DATA 609 Assignment 5: Discrete Probablistic Modeling

Dan Smilowitz

March 5, 2017

Section 6.1, Problem 1

A model is developed to calculate the values of the following variables:

- G_n: the percentage of students dining at Grease Dining Hall in period n
- S_n: the percentage of students dining at Sweet Dining Hall in period n

It is given that 25% of student dining at the Grease Dining Hall return to eat there again; since there are only two dining halls, this means that the remaining 75% of students eating at Grease Dining Hall eat at Sweet Dining Hall for their next meal (since the two probabilities must add to 1). Similarly, it is given that 93% of students dining at Sweet Dining Hall return to eat there; this means that the remaining 7% of students eat their next meal at Grease Dining Hall. With this information, the model can be modeled using the following two equations:

$$G_{n+1} = 0.25G_n + 0.07S_n$$
$$S_{n+1} = 0.75G_n + 0.93S_n$$

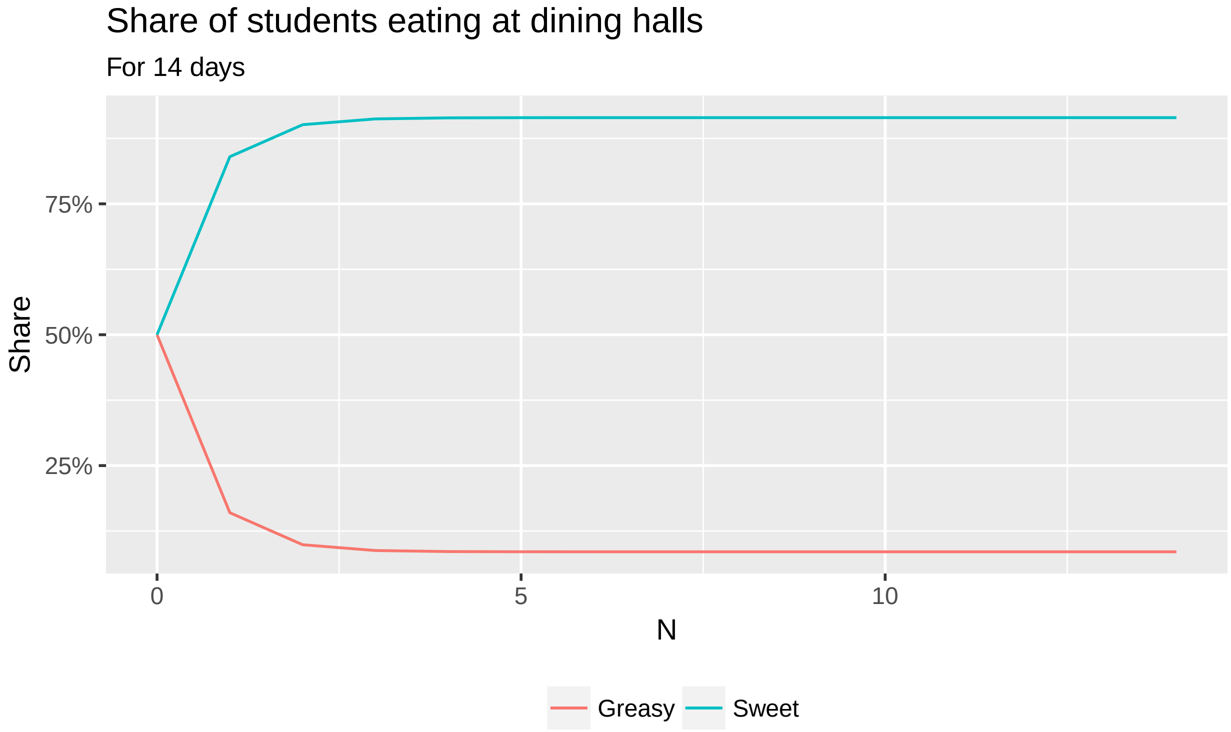
Assuming that initially, 50% of students eat at each dining hall, the long-term behavior of the system can be modeled:

```
# define function for steady state
# check if last n elements of vector are the same within given tolerance
steady_state <- function(x, n = 5, tol = .Machine$double.eps ^ 0.5) {
  # ignore items of length 1
  if (length(x) == 1) return(FALSE)
  # get last n items (or entire vector if shorter than n)
  x <- tail(x, n)
  # return steadiness
  return(abs(max(x) - min(x)) <= tol)
}

# set initial conditions
N <- 0
G <- 0.5
S <- 0.5

# calculate next state of each variable until steady state reached
while (!steady_state(G) & !steady_state(S)) {
  # increment counter
  N <- N + 1
  # get current states
  Gn <- G[length(G)]
  Sn <- S[length(S)]
  # calculate new values & append to vectors
  G <- c(G, 0.25 * Gn + 0.07 * Sn)
  S <- c(S, 0.75 * Gn + 0.93 * Sn)
}
```

The steady states for this system are

$$G = 0.085366$$
$$S = 0.914634$$

These values are achieved after 11 days.



Section 6.2, Problem 1

The model shown has a single component joined in series with two parallel subsystems. The reliability of the system can be given by the product of the reliabilities of the first component and each of the subsystems:

$$R(t) = R_{PA}(t) \times R_{CD-Radio}(t) \times R_{Speakers}(t)$$

The reliabilities of the two subsystems are given by the difference between the sums and the products of their components' reliabilities:

$$\begin{aligned} R_{CD-Radio}(t) &= R_{CD}(t) + R_{Radio}(t) - R_{CD}(t) \times R_{Radio}(t) \\ &= 0.98 + 0.97 - 0.98 \times 0.97 = 0.9994 \\ R_{Speakers}(t) &= R_{Speaker1}(t) + R_{Speaker2}(t) - R_{Speaker1}(t) \times R_{Speaker2}(t) \\ &= 0.99 + 0.99 - 0.99 \times 0.99 = 0.9999 \end{aligned}$$

Substituting these equations into the total system reliability equation above:

$$R(t) = 0.95 \times 0.9994 \times 0.9999 = 0.9493351$$

Assuming that the 5 individual components of the system are independent, the reliability of the system is 0.9493351.

Section 6.3

There are $m = 21$ observations to consider in the two problems in this section:

```
h <- c(60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70,
       71, 72, 73, 74, 75, 76, 77, 78, 79, 80)
w <- c(132, 136, 141, 145, 150, 155, 160, 165, 170, 175, 180,
       185, 190, 195, 201, 206, 212, 218, 223, 229, 234)
```

Problem 1

For the model $w \propto h$, the slope can be calculated as

$$a = \frac{21 \sum h_i w_i - \sum h_i \sum w_i}{21 \sum h_i^2 - \left(\sum h_i\right)^2} = 5.1364$$

The intercept is given by

$$b = \frac{\sum h_i^2 \sum w_i - \sum h_i w_i \sum h_i}{21 \sum h_i^2 - \left(\sum h_i\right)^2} = -178.4978$$

So the equation is given by

$$\boldsymbol{w = 5.1264h - 178.4978}$$

The error of the sum of squares is given by

$$SSE = \sum \left[w_i - (ah_i + b)\right]^2 = 24.6342$$

The total corrected sum of squares is given by

$$SST = \sum \left(w_i - \bar{w}\right)^2 = 20338.95$$

The regression sum of squares is given by

$$SSR = SST - SSE = 20338.95 - 24.6342 = 20314.32$$

The coefficient of determination is given by

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{24.6342}{20338.95} = 0.9988$$

These values can be verified using R's `lm` function:

```
mod1 <- lm(w ~ h)
summary(mod1)
anova(mod1)
```

	Estimate	Std. Error	t value	Pr(>|t|)
h	5.136	0.04103	125.2	3.528e-29
(Intercept)	-178.5	2.883	-61.91	2.198e-23

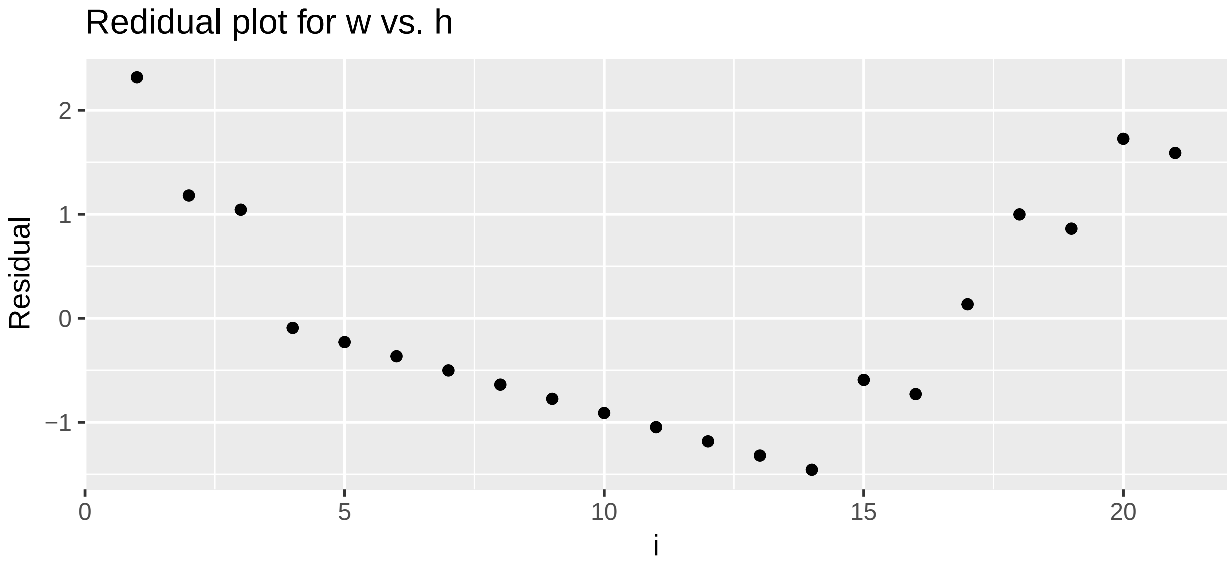
Table 2: Fitting linear model: w ~ h

Observations	Residual Std. Error	R^2	Adjusted R^2
21	1.139	0.9988	0.9987

Table 3: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
h	1	20314	20314	15668	3.528e-29
Residuals	19	24.63	1.297	NA	NA

The plot of the residuals of this fit are shown below:



Problem 2

For the model $w \propto h^3$, the slope can be calculated as

$$a = \frac{21\sum h_i^3 w_i - \sum h_i^3 \sum w_i}{21\sum h_i^6 - \left(\sum h_i^3\right)^2} = 0.0003467$$

The intercept is given by

$$b = \frac{\sum h_i^6 \sum w_i - \sum h_i^3 w_i \sum h_i^3}{21\sum h_i^6 - \left(\sum h_i^3\right)^2} = 59.4584$$

So the equation is given by

$$\boldsymbol{w = 0.0003467h^3 + 59.4584}$$

The error of the sum of squares is given by

$$SSE = \sum \left[w_i - (ah_i^3 + b)\right]^2 = 39.8620$$

The total corrected sum of squares is the same as above:

$$SST = \sum \left(w_i - \bar{w}\right)^2 = 20338.95$$

The regression sum of squares is given by

$$SSR = SST - SSE = 20338.95 - 39.8620 = 20299.09$$

The coefficient of determination is given by

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{39.8620}{20338.95} = 0.9980$$

These values are also verified in R:

```
mod2 <- lm(w ~ I(h^3))
```

	Estimate	Std. Error	t value	Pr(>|t|)
I(h^3)	0.0003467	3.525e-06	98.36	3.415e-27
(Intercept)	59.46	1.276	46.6	4.69e-21

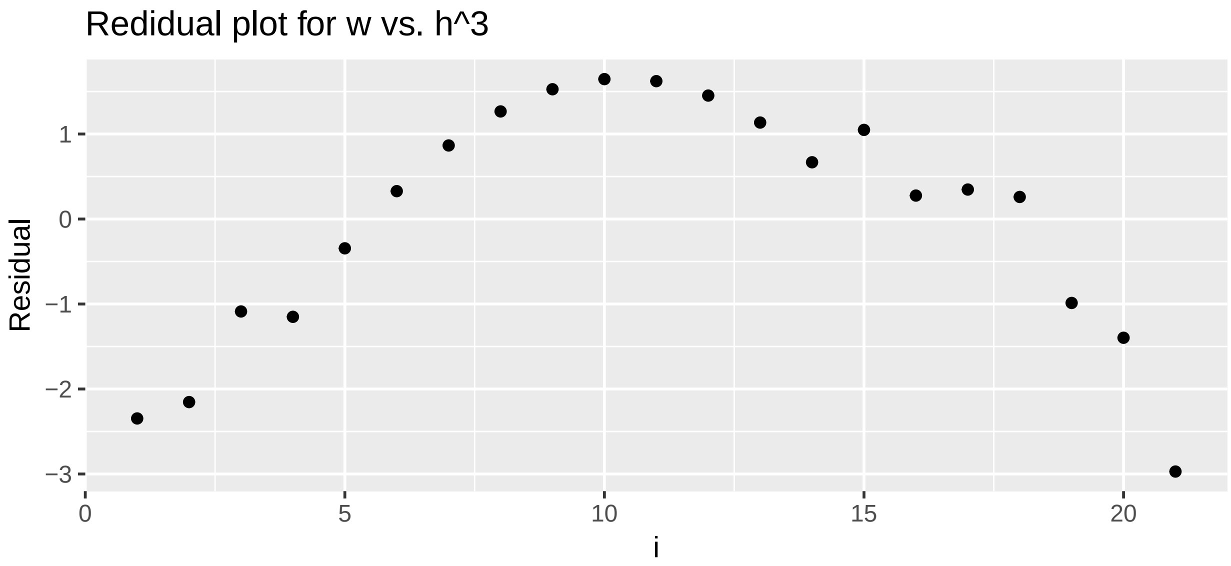
Table 5: Fitting linear model: w ~ I(h^3)

Observations	Residual Std. Error	R^2	Adjusted R^2
21	1.448	0.998	0.9979

Table 6: Analysis of Variance Table

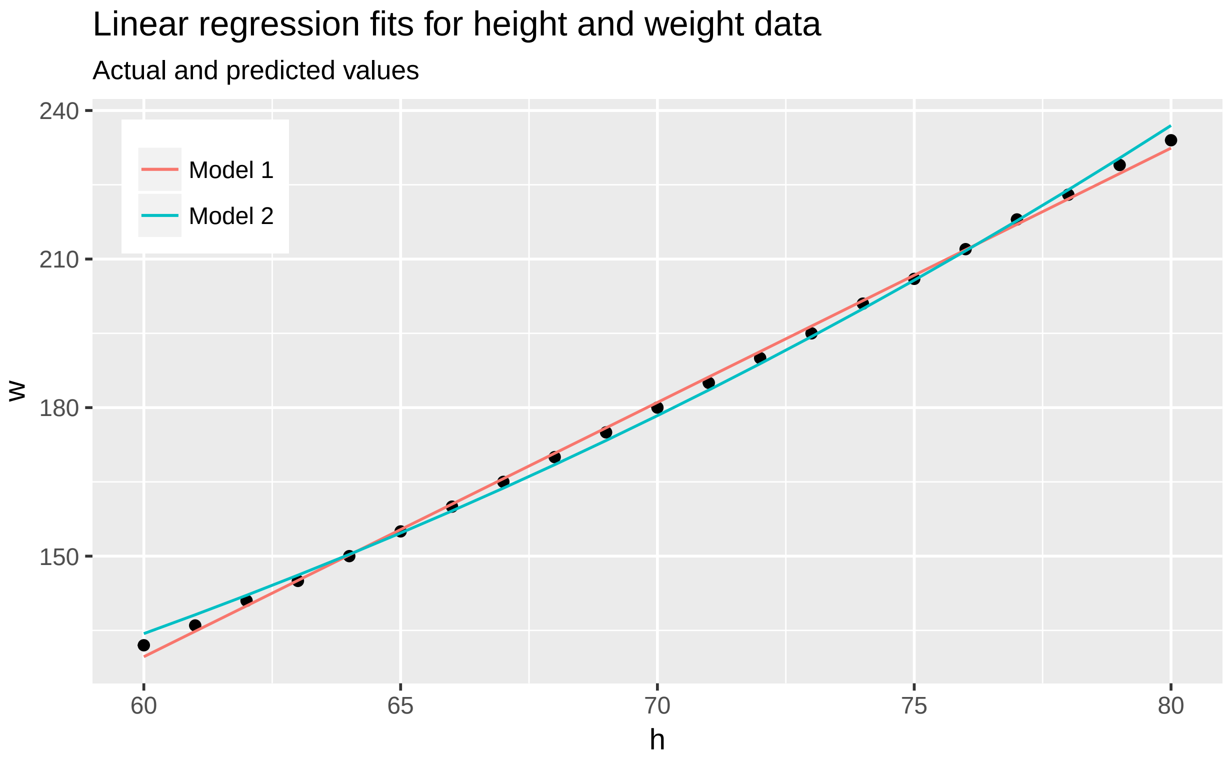
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(h^3)	1	20299	20299	9675	3.415e-27
Residuals	19	39.86	2.098	NA	NA

The plot of the residuals of this fit are shown below:



Comparison of Results

The models fit in Problems 1 and 2 ($w \propto h$ and $w \propto h^3$, respectively) both showed very strong R^2 results, with the first model having a very-slightly higher value (0.9988 vs. 0.9980). The residual plots for the two models both showed patterns - model 1's residuals decreased for the first 15 observations then increased, while model 2's residuals increased for the first 10 observations then decreased. Because of this, each of these models should be used with caution, especially outside the range of observed heights and weights. Due to the more parsimonious nature of Model 1, along with its slightly better fit, it is likely the preferable model.


Linear regression fits for height and weight data
Actual and predicted values
Model 1
Model 2
w
h