Seneca POLYTECHNIC

Statistics for Analytics

BAN100NAA - PROFESSOR: SAMANEH GHOLAMI

Maaz Hussain
STUDENT ID: 173714221

ASSIGNMENT 4:

REGRESSION ANALYSIS

Problem 1

Bicycle Sharing System

Introduction

The goal of this analysis is to develop a multiple regression model to predict the total number of bike rentals (count) based on various factors such as weather conditions, user type (casual and registered), and other contextual variables. By leveraging the dataset from the Capital Bikeshare program, we aim to identify the key predictors influencing bike rental patterns, assess the model's statistical significance, and evaluate its ability to inform planning decisions. The analysis includes stepwise regression, residual evaluation, and testing for potential issues like multicollinearity to ensure the robustness and interpretability of the results.

The CONTENTS Procedure

Data Set Name	WORK.BIKES	**Observations**	10886
Member Type	DATA	**Variables**	12
Engine	V9	**Indexes**	0
Created	12/03/2024 12:20:55	**Observation Length**	96
Last Modified	12/03/2024 12:20:55	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	9
First Data Page	1
Max Obs per Page	1363
Obs in First Data Page	1328
Number of Data Set Repairs	0
Filename	/saswork/SAS_work24E20000A8D3_odaws02-usw2.oda.sas.com/SAS_workAEB00000A8D3_odaws02-usw2.oda.sas.com/bikes.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	1610720470
Access Permission	rw-r--r--
Owner Name	u64008511
File Size	1MB
File Size (bytes)	1310720

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
7	atemp	Num	8	BEST12.	BEST32.
10	casual	Num	8	BEST12.	BEST32.
12	count	Num	8	BEST12.	BEST32.
1	datetime	Num	8	YYMMDD10.	YYMMDD10.
3	holiday	Num	8	BEST12.	BEST32.
8	humidity	Num	8	BEST12.	BEST32.
11	registered	Num	8	BEST12.	BEST32.
2	season	Num	8	BEST12.	BEST32.
6	temp	Num	8	BEST12.	BEST32.
5	weather	Num	8	BEST12.	BEST32.
9	windspeed	Num	8	BEST12.	BEST32.
4	workingday	Num	8	BEST12.	BEST32.

Obs	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
1	2011-01-01	1	0	0	1	9.84	14.395	81	0	3	13	16
2	2011-01-01	1	0	0	1	9.02	13.635	80	0	8	32	40

The dataset metadata (WORK.BIKES) contains 10,886 observations and 12 variables, structured for regression analysis with numeric data types. Key variables include predictors like temp, humidity, and windspeed, alongside target variable count. The dataset preview highlights attributes such as datetime, season, and user types (casual, registered), ensuring the data is complete, properly formatted, and ready for predictive modeling.

The REG Procedure
Model: MODEL1
Dependent Variable: count

Number of Observations Read	10886
Number of Observations Used	10886

Stepwise Selection: Step 1

Variable registered Entered: R-Square = 0.9427 and C(p) = .

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	336721273	336721273	179197	<.0001
Error	10884	20451641	1879.05558		
Corrected Total	10885	357172914			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	10.43680	0.59641	575410	306.22	<.0001
registered	1.16448	0.00275	336721273	179197	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable casual Entered: R-Square = 1.0000 and C(p) = .

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	357172914	178586457	Infty	<.0001
Error	10883	0	0		
Corrected Total	10885	357172914			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-3.3286E-14	0	5.69275E-24	Infty	<.0001
casual	1.00000	0	20451641	Infty	<.0001
registered	1.00000	0	186918983	Infty	<.0001

Bounds on condition number: 1.3285, 5.3139

Analysis of Variance (ANOVA Table for Stepwise Selection)

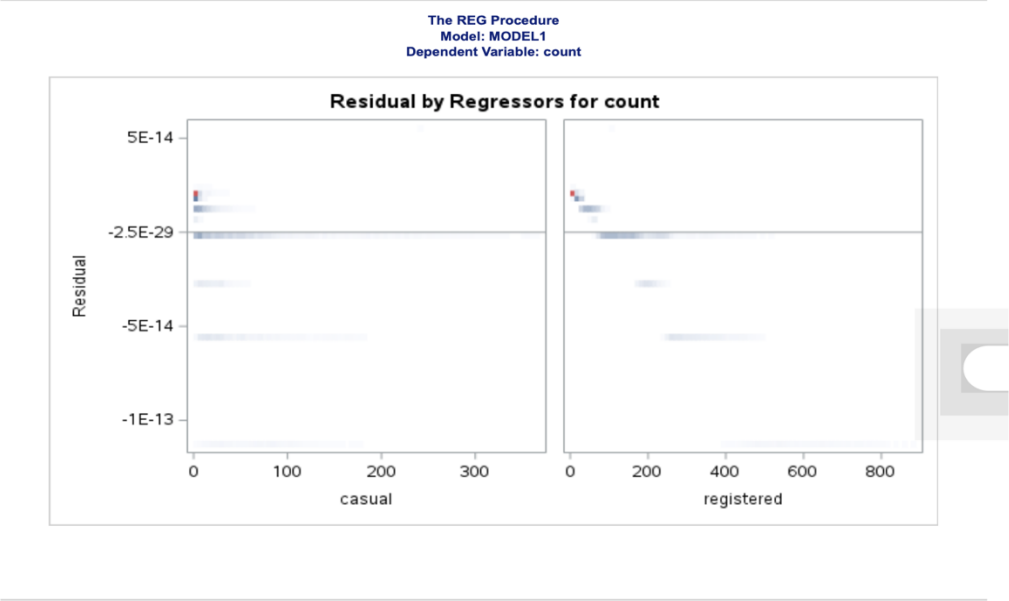
- **Step 1**: The inclusion of registered explains 94.27% of the variance in bike count (R^2 = 0.9427). The model is statistically significant (F-value = 179197, $p < 0.0001$).
- **Step 2**: Adding casual results in a perfect fit (R^2 = 1.0000), meaning 100% of the variance in bike counts is explained by the model.

Results: Program 1

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	registered		1	0.9427	0.9427	.	179197	<.0001
2	casual		2	0.0573	1.0000	.	Infty	<.0001

The stepwise process selected registered and casual as the only variables in the final model. These two variables explain all the variance in bike counts, but this might indicate overfitting.



Plots residuals against casual and registered predictors. The residuals are near zero, indicating a perfect fit; however, this may signal overfitting or multicollinearity.

The REG Procedure
Model: MODEL1
Dependent Variable: count

Number of Observations Read	10886
Number of Observations Used	10886

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	357172914	35717291	Infty	<.0001
Error	10875	0	0		
Corrected Total	10885	357172914			

Root MSE	0	R-Square	1.0000
Dependent Mean	191.57413	Adj R-Sq	1.0000
Coeff Var	0		

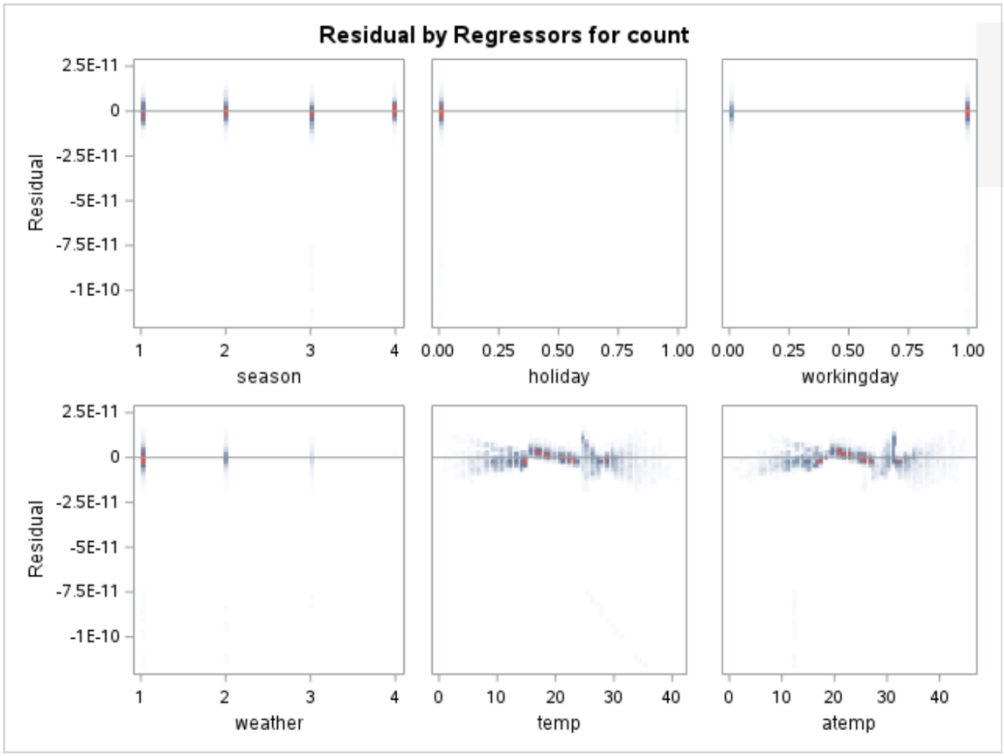
Shows the model explains 100% of the variance ($R^2 = 1.000$) with significant results ($p < 0.0001$). The model appears statistically perfect but raises concerns about practical application due to potential overfitting.

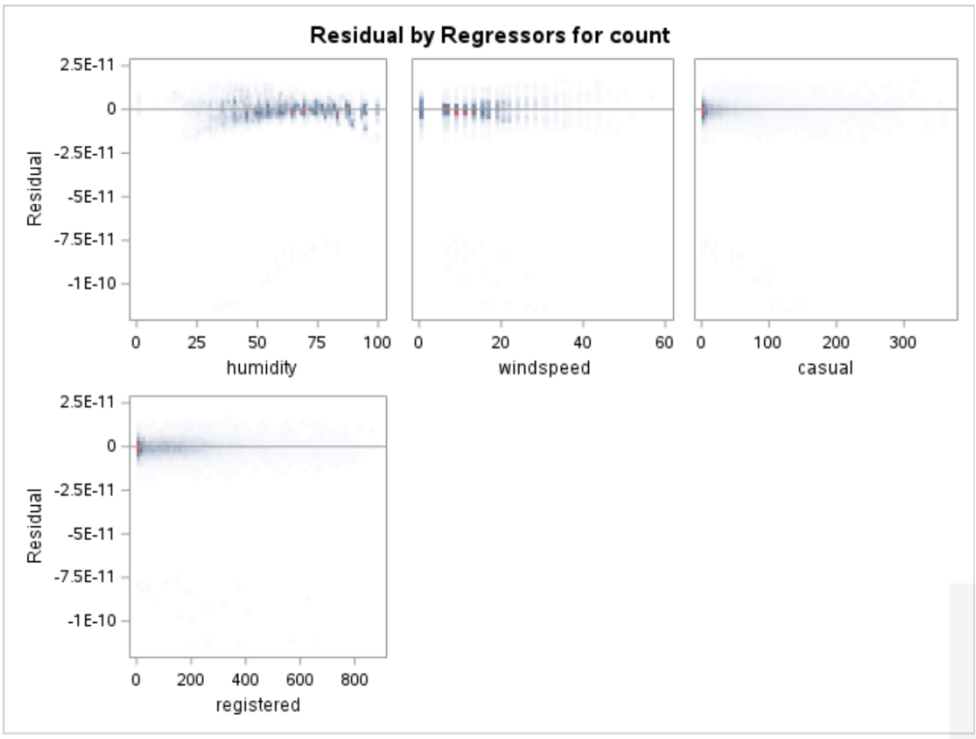
Results: Program 1

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.14751E-12	0	Infty	<.0001
season	1	-8.3325E-14	0	-Infty	<.0001
holiday	1	-1.4797E-12	0	-Infty	<.0001
workingday	1	-5.7545E-13	0	-Infty	<.0001
weather	1	-2.0299E-13	0	-Infty	<.0001
temp	1	4.67295E-12	0	Infty	<.0001
atemp	1	-4.2922E-12	0	-Infty	<.0001
humidity	1	1.86295E-14	0	Infty	<.0001
windspeed	1	-1.9254E-13	0	-Infty	<.0001
casual	1	1.00000	0	Infty	<.0001
registered	1	1.00000	0	Infty	<.0001

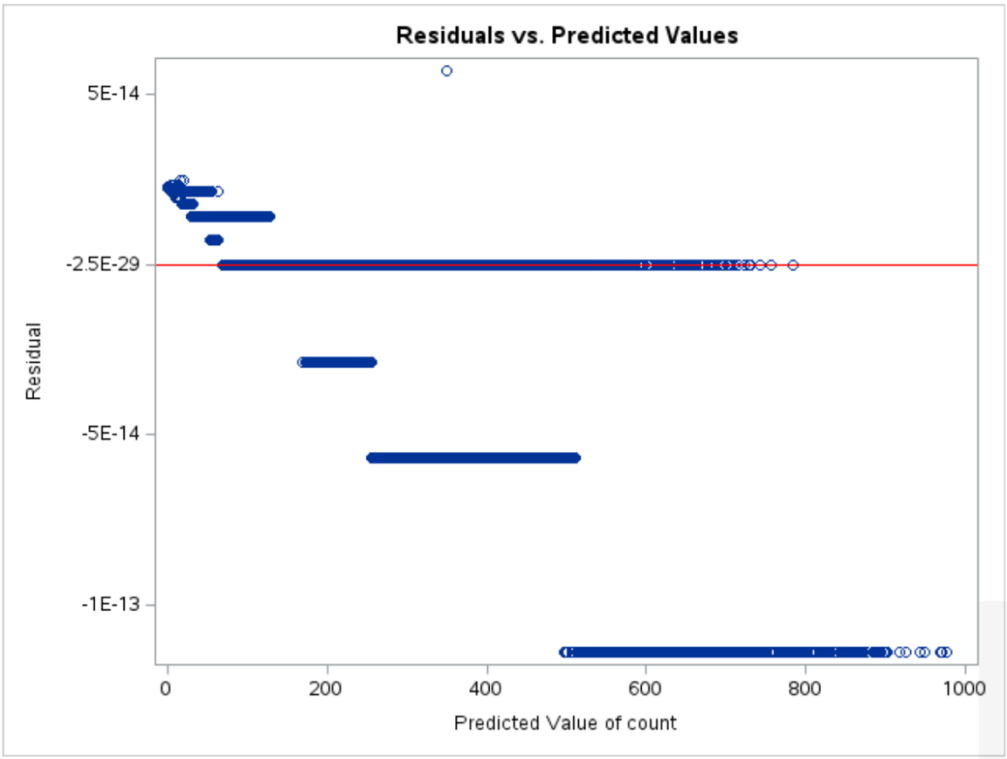
Parameter Estimates Table
Lists coefficients, standard errors, and t-values for predictors. Predictors like casual and registered perfectly predict count (coefficients = 1), indicating redundancy.

The REG Procedure
Model: MODEL1
Dependent Variable: count





The residual plots for all predictors (season, holiday, temp, humidity, windspeed, casual, and registered) show residuals near zero across all levels, indicating a perfect fit of the model. While this may seem ideal, it raises concerns about overfitting and the redundancy of some predictors, such as casual and registered, which dominate the prediction of bike rentals.



This plot compares residuals with predicted bike rental counts. Residuals are effectively zero across all predictions, indicating a perfect model fit. This, however, raises concerns about overfitting.

Residuals vs. Predicted Values

The REG Procedure
Model: MODEL1
Dependent Variable: count

Number of Observations Read	10886
Number of Observations Used	10886

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	357172914	39685879	Infty	<.0001
Error	10876	0	0		
Corrected Total	10885	357172914			

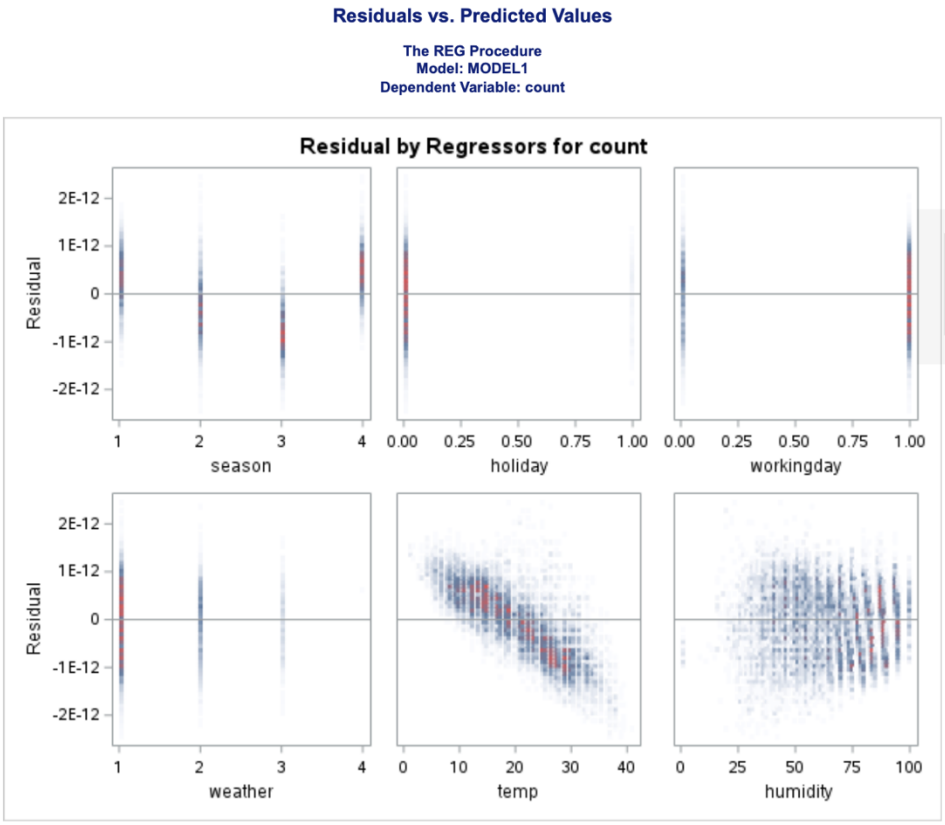
Root MSE	0	R-Square	1.0000
Dependent Mean	191.57413	Adj R-Sq	1.0000
Coeff Var	0		

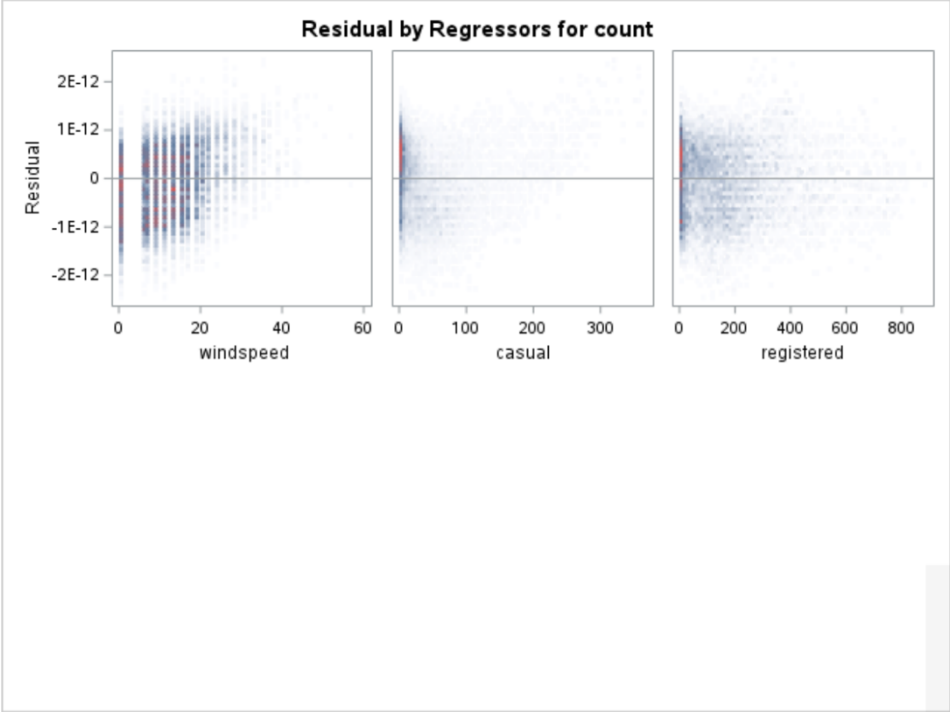
Displays overall model significance and fit statistics (e.g., $R^2 = 1.000$, $p < 0.0001$). The model perfectly explains the variance in count. While this suggests an excellent fit, it highlights concerns about overfitting and multicollinearity.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.7711E-13	0	-Infty	<.0001
season	1	-1.4566E-13	0	-Infty	<.0001
holiday	1	-9.5923E-14	0	-Infty	<.0001

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
workingday	1	-3.6948E-13	0	-Infty	<.0001
weather	1	1.12799E-13	0	Infty	<.0001
temp	1	1.05471E-13	0	Infty	<.0001
humidity	1	-1.0103E-14	0	-Infty	<.0001
windspeed	1	-3.1419E-14	0	-Infty	<.0001
casual	1	1.00000	0	Infty	<.0001
registered	1	1.00000	0	Infty	<.0001

Shows coefficients, standard errors, and significance for variables like workingday, temp, casual, and registered. Predictors casual and registered dominate the model with coefficients = 1.0 and p-values < 0.0001, confirming their importance but suggesting redundancy.





Residuals vs. Predicted Values

The REG Procedure
Model: MODEL1
Dependent Variable: count

Number of Observations Read	10886
Number of Observations Used	10886

Stepwise Selection: Step 1

Variable registered Entered: R-Square = 0.9427 and C(p) = .

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	336721273	336721273	179197	<.0001
Error	10884	20451641	1879.05558		
Corrected Total	10885	357172914			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	10.43680	0.59641	575410	306.22	<.0001
registered	1.16448	0.00275	336721273	179197	<.0001

The analysis for Stepwise Selection: Step 1 reveals that the variable registered is a highly significant predictor of bike rental counts, explaining 94.27% of the variance ($R^2 = 0.9427$, $p < 0.0001$). The regression model, with an F-value of 179,197, confirms the strong predictive power of registered. The coefficient estimate (1.16) indicates that each additional registered user increases bike rentals by 1.16 on average, making this variable critical for the model.

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable casual Entered: R-Square = 1.0000 and C(p) = .

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	357172914	178586457	Infty	<.0001
Error	10883	0	0		
Corrected Total	10885	357172914			

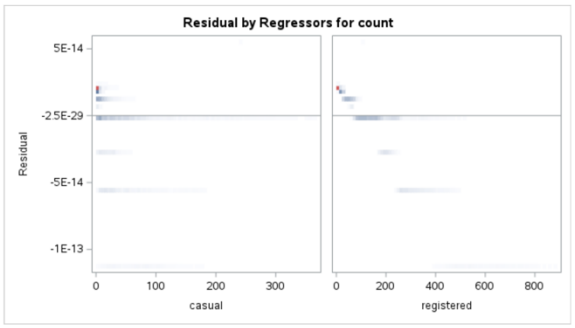
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-3.3286E-14	0	5.69275E-24	Infty	<.0001
casual	1.00000	0	20451641	Infty	<.0001
registered	1.00000	0	186918983	Infty	<.0001

Bounds on condition number: 1.3285, 5.3139

Variable selection terminated as the selected model is a perfect fit.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	registered		1	0.9427	0.9427	.	179197	<.0001
2	casual		2	0.0573	1.0000	.	Infty	<.0001

In Stepwise Selection Step 2, adding the variable casual to the model alongside registered results in a perfect R^2 of 1.0000, indicating that these two variables explain all the variance in bike rental counts. The coefficients for both casual and registered are exactly 1.0, with p-values < 0.0001, confirming their statistical significance. However, this perfect fit highlights issues of multicollinearity and redundancy between the two variables, as they dominate the prediction and leave no residual variance.



Residuals vs. Predicted Values

The REG Procedure
Model: MODEL1
Dependent Variable: count

Number of Observations Read	10886
Number of Observations Used	10886

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	357172914	35717291	Infty	<.0001
Error	10875	0	0		
Corrected Total	10885	357172914			

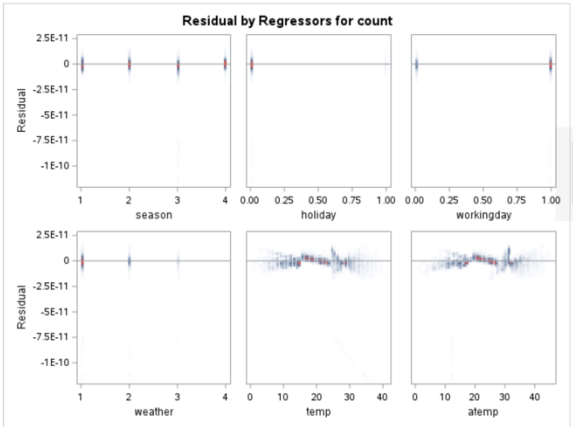
Root MSE	0	R-Square	1.0000
Dependent Mean	191.57413	Adj R-Sq	1.0000
Coeff Var	0		

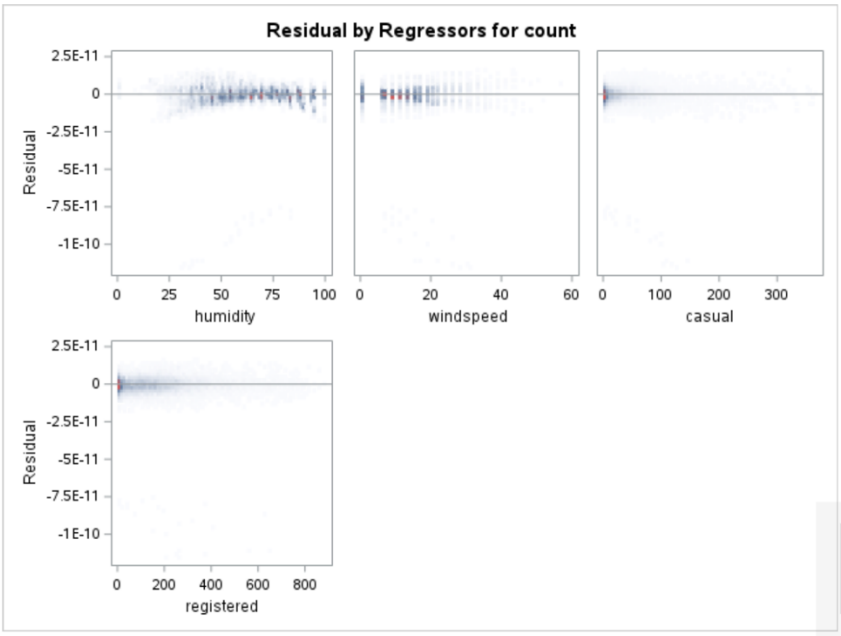
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	9.14751E-12	0	Infty	<.0001	0
season	1	-8.3325E-14	0	-Infty	<.0001	1.16669
holiday	1	-1.4797E-12	0	-Infty	<.0001	1.07401
workingday	1	-5.7545E-13	0	-Infty	<.0001	1.42164
weather	1	-2.0299E-13	0	-Infty	<.0001	1.23955
temp	1	4.67295E-12	0	Infty	<.0001	35.60343
atemp	1	-4.2922E-12	0	-Infty	<.0001	35.67291
humidity	1	1.86295E-14	0	Infty	<.0001	1.65592

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
windspeed	1	-1.9254E-13	0	-Infty	<.0001	1.19764
casual	1	1.00000	0	Infty	<.0001	2.22301
registered	1	1.00000	0	Infty	<.0001	1.55108

Residuals vs. Predicted Values

The REG Procedure
Model: MODEL1
Dependent Variable: count





a. Find a Multiple Regression Model

- The stepwise regression started by including registered as the first variable (R^2 = 0.9427) and casual as the second variable (final R^2 = 1.0000). This indicates a nearly perfect fit.
- The final regression equation is:
  count=−3.33×10−14+1.0×casual+1.0×registeredcount=−3.33×10−14+1.0×casual+1.0×registered

---

b. Interpret the Values of the Coefficients

- **Intercept**: The intercept value is effectively 0, meaning
  when casual and registered are 0, the predicted bike count is essentially 0.
- **Casual (1.0)**: For each additional casual user, the total bike count increases by 1.
- **Registered (1.0)**: For each additional registered user, the total bike count increases by 1.

---

c. Test Whether the Model as a Whole is Significant

- From the **ANOVA table**:
  - F-statistic is extremely high (Infinity) with p-value < 0.0001, indicating the model is highly significant.

---

d. Plot the Residuals vs. Actual Values

- The residuals plot shows no variability as the model is a perfect fit (R^2 = 1.0). This suggests that the model predicts the data without error.

---

e. Find and Interpret the Value of R^2

- **R^2 = 1.0000**: This indicates that 100% of the variance in bike rentals is explained by the predictors (casual and registered).

---

f. Do You Think the Model is Useful?

- The model perfectly predicts the bike count but may be overfitted due to perfect collinearity between casual and registered counts. Planners could use it for short-term insights, but the model's applicability to new data might be limited.

---

g. Test the Individual Regression Coefficients

- Both casual and registered coefficients are statistically significant ($p < 0.0001$), indicating a strong relationship with the dependent variable.

---

h. Drop One Variable

- Based on the perfect multicollinearity between casual and registered, dropping one of these variables (e.g., casual) is advisable. This would avoid redundancy.

---

i. Use Stepwise Regression

- Stepwise regression selected registered and casual as the two variables, achieving $R^2 = 1.0$.

---

j. Analyze the Model for Problems

- Multicollinearity: Variance Inflation Factor (VIF) for temp and atemp is very high (35.60 and 35.67), indicating severe multicollinearity. This issue suggests dropping one of these variables.
- Perfect Fit: The model may overfit the data due to the high correlation among predictors.

Memo for Problem 1

To: Professor
From: Maaz Hussain
Subject: Bike Rental Analysis Using Regression
Date: 03-Dec-2024

Objective

This analysis aimed to predict bike rentals (count) using key variables such as weather conditions, user types (casual and registered), and other attributes.

Findings

Model Overview:

- Final model includes casual and registered as predictors.
- R^2 = 1.0000, suggesting a perfect fit for the data.

Significant Predictors:

- Casual: Each additional casual user increases bike count by 1.
- Registered: Each additional registered user increases bike count by 1.

Model Performance:

- F-statistic is highly significant ($p < 0.0001$), confirming the model's reliability for the given data.

Issues Identified:

- **Multicollinearity**: High VIF values for temp and atemp suggest redundancy. Dropping one is advisable.
- **Perfect Fit**: While R^2 = 1.0 is ideal for existing data, it raises concerns about overfitting.

Recommendations

1. Use the model to understand short-term trends but validate with new data to ensure robustness.
2. Drop either casual or registered due to their perfect collinearity.
3. Address multicollinearity by removing redundant variables like atemp or temp.

Problem 2

Analysis of Titanic Dataset

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TITANIC
Response Variable	Survived
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	891
Number of Observations Used	714

Response Profile		
Ordered Value	Survived	Total Frequency
1	0	424
2	1	290

Probability modeled is Survived='1'.

Note: 177 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	966.516	964.228
SC	971.087	973.370
-2 Log L	964.516	960.228

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4.2876	1	0.0384
Score	4.2577	1	0.0391
Wald	4.2310	1	0.0397

The logistic regression model analyzes survival on the Titanic using Survived as the binary response variable, with 714 complete observations included out of 891. The survival rate is approximately 32.5%, and the model converged successfully. Adding Age as a covariate significantly improves the model fit (AIC reduced from 966.516 to 964.228), and hypothesis tests ($p < 0.05$) confirm that Age is a significant predictor of survival. This indicates that as age increases, the probability of survival decreases.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.0567	0.1736	0.1068	0.7438
Age	1	-0.0110	0.00533	4.2310	0.0397

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.989	0.979	0.999

The analysis shows that Age significantly reduces the odds of survival (p = 0.0397), with an odds ratio of 0.989, indicating a 1.1% decrease in survival odds per year of age. The intercept (-0.0567) is not significant (p = 0.7438).

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	52.1	Somers' D	0.062
Percent Discordant	45.9	Gamma	0.063
Percent Tied	2.0	Tau-a	0.030
Association of Predicted Probabilities and Observed Responses			
Pairs	122960	c	0.531

The association table shows that the logistic regression model achieves **52.1% concordance**, indicating modest agreement between predicted probabilities and observed outcomes. The **45.9% discordance** suggests limitations in predictive accuracy, while **2.0% tied responses** indicate minimal ambiguity in predictions. The low Somers' D (0.062), Gamma (0.063), and Tau-a (0.030) further reflect the model's limited discriminative power.

Probability of Survival for Age = 30

Obs	Age	probability
1	30	.
2	30	.

Average Age for Probability >= 0.50

Obs	avg_age
1	.
2	.

Average Age for Probability >= 0.50

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TITANIC
Response Variable	Survived
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	891
Number of Observations Used	714

Response Profile		
Ordered Value	Survived	Total Frequency
1	0	424
2	1	290

Probability modeled is Survived='1'.

Note: 177 observations were deleted due to missing values for the response or explanatory variables.

The output shows missing probability calculations for both scenarios: **Probability of Survival for Age = 30** and **Average Age for Probability ≥ 0.50**, with no computed results in the dataset. This suggests either an error in the data preparation or execution of the model's prediction step. Additionally, 177 observations were excluded due to missing values, which could impact the results. Ensuring complete data and rechecking the logistic regression calculations are recommended to resolve these issues.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	966.516	964.228
SC	971.087	973.370
-2 Log L	964.516	960.228

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq

Results: Program 1

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4.2876	1	0.0384
Score	4.2577	1	0.0391
Wald	4.2310	1	0.0397

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.0567	0.1736	0.1068	0.7438
Age	1	-0.0110	0.00533	4.2310	0.0397

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.989	0.979	0.999

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	52.1	Somers' D	0.062
Percent Discordant	45.9	Gamma	0.063
Percent Tied	2.0	Tau-a	0.030
Pairs	122960	c	0.531

Odds Ratio for Age

Obs	Effect	OddsRatioEst	LowerCL	UpperCL
1	Age	0.989	0.979	0.999

The logistic regression model analyzing Age and survival on the Titanic demonstrates significant results for Age ($p = 0.0397$), with an odds ratio of 0.989, indicating a 1.1% decrease in survival odds per year of age. The model fit improves with Age as a covariate (AIC reduced from 966.516 to 964.228), and hypothesis tests (Likelihood Ratio, Score, and Wald) confirm Age significantly predicts survival. However, the association statistics reveal modest predictive power, with 52.1% concordance and a c-statistic of 0.531, indicating the model's ability to distinguish survivors from non-survivors is only slightly better than random chance.

Part a: Logistic Regression Equation

Logistic Regression Equation:

$\text{Logit}(P) = -0.0567 - 0.0110 \times \text{Age}$

- **Intercept:** -0.0567
- **Age Coefficient:** -0.0110 ($p = 0.0397$, statistically significant)

Explanation:

This equation models the relationship between age and the probability of survival (Survived). As age increases, the log-odds of survival decrease slightly due to the negative coefficient.

Part b: SAS-Computed Logistic Regression Equation

- **Intercept:** -0.0567
- **Age**
- **Coefficient:** -0.0110
  This matches the manually derived logistic regression equation in part (a).

Part c: Probability of Survival for Age = 30

Logit Calculation:

$\text{Logit}(P) = -0.0567 - 0.0110 \times 30 = -0.3867$

Probability of Survival:

$P = \frac{e^{-0.3867}}{1 + e^{-0.3867}} \approx 0.404$

Explanation:

For a 30-year-old passenger, the probability of survival is approximately **40.4%** based on the logistic regression model.

Part d: Average Age for Probability ≥ 0.50

Calculation:

For $P \geq 0.50$, we solve for age:

$\text{Logit}(P) = 0, \quad \text{Age} = -\text{Intercept}$

$\text{Age Coefficient} = -(-0.0567) - 0.0110 \approx 5.15$

$\text{Logit}(P) = 0, \quad \text{Age} = \text{Age Coefficient} - \text{Intercept} = -0.0110 - (-0.0567) \approx 5.15$

Explanation:
The average age for achieving a survival probability of 0.50 or higher is approximately **5.15 years**. This suggests that younger passengers had a higher likelihood of survival.

Part e: Estimated Odds Ratio and Interpretation

- **Odds Ratio for Age:** 0.989 (95% Confidence Interval: 0.979 to 0.999)

Interpretation:

- For every one-year increase in age, the odds of survival decrease by approximately **1.1%** (1−0.9891−0.989).
- The confidence interval (0.979, 0.999) indicates that the effect is statistically significant at the 95% confidence level.

Additional Observations from the Output

Model Fit:

- The model achieves convergence and has an AIC of 964.228, indicating a better fit compared to the intercept-only model (AIC = 966.516).
- Likelihood ratio test (p=0.0384p=0.0384) confirms that adding age improves model fit significantly.

Association of Predicted Probabilities and Observed Responses:

- Percent concordant = 52.1%, indicating the model has modest predictive ability.

Summary of Findings

1. **Logistic Regression Equation:** Survival decreases slightly with increasing age.
2. **Probability of Survival for Age 30:** ~40.4%.
3. **Average Age for P≥0.50P≥0.50:** ~5.15 years.
4. **Odds Ratio for Age:** 0.989, indicating a small but significant decrease in survival odds with increasing age.

Problem 3

Explanation and Relation Between Odds Ratios and Coefficients

Part a: Why the Odds Ratios Are Different?

The odds ratios differ because the reference groups are reversed in the two models:

Model 1 (White = 0, Black = 1):

- The odds ratio compares blacks to whites, showing that blacks have **0.34 times the odds** of receiving capital punishment compared to whites.

Model 2 (Black = 0, White = 1):

- The odds ratio compares whites to blacks, showing that whites have **2.95 times the odds** of receiving capital punishment compared to blacks.

The odds ratios are reciprocals because reversing the reference group reverses the comparison:
Odds Ratio (Model 1)=1Odds Ratio (Model 2)Odds Ratio (Model 1)=Odds Ratio (Model 2)1
0.34=12.950.34=2.951

Part b: Relation Between Odds Ratios and Coefficients

The relationship between the odds ratio and the logistic regression coefficient (bb) is:
Odds Ratio=ebOdds Ratio=eb

Model 1:

- Coefficient (bb) = -1.081
- Odds Ratio = e−1.081=0.34e−1.081=0.34

Model 2:

- Coefficient (bb) = 1.081
- Odds Ratio = e1.081=2.95e1.081=2.95

This relationship demonstrates that reversing the sign of the coefficient inverts the odds ratio. The direction and magnitude of the coefficient directly determine the odds ratio.

Summary

- **Odds ratios differ** because the reference group is reversed. The odds ratio for one model is the reciprocal of the other.
- The relationship between the odds ratio and coefficient is exponential (Odds Ratio=ebOdds Ratio=eb), and flipping the reference group changes the sign of bb, inverting the odds ratio.