2 · *Fitting equations; predictable and unpredictable variation*

(1) Demand curves

Complete the case study on milk prices from class: use the available data to decide the profit-maximizing price as a function of the wholesale cost of milk, c.

Imagine that you work in the business-analytics office for a grocery store chain. Based on this case study, write a short summary of your pricing decision assuming that $c = 1$, and the evidence supporting that decision, that could be understood by a manager. You should assume that your manager is statistically literate, but that he or she wants you to focus on the conclusions and the evidence, rather than on the technical details of regression analysis.

(2) Polynomial regression and prediction intervals

For this question, you will return to the "utilities.csv" data set from class and the course packet. Recall that each row has information about a monthly utility bill for a house in Minnesota. The variable "gasbill" is the gas bill for that month, measured in dollars. The "temp" variable depicts the average temperature, in degrees F, for the billing period.

Compute the daily average gas bill, so that we can compare months with different numbers of days. Then fit first-order through third-order polynomial regression models for the daily average gas bill (Y) versus temperature (X). That is, fit the models

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i \\ y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i \\ y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_i \end{aligned}$$

For each model, report the fitted equation, together with the model's R^2 and residual standard deviation s_e.

Now pick your favorite model of these three, and briefly explain why you chose it. Use this favorite model to compute a prediction interval for the gas bill during a month in which the average temperature is 50 degrees Fahrenheit. Make sure you state the coverage level of your interval (i.e. quantify its likely forecasting accuracy).

(3) Should we aggregate or not?

For this question, you will need the "TenMileRace" data set from the `mosaic` package in R, which you will load using the command `data(TenMileRace)` after having loaded the `mosaic` package at the beginning of your R session. Quoting the data set description: "The Cherry Blossom 10 Mile Run is a road race held in Washington, D.C. in April each year. The name comes from the famous cherry trees that are in bloom in April in Washington.... This data frame contains the results from the 2005 race." If you type the command `?TenMileRace`, you will get a description of each variable in the data set.

(A) Fit a regression model to quantify the relationship between a runner's net finishing time (in seconds) and his or her age in years. What seems to be the effect of one additional year of age on finishing time?

(B) Now fit two separate linear models for finishing time versus age: one for men alone, and one for women alone. Within each subset, what seems to be the effect of one additional year of age on finishing time? Is this consistent with what you found in Part A? Describe what you think is going on here. Remember from the software walkthroughs: you can create a new data set from a subset of the original one using the subset command. For example:

```
women = subset(TenMileRace, sex=="F")
```

Notice the quotation marks and the double-equals sign, which is how we test for whether a variable takes a specific value.