Exercise 1: Analyzing E-commerce Shipping Performance

Suppose you have a dataset `shipping_data.csv` containing $n = 1500$ deliveries from a global online retailer. The goal is to predict the total **Transit Time** (the number of days from dispatch to arrival) using various shipment characteristics.

The dataset includes the following variables:

- `TransitTime`: Total days in transit (Outcome).
- `Weight`: Package weight in kilograms.
- `Distance`: Distance between the warehouse and the customer (km).
- `Carrier`: The shipping company used (e.g., DHL, FedEx, UPS).
- `ShipMode`: Economy, Standard, or Express.
- `Region`: Geographical region of the customer destination.
- `Holiday`: A binary indicator (1 if the transit occurred during a peak holiday season, 0 otherwise).

Tasks

(a) Read the data into R. Fit a multiple linear regression model with `TransitTime` as the outcome and all other variables as covariates. Provide a summary of the estimated coefficients, including their standard errors and p-values.

(b) Use R to calculate the 99% confidence intervals for all regression coefficients. Based on these intervals, determine which predictors have a statistically significant relationship with shipping speed at the $\alpha = 0.01$ level.

(c) Reproduce the least squares estimate of the error variance σ^2. *Hint: Use the* `residuals()` *function to obtain the residuals e_i, and calculate $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-p-1}$.*

(d) Use R to estimate both the coefficient of determination R^2 and the adjusted R^2_{adj}. Compare these values with the output provided in the model summary from part (a).

(e) Use backward selection by means of the `stepAIC` function from the `MASS` library to find the best model according to the Akaike Information Criterion (AIC).

(f) Obtain the R^2_{adj} from the reduced model identified in (e) and compare it to the full model from (a). Discuss whether the reduction in variables significantly impacted the model's explanatory power.

(g) Identify whether the fundamental model assumptions are satisfied. Generate and interpret a ”Residuals vs Fitted” plot and a ”Normal Q-Q” plot to check for heteroscedasticity and non-normality.

(h) Based on the model in (e), are specific `Carriers` or `Holiday` periods causing the delivery time to be significantly delayed? Interpret the magnitude of the `Holiday` coefficient in terms of days.

(i) Test whether it is useful to add a quadratic polynomial of `Weight` (i.e., $Weight^2$) to the model. Perform a partial F-test to compare the model with and without the quadratic term.

(j) Use the model identified in (e) to predict the transit time for a new delivery with: `Weight` = 5.2kg, `Distance` = 450km, `Carrier` = ”FedEx”, `ShipMode` = ”Standard”, and `Holiday` = 0. Use the `predict()` command to report a 95% prediction interval.

1 National Football League Team Performance (1976)

Table B.1 presents data on the performance of the 28 teams in the National Football League (NFL) during the 1976 season. The response variable and explanatory variables are defined as follows:

Variables Description

- y: Number of games won in a 14-game season
- x_1: Rushing yards (season)
- x_2: Passing yards (season)
- x_3: Punting average (yards per punt)
- x_4: Field goal percentage (field goals made divided by field goals attempted)
- x_5: Turnover differential (turnovers gained minus turnovers lost)
- x_6: Penalty yards (season)
- x_7: Percentage of rushing plays
- x_8: Opponents' rushing yards (season)
- x_9: Opponents' passing yards (season)

Exercises

Using the number of games won (y) as the response variable, answer the following:

1. Use **forward selection** to specify an appropriate subset regression model.
2. Use **backward elimination** to specify an appropriate subset regression model.
3. Use **stepwise regression** to specify an appropriate subset regression model.

4. Apply **all possible regressions** to the data. For each candidate model, evaluate the criteria R_p^2, Mallows' C_p, and the mean square of the residuals (MSRes). Based on these measures, recommend a final subset regression model.

5. Compare and contrast the models obtained in parts (a)–(d). Discuss similarities, differences, and the stability of the selected variables across the different selection methods.