveraged a 5-fold CV to select a near optimal regularization parameter (alpha = 54.62) while taking advantage of the Ridge squared coefficient penalty that shrinks them toward zero while **sharing weights across correlated predictors without removing any** as shown in Figure 13.

Similarly, in Table 15, two regularization parameters (alpha_min and alpha_1se) were derived using 5-fold cross-validation with MSE as the natural loss function. The alpha_min value selects the model that minimizes mean CV MSE, thereby **maximizing predictive accuracy**. In contrast, alpha_1se, defined as the largest regularization parameter whose mean CV MSE lies within one standard error of the minimum, is designed to **favor simpler and more stable models without a meaningful loss in predictive accuracy**.

As shown in Figure 14 and Figure 15, increasing the regularization strength leads to **progressive coefficient shrinkage and variable elimination**. The alpha_min model retains more predictors, with only one coefficient (workingday) shrunk exactly to zero, whereas the alpha_1se model applies stronger regularization and collapses four predictors to zero, as confirmed in Table 16 and Table 17. Together, these visuals illustrate the impact of increasing regularization on both predictive error (MSE) and coefficient sparsity.

Overall, Table 18 supports the following conclusions: 1) In comparison to our reduced/selected log-transformed OLS model (Selected OLS in Figure 14), we observe **marginal improvements of approximately 0.005 RMSE** in Ridge and Lasso (alpha_min). Differences in RMSE are small relative to their SDs with **no meaningful separation in predictive accuracy**. Regularization helps modestly, but predictive performance is not the main differentiator 2) However, in terms of interpretability, Lasso (alpha_1se) removed four predictors and produced a substantially **sparser and more interpretable model**, at the cost of a small RMSE increase 3) Lastly, in terms of regularization effects on collinearity, Ridge effectively shrank correlated coefficients together (e.g, the many correlated categorical predictors like C(hr)) **without removing any predictors**

7 Conclusion

7.1 Summary of findings

This project analyzed hourly bike-sharing demand using a combination of exploratory analysis, classical regression, and regularized models to **balance predictive accuracy, interpretability, and stability**. The goal was to identify key drivers for hourly bike rentals across a high-dimensional feature space dominated by categorical temporal and weather predictors.

Across all models, the time of day or hour emerged the strongest predictor of bike rental demand, reflecting clear commuting and leisure patterns. Temperature showed a significant positive association, while weather conditions and seasonal effects contributed secondary but meaningful adjustments. With time of day effects include, several calendar variables like workingday, weekday, or holiday provided limited additional explanatory power, likely due to redundancy among correlated predictors.

From a predictive accuracy standpoint, the reduced OLS, Ridge, and Lasso (alpha_min) achieved nearly identical CV RMSE. This highlights the fact that prediction alone does not fully justify regularization in this case. However, the **primary value of regularization was seen in stability and interpretability**. By applying the 1-SE rule to Lasso, our feature selection produced a substantially sparser model with little loss in RMSE, while improving interpretability.

In conclusion, the results show that bike-sharing demand is strongly associated to temporal structure and weather-related factors.