Homework 9

Aleck Zhao

December 9, 2016

Chapter 12: The Analysis of Variance

5. Derive the likelihood ratio for the null hypothesis of the one-way layout, and show that it is equivalent to the F test in the case $I = 2$.

Proof. In the case $I = 2$, we have two treatments. Suppose that X_i and Y_i are data from the two treatments, and there are J observations for each. We have $H_0 : \mu_X = \mu_Y = \mu_0$. Assuming X_i and Y_i are normal variables with common variance σ^2, the numerator of the likelihood ratio is evaluated at the value of μ_0 that maximizes the likelihood function

$$f(X_i, Y_i \mid \mu_0) = \prod_{i=1}^{J} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(X_i - \mu_0)^2}{2\sigma^2}\right) \prod_{i=1}^{J} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(Y_i - \mu_0)^2}{2\sigma^2}\right)$$

which is $\mu_0 = \frac{\bar{X}+\bar{Y}}{2}$.

The denominator of the likelihood ratio is the likelihood function using μ_X and μ_Y at the MLE, which are $\bar{X}$ and $\bar{Y}$, respectively. Thus, the denominator is

$$f(X_i, Y_i) = \prod_{i=1}^{J} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(X_i - \bar{X})^2}{2\sigma^2}\right) \prod_{i=1}^{J} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(Y_i - \bar{Y})^2}{2\sigma^2}\right)$$

The constants cancel in the ratio, and we are left with

$$\exp\left(-\frac{1}{2\sigma^2}\left[\left(\sum_{i=1}^{J}(X_i - \mu_0)^2 + \sum_{i=1}^{J}(Y_i - \mu_0)^2\right) - \left(\sum_{i=1}^{J}(X_i - \bar{X})^2 + \sum_{i=1}^{J}(Y_i - \bar{Y})^2\right)\right]\right)$$

Using $\mu_0 = \frac{\bar{X}+\bar{Y}}{2}$ and some algebra, this simplifies to

$$\Lambda = \exp\left(-\frac{J}{4\sigma^2}(\bar{X} - \bar{Y})^2\right)$$

The F test statistic in the case $I = 2$ is

$$F = \frac{SS_B/(I-1)}{SS_W/[I(J-1)]} = \frac{SS_B}{SS_W/[2(J-1)]}$$

Here, the grand mean is given by $\frac{\bar{X}+\bar{Y}}{2}$. Thus, we have

$$\begin{aligned} SS_B &= J\left[\left(\bar{X} - \frac{\bar{X}+\bar{Y}}{2}\right)^2 + \left(\bar{Y} - \frac{\bar{X}+\bar{Y}}{2}\right)^2\right] \\ &= 2J\left(\frac{\bar{X}-\bar{Y}}{2}\right)^2 = \frac{J}{2}(\bar{X} - \bar{Y})^2 \end{aligned}$$

Then

$$\frac{SS_W}{2(J-1)} = \frac{1}{2(J-1)}\left(\sum_{i=1}^{J}(X_i-\bar{X})^2 + \sum_{i=1}^{J}(Y_i-\bar{Y})^2\right) = s_p^2$$

so the F test is given by

$$F = \frac{J}{2\sigma_p^2}(\bar{X}-\bar{Y})^2$$

and the null hypothesis is rejected if F is large. If F is large, then $\exp(-F/2)$ is small, which is the same condition for rejecting the null hypothesis using the likelihood ratio. □

7. Show that, as claimed in Theorem B of Section 12.2.1, $SS_B/\sigma^2 \sim \chi^2_{I-1}$.

   Proof. We have

$$SS_B = J\sum_{i=1}^{I}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$
$$\implies \frac{SS_B}{\sigma^2} = \sum_{i=1}^{I}\frac{(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{\sigma^2/J}$$

   We know that $\mathrm{Var}(\bar{Y}_{i\cdot}) = \sigma^2/J$ since it is a sample mean, and

$$s^2 = \frac{1}{I-1}\sum_{i=1}^{I}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$
$$\implies \frac{(I-1)s^2}{\sigma^2/J} = \sum_{i=1}^{J}\frac{(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{\sigma^2/J} = \frac{SS_B}{\sigma^2}$$

   which by Theorem B of Section 6.3 follows a χ^2_{I-1} distribution, as desired. □

11. Consider a hypothetical two-way layout with four factors (A, B, C, D) each at three levels (I, II, III). Construct a table of cell means for which there is no interaction.

    Solution. The following table shows no interactions:

	A	B	C	D
I	1	2	3	4
II	5	6	7	8
III	9	10	11	12

□

12. Consider a hypothetical two-way layout with three factors (A, B, C) each at two levels (I, II). Is it possible for there to be interactions but no main effects?

    Answer. Yes, this is possible. The means may be the same, but they could still cross over.

21. During each of four experiments on the use of carbon tetrachloride as a worm killer, ten rats were infested with larvae. Eight days later, five rats were treated with carbon tetrachloride; the other five were kept as controls. After two more days, all the rats were killed and the numbers of worms were counted. The table below gives the counts of worms for the four control groups. Significant differences, although not expected, might be attributable to changes in experimental conditions. A finding of significant differences could result in more carefully controlled experimentation and thus greater precision in later work. Use both graphical techniques and the F test to test whether there are significant differences among the four groups.

34. Conduct a two-way analysis of variance to test the effects of the two main factors and their interaction.

Chapter 14: Linear Least Squares

1. Convert the following relationships into linear relationships by making transformations and defining new variables.

   a. $y = a/(b + cx)$

   Solution. Let $z = 1/y$. Then
   $$\frac{1}{z} = \frac{a}{b + cx} \implies z = \frac{b}{a} + \frac{c}{a}x$$
   which is a linear relation. □

   b. $y = ae^{-bx}$

   Solution. Let $z = \log y$. Taking the log of both sides, we have
   $$\log y = z = \log a - bx$$
   which is a linear relation. □

   c. $y = ab^x$

   Solution. Let $z = \log y$. Taking the log of both sides, we have
   $$\log y = z = \log a + x \log b$$
   which is a linear relation. □

   d. $y = x/(a + bx)$

   Solution. Let $w = 1/x$ and $z = 1/y$. Then we have
   $$\frac{1}{y} = z = \frac{a}{x} + b = aw + b$$
   which is a linear relation. □

   e. $y = 1/(1 + e^{bx})$

   Solution. Let $z = \log\left(\frac{1}{y} - 1\right)$. Then we have
   $$\frac{1}{y} - 1 = e^{bx} \implies \log\left(\frac{1}{y} - 1\right) = z = bx$$
   which is a linear relation. □

2. Plot y versus x :

   a. Fit a line $y = a + bx$ by the method of least squares, and sketch it on the plot.

   b. Fit a line $x = c + dy$ by the method of least squares, and sketch it on the plot.

   c. Are the lines in parts (a) and (b) the same? If not, why not?

3. Suppose that $y_i = \mu + e_i$, where e_i are independent errors with mean zero and variance σ^2. Show that $\bar{y}$ is the least squares estimate of μ.

Proof. Let $\hat{y}$ be the least squares estimate of μ, which minimizes

$$\sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}(\mu + e_i - \hat{y})^2$$

Taking the derivative with respect to $\hat{y}$, we have

$$\begin{aligned}\frac{\partial}{\partial \hat{y}}\sum_{i=1}^{n}(\mu + e_i - \hat{y})^2 &= -2\sum_{i=1}^{n}(\mu + e_i - \hat{y}) = 0 \\ \implies n\mu - n\hat{y} + \sum_{i=1}^{n} e_i = 0\end{aligned}$$

Solving for $\hat{y}$ we obtain

$$\hat{y} = \mu + \frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(\mu + e_i) = \bar{y}$$

as desired. □

6. Two objects of unknown weights w_1 and w_2 are weighed on an error-prone pan balance in the following way: (1) object 1 is weighed by itself, and the measurement is 3g; (2) object 2 is weighed by itself, and the result is 3g; (3) the difference of the weights (1-2) is 1g; (4) the sum of the weights measured as 7g. The problem is to estimate the true weights of the objects from these measurements.

   a. Set up a linear model, $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$.

   Solution. We have the system

$$\begin{aligned} w_1 + e_1 &= 3 \\ w_2 + e_2 &= 3 \\ w_1 - w_2 + e_3 &= 1 \\ w_1 + w_2 + e_4 &= 7 \end{aligned}$$

   which corresponds to the equation

$$\begin{bmatrix} 3 \\ 3 \\ 1 \\ 7 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$$

□

   b. Find the least squares estimates of w_1 and w_2.

   Solution. Let $\hat{w}_1$ and $\hat{w}_2$ be the least squares estimates of w_1 and w_2, respectively. Then the sum of squares is given by

$$(3 - \hat{w}_1)^2 + (3 - \hat{w}_2)^2 + (1 - \hat{w}_1 + \hat{w}_2)^2 + (7 - \hat{w}_1 - \hat{w}_2)^2$$

   Taking the derivatives with respect to $\hat{w}_1$ and $\hat{w}_2$, we get

$$\begin{aligned} -2(3 - \hat{w}_1) - 2(1 - \hat{w}_1 + \hat{w}_2) - 2(7 - \hat{w}_1 - \hat{w}_2) = 0 &\implies \hat{w}_1 = \frac{11}{3} \\ -2(3 - \hat{w}_2) + 2(1 - \hat{w}_1 + \hat{w}_2) - 2(7 - \hat{w}_1 - \hat{w}_2) = 0 &\implies \hat{w}_2 = 3 \end{aligned}$$

   as the least squares estimates. □

c. Find the estimate of σ^2.

Solution. With the two estimates above, we have

$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix} = \begin{bmatrix} -2/3 \\ 0 \\ 1/3 \\ 1/3 \end{bmatrix}$$

We know that e_i are iid with variance σ^2, so we calculate the sample variance s^2 to estimate σ^2, which is 2/9. □

d. Find the estimated standard errors of the least square estimates of part (b).

e. Estimate $w_1 - w_2$ and its standard error.

f. Test the null hypothesis $H_0 : w_1 = w_2$.

10. Show that the least squares estimate of the slope and intercept of a line may be expressed as

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Proof. Let $y = \hat{\beta}_0 + \hat{\beta}_1 x$ be the least squares line. Thus, the sum of squares

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

is minimized. Taking the derivative with respect to $\hat{\beta}_0$, we have

$$\begin{aligned} -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1) &= -2\left(\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i\right) = 0 \\ \implies n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1\bar{x} = 0 \\ \implies \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \end{aligned}$$

as desired.

Next, taking the derivative with respect to $\hat{\beta}_1$, we have

$$\begin{aligned} -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= -2\left(\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2\right) = 0 \\ \implies \sum_{i=1}^{n} x_i y_i - (\bar{y} - \hat{\beta}_1\bar{x})\sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \\ \implies \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} + n\hat{\beta}_1\bar{x}^2 - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \\ \implies \hat{\beta}_1\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \end{aligned}$$

We have

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}x_i^2-\bar{x}\sum_{i=1}^{n}x_i+n\bar{x}^2\\ &= \sum_{i=1}^{n}x_i^2-n\bar{x}^2\\ \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{n}x_iy_i-\bar{x}\sum_{i=1}^{n}y_i-\bar{y}\sum_{i=1}^{n}x_i+n\bar{x}\bar{y}\\ &= \sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}-n\bar{x}\bar{y}+n\bar{x}\bar{y}=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}\end{aligned}$$

Thus, solving for $\hat{\beta}_1$, we have

$$\hat{\beta}_1=\frac{\displaystyle\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\displaystyle\sum_{i=1}^{n}x_i^2-n\bar{x}^2}=\frac{\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

as desired. □

11. Show that if $\bar{x}=0$, the estimated slope and intercept are uncorrelated under the assumptions of the standard statistical model.

Proof. The covariance of $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by

$$\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_1)=\frac{\displaystyle -\sigma^2\sum_{i=1}^{n}x_i}{\displaystyle n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$$

from Theorem B Section 14.2.1. If $\bar{x}=0$, then the numerator is 0, so the covariance between the slope and intercept is 0, thus they are uncorrelated, as desired. □

12. Use the result of Problem 10 to show that the line fit by the method of least squares passes through the point $(\bar{x},\bar{y})$.

Proof. The least squares line is given by $y=\hat{\beta}_0+\hat{\beta}_1x$. From Problem 10, we know that $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$, so $\bar{y}=\hat{\beta}_0+\hat{\beta}_1\bar{x}$. Thus, the pair $(\bar{x},\bar{y})$ passes through the least squares line, as desired. □

13. Suppose that a line is fit by the method of least squares to n points, that the standard statistical model holds, and that we want to estimate the line at a new point, x_0. Denoting the value on the line by μ_0, the estimate is

$$\hat{\mu}_0=\hat{\beta}_0+\hat{\beta}_1x_0$$

a. Derive an expression for the variance of $\hat{\mu}_0$.

Solution. We have

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_0) &= \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\
&= \mathrm{Var}(\hat{\beta}_0) + x_0^2 \mathrm{Var}(\hat{\beta}_1) + 2x_0 \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\
&= \frac{\sigma^2 \sum x_i^2}{n \sum x_i^2 - \left(\sum x_i\right)^2} + x_0^2 \frac{n\sigma^2}{n \sum x_i^2 - \left(\sum x_i\right)^2} + 2x_0 \frac{-\sigma^2 \sum x_i}{n \sum x_i^2 - \left(\sum x_i\right)^2} \\
&= \frac{\sigma^2}{n \sum x_i^2 - \left(\sum x_i\right)^2} \left(\sum x_i^2 + n x_0^2 - 2x_0 \sum x_i \right) \\
&= \frac{\sigma^2}{n \sum (x_i - \bar{x})^2} \sum \left(x_i^2 + x_0^2 - 2x_0 x_i \right) \\
&= \frac{\sigma^2}{n} \cdot \frac{\sum (x_i - x_0)^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma^2}{n} \frac{\sum (x_i - \bar{x} + \bar{x} - x_0)^2}{\sum (x_i - \bar{x})^2} \\
&= \frac{\sigma^2}{n} \cdot \frac{\sum (x_i - \bar{x})^2 + n(\bar{x} - x_0)^2 + 2(x_0 - \bar{x}) \sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \\
&= \frac{\sigma^2}{n} \left(1 + \frac{n(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)
\end{aligned}$$

□

b. Sketch the SD of $\hat{\mu}_0$ as a function of $x_0 - \bar{x}$. The slope of the curve should be intuitively plausible.

Answer. As you can see above, my expression for the SD is a function of $x_0 - \bar{x}$. I'm not really sure how to go about sketching this without any sort of numbers.

c. Derive a 95% confidence interval for $\mu_0 = \beta_0 + \beta_1 x_0$ under an assumption of normality.

Solution. Under an assumption of normality, it holds that $\hat{\mu}_0$ is normally distributed with mean $\beta_0 + \beta_1 x_0$ and variance as found in part a. Thus, the 95% confidence interval is given by

$$(\beta_0 + \beta_1 x_0) \pm \sigma \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} z_{5/2}$$

□

14. Problem 13 dealt with how to form a CI for the value of a line of at a point x_0. Suppose that instead we want to predict the value of a new observation, Y_0, at x_0,

$$Y_0 = \beta_0 + \beta_1 x_0 + e_0$$

by the estimate

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

a. Find an expression for the variance of $\hat{Y}_0 - Y_0$, and compare it to the expression for the variance of $\hat{\mu}_0$ obtained in part (a) of Problem 13. Assume that e_0 is independent of the original observations and has the variance σ^2.

Solution. We have

$$\begin{aligned}
\mathrm{Var}(\hat{Y}_0 - Y_0) &= \mathrm{Var}(\hat{Y}_0) + \mathrm{Var}(Y_0) - 2\mathrm{Cov}(\hat{Y}_0, Y_0) \\
&= \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) + \sigma^2 \\
&= \sigma^2 \left(\frac{n+1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)
\end{aligned}$$

since β_0, β_1, x_0 are all constants, and e_0 is independent from all of the original observations. □

b. Assuming that e_0 is normally distributed, find the distribution of $\hat{Y}_0 - Y_0$. Use this result to find an interval I such that $P(Y_0 \in I) = 1 - \alpha$. This interval is called a $100(1-\alpha)\%$ prediction interval.

40. The following data come from the calibration of a proving ring, a device for measuring force.

    a. Plot load versus deflection. Does the plot look linear?

    b. Fit deflection as a linear function of load, and plot the residuals versus load. Do the residuals show any systematic lack of fit?

    c. Fit deflection as a quadratic function of load, and estimate the coefficients and their standard errors. Plot the residuals. Does the fit look reasonable?