DATA 505: Homework 3

Week 9-10 (Regression)

2025-11-04

Context

In this homework, you will reinforce concepts from week 9 on regression analysis. You will practice fitting simple and multiple linear regression models, interpreting coefficients, making predictions, checking model assumptions, and working with categorical predictors. This assignment will help you develop skills in both inference (understanding relationships) and prediction (forecasting outcomes). As with earlier homeworks, it will not be submitted, but a short quiz at the start of class on Tue Oct. 28 will test your understanding of the material.

References

- Lecture 9 and Lecture 10
- Thulin, M. (2024). Modern Statistics with R. Second edition. Chapman & Hall/CRC Press. ISBN 9781032512440. Section 8.1: Linear Models. Link

Coding Tasks

Create an R script and use it to answer the following questions.

Part 1: Simple Linear Regression

For this section, we'll use the built-in `mtcars` dataset, which contains information about various car models from 1974 Motor Trend magazine.

1. Exploring the Data

a. Load the `mtcars` dataset using `data(mtcars)` and inspect it using `head()` and `str()`.

b. Create a scatterplot of `mpg` (miles per gallon) vs. `wt` (weight in 1000 lbs) using base R's `plot()` function or `ggplot2`. Based on the plot, do you expect a positive or negative relationship?

c. Calculate the correlation between `mpg` and `wt` using `cor()`. Does this match your expectation from the plot?

2. Fitting a Simple Linear Regression Model

a. Fit a linear regression model predicting `mpg` from `wt` using the `lm()` function. Store the result as `model1`.

b. Use `summary(model1)` to view the model output. What are the estimated intercept and slope?

c. Write out the fitted regression equation in the form: mpg = + × wt

d. Interpret the slope coefficient in context: What does it tell you about the relationship between car weight and fuel efficiency?

3. Hypothesis Testing and Confidence Intervals

a. Based on the model summary, what is the p-value for the slope coefficient? Is there statistically significant evidence of a linear relationship between weight and mpg at the 0.05 significance level?

b. Use `confint(model1)` to obtain 95% confidence intervals for the intercept and slope. Interpret the confidence interval for the slope.

c. What proportion of the variance in `mpg` is explained by `wt`? (Hint: look at the R-squared value in the summary output.)

4. Making Predictions

a. Use `predict()` to estimate the mpg for a car that weighs 3.5 thousand pounds (i.e., `wt = 3.5`).

b. Calculate a 95% confidence interval for the **mean** mpg of cars weighing 3.5 thousand pounds using `predict()` with `interval = "confidence"`.

c. Calculate a 95% prediction interval for a **single new car** weighing 3.5 thousand pounds using `predict()` with `interval = "prediction"`.

d. Which interval is wider, and why?

5. Model Diagnostics

a. Create diagnostic plots for `model1` using `plot(model1)`. This will generate four diagnostic plots.

b. Based on the "Residuals vs Fitted" plot, does the linearity assumption appear to be satisfied?

c. Based on the "Normal Q-Q" plot, do the residuals appear to be approximately normally distributed?

d. Are there any observations that appear to be influential outliers based on the "Residuals vs Leverage" plot?

Part 2: Multiple Linear Regression

6. Adding More Predictors

a. Fit a multiple regression model predicting `mpg` from both `wt` and `hp` (horsepower). Store this as `model2`.

b. Use `summary(model2)` to view the results. Write out the fitted regression equation.

c. Interpret the coefficient for `wt` in this model. How does the interpretation differ from the simple regression model in Part 1?

d. Interpret the coefficient for `hp`. What does it tell you about the relationship between horsepower and fuel efficiency, holding weight constant?

7. Comparing Models

a. Compare the R-squared values of `model1` and `model2`. Which model explains more variance in `mpg`?

b. Use `anova(model1, model2)` to perform an F-test comparing the two models. Does adding `hp` significantly improve the model fit?

c. Based on these comparisons, which model would you prefer for predicting fuel efficiency? Explain your reasoning.

8. Predictions with Multiple Predictors

a. Use `model2` to predict the mpg for a car with `wt = 3.0` and `hp = 120`. Create a data frame with these values and use `predict()`.

b. Calculate a 95% prediction interval for this car using `predict()` with `interval = "prediction"`.

Part 3: Categorical Predictors

9. Including a Categorical Variable

a. Convert the `cyl` (number of cylinders) variable to a factor using `as.factor()` and update the `mtcars` dataset.

b. Fit a regression model predicting `mpg` from `cyl` (as a factor). Store this as `model3`.

c. Use `summary(model3)` to view the results. What is the reference category (baseline level) for `cyl`?

d. Interpret the coefficient for `cyl6`. What does it tell you about the difference in average mpg between 6-cylinder and 4-cylinder cars?

e. Interpret the coefficient for `cyl8`.

10. Multiple Regression with Categorical Predictors

a. Fit a model predicting `mpg` from both `wt` and `cyl` (as a factor). Store this as `model4`.

b. Interpret the coefficient for `wt` in this model.

c. Interpret the coefficient for `cyl8` in this model. How does this differ from the interpretation in model3?

d. Compare the R-squared of `model4` with that of `model2` (the model with `wt` and `hp`). Which model explains more variance?

Part 4: Challenge Problems

11. Exploring Interactions (Advanced)

a. Fit a model predicting `mpg` from `wt`, `hp`, and their interaction: `lm(mpg ~ wt * hp, data = mtcars)`. Store this as `model5`.

b. Use `summary(model5)` to view the results. Is the interaction term statistically significant?

c. What does a significant interaction term tell you about the relationship between weight, horsepower, and fuel efficiency?

12. Extrapolation Warning

a. What is the range of `wt` values in the `mtcars` dataset? (Use `range()` or `summary()`.)

b. Use `model1` to predict mpg for a car weighing 6 thousand pounds. Does this prediction seem reasonable? Why or why not?

c. Explain in your own words why extrapolation can be problematic in regression analysis.