کاربرد برآوردگرهای مؤلفه‌های واریانس در به‌نژادی گیاهان (مقاله مروری)

امیدعلی اکبرپور*

استادیار، گروه زراعت و اصلاح نباتات، دانشکده کشاورزی، دانشگاه لرستان، خرم‌آباد

(تاریخ دریافت: ١٣٩٥/١٠/١٩– تاریخ پذیرش: ١٣٩٦/٠٢/٢٧)

چکیده

برای اجرای هر برنامه به‌نژادی آگاهی از ساختار ژنتیکی صفت مورد بررسی، میزان تأثیر عوامل محیطی و اثر متقابل عوامل ژنتیکی و محیطی و همچنین اطلاع از تأثیر ثابت و تصادفی بودن فاکتورها بر تحلیل نتایج یک امر ضروری است. به طبع آن تجزیه و تحلیل مولفه‌های واریانس از اهمیت زیادی در به‌نژادی گیاه و دام برخوردار است. برای برآورد مولفه‌های واریانس از برآوردگرهای زیادی استفاده می‌شود که ANOVA یکی از مهمترین آنها است. این برآوردگر در برخی موقعیت ها که داده‌ها نامتعادل هستند و مولفه‌های واریانس منفی برآورد می‌شوند، ناکارآمدتر از برآوردگرهای حداکثر درستنمایی (Maximum Likelihood; ML) و حداکثر درستنمای محدود شده (Restricted Maximum Likelihood; REML) هستند. لذا هدف از این تحقیق بررسی مروری مدل‌های خطی مختلط و مقایسه برآورد مولفه‌های واریانس به روش‌های ANOVA، ML وREML با استفاده از داده‌های آزمایشی است.

واژگان کلیدی: برآورد، تجزیه واریانس، حداکثر درستنمایی، حداکثر درستنمایی محدود شده، مؤلفه‌های واریانس، رگرسیون

* نویسنده مسئول، آدرس پست الکترونیکی: akbarpour.oa@lu.ac.ir

مقدمه

ژنتیک صفات کمی، مبنای عملی و اصلی برای به‌نژادی گیاهان در ۱۰۰ سال اخیر بوده است. فیشر (Fisher, 1925) با ارائه مؤلفه‌های واریانس بنیان جدیدی برای ژنتیک کمی و به‌نژادی گیاهان پایه‌گذاری کرد. اکثر صفاتی که برای به‌نژادگران گیاه و دام که اهمیت اقتصادی دارند صفات قابل اندازه‌گیری (متریک) هستند و بیشتر تغییرات مربوط به تکامل ذره‌ای، تغییرات متریک هستند. تعداد صفات کمی قابل مطالعه در موجودات عالی بسیار زیاد است. اصولا هر صفتی که پیوسته و قابل اندازه‌گیری باشد به‌عنوان صفت کمی قابل مطالعه است (Falconer and Mackay, 1996).

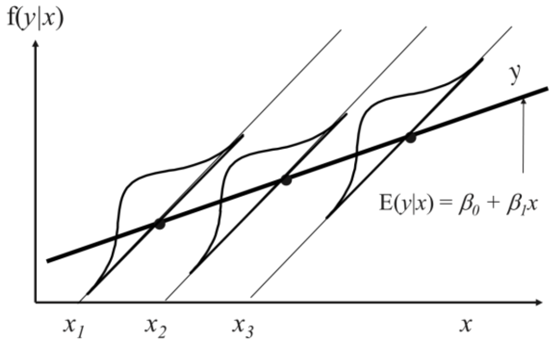
به‌نژادگران گیاه نیازمند یادگیری اصول رگرسیون و تجزیه واریانس برای اجرای مناسب آزمایش‌های مزرعه‌ای و آزمایشگاهی هستند تا بتوانند پارامترهایی نظیر وراثت‌پذیری، ترکیب‌پذیری‌های عمومی و خصوصی، همبستگی‌های ژنتیکی و سود ژنتیکی را به نحو شایسته محاسبه و تحلیل کنند. بنابراین آمار خطی و غیرخطی برای به‌نژادی گیاهان لازم و ضروری است (Acquaah, 2009).

در تحقیقات از برآوردگرهای مختلفی در محاسبه پارامترهای ژنتیکی دام و گیاه استفاده می‌شود؛ اما در تحقیقات داخلی مطالعات محدودی انجام شده است Akbarpour *et al.*, 2015a; Akbarpour *et al.*, 2015b) Ismaili *et al.*, 2016) و برای برآورد مولفه‌های واریانس، وراثت‌پذیری و همبستگی ژنتیکی به ندرت از برآوردگرهایی نظیر ML و REML به ویژه در بخش گیاه استفاده شده است. یکی از دلایل عدم استفاده از برآوردگرهای متعدد و جدید در محاسبه پارامترهای ژنتیکی در به‌نژادی گیاهان عدم اطلاع کافی محققین کشور نسبت به این برآوردگرها و یا عدم آگاهی از قدرت برخی از این برآوردگرها در تخمین درست و دقیق‌تر پارامترهای ژنتیکی است. لذا هدف از این مرور، تشریح و بسط برآوردگرهای آماری پرکاربرد در به‌نژادی گیاهان و محاسبه پارامترهای ژنتیکی با استفاده از آنها می‌باشد. در تشریح این روش‌ها فرض بر آن است که خواننده از جبر ماتریس‌ها اطلاع کافی دارد.

رگرسیون: در این قسمت مروری بر تجزیه داده‌های با توزیع نرمال توسط مدل‌های رگرسیون ارائه می‌شود. به طور کلی مدل خطی در رگرسیون به روش زیر ارائه می‌گردد:

$$y = X\beta + \varepsilon \qquad \text{رابطه (۱)}$$

که y بردار ارزش مشاهدات وابسته؛ X ماتریس طرح که مشاهدات متغیرهای مستقل را به عوامل ثابت یا تصادفی در بردار بتا مرتبط می‌کند؛ β بردار پارامترهایی که تخمین زده می‌شوند و بردار ε باقیمانده‌ها یا انحراف از مدل برازش شده می‌باشد.



شکل ۱- برازش رگرسیون خطی روی نقاط دارای توزیع نرمال

Figure 1. Linear regression fitting on the points with normal distribution

برآورد β به ساختار واریانس-کواریانس (ماتریس کواریانس) بردار باقیمانده‌ها یعنی ε بستگی دارد. واریانس y به صورت زیر محاسبه می‌گردد.

$$\sigma_y^2 = \begin{bmatrix} y_1 - E(y_1) \\ y_2 - E(y_2) \\ \vdots \\ y_n - E(y_n) \end{bmatrix} \begin{bmatrix} y_1 - E(y_1) \\ y_2 - E(y_2) \\ \vdots \\ y_n - E(y_n) \end{bmatrix}' \qquad \text{رابطه (۲)}$$

$$= \begin{bmatrix} \sigma^2(y_1) & \sigma(y_1, y_2) & .. & \sigma(y_1, y_n) \\ \sigma(y_2, y_1) & \sigma^2(y_2) & .. & \sigma(y_2, y_n) \\ \vdots & \vdots & .. & \vdots \\ \sigma(y_n, y_1) & \sigma(y_n, y_2) & .. & \sigma^2(y_n) \end{bmatrix}$$

واریانس خطا $\sigma^2(\varepsilon)=\sigma^2(e)=\sigma^2 I=\Sigma$ نیز به صورت زیر ارائه می‌شود. اگر n= 3 در نظر گرفته شود، پس

رابطه (٣) $$\sigma^2(\varepsilon)I=\Sigma=\sigma^2(e)\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}=\begin{bmatrix}\sigma^2&0&0\\0&\sigma^2&0\\0&0&\sigma^2\end{bmatrix}$$

که I ماتریس واحد است. چنانچه معادله از طریق حداقل مربعات حل شود (OLS[1]). باقیمانده‌ها دارای توزیع نرمال با میانگین صفر و واریانس $(0,\ \sigma^2 I)$ می‌باشند، یعنی باقیمانده‌ها همگن و مستقل می‌باشند (Neter *et al.*, 2004). چنانچه باقیمانده‌ها همبسته یا ناهمگن باشند، خطا دارای توزیع نرمال چند متغیره با میانگین صفر و واریانس–کوواریانس V می‌باشد: e ~ MVN(0, V). بردار β که بیانگر اثرات ثابت است با استفاده از معادله (٢) برآورد می‌شود (Lynch and Walsh, 1998).

روش حداقل مربعات بدین شرح است که خطای مدل حداقل می‌شود، لذا پارامترهای برآوردی نیز حداقل خطا را خواهند داشت (Graybill and Iyer, 1994; Neter *et al.*, 2004).

$$\sum e_i^2=Q=\sum[y_i-(\beta_0+\beta X_i)]^2$$

که روش ماتریسی آن به شکل زیر می‌باشد.

رابطه(٤) $$Q=(y-X\beta)'(y-X\beta)$$

که $(y-X\beta)'$ برگردان ماتریس $(y-X\beta)$ می‌باشد. با بسط فوق به شرح زیر،

$$Q=y'y-\beta'X'y-y'X\beta+\beta'X'X\beta$$

چون $(X\beta)'=\beta'X'$ است و $y'X\beta$ یک اسکالر است و با برگردان خود، یعنی با $\beta'X'Y$ برابر است، لذا،

رابطه (٥) $$Q=y'y-2\beta'X'y+\beta'X'X\beta$$

با مشتق گیری معادله نسبت به β و برابر با صفر قرار دادن آن می‌توان بردار β را برآورد نمود.

رابطه (٦) $$\frac{\partial Q}{\partial\beta}=-2X'y+2X'X\beta$$

در نتیجه بردار β با معادله زیر برآورد می‌گردد.

رابطه (٧) $$\beta=(X'X)^{-1}X'y$$

چنانچه ستون‌های ماتریس $X'X$ به صورت خطی به همدیگر وابسته باشند، یعنی مرتبه ماتریس[2] کمتر از رتبه آن باشد، یک معکوس یونیک برای آن وجود ندارد و متعاقبا برآوردهای منحصر به فردی نیز برای ضرائب در بردار b وجود نخواهد داشت. ماتریس با رتبه کامل[3] به معنای استقلال خطی سطرها و ستون‌های ماتریس از همدیگر است و هر تعداد از سطر و ستون‌های ماتریس که با یکدیگر رابطه خطی داشته باشند رتبه ماتریس کاهش می‌یابد (Graybill and Iyer, 1994). معادله (١) را می‌توان به صورت زیر بیان کرد.

رابطه (٨) $$y=X\beta\Leftrightarrow y=X(X'X)^{-1}X'y\Leftrightarrow y=Hy$$

که $H=X(X'X)^{-1}X'$ یک ماتریس خودتوان است و HH=H است. لذا،

رابطه (٩) $$e=y-y=y-Hy\Leftrightarrow y(I-H)$$

در مواقعی که یک عدد ثابت (A) در یک ماتریس متغیر (Y) ضرب می‌شود، روابط زیر حاکم است.

رابطه (١٠) $$W=AY \qquad E(A)=A$$

رابطه (١١) $$E(W)=E(AY)=AE(W)$$

رابطه (١٢) $$\sigma^2(W)=\sigma^2(AY)=A\sigma^2(Y)A'$$

برای تخمین ماتریس واریانس کواریانس خطای مدل که عبارت است از: e = (I-H)y؛ از معادله (١٢) استفاده می‌شود.

رابطه (١٣) $$\sigma^2(e)=(1-H)\sigma^2(Y)(1-H)'$$

چون $\sigma^2(Y)=\sigma^2(\varepsilon)=\sigma^2(I)$ است و $(I-H)=(I-H)'$ می‌باشد. همچنین $(I-H)(I-H)=(I-H)$ می‌باشد. لذا معادله (١٣) به شکل نهایی زیر نوشته می‌شود،

رابطه (١٤) $$\sigma^2(e)=\sigma^2(I-H)I(I-H)=\sigma^2(I-H)$$

تجزیه واریانس رگرسیون به دو شکل خطی و جبر ماتریس محاسبه می‌شود که جبر ماتریس نیز به دو روش کوادراتیک[4] و غیرکوادراتیک[5] ارائه می‌شود. مجموع مربعات به صورت کوادراتیک $(y'Ay)$ نوشته می‌شوند

1- Ordinary Least Squares
2- Rank
3- Full Rank

4- Quadratic
5- Non-Quadratic

که برای هر منبع تغییر میزان A متفاوت است (Graybill and Iyer, 1994; Neter *et al.*, 2004). به عنوان مثال مجموع مربعات کل به صورت زیر محاسبه می‌شود که برای این منبع تغییر $A = I - \frac{1}{n}J_n$ است.

$$SST = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$
$$= y'y - \frac{1}{n}y'Jy = y'(I - \frac{1}{n}J_n)y \qquad \text{رابطه (۱۵)}$$

که $J = 11'$؛ یعنی یک بردار یکان با ابعاد بردار y که در برگردان خود ضرب می‌شود. مجموع مربعات خطا نیز به شکل زیر است.

$$SSE = e'e = (y - Xb)'(y - Xb)$$
$$= y'y - 2b'X'y + b'X'Xb$$
$$= y'y - 2b'X'y + b'X'X(X'X)^{-1}X'$$
$$y = y'y - 2b'X'y + b'IX'y = y'y - b'X'y$$

با توجه به معادله (۸) $b'X' = (Xb)' = y' = (Hy)'$، $H' = H$ و لذا $b'X' = y'H$ خواهد بود و در نتیجه مجموع مربعات خطا به صورت زیر محاسبه می‌شود.

$$y'y - b'X'y = y'y - y'Hy = y'(I-H)y \qquad \text{رابطه (۱۶)}$$

و در نهابت برای محاسبه مجموع مربعات رگرسیون نیز از روابط زیر استفاده می‌شود.

$$SSR = b'X'y - y'\left(\frac{1}{n}J_n\right)y = y'Hy$$
$$- y'\left(\frac{1}{n}J_n\right)y = y'(H - \frac{1}{n}J_n)y \qquad \text{رابطه (۱۷)}$$

برای آزمون معنی‌داری ضرائب رگرسیون نیاز به خطای معیار هر یک از ضرائب می‌باشد، لذا برای محاسبه واریانس بردار b از معادله (۱۲) استفاده می‌شود. یعنی $b = (X'X)^{-1}X'y = Ay$ که $A = (X'X)^{-1}X'$ است و $A' = X(X'X)^{-1}$ بنابراین،

$$\sigma^2(b) = AyA' = (X'X)^{-1}X'\,\sigma^2(y)X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}X'\,(I)X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}X'X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}I = \sigma^2(X'X)^{-1} \qquad \text{رابطه (۱۸)}$$

رگرسیون به روش تعمیم یافته: کمترین مقدار مربعات تعمیم یافته[1] نیز با فرضیات زیر $e \sim MVN(0, V)$ به صورت رابطه (۳) ارائه می‌شود (Lynch and Walsh, 1998). اگر بخش تصادفی مدل، فقط باقیمانده باشد، $e \sim MVN(0, \sigma_e^2 R)$ است.

$$\beta = (X'V^{-1}X)^{-1}X'V^{-1}y \qquad \text{رابطه (۱۹)}$$

یک ساختار عمومی برای محاسبه بردار بتا وجود دارد که از طریق محاسبه معکوس تعمیم یافته[2] است، که این روش در زمانی که دترمینان ماتریس صفر می‌شود و ماتریس فاقد معکوس است، به شکل زیر ارائه می‌شود (Sahai and Ojeda, 2004).

$$\beta = (X'V^{-1}X)^{-}X'V^{-1}y \qquad \text{رابطه (۲۰)}$$

که $(X'V^{-1}X)^{-}$ معکوس تعمیم یافته $(X'V^{-1}X)$ می‌باشد. قابل ذکر است که حل معادلات به روش کمترین مقدار مربعات تعمیم یافته از طریق معکوس ماتریس، الزاما به روش معکوس تعمیم یافته نیست. استفاده از روش معکوس تعمیم یافته صرفا در زمان‌هایی است که دترمینان ماتریس صفر است. در رگرسیون به روش کمترین مقدار مربعات تعمیم یافته، محاسبه مجموع مربعات به شکل کوادراتیک فقط با جایگرینی y با $R^{-\frac{1}{2}}y$ و X با $R^{-\frac{1}{2}}X$ انجام می‌گردد (Henderson, 1984; Lynch and Walsh, 1998).

$$SST = y'R^{-\frac{1}{2}}(I - \frac{1}{n}J_n)R^{-\frac{1}{2}}y = y'(R^{-1} - \frac{1}{n}R^{-\frac{1}{2}}J_nR^{-\frac{1}{2}})y \qquad \text{رابطه (۲۱)}$$

$$SSE = e'R^{-1}e = y'\left[R^{-1} - R^{-1}X\left(X'R^{-1}X\right)^{-1}X'R^{-1}\right]y \qquad \text{رابطه (۲۲)}$$

$$SSR = y'\left[R^{-1} - R^{-1}X\left(X'R^{-1}X\right)^{-1}X'R^{-1} - \frac{1}{n}R^{-\frac{1}{2}}J_nR^{-\frac{1}{2}}\right]y \qquad \text{رابطه (۲۳)}$$

اثرات ثابت و تصادفی: یک اثر ثابت تکرارپذیر، قابل تعمیم به همان سطوح مطالعه شده در تحقیق است. به عنوان مثال وقتی عملکرد سه رقم سویا با همدیگر مقایسه می‌شود، رقم یک اثر ثابت است، زیرا علاوه بر خصوصیات بیان شده، هر رقم اثر یکسانی بر کلیه مشاهدات مربوط به خود دارد. مدلی که تمامی متغیرهای مستقل آن ثابت باشد، مدل ثابت نام دارد. اگر همه متغیرهای مستقل یک مدل تصادفی باشند، مدل را تصادفی می‌گویند. به عنوان مثال، انتخاب چند لاین گندم از بین لاین‌های نسل F_2 چنانچه

1- Generalized Least Squares

2- Generalized Inverse

در یک مدل که هم فاکتور ثابت و هم فاکتور تصادفی باشد آن مدل را مختلط[1] می‌نامند (Littell *et al.*, 2006).

مدل خطی یک طرفه[2]: مدل خطی تجزیه واریانس یک‌طرفه به صورت زیر می‌باشد.

رابطه (۲٤)
$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

که y_{ij}، مشاهده jام از تیمار iام؛ μ، میانگین کل جمعیت‌ها؛ α_i، اثر ثابت تیمار iام که $\sum_{i=1}^{k}\alpha_i = 0$

ε_{ij}، خطای تصادفی با توزیع نرمال، امیدریاضی صفر $E(\varepsilon_{ij}) = 0$ و واریانس $V(\varepsilon_{ij}) = \sigma^2$ است. با در نظر گرفتن اثر تیمار به صورت تصادفی امید ریاضی y_{ij} به صورت زیر تعریف می‌شود.

رابطه (۲۵)
$$E(y_{ij}) = E(\mu + \alpha_i + \varepsilon_{ij}) = E(\mu) + E(\alpha_i) + E(\varepsilon_{ij}) = E(\mu)$$

در مدل تصادفی $E(\alpha_i) = 0$ دارای واریانس $\sigma^2_{\alpha_i}$ و $E(\varepsilon_{ij}) = 0$ دارای واریانس $\sigma^2_{\varepsilon_{ij}}$ است. بنابراین

رابطه (۲٦)
$$\sigma^2_y = \sigma^2_{\alpha_i} + \sigma^2_{\varepsilon_{ij}}$$

که $\sigma^2_{\alpha_i}$ و $\sigma^2_{\varepsilon_{ij}}$ مولفه‌های واریانس نامیده می‌شوند. در مدل ثابت $E(y_{ij}) = \mu + \alpha_i$ می‌باشد اما در مدل تصادفی $E(y_{ij}) = \mu$ است و امید ریاضی α_iها صفر است.

فرض می‌شود که در شهرستان خرم‌آباد بیش از ۱۰۰ گاوداری وجود دارد که امکان مطالعه همه آنها وجود ندارد. برای یک تحقیق اگر تعداد ۱۰ گاوداری به تصادف از بین این ۱۰۰ گاوداری انتخاب شود و در هر گاوداری میزان شیر تولیدی برای ٦ راس گاو اندازه‌گیری شود. در این حال، فاکتور گاوداری می‌تواند به صورت تصادفی انتخاب شود. فاکتوری که تصادفی انتخاب شود دارای چند ویژگی است اول اینکه آن فاکتور دارای توزیع احتمال می‌باشد، دوم اینکه نتایج آن قابل تعمیم به سایر گاوداری‌ها می‌باشد و سوم اینکه اثر گاوداری قابل برآورد نیست بلکه پیش‌بینی می‌شود. اگرچه α_i مشابه زمانی که به عنوان اثر ثابت در نظر گرفته شود قابل برآورد نیست، اما اطلاعات اندکی درباره آن وجود دارد. به عنوان مثال $\bar{y}_{i.}$ میانگین تولید شیر گاوداری iام که دارای n_i مشاهده است. سوالی که در اینجا وجود دارد این است که چگونه بر اساس $\bar{y}_{i.}$ یک مقدار عددی به α_i به عنوان اثر تصادفی اختصاص دهیم؟ اگر مقدار اختصاص داده شده به α_i، α_i نامیده شود، آن‌را مقدار برآورد شده نمی‌گویند بلکه آن‌را مقدار پیش‌بینی شده می‌نامند، زیرا مقدار برآورد برای پارامتر است و α_i در مدل تصادفی پارامتر نیست. چون $\alpha_i = E(\alpha_i) = 0$ است و داده آن قابل استفاده نیست، لذا باید از میانگین شرطی $E(\alpha_i|\bar{y}_i)$ استفاده شود. سِرل و همکاران (Searle *et al.*, 2006) بیان داشتند که α_i و $\bar{y}_i$ به همدیگر وابسته‌اند و دارای تابع توزیع نرمال مشترک دو متغیره با میانگین و واریانس زیر هستند.

رابطه (۲۷)
$$E\begin{bmatrix}\alpha_i \\ \bar{y}_{i.}\end{bmatrix} = \begin{bmatrix}0 \\ \mu\end{bmatrix}, \quad Var\begin{bmatrix}\alpha_i \\ \bar{y}_{i.}\end{bmatrix} = \begin{bmatrix}\sigma^2_a & \sigma^2_a \\ \sigma^2_a & \sigma^2_a + \sigma^2_e / n_i\end{bmatrix}$$

با توجه به ویژگی شناخته شده تابع توزیع نرمال مشترک دو متغیر که از رابطه (۲۷) عمل می‌کند، می‌توان $E(\alpha_i|\bar{y}_i)$ را بسط داد (Bertsekas and Tsitsiklis, 2008).

رابطه (۲۸)
$$E(y \mid X) = \alpha + \beta X$$

از طرفی $E(y) = \alpha + \beta E(X)$ و $\alpha = E(y) - \beta E(X)$ است، در نتیجه،

رابطه (۲۹)
$$E(y \mid X) = E(y) + \beta[X - E(X)] \Leftrightarrow E(y) + \frac{Cov(X,y)}{Var(X)}[X - E(X)]$$

با توجه به رابطه (۲۷) و (۲۹) برای مقادیر α_i و $\bar{y}_i$ نیز رابطه زیر برقرار است.

رابطه (۳۰)
$$E(\alpha_i \mid \bar{y}_{i.}) = E(\alpha_i) + \frac{\sigma^2_\alpha}{\sigma^2_a + \sigma^2_e / n_i}(\bar{y}_{i.} - \mu) \Leftrightarrow \frac{\sigma^2_\alpha}{\sigma^2_a + \sigma^2_e / n_i}(\bar{y}_{i.} - \mu)$$

رابطه (۳۱)
$$\alpha_i = \frac{\sigma^2_\alpha}{\sigma^2_a + \sigma^2_e / n_i}(\bar{y}_{i.} - \mu)$$

رابطه (۳۰) و (۳۱) را بهترین پیش‌بینی نااریب خطی[3] (BLUP) از اثر تصادفی می‌نامند. همان‌طور که در این رابطه مشاهده می‌شود اثر تیمار در حالت ثابت از رابطه $\bar{y}_{i.} - \mu$ محاسبه می‌شود که یک مقدار ثابت است و آن را بهترین

1- Mixed Model
2- One-Way ANOVA
3- Best Linear Unbiased Prediction

برآورد نااریب خطی[1] (BLUE) می‌نامند. در روش BLUP یک ضریب از نسبت واریانس‌ها در BLUE پیش ضرب شده و آن را تعدیل می‌کند. اگر مدل خطی مناسب باشد و MSA از MSE بیشتر باشد، مقدار رابطه (۳۱) به سمت انقباضی شدن[2] میل می‌کند، یعنی اگر اثر تیمار در حالت ثابت مثبت باشد، در مدل تصادفی، میزان پیش‌بینی کمتر می‌شود و اگر در حالت ثابت اثر تیمار منفی باشد در مدل تصادفی میزان پیش‌بینی بیشتر از مقدار ثابت می‌شود. پیش‌بینی به روش BLUP عدم قطعیت‌های ناشی از توزیع احتمالی را تعدیل می‌نماید (Yang, 2010).

مدل خطی تجزیه واریانس خطی از فرضیات (۲۴) و (۲۵) تبعیت می‌کند.

رابطه (۳۲)
$$E(SSA)=E\left[n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})\right]=n\sum_{i=1}^{a}E[(\alpha_i-\bar{\alpha}_.)+(\bar{e}_{i.}-\bar{e}_{..})]^2$$
$$=n\sum_{i=1}^{a}E(\alpha_i-\bar{\alpha}_.)^2+(\bar{e}_{i.}-\bar{e}_{..})^2$$

مقدار $2(\alpha_i-\bar{\alpha})(\bar{e}_{i.}-\bar{e}_{..})^2=0$ می‌شود. بر اساس قانون امیدریاضی $\sigma^2=Ey^2-(E(y))^2$، و $E(\alpha_i)=E(e_{ij})=0$،

رابطه (۳۳)
$$E(SSA)=n\sum_{i=1}^{a}[var(\alpha_i-\bar{\alpha}_.)+var(\bar{e}_{i.}-\bar{e}_{..})]$$
$$=n\sum_{i=1}^{a}\left(\sigma_a^2+\frac{\sigma_a^2}{a}-\frac{2\sigma_a^2}{a}\right)+n\sum_{i=1}^{a}\left(\frac{\sigma_e^2}{n}+\frac{\sigma_e^2}{an}-\frac{2n\sigma_e^2}{nan}\right)$$
$$=n(a-1)\sigma_a^2+(a-1)\sigma_e^2=(a-1)(n\sigma_a^2+\sigma_e^2)$$

رابطه (۳۴)
$$E(MSA)=\frac{E(SSA)}{a-1}=n\sigma_a^2+\sigma_e^2$$

رابطه (۳۵)
$$E(MSE)=\frac{E(SSE)}{a(n-1)}=\frac{a(n-1)\sigma_e^2}{a(n-1)}=\sigma_e^2$$

فاکتور A دارای توزیع کای اسکوئر $SSA\sim\chi_{a-1}^2(\sigma_a^2+n\sigma_e^2)$ و خطا نیز دارای توزیع کای اسکوئر با درجه آزادی a(n-1) است ($SSA\sim\chi_{a(n-1)}^2(\sigma_e^2)$) (Sahai and Miguel, 2004).

تجزیه واریانس مدل خطی به روش ماتریس: تجزیه واریانس مدل خطی با روش ماتریس به صورت معادله (۱) نوشته می‌شود. در مدل ماتریسی $E(y)=\mu$ یک بردار از میانگین کل است، Var(y) = V است که به صورت y ~ N(μ , V) نوشته می‌شود. مدل ماتریسی محاسبه مجموع مربعات تجزیه واریانس به فرم کوادراتیک زیر است.

رابطه (۳۶)
$$SS=y'Qy$$

که Q بسته به نوع متعادل بودن و نامتعادل بودن آزمایش و منبع تنوع واریانس متفاوت می‌باشد. در یک آزمایش با داده‌های متعادل، مجموع مربعات کل، خطا و تیمار به ترتیب، مشابه روابط (۱۵)، (۱۶) و (۱۷) محاسبه می‌شوند، در صورتی که ماتریس $X'X$ دارای دترمینان صفر باشد. برای محاسبه معکوس ماتریس، از روش تعمیم‌یافته استفاده می‌گردد.

رابطه (۳۷)
$$SST=y'(I_n-\frac{1}{n}J_n)y$$

رابطه (۳۸)
$$SSE=y'\left(I-X(X'X)^{-}X'\right)y$$

رابطه (۳۹)
$$SSA=y'\left(X(X'X)^{-}X'-\frac{1}{n}J_n\right)y$$

با استفاده از قانون محاسبه اثر ماتریس[3] یا مجموع عناصر روی قطر ماتریس مربع، می‌توان نشان داد که $Y'QY=tr(y'Qy)=tr(Qy'y)$ بنابراین،

رابطه (۴۰)
$$E(y'Qy)=E[tr(y'Qy)]=E[tr(Qyy')]$$
$$=tr[E(Qyy')]=tr[Q(yy')]$$

که tr به معنی اثر ماتریس است. چون $E(y)=\mu$ و $Var(y)=V$ هست؛ لذا،

رابطه (۴۱)
$$E(yy')=V+\mu\mu'$$

با جایگزینی رابطه (۴۱) با (۴۰) ،

رابطه (۴۲)
$$E(y'Qy)=tr[Q(V+\mu\mu')]=tr[(QV)+(Q\mu\mu')]=tr(QV)+\mu'Q\mu$$

همانطور که قبلا بیان شد Q یک ماتریس خودتوان است.

در مدل ثابت تجزیه واریانس یک طرفه $E(y)=Xb$ و $Var(y)=\sigma_e^2I_n$ است. که I_n یک ماتریس مربع n بعدی به تعداد مشاهدات است، بنابراین رابطه (۴۲) به صورت زیر تعریف می‌شود.

رابطه (۴۳)
$$E(y'Qy)=b'X'QXb+\sigma_e^2tr(Q)$$

1- Best Linear Unbiased Estimator
2- Shrinkage
3- Trace

در صورت متعادل بودن طرح آزمایشی $Q = X(X'X)^{-1}X'$ معادله (٤٣) به صورت زیر نوشته می‌شود.

رابطه (٤٤)
$$E(y'Qy) = b'X'X(X'X)^{-1}X'Xb + \sigma_e^2 tr(X(X'X)^{-1}X') = b'X'Xb + \sigma_e^2 tr(X(X'X)^{-1}X')$$

در صورت عدم تعادل $Q = X(X'X)^{-}X'$ از معکوس تعمیم یافته استفاده می‌شود،

رابطه (٤٥)
$$E(y'Qy) = b'X'X(X'X)^{-}X'Xb + \sigma_e^2 tr(X(X'X)^{-}X') = b'X'Xb + \sigma_e^2 tr(X(X'X)^{-}X')$$

با توجه به اثبات (Searle (1971،

رابطه (٤٦)
$$X(X'X)^{-}X'X = X$$

رابطه (٤٧)
$$tr\left[X(X'X)^{-}X'\right] = tr\left[(X'X)^{-}XX'\right] = Rank\left[(X'X)^{-}XX'\right] = Rank(X)$$

در نتیجه،

رابطه (٤٨)
$$E(SSA) = E(y'Qy) = b'X'Xb + \sigma_e^2 rank(X)$$

اگر $Q = I_n$ باشد، آنگاه $E(y'Qy) = E(y'y)$، در نتیجه برای مجموع مربعات کل،

رابطه (٤٩)
$$E[y'y - SSA] = [N - Rank(X)]\sigma_e^2$$

(Sahai and Ojeda, 2004).

مدل تصادفی طرح کاملا تصادفی با روش ماتریس:

یادآوری می‌شود چنانچه تمامی اجزاء خطی یک مدل (به جز میانگین کل) تصادفی باشند، مدل را تصادفی می‌نامند. در طرح کاملا تصادفی به دلیل وجود یک منع تغییر و با فرض تصادفی گرفتن آن، کل مدل خطی تصادفی می‌شود. مدل تصادفی طرح کاملا تصادفی به شرح زیر است.

رابطه (٥٠)
$$y = Xb + Zu + e$$

که y بردار مشاهدات (n×1) است. b بردار اثرات ثابت (p×1) است. u بردار (q×1) اثرات تصادفی است که q برابر با تعداد سطوح برای اثرات تصادفی است. e بردار (n×1) بردار تصادفی اثرات باقیمانده است. X ماتریس طرح که استثناً در طرح کاملا تصادفی بردار یک می‌باشد (زیرا در این طرح اگر تیمار تصادفی باشد، X فقط بردار مربوط به میانگین را در خود دارد). Z ماتریس طرح (n×q) که مشاهدات را به اثرات تصادفی مرتبط می‌کند. متغیر $y \sim N(XB, V)$ است. در طرح کاملا تصادفی با در نظر گرفتن تیمار به عنوان فاکتور تصادفی بخش ثابت مدل، میانگین کل است و $Xb = \mu$ است (Henderson, 1984).

رابطه (٥١)
$$E(y) = E[Xb + Zu + e] \Leftrightarrow E(Xb) + E(Zu) + E(e) = Xb = \mu$$

رابطه (٥٢)
$$\begin{bmatrix} u \\ e \end{bmatrix} \sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} = \begin{pmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_e^2 \end{pmatrix}\right]$$

رابطه (٥٣)
$$Var(y) = V = Var(Zu + e) \Leftrightarrow Z\,Var(u)Z' + Var(e) + Cov(Zu, e) + Cov(e, Zu) \Rightarrow ZGZ' + R + Zcov(u, e) + Cov(e, u)Z' \Leftrightarrow Cov(u, e) = Cov(e, u) = 0 \Leftrightarrow \mathbf{V} = \mathbf{ZGZ'} + \mathbf{R}$$

امید ریاضی مجموع مربعات در طرح کاملا تصادفی برای فاکتور تیمار به صورت زیر تعریف می‌شود:

رابطه (٥٤)
$$E(SSA) = E(y'Qy) = tr(QV) + \mu'Q\mu = tr[Q(ZGZ' + R)] + \mu'Q\mu = tr[QZZ']\sigma_a^2 + tr[Q]\sigma_e^2 = (a-1)[n\sigma_a^2 + \sigma_e^2]$$

که $Q = (Z(Z'Z)^{-1}Z') - \frac{1}{N}J_n$ با توجه به رابطه (١٧) قسمت دوم معادله $\mu'Q\mu$ صفر می‌شود. امید ریاضی مجموع مربعات خطا نیز به شرح زیر می‌باشد:

رابطه (٥٥)
$$E(SSE) = E(y'Qy) = tr[QZZ']\sigma_a^2 + tr[Q]\sigma_e^2 + \mu'Q\mu = 0 + a(n-1)\sigma_e^2 + 0$$

که در اینجا $Q = I_n - (Z(Z'Z)^{-1}Z')$، با جایگذاری جمله اول و سوم صفر می‌گردد (Searle, 1971).

فرض کنید، σ^2 بردار مولفه‌های واریانس باشد که باید برآورد شوند، و s بردار مجموع مربعات منابع تغییر مدل باشد. امیدریاضی مجموع مربعات، با معادله خطی از مولفه‌های واریانس برابر است. یعنی E(s) یک بردار از معادلات خطی است، که آن را با $C\sigma^2$ نشان می‌دهند (Searle *et al.*, 2006).

رابطه (٥٦)
$$E(s) = C\sigma^2$$

برای مقادیر غیرمنفرد C، برآورد ANOVA از σ^2 بر اساس رابطه (٥٦) در معادله $s = C\sigma^2$ قابل تخمین است، یعنی

رابطه (٥٧)
$$\hat{\sigma}^2 = C^{-1}s$$

برای مدل یک طرفه با اثرات تصادفی از طریق $s = C\sigma^2$ می‌توان بیان داشت که،

$$\begin{bmatrix} (a-1)n & a-1 \\ 0 & a(n-1) \end{bmatrix} \begin{bmatrix} \hat{\sigma}_a^2 \\ \hat{\sigma}_e^2 \end{bmatrix} = \begin{bmatrix} SSA \\ SSE \end{bmatrix} \Rightarrow \hat{\sigma}^2 =$$

رابطه (٥٨)
$$C^{-1}s \Rightarrow \begin{bmatrix} \hat{\sigma}_a^2 \\ \hat{\sigma}_e^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{(a-1)n} & \frac{1}{an(n-1)} \\ 0 & \frac{1}{(a-1)n} \end{bmatrix} \begin{bmatrix} SSA \\ SSE \end{bmatrix}$$
$$= C^{-1}\begin{bmatrix} y'Q_a y \\ y'Q_e y \end{bmatrix} = \begin{bmatrix} \frac{MSA - MSE}{n} \\ MSE \end{bmatrix}$$

از رابطه (٥٧) و (٥٨) می‌توان بیان داشت، که هر جز از مولفه‌های واریانس یک ترکیب خطی از مجموع مربعات s است. علاوه بر این، هیچ ویژگی ذاتی خاص یا الزامی در روابط ذکر شده نیست که عناصر بردار σ^2 همیشه غیرصفر باشند. علاوه بر این، در برآوردگر ANOVA، زمانی که $MSA < MSE$ است، σ_a^2 منفی برآورد می‌شود. برآورد منفی مولفه‌های واریانس یکی از ویژگی‌های نامطلوب برآوردگر ANOVA است (Searle *et al.*, 2006). ویژگی مثبت این برآوردگر نااریبی[1] است. برآورد رابطه (٥٧) همیشه نااریب است زیرا،

رابطه (٥٩)
$$E(\hat{\sigma}^2) = C^{-1}E(s) = C^{-1}C\sigma^2 = \sigma^2$$

نااریبی، ویژگی برآوردگرهای ANOVA است. اگرچه نااریبی در ANOVA برای برآورد میانگین‌ها یک مزیت است، اما در برآورد مولفه‌های واریانس دو ایراد به این نااریبی وارد است. اول اینکه اگر نااریبی برآوردگر ANOVA قابل قبول است چرا مولفه‌های واریانس می‌توانند منفی برآورد شوند؟ بنابراین برآورد منفی از مولفه‌های واریانس برای این برآوردگر، یک عیب محسوب می‌شود. توضیح این مسئله در عمل برای برخی از محققین مختلف کار دشواری است. بنابراین استفاده از برآوردگرهایی که برآورد منفی از مولفه‌های واریانس نمی‌دهند، اگرچه اریب باشند ولی خوش‌آیند محققین است. دلیل دوم این است که مفهوم نااریبی زمانی معنا دارد که تکرارهای یادداشت برداری شده، از یک آزمایش نمونه برداری شوند. در بیشتر مواقع داده‌ها از آزمایش‌های کاملا کنترل شده ثبت نمی‌شوند بلکه در اکثر موارد داده‌ها حجیم بوده و تکرار به معنای دقیقی وجود ندارد. به عنوان مثال جمع‌آوری عملکرد شیر گاوها در یک منطقه که مورد دسترسی هستند و اغلب نامتعادل هستند، فاقد یک الگوی یکسان در طول زمان می‌باشند، لذا تکرار یک داده به معنای تکرار همان الگو در شروع یادداشت برداری نیست. در واقع، ماهیت نمونه‌برداری به گونه‌ای است که نااریب بودن برآوردگر، غیر عملی است. بنابراین افراد ممکن است برای توجیه مولفه‌های واریانس به جای نمونه برداری از داده‌های با حجم زیاد مانند ١٥٠٠٠٠٠ رکورد برای برآورد مولفه‌های واریانس استفاده کنند (Searle *et al.*, 2006). لذا همانطور که کمپتورن (Kempthorne, 1968) بیان داشتند، برآورد نااریب میانگین در تخمین اثرات ثابت ممکن است در روش ANOVA درست باشد زیرا باقیمانده حاصل از اثرات ثابت دارای روند سیستماتیکی نیستند، اما در برآورد مولفه‌های واریانس الزاماً پذیرفتنی نیست.

بهترین برآورد نااریب[2] یکی دیگر از ویژگی‌های روش ANOVA است. بهترین برآورد نااریب به این معناست که در بین برآوردگرهای نااریب پارامترها، برآوردگری بهترین است که کمترین واریانس ممکن را داشته باشد (Casella and Berger, 1990). با رعایت فرضیات نرمال بودن (Graybill and Wortham, 1956) نشان دادند که برای هر مدل تصادفی، برآوردگر ANOVA یکی از بهترین برآوردگرهای نااریب به همراه آمار بسنده کافی[3] است. بر

1- Unbiasedness
2- Best unbiasedness
3- Complete sufficient statistics

طبق تعریف، آماره‌ای بسنده است که دانستن آن، محقق را از داشتن کلیه نمونه‌های ممکن برای تخمین پارامترهای توزیع نمونه‌ها بی‌نیاز سازد.

از دیگر ویژگی‌های مثبت برآوردگر ANOVA این است که در داده‌های متعادل دارای آماره بسنده کمینه[1] است. این مفهوم در کتاب‌های استباط آماری (Casella and Berger, 1990; Mood *et al.*, 1974) به طور مفصل شرح داده شده است. یعنی برای یک مدل تصادفی، آماره بسنده کمینه بر پایه فرضیات نرمال، میانگین حسابی داده‌ها و مجموع مربعات تجزیه واریانس تعریف می‌شود. بنابراین برآورد مولفه‌های واریانس با روش ANOVA که توابع خطی از مجموع مربعات است آماره‌های بسنده کمینه هستند، یعنی در عین حال که اطلاعات کافی از جمعیت ارائه می‌دهند، واریانس کمتری نسبت به سایر برآوردگرها دارند.

یکی از ویژگی‌ها منفی ANOVA، عدم یکتایی[2] در برآورد مولفه‌های واریانس به خصوص در داده‌های نامتعادل است (Aitkin, 1978; Nelder, 1977). در حقیقت زمانی که داده‌ها نامتعادل هستند، از معکوس تعمیم یافته برای پیدا کردن ماتریس‌های ویژه استفاده می‌شود، بنابراین یک مسیر منحصر به فردی، برای برآورد اثرات اصلی و متقابل وجود ندارد بلکه راه‌های مختلفی وجود دارد زیرا اثرات ساده و متقابل مستقل از هم نیستند و تا به حال هیچ مدلی برای تفکیک واریانس متغیر وابسته به مجموع مربعات غیرهمپوشان و مستقل از هم ارائه نشده است. به عنوان مثال، در برآورد مجموع مربعات سه فاکتور جنس، سن و اثر متقابل آنها در داده‌های نامتعادل اثر متقابل نسبت به اثرات اصلی واریانس کمتری از متغیر پاسخ توجیه می‌کند. این نتیجه به این معناست که اختصاص میزان مجموع مربعات به ترتیب قرار گرفتن فاکتورها در مدل بستگی دارد و این مشکل عدم یکتایی باعث سردرگمی محققین و حتی آماردانان برای انتخاب بهترین روش تجزیه طرح‌ها شده است (Der and Everitt, 2002). به طور کلی از زمانی که سه روش محاسبه تجزیه واریانس هندرسون ارائه شده است (I, II, III)، تحقیقات زیادی از این روش‌ها برای برآورد مولفه‌های واریانس استفاده کرده‌اند، با این وجود مشکلات و نقاط ضعف برآوردگر ANOVA نظیر برآورد منفی، فقدان یکتایی در برآوردها، نبود خصوصیات توزیعی مناسب و عدم روشی مفید برای مقایسه انواع برآوردهای مختلف آن (Henderson *et al.*, 1974) وجود دارد که باعث شده است روش‌های متعدد دیگری مانند روش حداکثر درستنمایی[3] (ML) و حداکثر درستنمایی محدود شده[4] (REML) بررسی و استفاده شوند (Searle *et al.*, 2006).

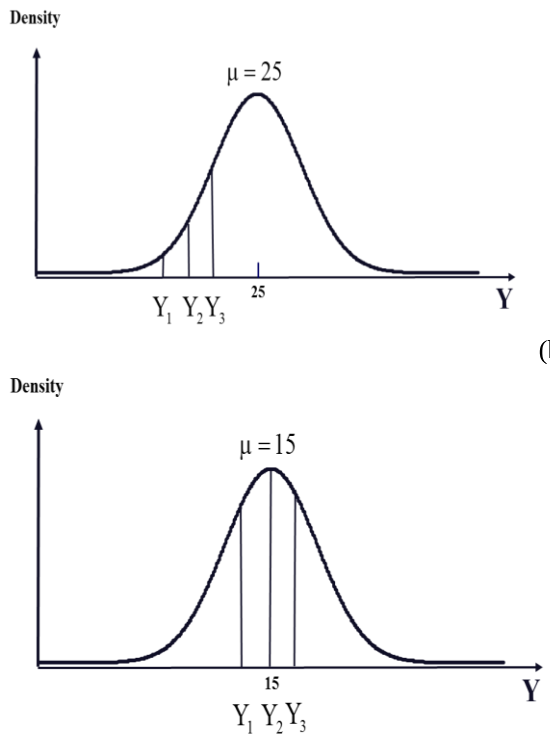
روش حداکثر درستنمایی(ML): حداکثر درستنمایی، تکنیکی است که در آن پارامترهای داده با حداکثر احتمال ممکن برآورد می‌شوند. به عبارتی پارامتر برآورد شده بیشترین سازگاری ممکن را با نمونه داده دارند (Neter *et al.*, 2004). به عنوان مثال، فرض کنید یک جامعه نرمال دارای انحراف معیار $\sigma=5$ و میانگین ناشناخته است. سه نمونه تصادفی با مقادیر $y_1=11$، $y_2=15$ و $y_3=18$ از جامعه مورد نظر به صورت تصادفی انتخاب می‌شوند. فرض می‌شود که میانگین فرضی این سه نمونه $\mu = 25$ است. شکل ٢-a دارای توزیع نرمال با میانگین ٢٥ و انحراف معیار ٥ است. موقعیت سه مشاهده تصادفی نیز در شکل ارائه شده است. این سه مشاهده در سمت چپ توزیع نرمال به شرط $\mu = 25$ هستند. بنابراین $\mu = 25$ مطابقت کمتری با داده‌های مذکور دارد. در شکل ٢-b سه مشاهده در اطراف مرکز توزیع قرار دارند، بنابراین مطابقت بیشتری با میانگین داده‌ها یعنی $\mu=15$ دارند. روش حداکثر درستنمایی از تراکم یا تجمع توزیع احتمال[5] (pdf) در y_i (برای مثال ارتفاع نمودار در محل y_i) به عنوان معیار مطابقت مشاهدات با پارامتر استفاده می‌کند. اگر y_i در قسمت انتهای توزیع نرمال باشد، میزان ارتفاع مشاهده کم

1- Minimal sufficient statistics
2- Lack of uniqueness
3- Maximum likelihood
4- Restricted maximum likelihood
5- Probability density function

است (شکل ۲-a) و اگر در مرکز باشد میزان ارتفاع آن زیاد است (شکل ۲-b).

(a)

(b)



شکل ۲- تابع تراکم برای دو مقدار μ در Y_i=11, 15, 18

Figure 2. The density of the probability distribution for two μ at Y_i = 11, 15 , 18

با استفاده از تابع تراکم برای توزیع احتمال نرمال، می‌توان میزان تراکم یا به عبارتی میزان احتمال برای y_1 که با f_1 بیان می‌شود را در دو مقدار میانگین μ به صورت زیر محاسبه کرد:

$$\mu = 25 : f\,(y \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]$$

$$= f_1 = \frac{1}{\sqrt{2\pi}(5)}\exp\left[-\frac{1}{2}\left(\frac{11-25}{5}\right)^2\right] = 0.007915$$

$$\mu = 15 : f_1 = \frac{1}{\sqrt{2\pi}(5)}\exp\left[-\frac{1}{2}\left(\frac{11-15}{5}\right)^2\right]$$

$$= 0.289692$$

تراکم یا میزان احتمال سه مشاهده برای دو میانگین ۱۵ و ۲۵ در جدول (۱) ارائه شده‌اند.

جدول ۱- تراکم یا میزان احتمال سه مشاهده در مقدار متفاوت μ

Table 1. Density function or probability value of three observations for different μ

	μ = 15	μ = 25
f_1	0.289692	0.007915
f_2	0.398942	0.053991
f_3	0.333225	0.149727

فرض می‌شود که $y_1; y_2; : : : ; y_n$ نمونه‌های تصادفی از یک جمعیت است که از یک توزیع تجمعی احتمالی تبعیت می‌نمایند. پارامتر θ ممکن است یک بردار باشد $\theta = (\theta_1, \theta_2, \ldots \theta_p)$، تابع حداکثر درستنمایی به صورت زیر است.

رابطه (۶۰) $$L(\theta \mid y_1 \ldots y_n) = P(y_1 \ldots y_n \mid \theta) = \prod_{i=1}^{n} P(y_i \mid \theta)$$

در این تابع، حداکثر درستنمایی پارامتر از حاصلضرب احتمال وقوع هر یک از مشاهدات تصادفی به شرط وقوع پارامتر مذکور محاسبه می‌گردد. بنابراین در مثال مورد نظر،

$$L(\mu = 25) = (0.007915)(0.053991)(0.149727) = 6.39842\times10^{-5}$$

$$L(\mu = 15) = (0.289692)(0.398942)(0.333225) = 0.038511$$

همانطور که مشاهده می‌شود، مقدار درستنمایی $\mu = 25$ یک عدد خیلی کمتر از $\mu = 15$ است. روش حداکثر درست‌نمایی مقدار پارامتر جامعه را در زمانی که بیشترین مقدار درست‌نمایی (L) بدست آید، برآورد می‌کند.

به طور کلی سه روش برای برآورد کردن پارامترها با روش حداکثر درستنمایی وجود دارد که شامل روش تحلیلی، جستجوی سیستماتیک و روش‌های عددی هستند. در این مثال از روش سیستماتیک استفاده شده است. در روش تحلیلی، مشتق تابع درست‌نمایی نسبت به پارامتر مد نظر گرفته و برابر با صفر قرار داده می‌شود و معادلات برای بدست آوردن پارامتر حل می‌شوند. روش‌های جستجوی شبکه‌ای و الگوریتم‌های صعود از تپه مانند الگوریتم Newton-Raphson، BHHH[۱]، DFP[۲] و سایر روش‌های دیگر در منابع متعدد وجود دارند (King, 1998).

چون لگاریتم طبیعی یک تابع ضرب‌پذیر، آهنگ متناسبی با کاهش و افزایش تابع ضرب پذیر L دارد (با افزایش تابع L افزایش در Log L و با کاهش L کاهش در Log L دیده

1- Berndt, Hall, Hall, and Hausman
2- Davidson-Fletcher-Powell

می‌شود)، از توابع ضرب‌پذیر لگاریتم گرفته شده و به صورت لگاریتم مجموع بررسی می‌گردند (Sorensen and Gianola, 2007).

رابطه (٦١)
$$L(\theta) \Rightarrow \operatorname{Log} L(\theta) = \sum_{i=1}^{n} \operatorname{Log} p(y_i \mid \theta)$$

اگر چندین پارامتر ناشناخته در هر تابع وجود داشته باشد، تابع درستنمایی لگاریتمی برای هر پارامتر مشتق گرفته شده و برابر با صفر قرار داد شده و در نتیجه مقدار پارامتر محاسبه می‌شود.

رابطه (٦٢)
$$\frac{\partial L(\theta)}{\partial \theta_1} = \sum_{i=1}^{n} \frac{\operatorname{Log} f(x_i \mid \theta)}{\partial \theta_1} = 0, \frac{\partial L(\theta)}{\partial \theta_2} = \sum_{i=1}^{n} \frac{\operatorname{Log} f(x_i \mid \theta)}{\partial \theta_2} = 0, \ldots, \frac{\partial L(\theta)}{\partial \theta_n} = \sum_{i=1}^{n} \frac{\operatorname{Log} f(x_i \mid \theta)}{\partial \theta_m} = 0$$

که m تعداد پارامتر ناشناخته است. به عنوان مثال، در توزیع نرمال، تابع درستنمایی به صورت زیر است.

رابطه (٦٣)
$$L(\mu, \sigma^2 \mid y_1, y_2, .., y_n) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right)\right) = -n \ln(\sqrt{2\pi}) - \frac{n}{2}\ln(\sigma^2) - \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2}$$

که با مشتق‌گیری تابع درستنمایی به پارامتر میانگین و واریانس، برآورد آنها به صورت زیر محاسبه می‌شود.

رابطه (٦٤)
$$\frac{\partial L}{\partial \mu} = \frac{\sum_{i=1}^{n}(y_i - \mu)}{\sigma^2} = 0 \iff \hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} \iff \hat{\mu} = \bar{y}$$

رابطه (٦٥)
$$\frac{\partial L}{\partial(\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4} \sum_{i=1}^{n}(y_i - \mu)^2 = 0 \iff \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu})^2}{n}$$

با فرض اینکه واریانس برابر با $V = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$، روش درستنمایی تخمین اریبی از واریانس در نمونه‌های با تعداد کم می‌دهد، بنابراین هرچه تعداد نمونه بیشتر باشد تخمین روش درستنمایی به واقعیت نزدیک‌تر است (Lynch and Walsh, 1998). اگر واریانس به صورت $V = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ برآورد شود،

رابطه (٦٦)
$$\sum_{i=1}^{n}(y_i - \hat{\mu})^2 = \sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - \hat{\mu})^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}(\bar{y} - \hat{\mu})^2 + 2(\bar{y} - \hat{\mu}) \sum_{i=1}^{n}(y_i - \bar{y}) = n[V + (\bar{y} - \mu)^2]$$

در نتیجه،

رابطه (٦٧)
$$\sigma^2 = V + (\bar{y} - \mu)^2$$

با توجه به رابطه (٦٧) می‌توان بیان داشت که برآورد واریانس با روش حداکثر درستنمایی به ویژه در نمونه‌های کوچک متاثر از مقدار میانگین است و حداکثر درستنمایی σ^2 با این فرض برآورد می‌شود که میانگین تخمین زده شده $\bar{y}$ بدون اشتباه برآورد شده و با μ برابر است، لذا قسمت دوم عبارت که الزاما مثبت است، چشم‌پوشی شده است، بنابراین واریانس با کمی اریبی بیش از مقدار واقعی برآورد می‌گردد (Lynch and Walsh, 1998).

رابطه (٦٨)
$$\sigma^2 = V$$

در مرور منابع، ابهامات زیادی در برآورد به روش حداکثر درست‌نمایی وجود دارد. اگر به طور دقیق تعریفی از برآورد ML باشد، باید یک نقطه از فضای پارامتری باشد و اگر چنین نقطه‌ای وجود داشته باشد، در همان نقطه تابع درستنمایی دارای بیشترین مقدار است. در برخی از مسائل، برآوردهای روش ML که یک نتیجه منحصر به فرد می‌دهند و شناخته شده هستند، برآوردهای درستی هستند. اما در برخی از مسائل پیچیده، معادله درستنمایی ممکن است دارای چند جواب باشد و فضای پارامتر ممکن است به یک همگرایی نرسد. در این مسائل، تعیین برآوردهای درست به روش ML صرفا به راحتی حل کردن معادلات درستنمایی نیست (Sahai and Miguel, 2004). راه حل هر دوی این موقعیت‌ها در کتاب مولفه‌های واریانس سیرل و همکاران (Searle *et al.*, 2006) ارائه شده است. به عنوان مثال، به استثنا طرح‌های نرمال، الگوریتم‌های تکرار شونده برای

حداکثر کردن تابع درستنمایی نیاز است و بر حسب نوع ماتریس طرح، مقادیر پارامترهای مورد نظر و نوع الگوریتم استفاده شده ممکن است همگرایی یا عدم همگرایی در فضای پارامتر به وجود بیاید. اگر این همگرایی وجود داشته باشد، نمی‌توان مطمئن بود که حداکثر درستنمایی کلی است و یا مکانی و مقطعی. برای فائق شدن بر مشکل چند ریشه بودن برآوردهای درست حداکثر درست‌نمایی و عدم تطابق بین آنها و همچنین برآورد درست پارامترها، برخی از آمارادانان راه‌حل‌های خاصی پیشنهاد داده‌اند (Lehmann and Casella, 2006; Small *et al.*, 2000). در تحقیقی (Barnett, 1966) از روش‌های عددی برای یافتن ریشه معادلات درستنمایی استفاده کرده است و پنج روش عددی برای یافتن ریشه معادلات درستنمایی پیشنهاد داده است.

تحت فرضیات نرمال، برآورد پارامترهای برخی مدل‌ها با اثرات ثابت با استفاده از روش درست‌نمایی مشابه روش کمترین مقدار مربعات باشد. بنابراین ممکن است انتظار نیز این باشد که مولفه‌های واریانس برآورد شده در روش ANOVA و ML یکسان باشد که این انتظار اشتباه اشت. برآوردگر ANOVA می‌تواند دارای برآوردهای منفی از مولفه‌های واریانس باشد، در حالیکه برآوردگر ML و در ادامه برآوردگر REML فاقد این ویژگی هستند و برآورد منفی از مولفه‌های واریانس ندارد. برآوردهای مولفه‌های واریانس و همچنین اثرات ثابت توسط روش ML ساده و مستقیم نیستند. همچنین در داده‌های نامتعادل یک روش مشخص برای تخمین مولفه‌های واریانس وجود ندارد. در داده‌های متعادل، طرح یک طرفه معادله درستنمایی به صورت زیر است (Sahai and Miguel, 2004).

$$L=\frac{\exp\left[-\frac{1}{2}\left\{\frac{SSE}{\sigma_e^2}+\frac{SSA}{\sigma_e^2+n\sigma_a^2}+\frac{an(\bar{y}_{..}-\mu)^2}{\sigma_e^2+n\sigma_a^2}\right\}\right]}{(2\pi)^{\frac{1}{2}an}(\sigma_e^2)^{\frac{1}{2}a(n-1)}(\sigma_e^2+n\sigma_a^2)^{\frac{1}{2}a}} \quad \text{رابطه (۶۹)}$$

با استفاده از لگاریتم تابع درستنمایی معادله (۶۹) به صورت زیر نوشته می‌شود.

$$\ln(L)=-\frac{1}{2}\Big[(an)\ln(2\pi)+a(n-1)\ln(\sigma_e^2)+a\ln(\sigma_e^2+n\sigma_a^2)+\frac{SSE}{\sigma_e^2}+\frac{SSA}{\sigma_e^2+n\sigma_a^2}+\frac{an(\bar{y}_{..}-\mu)^2}{\sigma_e^2+n\sigma_a^2}\Big] \quad \text{رابطه (۷۰)}$$

با مشتق جزیی $\ln(L)$ نسبت به μ، σ_e^2 و σ_a^2 معادلات درستنمایی و برابر با صفر قرار دادن معادله تخمین پارامترها به صورت زیر است.

$$\frac{\partial\ln(L)}{\partial\mu}=\frac{1}{2}\frac{an(\bar{y}_{..}-\mu)}{\sigma_e^2+n\sigma_a^2}=0\Leftrightarrow\hat{\mu}=\bar{y}_{..} \quad \text{رابطه (۷۱)}$$

$$\frac{\partial\ln(L)}{\partial\sigma_e^2}=-\frac{1}{2}\left[\frac{a(n-1)}{\sigma_e^2}+\frac{a}{\sigma_e^2+n\sigma_a^2}-\frac{SSE}{\sigma_e^4}-\frac{nSSA}{(\sigma_e^2+n\sigma_a^2)^2}-\frac{an(\bar{y}_{..}-\mu)^2}{(\sigma_e^2+n\sigma_a^2)^2}\right]=0\Leftrightarrow\hat{\sigma}_e^2=\frac{SSE}{a(n-1)} \quad \text{رابطه (۷۲)}$$

$$\frac{\partial\ln(L)}{\partial\sigma_a^2}=-\frac{1}{2}\left[\frac{an}{\sigma_e^2+n\sigma_a^2}-\frac{nSSA}{(\sigma_e^2+n\sigma_a^2)^2}-\frac{an^2(\bar{y}_{..}-\mu)^2}{(\sigma_e^2+n\sigma_a^2)^2}\right]=0\Leftrightarrow\hat{\sigma}_a^2=\frac{1}{n}\left(\frac{SSA}{a}-\frac{SSE}{a(n-1)}\right) \quad \text{رابطه (۷۳)}$$

در برآوردهای مذکور برای مولفه‌های واریانس، حداکثر کردن L ممکن است منجر به مقادیر منفی σ_e^2 و σ_a^2 شود. بنابراین این معادلات برآورد درستی از مولفه‌های واریانس نمی‌دهند، لذا سهای و تامپسون (Sahai and Thompson, 1973) یک روش ساده برای یافتن مولفه‌های واریانس با محدودیت غیرمنفی σ_e^2 و σ_a^2 ارائه دادند. برآورد غیرمنفی ML با روش ارائه شده به صورت زیر است.

$$\sigma_e^2,ML=\min\left(\frac{SSE}{a(n-1)},\frac{SSA+SSE}{an}\right) \quad \text{رابطه (۷۴)}$$

$$\sigma_a^2,ML=\max\left[\frac{1}{n}\left(\frac{SSA}{a}-\frac{SSE}{a(n-1)}\right),0\right] \quad \text{رابطه (۷۵)}$$

همانطور که در رابطه (٧٥) مشاهده می‌شود، اگر MSA کمتر از MSE باشد، مقدار واریانس σ_a^2 برابر با صفر در نظر گرفته می‌شود.

روش ماتریسی حل معادلات حداکثر درستنمایی: در مدل خطی با اثرات ثابت و تصادفی تیمار به صورت رابطه (٥٠) تابع چگالی احتمال داده y مشابه معادله (٦٣) به روش ماتریسی زیر ارائه می‌شود (Hartley and Rao, 1967; Harville, 1977; Henderson, 1984; Lynch and Walsh, 1998; Schaeffer, 1998; Searle *et al.*, 2006; Sorensen and Gianola, 2007).

رابطه (٧٦)
$$p(y|X\beta,V)=(2\pi)^{-n/2}|V|^{-1/2}\exp\left[-\frac{1}{2}(y-X\beta)'V^{-1}(y-X\beta)\right]$$

لگاریتم طبیعی معادله مذکور به شرح زیر است.

رابطه (٧٧)
$$L(\beta,V|X,y)=-\frac{n}{2}\ln(2\pi)-\frac{1}{2}\ln|V|-\frac{1}{2}(y-X\beta)'V^{-1}(y-X\beta)$$

در این معادله چندین پارامتر ناشناخته وجود دارد که شامل بردار ثابت β و اجزا مولفه V شامل $\sigma_1^2,\sigma_2^2,...,\sigma_i^2$ و σ_e^2 هستند. اگر مشتق جزئی برای معادله لگاریتم درستنمایی نسبت به هر جز گرفته شود و معادله برابر با صفر قرار داده شود هر پارامتر برآورد می‌شود که برای بردار ثابت β به صورت زیر است.

رابطه (٧٨)
$$\frac{\partial[(y-X\beta)'V^{-1}(y-X\beta)]}{\partial\beta}=-2X'V^{-1}(y-X\beta)=0\Rightarrow X'V^{-1}(y-X\beta)=X'V^{-1}y-X'V^{-1}X\beta$$

که با بازترتیبی، معادله نهایی برآورد پارامترهای ثابت به شکل زیر می‌باشد.

رابطه (٧٩)
$$\beta=(X'V^{-1}X)^{-1}X'V^{-1}y$$

برای بدست آوردن مشتق معادله L به اجزا تصادفی مولفه واریانس لازم است که چندین رابطه و مشتق‌گیری ماتریسی یادآوری شود. اگر ماتریسی بنام M باشد و عناصر آن تابعی از عناصر X باشد، روابط زیر برای مشتق این ماتریس برقرار است.

رابطه (٨٠)
$$\frac{\partial\ln|M|}{\partial x}=tr(M^{-1}\frac{\partial M}{\partial x})$$

رابطه (٨١)
$$\frac{\partial M^{-1}}{\partial x}=-M^{-1}\frac{\partial M}{\partial x}M^{-1}$$

با توجه به رابطه میانگین مربعات انحرافات افراد از میانگین جمعیت در معادله (٦٦) که عبارت بود از $(y_i-\bar{y}+\bar{y}-\mu)\Leftrightarrow(y_i-\bar{y})+(\bar{y}-\mu)$، مشابه این حالت در شکل ماتریسی نیز قابل انجام است که به صورت زیر انجام می‌شود.

رابطه (٨٢)
$$(y-X\beta)'V^{-1}(y-X\beta)=(y-X\hat{\beta})'V^{-1}(y-X\hat{\beta})+(\hat{\beta}-\beta)'X'V^{-1}X(\hat{\beta}-\beta)$$

به طوریکه β برآوردی از β است. مشتق معادله به مولفه‌های ساده واریانس σ_i^2 و σ_e^2 گرفته می‌شود. عبارت V در اینجا حاوی اجزا تصادفی و خطای آزمایش $V=\sum_{i=1}^{m}Z_iZ_i'\sigma_i^2+\sigma_e^2I$ است. با استفاده از ترمینولوژی σ_i^2 برای مولفه‌های واریانس، عبارات زیر از مشتق V حاصل می‌گردند.

رابطه (٨٣)
$$\frac{\partial V}{\partial\sigma_i^2}=\begin{cases}I & \text{اگر}\quad \sigma_i^2=\sigma_e^2\\ Z_iZ_i' & \text{اگر}\quad \sigma_i^2\neq\sigma_e^2\end{cases}$$

با جایگزینی معادله (٨٢) به جای معادله (٧٧) و استفاده از معادلات (٨٠) و (٨١) پس از مشتق‌گیری نسبت به σ_i^2 معادله عمومی زیر حاصل می‌گردد.

رابطه (٨٤)
$$\frac{\partial L(\beta,V|X,y)}{\partial\sigma_i^2}=-\frac{1}{2}tr(V^{-1}Z_iZ_i')+\frac{1}{2}(y-X\hat{\beta})'V^{-1}Z_iZ_i'V^{-1}(y-X\hat{\beta})+\frac{1}{2}(\beta-\beta)'X'V^{-1}Z_iZ_i'V^{-1}X(\beta-\beta)$$

قابل ذکر است که $V=\sum_{i=1}^{m}Z_iZ_i'\sigma_i^2+\sigma_e^2I$ تابعی از مولفه‌های واریانس برآورد شده است. برآوردگر حداکثر درست‌نمایی، مولفه‌های واریانس را با قرار دادن $\beta=\beta$ در معادله (٨٤) و با حذف قسمت انتهایی معادله به صورت زیر برآورد می‌کند:

رابطه (٨٥)
$$tr(V^{-1}Z_iZ_i')=(y-X\beta)'V^{-1}Z_iZ_i'V^{-1}(y-X\beta)$$

همان‌طور که در این معادله مشاهده می‌شود هم اثرات تصادفی و هم اثرات ثابت در این معادله مشاهده می‌شوند. برای ساده تر شدن محاسبات، ابتدا یک ماتریسی با عنوان

ماتریس P در زیر تعریف می‌شود که معادله بالا بر حسب P در ادامه محاسبه می‌شود:

رابطه (۸۶) $\mathbf{P}=\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$

همچنین به طور خاص، نتیجه کاربردی‌تر حاصل از این تغیرات عبارت است از،

رابطه (۸۷) $Py=V^{-1}y-V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}y=V^{-1}(y-X\beta)$

با استفاده معادله (۸۷) می‌توان معادله (۸۵) را به صورت زیر نوشت.

رابطه (۸۸) $tr(V^{-1}Z_iZ_i')=y'PZ_iZ_i'Py$

به جای P از P استفاده می‌شود که تابعی از V یا مولفه‌های واریانس است، لذا این ماتریس نیز برآورد شده است. اگرچه کاملا واضح نیست، اما این تخمین مشابه تخمین حداکثر درستنمایی معادله (٦٨) است. مولفه‌های واریانس‌های σ_i^2 و σ_e^2 توسط رابطه (۸۸) برآورد می‌شوند، که در هر دو طرف معادله قرار دارند. همچنین معکوس ماتریس V^{-1} در P مستتر است. به طور کلی برآوردهای درست‌نمایی برای مدل‌های افزایشی و مولفه‌های واریانس به صورت زیر است.

رابطه (۸۹) $tr(V^{-1})=y'PPy$ برای σ_E^2

رابطه (۹۰) $tr(V^{-1}Z_iZ_i')=y'PZ_iZ_i'Py$ برای σ_i^2

با فرض،

رابطه (۹۱) $y=Xb+\sum_{i=1}^{m}Z_iu_i+e\Leftrightarrow y=Xb+\sum_{j=0}^{m}Z_ju_j$

اندیس i از صفر تا m به این معناست که خطا در قسمت چپ معادله با جز تصادفی مدل ادغام شده است و رابطه (۹۲) برقرار است.

رابطه (۹۲)
$$\begin{aligned}tr(V^{-1}Z_iZ_i')&=tr(V^{-1}Z_iZ_i'V^{-1}V)\\&=tr(V^{-1}Z_iZ_i'V^{-1}\sum_{i=0}^{m}Z_jZ_j'\sigma_j^2)\\&=\sum_{j=0}^{m}tr(V^{-1}Z_iZ_i'V^{-1}Z_jZ_j')\sigma_j^2\\&=y'PZ_iZ_i'Py\end{aligned}$$

برای معادلات تصادفی بیش از یک فاکتور، این فرمول‌ها قابل تعمیم هستند و m+1 مولفه واریانس (فاکتور تصادفی به همراه خطای آزمایش) قابل برآورد است. حل این معادلات دارای دو ویژگی مهم است. اول اینکه برخلاف مدل ساده در ابتدای این بخش که یک حالت بسته و بدون وابسته به واریانس برای برآورد اثرات ثابت μ استفاده شد، در حداکثر درست‌نمایی اثرات ساده β تابعی از ماتریس V هست که V نیز دارای چندین مولفه واریانس است که باید تخمین زده شوند. دوم اینکه در این معادلات معکوس V لحاظ شده است، بنابراین V نیز تابعی از توابع غیرخطی مولفه‌های واریانس است. در نتیجه یک راه حل ساده برای حل این معادلات وجود ندارد. تخمین حداکثر درست‌نمایی β، σ_i^2 و σ_e^2 نیازمند الگوریتم‌های تکرارشونده[1] هستند که در نهایت منجر به تخمین مولفه‌های واریانس می‌شوند (Lynch and Walsh, 1998). برای محاسبه دقیق مولفه‌های واریانس نیاز است که ماتریس مجانب یا واریانس مولفه‌های واریانس که عبارت است از مشتق دوم معادله (۷۷)، محاسبه شود.

رابطه (۹۳)
$$\begin{aligned}&\frac{\partial^2\ln L}{\partial\sigma_i^2\partial\sigma_j^2}\\&=\frac{\partial\left(-\frac{1}{2}tr\left(V^{-1}Z_iZ_i'\right)+\frac{1}{2}y'PZ_iZ_i'Py\right)}{\partial\sigma_i^2\partial\sigma_j^2}\\&=\frac{1}{2}tr(V^{-1}Z_iZ_i'V^{-1}Z_jZ_j')\\&-y'PZ_iZ_i'PZ_jZ_j'y\end{aligned}$$

بر اساس اثبات بیشاپ (Bishop, 1992) ماتریس داده فیشر[2] از طریق معادله (۹٤) محاسبه می‌شود.

رابطه (۹٤) $F=-E\left[\frac{\partial^2\ln L}{\partial\sigma_i^2\partial\sigma_j^2}\right]=-E[H]$

که H ماتریس هسین[3] می‌باشد. ماتریس H مشتق دوم تابع حداکثر درستنمایی به مولفه‌های واریانس است.

$$-E\left[\frac{1}{2}tr(\hat{V}^{-1}Z_iZ_i'\hat{V}^{-1}Z_jZ_j')-y'\hat{P}Z_iZ_i'\hat{P}Z_jZ_j'y\right]$$

1- Iterative
2- Fisher information matrix
3- Hessian matrix

$$=\frac{1}{2}\mathrm{tr}(\widehat{V}^{-1}Z_iZ_i'\widehat{V}^{-1}Z_jZ_j')$$

رابطه (٩٥)

$$\mathrm{Var}\begin{bmatrix}\sigma_1^2\\ \sigma_2^2\\ \vdots\\ \sigma_m^2\end{bmatrix}=(F)^{-1}=$$

رابطه (٩٦)

$$\left(\frac{1}{2}\begin{bmatrix}\mathrm{tr}\left(V^{-1}Z_1Z_1'V^{-1}Z_1Z_1'\right) & \mathrm{tr}\left(V^{-1}Z_1Z_1'V^{-1}Z_2Z_2'\right) & . & \mathrm{tr}\left(V^{-1}Z_1Z_1'V^{-1}\right)\\ \mathrm{tr}\left(V^{-1}Z_2Z_2'V^{-1}Z_1Z_1'\right) & \mathrm{tr}\left(V^{-1}Z_1Z_1'V^{-1}Z_1Z_1'\right) & . & \mathrm{tr}\left(V^{-1}Z_2Z_2'V^{-1}\right)\\ \vdots & \vdots & . & \vdots\\ \mathrm{tr}\left(V^{-1}V^{-1}Z_1Z_1'\right) & \mathrm{tr}\left(V^{-1}V^{-1}Z_2Z_2'\right) & . & \mathrm{tr}\left(V^{-1}V^{-1}\right)\end{bmatrix}\right)^{-1}$$

برای محاسبه واریانس از الگوریتم نیوتن رافسون[1] استفاده می‌شود. این الگوریتم اگرچه کند است ولی در برخی معادلات روش بسیار مفیدی برای برآورد مولفه‌های واریانس به روش ML و REML است. این الگوریتم تا همگرا شدن و برابری $\theta^{(t)}\approx\theta^{(t+1)}$ ادامه می‌یابد.

$$\theta^{(t+1)}=\theta^{(t)}-\left(S^{(t)}\times\left[H^{(t)}\right]^{-1}\right)\Leftrightarrow$$

$$\begin{bmatrix}\sigma_1^2\\ \sigma_2^2\\ \vdots\\ \sigma_m^2\end{bmatrix}^{(t+1)}=\begin{bmatrix}\sigma_1^2\\ \sigma_2^2\\ \vdots\\ \sigma_m^2\end{bmatrix}^{(t)}-\begin{bmatrix}\left(y'\hat{P}Z_1Z_1'\hat{P}y\right)\\ \left(y'\hat{P}Z_2Z_2'\hat{P}y\right)\\ \vdots\\ \left(y'\hat{P}\hat{P}y\right)\end{bmatrix}^{(t)}\times\left[F^{-1}\right]^{(t)}$$

رابطه (٩٧)

که θ عبارت است از مولفه‌های واریانس، H ماتریس هسین و S بردار محاسبه[2] و مجموع مربعات فاکتورها است (Lynch and Walsh, 1998; Searle *et al.*, 2006).

حداکثر درستنمایی محدود شده: ایده روش حداکثر درستنمایی محدود شده در ابتدا توسط اندرسون و بانکرافت (Anderson and Bancroft, 1952) و بعد از آن توسط تامپسون و مور (Thompson and Moore, 1963) برای داده‌های متعادل تصادفی ارائه شد. روش REML برای داده‌های نامتعادل توسط پترسون و تامپسون (Patterson and Thompson, 1971) ارائه شد. روش REML بخش غیر ثابت تابع درستنمایی را حداکثر می‌کند. دو مزیت عمده روش REML این است که به بخش ثابت مدل، درجه آزادی اختصاص می‌دهد. به عنوان مثال برای مشاهدات y_1، y_2، و y_n با مدل خطی $y_{ij}=\mu+e_i$ واریانس خطا در روش ML به صورت $\sum_{i=1}^{n}(y_i-\bar{y})^2/n$ برآورد می‌گردد که اریب می‌باشد. در حالی‌که در روش REML برآورد واریانس $\sum_{i=1}^{n}(y_i-\bar{y})^2/n-1$ است. مزیت دوم روش REML این است که برآوردگرهای سنتی مانند ANOVA را رد نمی‌کند و با روش ANOVA در حالت‌هایی که داده‌ها متعادل هستند و واریانس منفی برآورد نمی‌شود، یکسان است. در مواردی که داده‌ها نامتعادل هستند و داده گم شده در آزمایش وجود دارد و همچنین واریانس منفی برآورد می‌شود، روش REML یک روش مناسب‌تر از ANOVA است (Sahai and Miguel, 2004).

در روش حداکثر درستنمایی محدود شده (REML) اریب واریانس در روش حداکثر درستنمایی را با در نظر گرفتن اشتباه حاصل از تخمین μ برطرف می‌نماید (Lynch and Walsh, 1998). با توجه به معادله (٦٦) میزانی که σ^2 از مقدار واقعی σ^2 فاصله دارد و کمتر برآورد می‌شود برابر با ارزش مورد انتظار $(\bar{y}-\mu)^2$ می‌باشد که با در نظر گرفتن نمونه‌های تصادفی و مستقل از هم، این امید ریاضی، برابر با واریانس میانگین نمونه می‌باشد (σ^2/n). بنابراین برآورد واریانس به روش REML عبارت است از:

$$\sigma^2=V+E[(\bar{y}-\mu)^2]=V+\frac{\sigma^2}{n}$$

رابطه (٩٨)

در اینجا به طور ملموس میزان اریب مشخص نیست، زیرا میزان قطعی σ^2 مشخص نیست. در این مواقع به روش‌های تکرارشونده برای برآورد واریانس نیاز است. در معادله (٩٨) میزان اریب فقط قابل تخمین است. مقدار اولیه برآورد واریانس با استفاده از حداکثر درست‌نمایی V است که این ارزش $\sigma^2_{(0)}=V$ شروع اولیه برای برآورد بهتر

1- Newton Raphson
2- Score vector

واریانس است. در دور دوم تخمین، واریانس عبارت است از:

$$\sigma^2_{(1)} = V + \frac{\sigma^2_{(0)}}{n} = V + \frac{V}{n}$$

و در صورت تکرار،

$$\sigma^2_{(2)} = V + \frac{\sigma^2_{(1)}}{n} = V + \frac{V+(V/n)}{n}$$

در نهایت،

رابطه (۹۹)
$$\sigma^2_{(t+1)} = V + \frac{\sigma^2_{(t)}}{n}$$

آخرین مرحله برای برآورد σ^2 به جایی ختم می‌شود که الگوریتم، پیشرفت زیادی نداشته باشد و مقدار $\sigma^2_{(t+1)} = \sigma^2_{(t)}$ برابر شود، در نتیجه با یک جایگزینی،

رابطه (۱۰۰)
$$\sigma^2 = \frac{n}{n-1}V = \frac{\sum_{i=1}^{n}(y_i - \bar{y})}{n-1}$$

که این مقدار، میزان برآورد نااریب واریانس است (Lynch and Walsh, 1998).

روش REML برای بهبود روش ML به منظور محاسبه درجه آزادی در برآورد اثرات ثابت پیشنهاد شده است. روش REML بر اساس تبدیل خطی داده‌های مشاهده شده y که اثرات ثابت از مدل حذف شده‌اند انجام می‌گیرد. یکی از ساده‌ترین روش‌ها برای این تبدیل، ضرب ماتریس K در ماتریس طرح X با شرط KX=0 برقرار است (Searle *et al.*, 2006). استفاده از این تبدیل در مدل مختلط (۵۰) به صورت زیر است.

رابطه (۱۰۱)
$$y^* = Ky = K(Xb + Zu + e) = KZu + Ke$$

تابع درستنمایی استفاده شده در REML برای باقیماندها با فرض توزیع نرمال چند متغیره است. به طور کلی در برآوردهای بهترین پیش‌بینی نااریب این مفهوم وجود دارد به ازائ هر ضریب از متغیرهای مستقل، متغیر وابسته y قابل برآورد است. این ضریب در اینجا K تعریف شده است که برابر است با $I - X(X'X)^{-1}X'$. معادله درستنمایی روش REML با محاسبه ضریب K در معادله (۷۷) و در نهایت با جایگزینی $y = K'y$، $Z = K'Z$، $X =$ $K'X = 0$، $V = K'VK$، $P = PVP$ بر اساس اثبات سیرل و همکاران (Searle *et al.*, 2006) می‌توان بیان داشت که،

رابطه ۱۰۲
$$L(V \mid K'y) = -\frac{1}{2}\{N - \text{rank}(x)\}\ln(2\pi) - \frac{1}{2}\ln|K'VK| - \frac{1}{2}y'K(K'VK)^{-1}K'y$$

که rank(x) میزان درجه آزادی مربوط به بخش ثابت است که از درجه آزادی کل کم می‌گردد. اجزا مدل REML به دو صورت زیر

$$\ln|K'VK| = \ln|V| + \ln|X'VX|$$
$$y'K(K'VK)^{-1}K'y = y'Py$$

می‌باشند که P به صورت معادله زیر ارائه می‌شود.

$$P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1} = K(K'VK)^{-1}K'$$

رابطه (۱۰۳)
$$\frac{\partial \ln L}{\partial \sigma_i^2} = \frac{\partial\left[-\frac{1}{2}\{N-\text{rank}(x)\}\ln(2\pi) - \frac{1}{2}\ln|K'VK| - \frac{1}{2}y'K(K'VK)^{-1}K'y\right]}{\partial \sigma_i^2}$$

رابطه (۱۰٤)
$$= \frac{\partial\left[-\frac{1}{2}\ln|K'VK| - \frac{1}{2}y'Py\right]}{\partial \sigma_i^2}$$
$$= -\frac{1}{2}\text{tr}\left[(K'VK)^{-1}K'\frac{\partial V}{\partial \sigma_i^2}K\right] + \frac{1}{2}y'P\frac{\partial V}{\partial \sigma_i^2}Py$$
$$= -\frac{1}{2}\text{tr}\left(P\frac{\partial V}{\partial \sigma_i^2}\right) + \frac{1}{2}y'P\frac{\partial V}{\partial \sigma_i^2}Py$$
$$= -\frac{1}{2}\text{tr}\left(PZ_iZ_i'\right) + \frac{1}{2}y'PZ_iZ_i'Py$$

در نهایت می‌توان مشابه با رابطه ۹۰ در ML معادله برای برآورد REML را نیز به صورت زیر نوشت.

رابطه (۱۰۵)
$$\text{tr}(PZ_iZ_i') = y'PZ_iZ_i'Py$$

تنها اختلاف معادله ML (۸۸) با REML (۱۰۵) این است که رابطه (۱۰۵) دارای P است اما رابطه (۸۸) دارای V^{-1} است. سایر محاسبات روش REML مشابه ML است با این تفاوت P به جای V^{-1} قرار می‌گیرد (Searle *et al.*, 2006). برآورد اثرات ثابت و تصادفی از طریق مدل مختلط به صورت زیر می‌باشد (Henderson, 1984).

رابطه (۱۰٦)
$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix}\begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} X'R^{-1}Y \\ Z'R^{-1}Y \end{bmatrix}$$

مثال کاربردی: در این تحقیق از دو صفت وزن دانه در سنبله و عملکرد دانه گندم برای محاسبات مولفه‌های

واریانس استفاده شد. آزمایش به صورت طرح بلوک‌های کامل تصادفی با ۳٦ ژنوتیپ در سه تکرار اجرا شد (Valizadeh, 2014). داده‌ها مربوط به سال زراعی ۱۳۹۳ در مزرعه تحقیقاتی دانشکده کشاورزی دانشگاه لرستان واقع در شهرستان خرم آباد، کیلومتر ۱۲ می‌باشد که در دو آزمایش نرمال و تنش خشکی اجرا شد. در این تحقیق فقط از داده‌های نرمال استفاده شد. در این داده‌ها ۳٦ ژنوتیپ به صورت تصادفی از بین ۱۱۲ ژنوتیپ گندم انتخاب و کشت شدند. بنابراین ژنوتیپ به صورت تصادفی در نظر گرفته شد. تکرار نیز به صورت تصادفی در نظر گرفته شد، لذا به دلیل تصادفی بودن تمامی فاکتورها، مدل این آزمایش، تصادفی بود. از هر سه روش ANOVA، ML و REML برای برآورد مولفه‌های واریانس دو صفت مذکور استفاده شد. نتایج برآورد مولفه‌های واریانس در جدول ۲ ارائه شد. صفت عملکرد فاقد داده گم شده در آزمایش بود ولی صفت وزن دانه در سنبله دارای مشاهدات گم شده بود بنابراین یک طرح نامتعادل بود. در این آزمایش از واریانس‌های برآورد شده برای محاسبات واریانس فنوتیپی و وراثت‌پذیری عمومی بر مبنای تک بوته، استفاده شد.

نتایج تجزیه به سه روش برآوردگر برای دو صفت عملکرد و وزن دانه در سنبله ویژگی‌های بیان شده سه برآوردگر مذکور را تا حدودی مشخص کرد. همان‌طور که مشاهده می‌شود میزان واریانس منابع تغییر در دو روش ANOVA و REML در صفت عملکرد مشابه هم می‌باشد.

در واقع اگر آزمایشات فاقد داده گم شده باشند و طرح متعادل باشد، نتایج این دو برآوردگر مشابه هم می‌باشد. همچنین نتایج دو برآوردگر ANOVA و REML با نتیجه ML متفاوت است که این دلیل ناشی از اریب واریانس در روش حداکثر درستنمایی است که در رابطه (۹۸) به صورت تئوری بیان شد. در صفت وزن دانه در سنبله به دلیل داده گم شده، نتایج هر سه برآوردگر برای واریانس منابع تغییر متفاوت بود. همان‌طور که بیان شد در داده‌های گم شده برآوردگرهای REML و ML نسبت به روش ANOVA برتری دارند و از بین ML و REML نیز در داده‌های گم شده روش REML برتری دارد. نوع برآوردگر بر مقدار پارامترهای ژنتیکی مانند واریانس فنوتیپی و وراثت‌پذیری در داده‌های گم شده تاثیر گذار است، اگرچه در این تحقیق این تفاوت اندک است ولی در تحقیقاتی که حجم داده‌ها زیاد است و نیاز به محاسبات دقیق دارند، نتایج این برآوردگرها می‌توانند به محققین در گزینش و برنامه‌های به‌نژادی با دقت بیشتری کمک نمایند.

جهت راحتی و درک بهتر محاسبه مولفه‌های واریانس به روش ML و REML یک برنامه SAS ارائه شده است که به دو روش دستی و مرحله به مرحله در قالب ماتریس که در یک مدول تکرار شونده IML قرار گرفته و فرمان Proc Mixed ، مولفه‌های واریانس را محاسبه می‌کند. این برنامه در ضمیمه A این مقاله ارائه شده است. در این برنامه ماتریس‌های طرح بر اساس تکرار و ژنوتیپ طراحی شدند و برای برآورد میانگین از بردار یکان به جای ماتریس X استفاده شد. نتایج هر دو برنامه SAS مشابه هم می‌باشد.

نتیجه‌گیری

به‌نژادگران گیاه و متخصصین ژنتیک کمی وراثت‌پذیری، همبستگی‌های ژنتیکی و مولفه‌های واریانس را به طور عمومی بر اساس تجزیه واریانس به روش حداقل مربعات برآورد می‌کنند. سپس میانگین مربعات را با امید ریاضی آن (یک ترکیب خطی از مولفه‌هی واریانس) برابر قرار داده و مولفه‌های واریانس به صورت توابع جبری برآورد می‌شوند. این روش برآورد را روش گشتاوری نیز می‌نامند (Milliken and Johnson, 1992). قوانین استخراج مولفه‌های واریانس از امید ریاضی میانگین مربعات توسط استیل و همکاران (Steel and Torrie, 1997) و به طور اختصاصی در طرح‌های ژنتیکی توسط هالور (Hallauer *et al.*, 2010) ارائه شده است.

در مواردی که داده‌ها متعادل هستند، خطای استاندارد مولفه‌های واریانس به راحتی قابل برآورد است و روش‌های زیادی برای برآورد میزان وراثت‌پذیری و حدود اطمینان پارامترهای ژنتیکی ارائه شده است (Holland *et al.*, 2003). به طور عمومی، طرح‌های آزمایشات در کشاورزی و به‌نژادی گیاهان متعادل هستند، اما واقعیت این است که در خیلی از موارد به ویژه در طرح‌های ژنتیکی، محققین

دچار کمبود بذر در برخی از لاین‌ها یا خانواده می‌شوند، یا ممکن است یک کرت را از دست بدهند و یا داده گم شده

جدول ۲- برآورد واریانس تکرار، ژنوتیپ و باقیمانده با سه روش ANOVA، ML و REML در دو صفت گندم

Table 2. Estimation of replication, genotype and residual with ANOVA, ML and REML estimators in two characteristics of wheat

منبع تغییر Source of variation	واریانس عملکرد دانه Grain yield variance			واریانس وزن دانه در سنبله Kernel weight per spike variance		
برآوردگر Estimator	آنوا ANOVA	حداکثر درستنمایی ML	حداکثر درستنمایی محدود شده REML	آنوا ANOVA	حداکثر درستنمایی ML	حداکثر درستنمایی محدود شده REML
تکرار REP	39724	25011	39724	0.01	0.00	0.00
ژنوتیپ Genotype	642759	619548	642759	0.83	0.75	0.80
باقیمانده Residual	1105858	1110443	1105858	2.63	2.65	2.65
واریانس فنوتیپی σ_p^2	1748617	1729991	1748617	3.460	3.400	3.450
وراثت‌پذیری h_b^2	36.758	35.812	36.758	23.988	22.059	23.188

به دلایل مختلف در مرحله کاشت، داشت و یا برداشت داشته باشند. این نامتعادلی داده باعث تغییر ضرایب مولفه‌های واریانس در امیدریاضی میانگین مربعات می‌شود و منجر به از دست رفتن عدم استقلال بین میانگین مربعات فاکتورها شود. اگرچه تغییرات ضرایب امیدریاضی میانگین مربعات با استفاده از روش میلیکن و جوهانسون (Milliken and Johnson, 1992) قابل تصحیح هستند و ضرایب درست مولفه‌های واریانس در امیدریاضی میانگین مربعات با استفاده از نرم‌افزارهایی مانند SAS قابل محاسبه هستند (Rawlings *et al.*, 2001). علاوه بر این روش‌های تخمینی برای کنترل داده‌های گم شده در طرح‌های آزمایشات با استفاده از تجزیه کوواریانس توسط محققین ارائه شده است (Nyquist and Baker, 1991; Steel and Torrie, 1997). همه این روش‌ها ممکن است در داده‌های گم شده بهترین برآورد نااریب را بدست بدهند؛ اما منجر به کمترین واریانس برآورد شده نشوند. علاوه بر این ویژگی، توزیع آنها شناخته شده نیست به طوری که برآوردهای دقیق مولفه‌های واریانس و وراثت‌پذیری آنها قابل اتکا نیست (Milliken and Johnson, 1992). از طرفی حذف خانواده‌هایی که دارای داده گم شده هستند جهت رسیدن به داده متعادل بازدهی و دقت آزمایش ها را پایین می‌آورد. به عقیده سیرل و همکاران (Searle *et al.*, 2006) روش‌های ML و REML برای برآورد مولفه‌های واریانس داده‌های نامتعادل مفیدتر هستند، زیرا اصل درستنمایی که در پیش‌فرض این دو روش قرار دارد دارای خصوصیات، یکنواختی، نرمال بودن مجانبی برآوردها است. این ویژگی برای برآورد حدود اطمینان‌ها و آزمون فرضیات پارامترها مناسب است. از طرفی برآوردگرهای ML و REML بر مبنای نرمال بودن داده‌ها استوارند، اما در مواردی که این فرضیات صادق نیستند و تعداد داده‌ها زیاد نیست، مزیت این روش‌ها خنثی می‌شود، زیرا ویژگی مجانبی واریانس کواریانس زمانی اعتبار بیشتر دارد که تعداد داده‌ها زیاد باشند.

روش‌های ML و REML نیازمند محاسبات زیادی هستند و عمدتا از الگوریتم‌های تکرارشونده برای برآورد مولفه‌های واریانس استفاده می‌شود که این یکی از معایب این دو روش محسوب می‌شد. ولی امروزه به دلیل نرم افزارهای پیشرفته این عیب وجود ندارد. در بین روش ML

و REML هر دو دارای ویژگی‌های مثبت یکسانی هستند که بر مبنای حداکثر درستنمایی می‌باشند. در روش ML اثرات ثابت نیز برآورد می‌شوند، اما در روش REML حداکثر درستنمایی در بخش تصادفی مدل حاصل می‌شود. اما به طور کلی برآورد پارامترها در داده‌های کوچک به روش ML با اریب همراه است ولی در روش REML برآوردها در داده‌های متعادل مشابه ANOVA است یعنی کمترین واریانس ممکن را دارا هستند. به طور کلی روش‌های REML و ML در زمانی که مولفه‌های واریانس به روش ANOVA منفی برآورد می‌شوند توصیه می‌گردند. همچنین در داده‌های بزرگ و نامتعادل توصیه می‌شود که از روش‌های درستنمایی مانند ML و REML استفاده شود.

ضمیمه

Appendix

```
data exam;
input REP    Gen    KWS    GY;
cards;
1    1    4.5    1808.33
2    1    .      1786.84
3    1    5      2405
1    2    7.1    3792.5
2    2    6.3    4684.21
3    2    5.1    1847.5
1    3    5.1    1991.94
2    3    6.1    2405.26
3    3    6.7    3237.5
more data lines....
;
proc iml;
start reml(x,y,rand,init,variance,nr,n,cov,z1,z2,z3,model,neg);
if nrow(neg)=0 then neg=j(nr,1,0);
do i = 1 to nr;
if init[i,]<0||neg[i,]=1 then do;
init[i,]=0;
neg[i,]=1;
end;
end;
V=z1*init[1]*z1`+z2*init[2]*z2`+z3*init[3]*z3`;
vi=inv(V);
PY=vi*(y-x*sweep(x`*vi*x)*x`*vi*y);
if model=1 then p=vi;
else p=vi-vi*x*sweep(x`*vi*x)*x`*vi;
info=j(nr,nr,0);
ss=j(nr,1,0);
```

```
do i=1 to nr;
Zi=design(rand[1:n,i]);
do j=1 to nr;
Zj=design(rand[1:n,j]);
info[i,j]=trace(P*Zi*Zi`*P*Zj*zj`);
info[j,i]=info[i,j];
end;
ss[i]=(PY`*zi*zi`*PY);
end;
do i=1 to nr;
if neg[i]=1 then do;
info[i,]=j(1,nr,0);
info[,i]=j(nr,1,0);
ss[i]=0;
end;
end;
cov=sweep(info);
variance=cov*ss;
finish reml;
store module=(reml);
use exam;
read all;
READ ALL VAR {GY} INTO y;
READ ALL VAR {Rep} INTO Rep;
READ ALL VAR {GEN} INTO GEN;
n=nrow(y);
x=j(n,1,1);
e=(1:n)`;
z1=design(rep);
z2=design(gen);
z3=i(n);
rand=(rep||gen||e);
nr=ncol(rand);
init={1.5,.1,.5};
do iter=1 to 20;
call reml(x,y,rand,init,variance,nr,n,cov,z1,z2,z3,2,neg);
init=variance;
end;
covariace=cov*2;
print variance,covariace;
run;
```

```
quit;
proc mixed data=exam method=REML asycov;/*REML can be changed to type1-3 or ML*/
class rep Gen;
model gy=;
random rep gen;
run;
```

References

Acquaah, G. (2009). *Principles of Plant Genetics and Breeding*. John Wiley & Sons, New Jersey, USA.

Aitkin, M. (1978). The analysis of unbalanced cross-classifications. *Journal of the Royal Statistical Society Series A (General)*, **141**: 195-223.

Akbarpour, O., Dehghani, H., Rousta, M., J, and Amini, A. (2015a). Evaluation of some properties of Iranian wheat genotypes in normal and salt-stressed conditions using Restricted Maximum Likelihood (REML). *Iranian Journal of Field Crop Science*, **46**: 57-69 (In Persian).

Akbarpour, O., Dehghani, H. and Rousta, M.J. (2015b). Evaluation of salt stress of Iranian wheat germplasm under field conditions. *Crop and Pasture Science*, **66**: 770-781.

Anderson, R.L. and Bancroft, T.A. (1952). *Statistical Theory in Research*. McGraw-Hill Book Company, Inc, New York, USA.

Barnett, V. (1966). Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots. *Biometrika*, **53**: 151-165.

Bertsekas, D.P. and Tsitsiklis, J.N. (2008). *Introduction to Probability*. American Mathematical Society Press, Providence, USA.

Bishop, C. (1992). *Exact Ccalculation of the Hessian matrix for the Multilayer Perceptron*. MIT Press, Massachusetts, USA.

Casella, G. and Berger, R.L. (1990). *Statistical Inference*. Cole Advanced Books & Software, Pacific Grove, California, USA.

Der, G. and Everitt, B. (2002). *A handbook of statistical analyses using SAS*, Chapman and Hall, London, UK.

Falconer, D. and Mackay, T. (1996). *Introduction to Quantitative Genetics*. Longman, London, UK.

Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, London, UK.

Graybill, F.A. and Iyer, H.K. (1994). *Regression analysis: Concepts and Applications*, Belmont, California, USA.

Graybill, F.A. and Wortham, A. (1956). A note on uniformly best unbiased estimators for variance components. *Journal of the American Statistical Association*, **51**: 266-268.

Hallauer, A.R., Carena, M.J. and Miranda Filho, J.d. (2010). *Quantitative Genetics in Maize Breeding*, Springer Science & Business Media, Berlin, DE.

Hartley, H.O. and Rao, J.N. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, **54**: 93-108.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**: 320-338.

Henderson, C.R. (1984). *Application of Linear Models in Animal Breeding*. University of Guelph, Ontario, California, USA.

Henderson, C., Searle, S. and Schaeffer, L. (1974). The invariance and calculation of method 2 for estimating variance components. *Biometrics*, **30**: 583-588.

Holland, J.B., Nyquist, W.E. and Cervantes-Martínez, C.T. (2003). Estimating and interpreting heritability for plant breeding: an update. *Plant Breeding Reviews*, **22**: 9-112.

Ismaili, A., Karami, F., Akbarpour, O. and Rezaei Nejad, A. (2016). Estimation of Genotypic Correlation and Heritability of Apricot Traits, Using Restricted Maximum Likelihood in Repeated Measures Data. *Canadian Journal of Plant Science*, **96**: 439-447.

Kempthorne, O. (1968). Discussion of Searle. *Biometrics*, **24**: 782-784.

King, G. (1998). *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Cambridge University Press, Londan, UK.

Lehmann, E.L. and Casella, G. (2006). *Theory of Point Estimation*. Springer Science & Business Media, Berlin, DE.

Littell, R., Milliken, G., Stroup, W. and Wolfinger, R. (2006). *SAS system for Mixed Models*. SAS Institute Inc, Cary, North Carolina, USA.

Lynch, M. and Walsh, B. (1998). *Genetics and analysis of Quantitative Traits.* Sinauer Associates, Massachusetts, USA.

Milliken, G.A. and Johnson, D.E. (1992). *Analysis of Messy Data. Volume I: Designed experiments.*, Chapman & Hall, New York, USA.

Mood, A.M., Graybill, F.A. and Boes, D.C. (1974). *Introduction to the Theory of Statistics. 3rd ed*, USA.

Nelder, J. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society Series A (General)*, **140**: 48-77.

Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W. (2004). *Applied Linear Statistical Models*, Irwin Chicago, USA.

Nyquist, W.E. and Baker, R. (1991). Estimation of heritability and prediction of selection response in plant populations. *Critical Reviews in Plant Sciences*, **10**: 235-322.

Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58(3)**: 545-554.

Rawlings, J.O., Pantula, S.G. and Dickey, D.A. (2001). *Applied Regression Analysis: A Research Tool*. Springer Science & Business Media, Berlin, DE.

Sahai, H. and Miguel , M.O. (2004). *Analysis of Variance for Random Models Volume I: Balanced Data Theory, Methods, Applications and Data Analysis*. Business Media New York, USA.

Sahai, H. and Ojeda, M.M. (2004). *Analysis of Variance for Random Models, Volume 2: Unbalanced Data: Theory, Methods, Applications, and Data Analysis*, Springer Science & Business Media, Berlin, DE.

Sahai, H. and Thompson, W.O. (1973). The Teacher's Corner: Non-Negative Maximum Likelihood Estimators of Variance Components in a Simple Linear Model. *The American Statistician*, **27**: 112-113.

Schaeffer, L. (1998). *Variance Component Estimation Course Notes*. University of New England, Armidale, NSW.

Searle, S., Casella, G. and McCulloch, C. (2006). *Variance Components*. John Wiley and Sons, New York, USA.

Searle, S.R. (1971). *Linear Models*, John Wiley & Sons, New Jersey, USA.

Small, C.G., Wang, J. and Yang, Z. (2000). Eliminating multiple root problems in estimation *Statistical Science*, **15**: 313-341.

Sorensen, D. and Gianola, D. (2007). *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*, Springer Science & Business Media, Berlin, DE.

Steel, R. and Torrie, J. (1997). *Principles and Procedures of Statistics. A Biometrical Approach*, McGraw-Hill Book Company In Company, New York, USA.

Thompson, W. and Moore, J.R. (1963). Non-negative estimates of variance components. *Technometrics*, **5**: 441-449.

Valizadeh, S. (2014). *Evaluation of genotypic variation of wheat genotypes under low water stress in Khorramabad climate conditions*. Lorestan Univarsity, Lorestan, IR.

Yang, R.C. (2010). Towards understanding and use of mixed-model analysis of agricultural experiments. *Canadian Journal of Plant Science*, **90**: 605-627.

Plant Genetic Researches, Vol. 4, No. 1, 2017

Application of Variance Components Estimators in Plant Breeding

(Review Article)

Omidali Akbarpour*

Assistant Professor, Department of Agronomy and Plant Breeding, Faculty of Agriculture, Lorestan University, Khorramabad, Iran

(Received: January 12, 2017– Accepted: May 17, 2017)

Abstract

To conduct any breeding program, understanding of the genetic structure of traits and effect of environment and genetic by environment interaction as well as effects of random or fixed in the analysis of results is essential. Subsequently, analysis of variances and variance components are important in plant and animal breeding. The ANOVA is one of the best estimators for variance components. But this estimator is not preferred to maximum likelihood (ML) and Restricted Maximum Likelihood (REML) methods when variance components are negatively estimated and unbalanced datasets arise. Therefore, the objective of this research is a review of comparison of estimates of variance components using ANOVA, ML and REML method in linear mixed models using experimental data.

Keywords: Estimation, Analysis of Variance, Maximum Likelihood, Restricted Maximum Likelihood, Variance components, Regression

* **Corresponding Author, E-mail: akbarpour.oa@lu.ac.ir**