Stat 539 Homework 1

Kenny Flagg

January 26, 2017

1. *Agresti Exercise 1.4 (p. 21) Extend the model in Section 1.2.1 relating income to racial-ethnic status to include education and interaction explanatory terms. Explain how to interpret the parameters when software constructs the indicators using (**a**) first-category-baseline coding, (**b**) last-category-baseline coding.*

   If x_{i3} is the number of years of schooling for the ith individual, the model becomes

   $$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i3} + \beta_5 x_{i2} x_{i3}.$$

   (**a**) With first-category-baseline, black is the baseline racial-ethnic status, so parameters for Hispanics and whites are interpreted relative to blacks.

   - β_0: The mean income for blacks with no schooling is β_0.
   - β_1: The mean income for Hispanics with no schooling is β_1 greater than the mean income for blacks with no schooling.
   - β_2: The mean income for whites with no schooling is β_2 greater than the mean income for blacks with no schooling.
   - β_3: The mean income for blacks with a given number of years of schooling is β_3 greater than the mean income for blacks with one fewer year of schooling.
   - β_4: The difference in mean income between Hispanics with a given number of years of schooling and Hispanics with one fewer year of schooling is β_4 greater than the difference in mean income between blacks with a given number of years of schooling and blacks with one fewer year of schooling.
   - β_5: The difference in mean income between whites with a given number of years of schooling and whites with one fewer year of schooling is β_5 greater than the difference in mean income between blacks with a given number of years of schooling and blacks with one fewer year of schooling.

   (**b**) With first-category-baseline, white is the baseline racial-ethnic status, so parameters for blacks and Hispanics are interpreted relative to whites.

   - β_0: The mean income for whites with no schooling is β_0.
   - β_1: The mean income for blacks with no schooling is β_1 greater than the mean income for whites with no schooling.

- β_2: The mean income for Hispanics with no schooling is β_2 greater than the mean income for whites with no schooling.
- β_3: The mean income for whites with a given number of years of schooling is β_3 greater than the mean income for whites with one fewer year of schooling.
- β_4: The difference in mean income between blacks with a given number of years of schooling and blacks with one fewer year of schooling is β_4 greater than the difference in mean income between whites with a given number of years of schooling and whites with one fewer year of schooling.
- β_5: The difference in mean income between Hispanics with a given number of years of schooling and Hispanics with one fewer year of schooling is β_5 greater than the difference in mean income between whites with a given number of years of schooling and whites with one fewer year of schooling.

2. *Agresti Exercise 1.8 (p. 22) A model M has model matrix $\mathbf{X}$. A simpler model M_0 results from removing the final term in M, and hence has model matrix $\mathbf{X}_0$ that deletes the final column from $\mathbf{X}$. From the definition of a column space, explain why $C(\mathbf{X}_0)$ is contained in $C(\mathbf{X})$.*

   Intuitively, $C(\mathbf{X}_0)$ is a subset of $C(\mathbf{X})$ because its basis (the columns of $\mathbf{X}_0$) is a subset of the basis of $C(\mathbf{X})$ (the columns of $\mathbf{X}$). Formally, if $\boldsymbol{\mu} \in C(\mathbf{X}_0)$ and $\mathbf{x}_m$ is the final column of $\mathbf{X}$ then

$$\begin{aligned}
\boldsymbol{\mu} &= \mathbf{X}_0\boldsymbol{\beta}_0 \\
&= \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{x}_m \times 0 \\
&= \begin{bmatrix} \mathbf{X}_0 & \mathbf{x}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_0 \\ 0 \end{bmatrix} \\
&= \mathbf{X}\boldsymbol{\beta} \in C(\mathbf{X})
\end{aligned}$$

   with $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_0 \\ 0 \end{bmatrix}$, so $C(\mathbf{X}_0) \subset C(\mathbf{X})$.

3. *Agresti Exercise 2.10 (p. 72) For a projection matrix $\mathbf{P}$, for any $\mathbf{y}$ in $\mathbb{R}^n$ show that $\mathbf{Py}$ and $\mathbf{y} - \mathbf{Py}$ are orthogonal vectors.*

   Using the fact that $\mathbf{P}$ is a projection matrix if and only if $\mathbf{P}$ is symmetric and idempotent (proved in Section 2.2.1),

$$\begin{aligned}
(\mathbf{Py})^T(\mathbf{y} - \mathbf{Py}) &= \mathbf{y}^T\mathbf{P}^T(\mathbf{y} - \mathbf{Py}) \\
&= \mathbf{y}^T\mathbf{P}(\mathbf{y} - \mathbf{Py}) \\
&= \mathbf{y}^T\mathbf{Py} - \mathbf{y}^T\mathbf{P}^2\mathbf{y} \\
&= \mathbf{y}^T\mathbf{Py} - \mathbf{y}^T\mathbf{Py} \\
&= 0
\end{aligned}$$

   so $\mathbf{Py}$ and $\mathbf{y} - \mathbf{Py}$ are orthogonal.

4. *Consider the following model: For* $i = 1, \ldots, n$,

$$y_i = \beta_1 x_i + e_i,$$

where $E(e_i) = 0$ *and* $Var(e_i) = \sigma^2$.

(a) *Write this model in the matrix form* $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, *writing out the dimensions and elements of* $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, *and* $\mathbf{e}$.

The matrix form of this model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ is $n \times 1$,

$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ is $n \times 1$, $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \end{bmatrix}$ is 1×1, and $\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$ is $n \times 1$.

(b) *Derive the least squares estimate for* β_1, $\hat{\beta}_1$.

We need to minimize

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2 .$$

Taking the derivative,

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - \hat{y}_i) = -2 \sum_{i=1}^{n} x_i (y_i - \beta_1 x_i) .$$

Now we set the derivative equal to zero and solve,

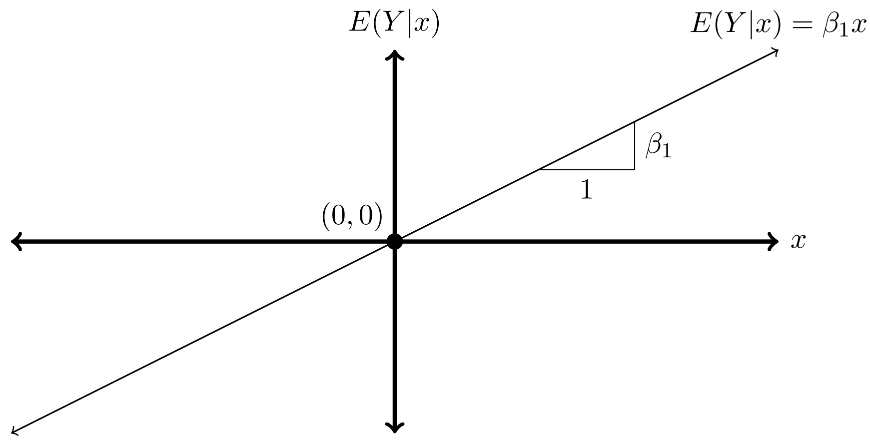
$$\begin{aligned} -2 \sum_{i=1}^{n} x_i (y_i - \beta_1 x_i) &= 0 \\ \sum_{i=1}^{n} x_i y_i - \beta_1 \sum_{i=1}^{n} x_i^2 &= 0 \end{aligned}$$

so the OLS estimator is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

(For thoroughness, I'll point out that the second derivative is positive for all β_1 as long as some $x_i \neq 0$, so this is in fact a minimum.)

(c) *Draw a plot (either by hand or with software) of $E(Y|x)$ versus x.*



5. *Patients with partial seizures were enrolled in a randomized clinical trial of the anti-epileptic drug, progabide. Participants were randomized to either progabide or a placebo. Prior to receiving treatment, baseline data on the number of seizures during the preceding 8-week interval were recorded. Counts of epileptic seizures during 2-week intervals before each of four successive post-randomization clinic visits were recorded. The main goal of the study was to compare the changes in the average rates of seizures in the two groups.*

   Displayed below are the data for the first three and last three individuals in the data set, where the response listed in columns 4–8 is the rate of seizures per week for that individual at baseline (week 0), week 2, week 4, week 6, and week 8 (rate = count/8 for baseline measurement; rate = count/2 for post-baseline measurements):

```
> head(epilepsy,3)
   ID     trt  age  Week0 Week2 Week4 Week6 Week8
1   1  Placebo   31  1.375   2.5   1.5   1.5   1.5
2   2  Placebo   30  1.375   1.5   2.5   1.5   1.5
3   3  Placebo   25  0.750   1.0   2.0   0.0   2.5
> tail(epilepsy,3)
    ID       trt  age  Week0 Week2 Week4 Week6 Week8
57  57  Progabide   21  3.125   1.0   1.5   0.0   0.5
58  58  Progabide   36  1.625   0.0   0.0   0.0   0.0
59  59  Progabide   37  1.500   0.5   2.0   1.5   1.0
```

   Define variables: Y_{ij} = jth rate measurement on the ith individual,

$$T_i = \begin{cases} 1 & \text{individual } i \text{ was randomly assigned to Progabide} \\ 0 & \text{individual } i \text{ was randomly assigned to Placebo} \end{cases}$$

   and

$$w_j = \text{week of the } j\text{th measurement (e.g., } w_1 = 0).$$

   where $i = 1, \ldots, 59$, and $j = 1, \ldots, n$. (Note that we are treating week as a quantitative variable.)

(a) *Is this study exploratory or confirmatory? Explain.*

This study is confirmatory because the researchers selected one drug (progabide) for the treatment group, implying that they already had a hypothesis they wanted to test about progabide's effectiveness relative to a placebo.

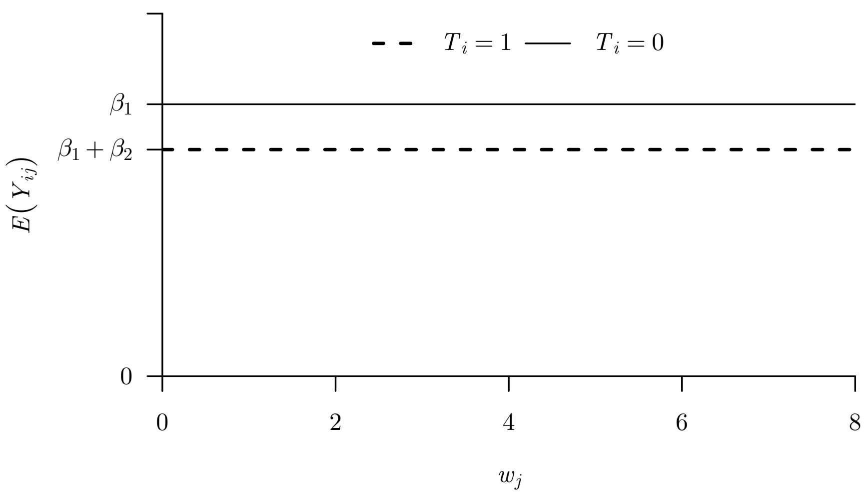
(b) *For each of the following mean models expressed in scalar notation below,*

1. *write out the vector of regression parameters* $\boldsymbol{\beta}$ *(the elements will be symbols),*
2. *write out the rows of the design matrix* $\mathbf{X}$ *that correspond to the observations on individuals 3 and 59 (the elements will be numbers),*
3. *on a single well-labeled plot, draw two hypothetical mean response trajectories (assuming all elements of* $\boldsymbol{\beta}$ *are non-zero), one for each treatment (Placebo and Progabide), with time (in weeks) on the x-axis, and the mean rate of seizures on the y-axis.*

i. $E(Y_{ij}) = \beta_1 + \beta_2 T_i$

$$\boldsymbol{\beta} = \begin{bmatrix}\beta_1 \\ \beta_2\end{bmatrix}, \qquad \mathbf{X}_3 = \begin{bmatrix}1 & 0\end{bmatrix}, \qquad \mathbf{X}_{59} = \begin{bmatrix}1 & 1\end{bmatrix}$$

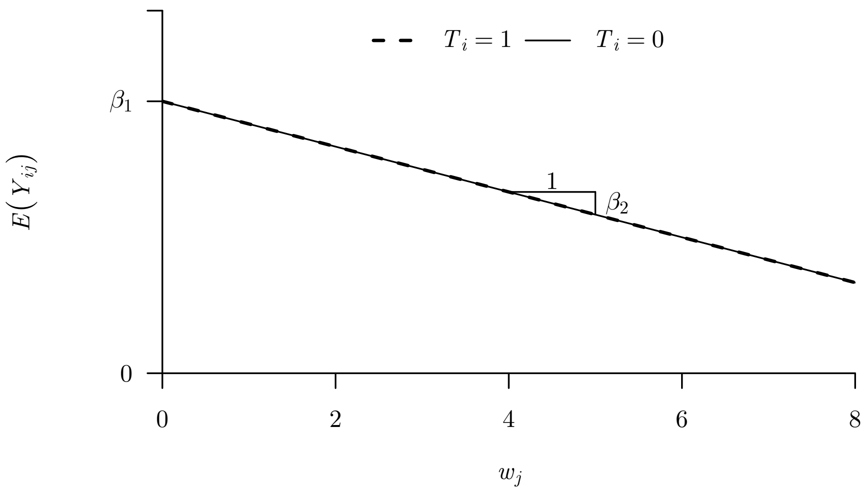
Plot assuming $\beta_2 < 0$:



ii. $E(Y_{ij}) = \beta_1 + \beta_2 w_j$

$$\boldsymbol{\beta} = \begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}, \qquad \mathbf{X}_3 = \begin{bmatrix}1 & 0\\ 1 & 2\\ 1 & 4\\ 1 & 6\\ 1 & 8\end{bmatrix}, \qquad \mathbf{X}_{59} = \begin{bmatrix}1 & 0\\ 1 & 2\\ 1 & 4\\ 1 & 6\\ 1 & 8\end{bmatrix}$$

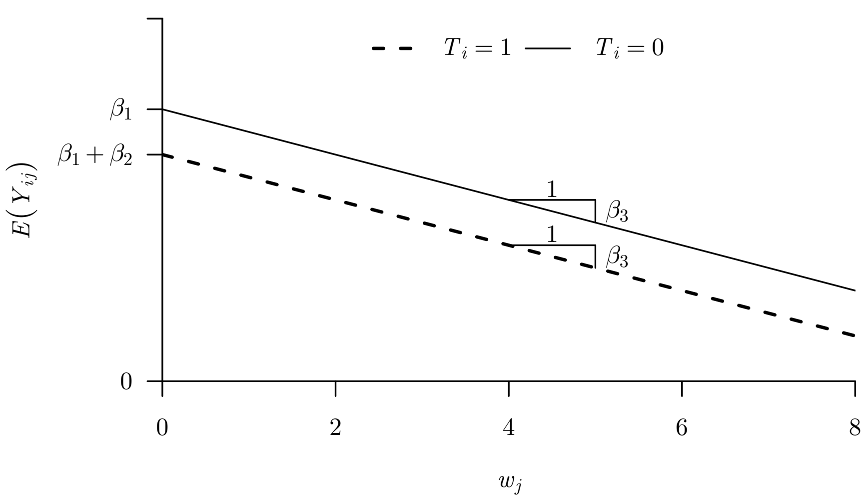
Plot assuming $\beta_2 < 0$:



iii. $E(Y_{ij}) = \beta_1 + \beta_2 T_i + \beta_3 w_j$

$$\boldsymbol{\beta} = \begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix}, \qquad \mathbf{X}_3 = \begin{bmatrix}1 & 0 & 0\\ 1 & 0 & 2\\ 1 & 0 & 4\\ 1 & 0 & 6\\ 1 & 0 & 8\end{bmatrix}, \quad \mathbf{X}_{59} = \begin{bmatrix}1 & 1 & 0\\ 1 & 1 & 2\\ 1 & 1 & 4\\ 1 & 1 & 6\\ 1 & 1 & 8\end{bmatrix}$$

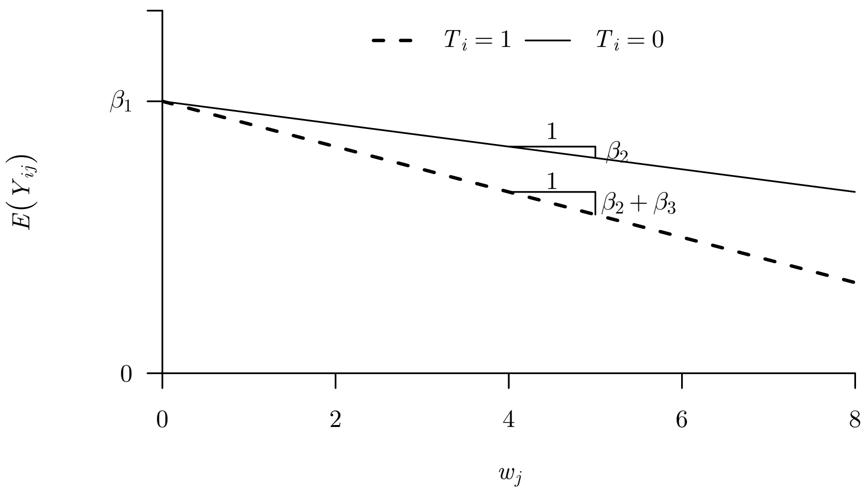
Plot assuming $\beta_2 < 0$ and $\beta_3 < 0$:



iv. $E(Y_{ij}) = \beta_1 + \beta_2 w_j + \beta_3 T_i \times w_j$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \qquad \mathbf{X}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 4 & 0 \\ 1 & 6 & 0 \\ 1 & 8 & 0 \end{bmatrix}, \qquad \mathbf{X}_{59} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 2 \\ 1 & 4 & 4 \\ 1 & 6 & 6 \\ 1 & 8 & 8 \end{bmatrix}$$

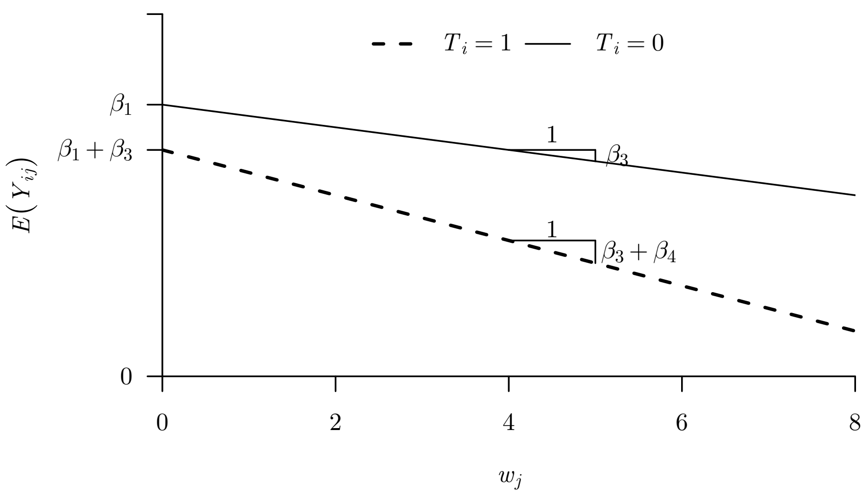
Plot assuming $\beta_2 < 0$ and $\beta_3 < 0$:



v. $E(Y_{ij}) = \beta_1 + \beta_2 T_i + \beta_3 w_j + \beta_4 T_i \times w_j$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}, \qquad \mathbf{X}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 4 & 0 \\ 1 & 0 & 6 & 0 \\ 1 & 0 & 8 & 0 \end{bmatrix}, \qquad \mathbf{X}_{59} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 2 & 2 \\ 1 & 1 & 4 & 4 \\ 1 & 1 & 6 & 6 \\ 1 & 1 & 8 & 8 \end{bmatrix}$$

Plot assuming $\beta_2 < 0$, $\beta_3 < 0$, and $\beta_4 < 0$:



(c) *Which of the models in part (b) is most appropriate for the main goal of the study? Explain.*

Model iv. is the most appropriate model for their goal. There should not be a difference between the groups at week zero because the treatment was not yet applied. Any difference between progabide and the placebo would appear as a difference in the trend over time, so we should allow the groups to have different slopes, but the varying intercept is not necessary.

(d) *Suppose you fit a linear model to these data using ordinary least squares. What is the primary reason why inference on the linear model coefficients would be invalid? Explain.*

The only possible values of the response variable are multiples of 0.125 (for week 0) or 0.5 (for the other weeks), so it is not appropriate to use the normal distribution for inference.

(e) *In mean model v. in part (b), if we treated week as a categorical variable, how many additional coefficients would we add?*

We would need 6 additional coefficients: 3 more week main effects coefficients and 3 more week by treatment interaction coefficients.