STAT 151A: Lab 10
Review for Midterm 2

Billy Fang

3 November 2017

Feedback form is at the same place: https://goo.gl/forms/fKjLeKItix2Djg5l2. Please leave comments and suggestions for lab, office hours, etc.

1 One-way ANOVA

Relevant reading: lecture notes, Lab 5 notes, Fox 8.1.

1.1 Setting up the model

The purpose of this section is to very slowly clarify the various notations used in this topic. If you are comfortable with this, you can skip to the next subsection.

Example dataset. Each weekday that you wait for the bus, you keep track of how many minutes later than schedule it arrives. (Negative values correspond to cases where the bus arrives earlier than scheduled.) Some days you don't ride the bus, so you don't have data for every day. You also ride the bus at most once per day.

Minutes late	Day of the week
1	M
-1.5	T
3	Th
-2	W
10	F
$\vdots$	$\vdots$
12	Th
-1	T

Assumptions. In one-way ANOVA we assume the following.

- Each data point is independent of the others.
- The number of minutes late on a Monday follows the $N(\mu^*_{\text{M}}, \sigma^2)$ distribution where μ^*_{M} is some number representing the true mean number of minutes late on Mondays. Similarly, we assume the number of minutes late on Tuesday follows the $N(\mu_{\text{T}}, \sigma^2)$ distribution, and so on.

An important assumption is that σ^2 is the same across all data points (and thus the same across all the groups).

Writing the model in $y = X\beta^* + \epsilon$ form. We can write this model as[1]

$$y_i = \mu^*_{\text{M}} I(\text{Day}_i = \text{M}) + \mu^*_{\text{T}} I(\text{Day}_i = \text{T}) + \cdots + \mu^*_{\text{F}} I(\text{Day}_i = \text{F}) + \epsilon_i, \qquad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \tag{1}$$

[1] I will use asterisks $*$ for the true parameters (true group mean, etc.) to emphasize that they are the true parameter. Your lecture notes omit the asterisks.

where $I(\text{Day}_i = \text{M})$ equals 1 if the ith datapoint was for a Monday, and otherwise equals 0. From here we can write this in $y = X\beta^* + \epsilon$ form.

$$y = \begin{bmatrix} 1 \\ -1.5 \\ 3 \\ -2 \\ 10 \\ \vdots \\ 12 \\ -1 \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}, \qquad \mu^* = \beta^* = \begin{bmatrix} \mu^*_{\text{M}} \\ \mu^*_{\text{T}} \\ \mu^*_{\text{W}} \\ \mu^*_{\text{Th}} \\ \mu^*_{\text{F}} \end{bmatrix}, \qquad \epsilon \sim N(0, \sigma^2 I_n). \tag{2}$$

Here, the first column of the design matrix corresponds to $I(\text{Day}_i = \text{M})$, and so on. *Check for yourself that $y = X\beta^* + \epsilon$ indeed describes the same model as* (1).

Note that this is exactly the same as encoding a linear model for y against a single categorical variable using dummy variables, but keeping all the indicator columns and removing the intercept column.

We will use μ instead of β to stick with the notation in the lecture notes and textbook, but you should recognize that *this model is just a special case of the general setting $y = X\beta^* + \epsilon$* that we studied a lot before the first midterm.

Re-indexing. Let us forget the names of the days of the week, and just think of them as "group 1" to "group 5," with true group means $\mu_1, \dots, \mu_5$. Let n_1 be the number of datapoints in group 1 (Monday), and so on for $n_2, \dots, n_5$. In one-way ANOVA, it is convenient to change the indexing. For example, instead of $y_1, \dots, y_n$ where n is the number of datapoints, your lecture notes and the textbook indexes by group. In the above example, if we arrange y by group, we can make a table that looks something like

Group 1	1	⋯	
Group 2	-1.5	⋯	-1
Group 3	-2	⋯	
Group 4	3	⋯	12
Group 5	10	⋯	

The first row has the y values for the n_1 datapoints in group 1, and so on. In your lecture notes, the y values are now indexed as

$$y_{ij}, \qquad i \in \{1, \dots, 5\}, \quad j \in \{1, \dots, n_i\},$$

where i is the group that y_{ij} belongs to, and j is its position in the above table. (That is, y_{ij} is the jth element in the ith row of the above table.) We can re-index $\epsilon_1, \dots, \epsilon_n$ as ϵ_{ij} similarly. Thus, we can rewrite the model (1) as

$$y_{ij} = \mu^*_i + \epsilon_{ij}, \qquad \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2). \tag{3}$$

Check that this indeed is the same as (1). Similarly, we can convert from this indexing back to something like (2). (The rows in (2) might be in a different order, but that does not ultimately matter.)

We have seen three ways of writing the model (1). You should be flexible switching between models (2) and (3).

1.2 Not-so-scary formulas

We have see a lot of formulas in the general situation $y = X\beta^* + \epsilon$. However, in this special case, a lot of the formulas can be written down explicitly in terms of natural quantities.

Let us fully wean ourselves off of the concrete example above. Suppose we have t groups (instead of just 5) and we get observations y_{ij} from the model (3). Let n be the total number of observations, and let $n_1, \dots, n_t$ be the number of observations in each group respectively.

Exercise 1.1. *If we write our design matrix as in* (2)*, check that the dimension of X is $n \times t$. Describe what each row of X represents, and what each column represents.* ■

1.2.1 Least squares coefficients

Let us start with the least squares estimate of μ. We provide three different ways to compute it.

Approach 1: use the general formula. The general formula is

$$\widehat{\mu} = (X^\top X)^{-1} X^\top y.$$

But our X matrix has a very special form, which allows us to write down a very simple expression for $\widehat{\mu}$.

Exercise 1.2.

(a) Show

$$X^\top X = \begin{bmatrix} n_1 & & & \\ & n_2 & & \\ & & \ddots & \\ & & & n_t \end{bmatrix}.$$

(b) Show

$$X^\top y = \begin{bmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{j=1}^{n_2} y_{2j} \\ \vdots \\ \sum_{j=1}^{n_t} y_{tj} \end{bmatrix}.$$

In plain words, describe the elements of this vector.

(c) Show

$$\widehat{\mu} = \begin{bmatrix} \overline{y}_1 \\ \overline{y}_2 \\ \vdots \\ \overline{y}_t \end{bmatrix},$$

where $\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$ is the sample mean for group i (mean of the datapoints in group i).

■

Thus, we obtain an intuitive result: the least squares estimate of the true group means μ_t is the sample group mean.

Approaches 2 and 3: choose μ to minimize the residuals. The other two approaches directly minimize the function

$$S(\mu) = \|y - X\mu\|^2 = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2.$$

over all vectors μ. (Recall the definition of "least squares." In the general setting $y = X\beta^* + \epsilon$, our function was $S(\beta) = \|y - X\beta\|^2 = \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2$. Check that you understand that the above double sum is essentially the same thing.)

Here, S is a function whose argument is a vector μ, and the least squares estimate $\widehat{\mu}$ is the vector that minimizes S. Note that μ is some arbitrary vector, and is *not* the true mean μ^*. (This was a source of confusion for many of you, which is why I like to use the asterisk μ^* notation for the true mean.)

One way to minimize S is to compute the gradient with respect to the vector μ and set it to zero. Equivalently, compute the partial derivative with respect to each μ_i and set them all to zero.

Exercise 1.3. *Compute*

$$\frac{\partial}{\partial \mu_i} S(\mu).$$

When you set this to zero, you should get $\widehat{\mu}_i = \overline{y}_i$. ■

Finally, the last approach is the approach taken in your lecture notes.

$$\begin{aligned}
S(\mu) &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2 \\
&= \sum_{i=t}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i+\overline{y}_i-\mu_i)^2 && \text{add and subtract } \overline{y} \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 + 2\sum_{i=1}^{t}(\overline{y}_i-\mu_i)\underbrace{\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)}_{=0} + \sum_{i=1}^{t}n_i(\overline{y}_i-\mu_i)^2 && \text{expand the square} \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 + \sum_{i=1}^{t}n_i(\overline{y}_i-\mu_i)^2.
\end{aligned}$$

Again, $\widehat{\mu}$ is the vector that minimizes the above expression.

Exercise 1.4. *Based on the last expression above, why does choosing $\widehat{\mu}_i = \overline{y}_i$ minimize $S(\mu)$?* ■

Now that we have $\widehat{\mu}$, we can automatically compute the RSS [of the full model M] by simply plugging $\widehat{\mu}$ into S. (Recall in the general situation $y = X\beta^* + \epsilon$, we have RSS $= \|y - X\widehat{\beta}\|^2 = S(\widehat{\beta})$. This is the same thing.)

$$\text{RSS}(M) = S(\widehat{\mu}) = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2.$$

The inner sum measures variability *within* group i, and the outer sum adds all these quantities across the t groups.

1.2.2 Variance decomposition, F-statistic for $H_0 : \mu_1 = \cdots = \mu_t$.

The typical hypothesis that is tested in one-way ANOVA is

$$H_0 : \mu_1 = \cdots = \mu_t, \tag{4}$$

i.e. the null hypothesis is that all group means are the same. For example, in the case of our earlier dataset, we want to test if the average lateness/earliness of the bus does not vary with the day of the week.

One quantity that will be useful for us is RSS(m), where m is the model with the hypothesis's constraints imposed. Then the model is simply $y_{ij} = \mu + \epsilon_{ij}$, which is like an intercept-only model. At the very beginning of the course you showed that then the least squares estimate in the intercept-only model is the mean of the y_{ij}, which in our case is the grand mean

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{t}\sum_{j=1}^{n_i}y_{ij}.$$

Then you can use the above work (substitute μ_i with $\overline{y}$ for all i) to get

$$\begin{aligned}
\text{RSS}(m) &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 \\
&= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 + \sum_{i=1}^{t}n_i(\overline{y}_i-\overline{y})^2.
\end{aligned}$$

Note that this is also the TSS! Thus, the decomposition TSS $=$ RSS $+$ RegSS looks like

$$\underbrace{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2}_{\text{RSS}(M)} + \underbrace{\sum_{i=1}^{t}n_i(\overline{y}_i-\overline{y})^2}_{\text{RegSS}(M)}.$$

This has a nice interpretation. The left-hand side measures variability of the y_{ij} around the grand mean $\overline{y}$. The RSS term measures variability within each group (with respect to the sample group means) and sums this over all groups. The RegSS term measures variability of the [sample] group means around the grand mean. (See Lab 5 for nice visualizations.)

Let us now compute the F-statistic. Recall in the $y = X\beta^* + \epsilon$ setting, when testing a linear hypothesis with $q \leq p$ constraints, the general F-statistic is $\frac{(\mathrm{RSS}(m)-\mathrm{RSS}(M))/q}{\mathrm{RSS}(M)/(n-p-1)}$, where m is the model with the constraints in the null hypothesis, M is the full model, q is the number of constraints, and $n-p-1$ is "$n -$ number of parameters" or "$n -$ number of columns of X". Essentially the same formula holds here, provided we compute the degrees of freedom and the RSS terms correctly.

Exercise 1.5.

(a) How many linear constraints are in the hypothesis (4)*?*

(b) How many columns does X have?

(c) Plug in our expressions for $\mathrm{RSS}(M)$ *and* $\mathrm{RSS}(m)$ *to show that the F-statistic is*

$$\frac{\sum_{i=1}^{t} n_i(\overline{y}_i - \overline{y})^2/(t-1)}{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2/(n-t)}.$$

■

This is also interpretable. If the group means are very different (contrary to the hypothesis), then the numerator ought to be very large relative to the denominator, making the F-statistic large and more likely to be rejected.

1.2.3 Other stuff

Exercise 1.6. *Show that $\widehat{\mu} \sim N(\mu, \sigma^2(X^\top X)^{-1})$. What is $\mathrm{Var}(\widehat{\mu}_i)$? What is $\mathrm{Cov}(\widehat{\mu}_i, \widehat{\mu}_j)$ for $i \neq j$?* ■

We can also compute the relevant quantities for t-tests. The unbiased estimate for σ^2 is

$$\widehat{\sigma}^2 := \frac{\mathrm{RSS}(M)}{n-t} = \frac{1}{n-t}\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2.$$

Having computed $\widehat{\sigma}$, the t-statistic for testing $H_0 : \mu_2 = 0$ is then

$$\frac{\widehat{\mu}_2}{\widehat{\sigma}\sqrt{((X^\top X)^{-1})_{2,2}}} = \frac{\widehat{\mu}_2}{\widehat{\sigma}/\sqrt{n_2}}.$$

The left-hand side is essentially the same as the t-statistic $\frac{\widehat{\beta}_2}{\widehat{\sigma}\sqrt{((X^\top X)^{-1})_{2,2}}}$ for $\beta_2 = 0$ in the general setting $y = X\beta^* + \epsilon$. (Recall that this is derived by noting $\widehat{\beta}_2 \sim N(0, \sigma^2((X^\top X)^{-1})_{2,2})$ under this hypothesis, standardizing it to get $\frac{\widehat{\beta}}{\sigma\sqrt{((X^\top X)^{-1})_{2,2}}} \sim N(0,1)$, and then replacing σ with $\widehat{\sigma}$.)

In our current situation, $X^\top X$ has a special form (see Exercise 1.2), which allows us to arrive at the right-hand side.

Exercise 1.7. *Compute the t-statistic for $H_0 : \mu_2 - 3\mu_3 = 0$. (You will be able to write it in terms of $\widehat{\mu}_2$, $\widehat{\mu}_3$, $\widehat{\sigma}$, n_2, and n_3.)* ■

2 Variance-stabilizing transformations

In our linear model

$$y_i = x_i^\top \beta^* + \epsilon_i, \qquad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2),$$

we have $\mathrm{Var}(y_i) = \sigma^2$ and $\mathbb{E}[y_i] = x_i^\top \beta^*$. In particular, the variance does not vary with the expectation. This is one of the properties of the linear model that may be violated.

Contrast this with the following model

$$y_i \sim \mathrm{Poisson}(x_i),$$

where $\mathrm{Var}(y_i) = \mathbb{E}[y_i] = x_i$. The variance clearly depends on the expectation.

Our goal here is to find a function h such that $h(y)$ has variance not varying with its expectation. That is, we want h to satisfy

$$\mathrm{Var}(h(y)) = c$$

for some constant c that does not depend on $\mathbb{E}[h(y)]$. We call this h a variance-stabilizing transformation.

By a Taylor series expansion, a rough approximation yields

$$\begin{aligned} h(y) &\approx h(\mathbb{E}y) + h'(\mathbb{E}y) \cdot (y - \mathbb{E}y) \\ \mathrm{Var}(h(y)) &\approx (h'(\mathbb{E}y))^2 \,\mathrm{Var}(y). \end{aligned}$$

We want the right-hand side to be constant (with respect to $\mathbb{E}y$), so we would like

$$h'(\mathbb{E}y) = \frac{\widetilde{c}}{\sqrt{\mathrm{Var}(y)}} \tag{5}$$

for some constant $\widetilde{c}$ that does not depend on $\mathbb{E}y$.

A rough template for finding the variance-stabilizing h is the following: write $\mathrm{Var}(y)$ as a function of $\mathbb{E}y$ and plug it into the desired equation (5), then find some h that satisfies this equation for any value of $\mathbb{E}y$.

As an example, consider the Poisson case where $\mathrm{Var}(y) = \mathbb{E}(y)$. The desired equation (5) becomes

$$h'(\mathbb{E}y) = \frac{\widetilde{c}}{\sqrt{\mathbb{E}y}}.$$

Perhaps replacing $\mathbb{E}y$ with a dummy variable will make our task clearer.

$$h'(z) = \frac{\widetilde{c}}{\sqrt{z}}.$$

From here, we see that $h(z) = \sqrt{z}$ is the desired transformation, since $h'(z) = \frac{1}{\sqrt{z}}$.

Finally, you may have seen the notation $\propto$ in place of "$=$" in the lecture notes. This is mainly for convenience to hide multiplicative constants (the $\widetilde{c}$ above).

Exercise 2.1. *Find the variance stabilizing h in the following situations.*

(a) $\mathrm{Var}(y) \propto (\mathbb{E}y)^2$.

(b) $\mathrm{Var}(y) \propto (\mathbb{E}y)^{2b}$ *for* $b \neq 1$.

■

Exercise 2.2. *Question 3 on "Optional problems (M2)."* ■

3 Added variable plot proof intuition

Relevant reading: lecture notes, Fox 11.6.1, `misc/coeff_comp.pdf` on bCourses.

Not realizing that this proof would be covered later in the course, I posted a proof of the first half of the AV plot theorem (stated below) on bCourses before the first midterm. You may check it out if you are curious. I think it is shorter and easier to understand, but it may potentially confuse you because the notation is a bit different from the presentation in the lecture notes. The ingredients of both proofs are essentially the same. Below, I will follow the lecture notes.

I uploaded a GeoGebra file on bCourses called `avplot_intuition.ggb`. You can play around with a 3D model of the geometric intuition I want to convey. GeoGebra is free, but if you do not want to download it, I have also uploaded some static images.

Let us restate the theorem again.

Theorem 3.1. Let $\widehat{\beta}$ be the vector of coefficients from the regression of Y on X with residuals $\widehat{e}$. Let $X(p)$ be the pth column of X.

1. Let $X^{(p)}$ be the residuals from regressing $X(p)$ on the other variables (including intercept).
2. Let $Y^{(p)}$ be the residuals from regressing Y on all variables except $X(p)$.
3. Let $b^{(p)}$ be the slope from the simple regression of $Y^{(p)}$ on $X^{(p)}$, with residuals $e^{(p)}$.

Then

$$\begin{aligned} b^{(p)} &= \widehat{\beta}_p, \\ e^{(p)} &= \widehat{e}. \end{aligned}$$

Note that above, p was chosen arbitrarily, and can be replaced with any $j = 1, \dots, p$.

Proof of first claim. Recall that the fitted values from a regression can be viewed as the projection of the response variable onto the span of the explanatory variables. Specifically, $\widehat{Y} = HY$ where $H = X(X^\top X)^{-1}X^\top$ is the projection onto $C(X)$. Therefore, the residuals from the regression of Y on X can be written as

$$\widehat{e} = Y - \widehat{Y} = (I - H)Y.$$

Let V be the matrix formed by dropping the column $X(p)$ from X. Since $X^{(p)}$ and $Y^{(j)}$ are each residuals from regressions onto V,

$$\begin{aligned} X^{(p)} &= (I - H(-p))X(p), \\ Y^{(p)} &= (I - H(-p))Y, \end{aligned}$$

where $H(-p) = V(V^\top V)^{-1}V^\top$ is the projection matrix onto $C(V)$.

Geometrically, what is going on so far? Our original regression was projecting Y onto $C(X)$. But the regressions in steps 1 and 2 above are essentially projections of Y and $X(p)$ onto $C(V)$, which is a subspace of $C(X)$.

The residuals $X^{(p)}$ sum to zero, as do the residuals $Y^{(p)}$. (This is because they are residuals from a regression on V, which has an intercept column.) Therefore, the intercept term in the simple regression in step 3 is zero (recall the formula for the intercept term.) The slope term can thus be written as

$$b^{(p)} = \frac{(X^{(p)})^\top Y^{(p)}}{\|X^{(p)}\|^2}. \tag{6}$$

We want to show that $\widehat{\beta}_p$ is equal to this slope.

Now recall the definition of least squares. The coefficient $\widehat{\beta}$ is the minimizer for the optimization problem

$$\min_{\beta} \|Y - X\beta\|^2 = \min_{\beta_p} \min_{\beta_0, \dots, \beta_{p-1}} \|Y - \beta_0 \vec{1} - \beta_1 X(1) - \beta_2 X(2) - \cdots - \beta_{p-1} X(p-1) - \beta_p X(p)\|^2.$$

If $\widehat{\beta}$ is indeed the minimizer, then $\widehat{\beta}_0, \dots, \widehat{\beta}_{p-1}$ must be the minimizer of the inner minimization problem on the right-hand side, with $\widehat{\beta}_p$ plugged in for β_p. That is, $\widehat{\beta}_0, \dots, \widehat{\beta}_{p-1}$ must be the minimizers for

$$\min_{\beta_0, \dots, \beta_{p-1}} \|(Y - \widehat{\beta}_p X(p)) - \beta_0 \vec{1} - \beta_1 X(1) - \beta_2 X(2) - \cdots - \beta_{p-1} X(p-1)\|^2 = \min_{\alpha \in \mathbb{R}^p} \|(Y - \widehat{\beta}_p X(p)) - V\alpha\|, \tag{7}$$

which is just another linear regression problem, specifically the regression of $Y - \widehat{\beta}_p X(p)$ onto V. Thus we know from the normal equation[2] for this problem (7) that the minimizing α is

$$\widehat{\alpha} = (V^\top V)^{-1} V^\top (Y - \widehat{\beta}_p X(p)).$$

[This is called "$\alpha(\widehat{\beta}_p)$" in the lecture notes.] Since we said that $\widehat{\beta}_0, \dots, \widehat{\beta}_{p-1}$ are also the minimizers for (7), we have $\widehat{\alpha}^\top = (\widehat{\beta}_0, \dots, \widehat{\beta}_{p-1})$. This reasoning skips over some arguments in the lecture notes (expanding the square, computing gradient, etc.) that are needlessly re-doing the computation for the normal equation that you already know.

Thus,

$$\begin{aligned}
HY &= X\widehat{\beta} \\
HY &= V\widehat{\alpha} + \widehat{\beta}_p X(p) \\
HY &= V(V^\top V)^{-1}V(Y - \widehat{\beta}_p X(p)) + \widehat{\beta}_p X(p) && \text{plug in expression for } \widehat{\alpha} \\
HY &= H(-p)(Y - \widehat{\beta}_p X(p)) + \widehat{\beta}_p X(p) && H(-p) = V(V^\top V)^{-1}V^\top \\
HY &= H(-p)Y + \widehat{\beta}_p (I - H(-p))X(p) \\
(I - H(-p))Y &= (I - H)Y + \widehat{\beta}_p (I - H(-p))X(p) \\
Y^{(p)} &= (I - H)Y + \widehat{\beta}_p X^{(p)} \\
\frac{(X^{(p)})^\top Y^{(p)}}{\|X^{(p)}\|^2} &= 0 + \widehat{\beta}_p.
\end{aligned}$$

In the last step we projected both sides onto $\mathrm{span}\{X^{(p)}\}$, i.e. we multiplied both sides by $\frac{(X^{(p)})^\top}{\|X^{(p)}\|^2}$. The first term on the right-hand side is zero because $(I - H)Y$ is orthogonal to $C(X)$ (which contains $X^{(p)} = X(p) - H(-p)X(p)$ since $X^{(p)}$ is the sum of two vectors in $C(X)$).

[Note that I ended the proof slightly differently than the lecture notes did. Both are correct, but this presentation shows $\widehat{\beta}$ is equal to (6) directly, rather than to the longer expression $\frac{Y^\top (I - H(-p))X(p)}{X(p)^\top (I - H(-p))X(p)}$, which I avoided altogether.] □

Exercise 3.2. *Prove the second claim, that $\mathrm{e}^{(p)} = \widehat{e}$. [Hint: It follows directly from the previous claim $\mathrm{b}^{(p)} = \widehat{\beta}_p$ combined with one of the lines I've written above. But on your homework you should base your work on the lecture notes; this might only add one more line of work.]* ■

[2] The formula for α is just adapting the familiar formula “$\widehat{\beta} = (X^\top X)^{-1}X^\top y$” to this problem, i.e. y is $Y - \beta_p X(p)$ and X is V.