Due Date: No Later than Dec 23 2025

Regression Project Instructions (3-Page Report)

Dataset Source: UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/)

Your task is to pick **one dataset** from the UC Irvine Repository and conduct a **multiple regression analysis**. You will clean the data, run a regression model, interpret results, check assumptions, and write a short report summarizing your findings.

Your final deliverable is a **3-page written report (excluding tables/figures)**.

Follow the steps below.

STEP-BY-STEP INSTRUCTIONS

---

1. Choose a Dataset

Select a dataset from the UC Irvine Machine Learning Repository that contains **at least one continuous dependent variable**.

Examples:

- *Wine Quality* → Predict wine quality from chemical properties
- *Auto MPG* → Predict miles per gallon from engine features
- *Concrete Strength* → Predict concrete compressive strength from mixture components

Be sure to cite the dataset.

---

2. State the Research Question

Clearly describe the problem your regression will investigate.

Example:

> “I will examine how engine displacement, weight, and horsepower predict automobile fuel efficiency (mpg).”

This is your guiding objective.

---

3. Load and Inspect the Data

Import your dataset into R and conduct an initial inspection.

Required actions:

- Check variable names
- Look for missing values
- Identify non-numeric or irrelevant columns

R example:

```r
data <- read.csv("filename.csv")
str(data)
summary(data)
```

---

4. Clean and Prepare the Data

Depending on the dataset, you may need to:

- Remove rows with missing values (or impute appropriately)
- Convert categorical variables to factors
- Create new variables if justified

- Drop variables not needed for analysis

Example:
If “horsepower” contains missing `NA` values:

```
data <- na.omit(data)
```

---

5. Provide Summary Statistics

Include descriptive statistics for all variables used in the regression:

- Mean
- Standard deviation
- Minimum
- Maximum
- Histogram or boxplot (optional)

R example:

```
summary(data)
```

---

6. Define the Model

Clearly state:

- **Dependent variable (Y)**
- **Independent variables (X’s)**

Example:

Y = Wine Quality
X = Alcohol content, Acidity, Sugar level

---

7. Run the Multiple Regression Model

Estimate the model using R.

R example:

```
model <- lm(Y ~ X1 + X2 + X3, data = data)
summary(model)
```

---

8. Interpret the Coefficients

For each predictor, explain:

- Direction (positive/negative effect)
- Magnitude (size of effect)
- Practical meaning

Example:

> "Holding all else constant, a 1% increase in alcohol content is associated with a 0.28-point increase in wine quality."

---

9. Assess Statistical Significance

For each coefficient:

- Check **p-values**
- Identify which predictors are statistically significant
- Discuss what that means in context

---

10. Calculate and Interpret Confidence Intervals

Provide 95% confidence intervals for the coefficients.

R example:

```
confint(model)
```

Explain what the intervals imply:

> “We are 95% confident that the impact of weight on mpg lies between –0.008 and –0.005.”

---

11. Report the Coefficient of Determination (R^2)

Explain how well the model explains the variation in the dependent variable.

Example:

> “The model explains 82% of the variation in concrete strength, suggesting excellent predictive capability.”

---

12. Examine Model Residuals

Check whether residuals appear:

- Normally distributed
- Homoscedastic (equal variance)
- Randomly scattered (no pattern)

R example:

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

Discuss any violations.

---

13. Make at Least One Prediction

Choose one hypothetical or real observation and generate a predicted value.

R example:

```
new_obs <- data.frame(X1=..., X2=..., X3=...)
predict(model, new_obs, interval="prediction")
```

Explain your prediction in practical terms.

---

14. Write a Clear, Organized 3-Page Report

Your report **must include**:

1. **Title and Research Question**
2. Dataset Description (source, size, variables)
3. Summary Statistics
4. Regression Model Specification
5. Interpretation of Coefficients
6. Statistical Significance Results
7. Confidence Intervals
8. R² and Model Fit Evaluation
9. Residual Analysis
10. Predictions
11. **Conclusion**

---

15. Conclusion Section

Summarize the key insights in plain language.

Example:

> “ TV and radio advertising significantly increase sales, while newspaper advertising does not. The model fits well ($R^2 = 0.89$). Predictions suggest that increasing TV spending by $10,000 would increase sales by approximately 0.5 units.”

Also note any limitations:

- Missing variables
- Nonlinear patterns
- Small sample size

---

16. Include an Appendix

Not counted in the 3 pages:

- R code
- Tables
- Figures
- Diagnostic plots

Summary Checklist

Your project must include:

- Dataset chosen from UCI Repository
- Clear research question
- Cleaned dataset

- Summary statistics
- Defined dependent and independent variables
- Regression model in R
- Coefficient interpretation
- Statistical significance
- Confidence intervals
- R^2 explained
- Residual diagnostics
- At least one prediction
- 3-page written report
- Conclusion and limitations

Example: Regression Project Guidance

Fit a multiple regression model:

$Sales=\beta 0+\beta 1(TV)+\beta 2(Radio)+\beta 3(Newspaper)+\varepsilon$

R Code for the Full Analysis

1. Load data (in this example, the dataset is called “Advertising.csv”)

```
data <- read.csv("Advertising.csv")
head(data)
```

---

2. Run Multiple Regression

```
model <- lm(Sales ~ TV + Radio + Newspaper, data = data)
summary(model)
```

---

3. Interpretation of Output

From the published results:

- **TV coefficient** is **positive and highly significant**
  → For every $1,000 increase in TV advertising, sales increase by about **0.05 units** ($p < 0.001$).

- **Radio coefficient** is also significant
  → Radio ads increase sales by **0.19 units** per $1,000 ($p < 0.001$).

- **Newspaper coefficient** is *not* statistically significant
  → Newspaper spend does **not** reliably increase sales.

This lets you discuss:

- Statistical significance

- Coefficient size
- Practical interpretation

---

4. Check Confidence Intervals

```
confint(model)
```

Example interpretation (from values):

- TV CI might look like (0.04, 0.06)
- Radio CI might look like (0.17, 0.21)
- Newspaper CI includes 0 → not significant

---

5. Make Predictions

```
newdata <- data.frame(TV = 100, Radio = 25, Newspaper = 10)
predict(model, newdata, interval = "prediction")
```

This produces:

- Predicted sales
- Lower 95% prediction interval
- Upper 95% prediction interval

---

6. Assess Model Fit

R² from the published analysis:

- **R^2 ≈ 0.897**
  Meaning: ~90% of variation in Sales is explained by advertising variables.

---

7. Diagnostic Plots

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

You should explore the following:

- Residual patterns
- Homoscedasticity
- Normality of residuals
- Influence points