For this homework, write a brief but comprehensive report for each problem. Your reports for problems 1 and 2 may be brief, but your report for problem 3 should be long enough that it is clear WHY you are making the modeling choices you are.

Start each new problem on a new page. Include relevant plots and model statements. Make sure to clearly label your plots, as there will be multiple plots for each problem. Make sure your plots are big enough to read.

You can, but do not need to, include your R code in your report.

1. **Simple Linear Regression: Car Stopping Distance**. Fit a simple linear model to the "cars" data in R. Read in the data using the following commands:

   ```
   data(cars)
   ?cars
   str(cars)
   ```

   Use the stopping distance (measured in feet) as the response variable, and the speed as the predictor variable. Address the following points:

   - Conduct an appropriate exploratory data analysis and comment on anything noteworthy about the data.
   - Write down the simple linear regression model that will be fit.
   - Use "lm" in R to estimate model parameters and residuals.
   - Examine the model fit and residuals. Comment on any possible violations of the modeling assumptions.
   - If necessary (and if possible), modify your model so that the modeling assumptions are valid. Consider transforming the response and/or predictor variables using a power transform or a log transform.
   - Report the final model fit, the estimates of all model parameters, and their interpretation.

2. **Simple Linear Regression: Predicting Tree Girth from Tree Height**. Consider modeling tree girth (DBH) as a function of tree height. The motivation is that tree height can be observed remotely (e.g., via LIDAR) but tree girth is harder to measure remotely. Read in the "trees" data and examine it using the following commands:

   ```
   data(trees)
   ?trees
   str(trees)
   ```

   Consider a simple linear regression model with girth as the response variable and height as the predictor variable. Address the six bullet points given in Problem 1 for this data set.

3. **Multiple Linear Regression: Munich Rent**. In this problem you will model apartment rent in Munich in 1999 (rent) as a function of the size in square meters of the apartment (area) and the year the dwelling was constructed (yearc). Read in the Munich rent data by doing

   ```
   Munich=read.csv("rent99.raw",sep=" ")
   ```

   (a) Conduct an appropriate exploratory data analysis of the *rent*, *area*, and *yearc* variables in the Munich data set and comment on anything noteworthy about the three variables we are modeling.

   (b) Fit the linear regression model:

   $$\text{rent}_i = \beta_0 + \beta_1 \text{area}_i + \beta_2 \text{yearc}_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

   and examine the model fit and residuals. Examine plots of model residuals vs fitted values, partial residual plots for each predictor variable, and QQ-plots of residuals. Describe the violations of the modeling assumptions that are evident in the residual analysis.

(c) Now fit the following linear regression model

(d) Fit the linear regression model

$$log(\text{rent}_i) = \beta_0 + \beta_1 \text{area}_i + \beta_2 \text{yearc}_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

and examine the model fit and residuals. Examine plots of model residuals vs fitted values, partial residual plots for each predictor variable, and QQ-plots of residuals. Describe any violations of the modeling assumptions that are evident in the residual analysis.

(e) Now consider modeling rent per square meter (*rentsqm* in the Munich data set) instead of *rent*. The *rentsqm* variable in Munich is defined by:

$$\text{rentsqm}_i = \frac{\text{rent}_i}{\text{area}_i}.$$

Fit the linear regression model

$$\text{rentsqm}_i = \beta_0 + \beta_1 \text{area}_i + \beta_2 \text{yearc}_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

and examine the model fit and residuals. Examine plots of model residuals vs fitted values, partial residual plots for each predictor variable, and QQ-plots of residuals. Describe the violations of the modeling assumptions that are evident in the residual analysis.

(f) Now consider models of the form:

$$\text{rentsqm}_i = \mu + f(\text{area}_i) + g(\text{yearc}_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

where $f(\cdot)$ and $g(\cdot)$ are unknown functions. Find forms for f and g expressed as linear functions of basis vectors which are functions of area and year constructed. Example 3.3-3.5 in the text may be helpful. You may follow the book or consider variations. You do NOT need to consider any interaction effects.

- Write out the model you will fit, and express f and g as functions of terms in your model. Report the final model fit and the estimates of all model parameters.
- Comment on model assumptions for your final model. Include any plots you think are helpful.
- Interpret the estimated model by describing in words the relationship between apartment size, apartment rent per square meter, and construction year as you have modeled them. Include any plots you feel aid in the interpretation of your model results.
- If you try multiple models before settling on a final model, describe briefly what you tried and why you abandoned it in favor of your final model. You do not need to provide plots or model output for any intermediate models, just describe some things you tried and why they are inferior to your final model.

(g) Your final model from (f) is a linear regression model with (rent per area) as the response variable. Use this model to specify an equivalent model for (rent). Write out the model, including estimated parameter values (which have been calculated in your model for rent per area). What statements can you make about the effect that the size of the apartment and the construction year have on the mean and variance of the rent?

(h) Comment on the relative strengths and weaknesses of the four models that you fit to the rent data. If you had to choose only one of these to explain variation in rent in Munich, which would you choose and why?