Chapter 4. Gauss-Markov Model

4.1 Model Assumptions

So far we've approached the linear model only as a method of mathematical approximation. In this chapter, we pose the Gauss-Markov model which embodies the most common assumptions for the statistical approach to the linear model, leading to the Gauss-Markov Theorem. The Gauss-Markov model takes the form

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{e} \tag{4.1}$$

where **y** is the (N by 1) vector of observed responses, and **X** is the (N by p) known design matrix. As before, the coefficient vector **b** is unknown and to be determined or estimated. The main features of the Gauss-Markov model are the assumptions on the error **e**:

$$E(\mathbf{e}) = \mathbf{0} \quad \text{and } \mathrm{Cov}(\mathbf{e}) = \sigma^2\mathbf{I}_N. \tag{4.2}$$

The notation for expectation and covariances above can be rewritten component by component:

$(\ E(\mathbf{e})\ )_i = i^{th}$ component of $E(\mathbf{e}) = E(e_i)$

$(\ \mathrm{Cov}(\mathbf{e})\ )_{ij} = i,j^{th}$ element of covariance matrix $= \mathrm{Cov}(e_i, e_j)$,

so that the Gauss-Markov assumptions can be rewritten as

$E(e_i) = 0,\ i = 1, ..., N$

$$\mathrm{Cov}(e_i, e_j) = \begin{cases} \sigma^2 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

that is, the errors in the model have a zero mean, constant variance, and are uncorrelated. An alternative view of the Gauss-Markov model does not employ the error vector **e**:

$$E(\mathbf{y}) = \mathbf{X}\,\mathbf{b},\ \mathrm{Cov}(\mathbf{y}) = \sigma^2\mathbf{I}_N.$$

The assumptions in the Gauss-Markov model are easily acceptable for most practical problems and deviations from these assumptions will be considered in more detail later.

Before we get to the Gauss-Markov Theorem, we will need some simple tools for conveniently working with means and variances of vector random variables, in particular, a linear combination of variables $\mathbf{a}^T\mathbf{y}$ where **a** is a fixed vector. The rules are:

i) $E(\mathbf{a}^T\mathbf{y}) = \mathbf{a}^T E(\mathbf{y})$

ii) $\mathrm{Var}(\mathbf{a}^T\mathbf{y}) = \mathbf{a}^T\mathrm{Cov}(\mathbf{y})\,\mathbf{a}$

iii) $\mathrm{Cov}(\mathbf{a}^T\mathbf{y}, \mathbf{c}^T\mathbf{y}) = \mathbf{a}^T\mathrm{Cov}(\mathbf{y})\,\mathbf{c}$, for fixed **a, c**

iv) $\mathrm{Cov}(\mathbf{A}^T\mathbf{y}) = \mathbf{A}^T\mathrm{Cov}(\mathbf{y})\,\mathbf{A}$, for fixed matrix **A**.

The key result is (iii), from which (ii) and (iv) follow easily, and just algebraic bookkeeping is necessary:

$$\mathrm{Cov}(\mathbf{a}^T\mathbf{y}, \mathbf{c}^T\mathbf{y}) = \mathrm{Cov}\left(\sum_i a_i y_i, \sum_j c_j y_j\right)$$

$$= \sum_i a_i \mathrm{Cov}\left(y_i, \sum_j c_j y_j\right) = \sum_i \sum_j a_i c_j \mathrm{Cov}(y_i, y_j) = \mathbf{a}^T\mathrm{Cov}(\mathbf{y})\,\mathbf{c}.$$

Example 4.1. Variance and covariance calculations.

Let $\mathrm{Cov}(\mathbf{y}) = \begin{bmatrix} 4 & 2 & 4 \\ 2 & 5 & 0 \\ 4 & 0 & 25 \end{bmatrix}$, $\mathbf{c} = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$, $\mathbf{a} = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}$, then we have

$$\text{Var}(y_1 - y_2 + 2y_3) = \text{Var}(\mathbf{c}^T\mathbf{y}) = \mathbf{c}^T\text{Cov}(\mathbf{y})\,\mathbf{c} =$$
$$[\,1 \quad -1 \quad 2\,]\begin{bmatrix} 4 & 2 & 4 \\ 2 & 5 & 0 \\ 4 & 0 & 25 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} = [\,1 \quad -1 \quad 2\,]\begin{bmatrix} 10 \\ -3 \\ 54 \end{bmatrix} = 121.$$ Also

$$\text{Cov}(2y_1 + y_3, y_1 - y_2 + 2y_3)\ \text{Cov}(\mathbf{a}^T\mathbf{y}, \mathbf{c}^T\mathbf{y}) = \mathbf{a}^T\text{Cov}(\mathbf{y})\,\mathbf{c}$$
$$= [\,2 \quad 0 \quad 1\,]\begin{bmatrix} 4 & 2 & 4 \\ 2 & 5 & 0 \\ 4 & 0 & 25 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} = [\,2 \quad 0 \quad 1\,]\begin{bmatrix} 10 \\ -3 \\ 54 \end{bmatrix} = 74.$$

Example 4.2. Variance of a least squares estimator. Let $\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{e}$, with the Gauss-Markov assumptions on $\mathbf{e}$, so that $\text{Cov}(\mathbf{y}) = \sigma^2\mathbf{I}_N$, and let $\boldsymbol{\lambda}^T\mathbf{b}$ be an estimable function. Then the variance of the least squares estimator follows the calculation (see Exercise 4.2)

$$\text{Var}(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) = \text{Var}(\,\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y}) = \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\text{Cov}(\mathbf{y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{gT}\boldsymbol{\lambda}$$
$$= \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T(\sigma^2\mathbf{I}_N)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{gT}\boldsymbol{\lambda} = \sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\boldsymbol{\lambda}$$

Example 4.3. In the simple linear regression case in Example 3.1, we solved the normal equations and found the usual slope estimate

$$\hat{b}_2 = \sum_{i=1}^{N}(x_i - \bar{x})y_i / \sum_{i=1}^{N}(x_i - \bar{x})^2.$$

To compute the mean and variance of $\hat{b}_2$, we have two routes. One route is to treat $\hat{b}_2$ as a linear combination of the y_i's:

$$\hat{b}_2 = \sum_{i=1}^{N}(x_i - \bar{x})y_i / \sum_{i=1}^{N}(x_i - \bar{x})^2 = \sum_{i=1}^{N}[(x_i - \bar{x})/S_{xx}]\,y_i \text{ where } S_{xx} = \sum_{i=1}^{N}(x_i - \bar{x})^2.$$

Then $E(\hat{b}_2) = \sum_{i=1}^{N}[(x_i - \bar{x})/S_{xx}]\,E(y_i) = \sum_{i=1}^{N}[(x_i - \bar{x})/S_{xx}]\,(b_1 + b_2x_i) = \sum_{i=1}^{N}(x_i - \bar{x})x_i/S_{xx}b_2 = b_2$ and the usual slope estimate is unbiased. Its variance can be found from the following algebra:

$$\text{Var}(\hat{b}_2) = \sum_{i=1}^{N}[(x_i - \bar{x})/S_{xx}] \sum_{j=1}^{N}[(x_j - \bar{x})/S_{xx}]\,\text{Cov}(y_i, y_j)$$
$$= \sum_{i=1}^{N}[(x_i - \bar{x})/S_{xx}]^2\text{Var}(y_i) = \sigma^2\sum_{i=1}^{N}(x_i - \bar{x})^2/S_{xx}^2 = \sigma^2/S_{xx},$$

employing the usual assumptions of constant variance and uncorrelated observations. The other route would be to follow the algebra in Example 4.2 above, and find the (2, 2) element of

$$\sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2\begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} = \frac{\sigma^2}{NS_{xx}}\begin{bmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & N \end{bmatrix}.$$

4.2. The Gauss-Markov Theorem

The goal throughout this chapter has been to show that the least squares estimators derived back in Section 3.2 are the 'best' estimators in some sense. The Gauss-Markov Model that we've been talking about consists of just those set of assumptions that are sufficient. Recall that a linear estimator takes the form $c + \mathbf{a}^T\mathbf{y}$, and will be unbiased for estimable $\boldsymbol{\lambda}^T\mathbf{b}$ when

$$E(\,c + \mathbf{a}^T\mathbf{y}) = \boldsymbol{\lambda}^T\mathbf{b}$$

for all $\mathbf{b}$, which leads to $c = 0$, and $\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a}$. In our One-Way ANOVA Model, $y_{ij} = \mu + \alpha_i + e_{ij}$, then y_{11}, $(y_{11} + 2y_{12})/3$ and $\overline{y}_{1\cdot}$ are all unbiased estimators of $\mu + \alpha_1$. In Simple Linear Regression, $y_i = \beta_0 + \beta_1 x_i + e_i$, with $x_i = i$, say, then $y_2 - y_1$ or $y_3 - y_2$ are both unbiased estimators of β_1. Which is a better estimator and how shall we measure? When two estimators are both unbiased, then the estimator with smaller variances is better, since variance is a measure of variability around its mean. Our least squares estimator of $\boldsymbol{\lambda}^T\mathbf{b}$ is $\boldsymbol{\lambda}^T\hat{\mathbf{b}}$ where $\hat{\mathbf{b}}$ is a solution to the normal equations. Recall that if we construct all solutions to the normal equations,

$$\hat{\mathbf{b}}(\mathbf{z}) = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y} + (\mathbf{I} - (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X})\mathbf{z}$$

then for estimable $\boldsymbol{\lambda}^T\mathbf{b}$, $\boldsymbol{\lambda}^T\hat{\mathbf{b}}(\mathbf{z})$ is constant for all values of $\mathbf{z}$ -- all solutions to the normal equations lead to the same least squares estimator.

Theorem 4.1. (Gauss-Markov Theorem) Under the assumptions of the Gauss-Markov Model,

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{e}, \text{ where } E(\mathbf{e}) = \mathbf{0} \text{ and } \mathrm{Cov}(\mathbf{e}) = \sigma^2\mathbf{I}_N,$$

if $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable, then $\boldsymbol{\lambda}^T\hat{\mathbf{b}}$ is the best (minimum variance) linear unbiased estimator (BLUE) of $\boldsymbol{\lambda}^T\mathbf{b}$, where $\hat{\mathbf{b}}$ solves the normal equations

$$\mathbf{X}^T\mathbf{X}\,\mathbf{b} = \mathbf{X}^T\mathbf{y}.$$

Proof: Suppose $c + \mathbf{d}^T\mathbf{y}$ is another unbiased estimator of $\boldsymbol{\lambda}^T\mathbf{b}$. Then $c = 0$ and $\mathbf{d}^T\mathbf{X} = \boldsymbol{\lambda}^T$ since $E(c + \mathbf{d}^T\mathbf{y}) = c + \mathbf{d}^T\mathbf{X}\mathbf{b} = \boldsymbol{\lambda}^T\mathbf{b}$ for all $\mathbf{b}$. Now,

$$\begin{aligned}
\mathrm{Var}(c + \mathbf{d}^T\mathbf{y}) &= \mathrm{Var}(\mathbf{d}^T\mathbf{y}) = \mathrm{Var}(\boldsymbol{\lambda}^T\hat{\mathbf{b}} + \mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) \\
&= \mathrm{Var}(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) + \mathrm{Var}(\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) + 2\,\mathrm{Cov}(\boldsymbol{\lambda}^T\hat{\mathbf{b}}, \mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) \\
&= \mathrm{Var}(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) + \mathrm{Var}(\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) + 2\,\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\,(\sigma^2\mathbf{I}_N)\,(\,\mathbf{d} - \mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{gT}\boldsymbol{\lambda}\,) \\
&= \mathrm{Var}(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) + \mathrm{Var}(\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) + 2\sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g(\mathbf{X}^T\mathbf{d} - \mathbf{X}^T\mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{gT}\boldsymbol{\lambda}\,) \\
&= \mathrm{Var}(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) + \mathrm{Var}(\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) + 2\,\sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g(\,\boldsymbol{\lambda} - \boldsymbol{\lambda}\,) \\
&= \mathrm{Var}(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) + \mathrm{Var}(\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}})
\end{aligned}$$

Since $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable, we have $\boldsymbol{\lambda} \in \mathcal{C}(\mathbf{X}^T)$ and a projection $\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{gT}$ onto it, hence $\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{gT}\boldsymbol{\lambda} = \boldsymbol{\lambda}$. So $\mathrm{Var}(\mathbf{d}^T\mathbf{y}) \geq \mathrm{Var}(\boldsymbol{\lambda}^T\hat{\mathbf{b}})$, with equality iff

$$\mathrm{Var}(\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) = \mathrm{Var}(\,(\,\mathbf{d} - \mathbf{X}\,(\mathbf{X}^T\mathbf{X})^g\boldsymbol{\lambda}\,)^T\mathbf{y}\,) = \sigma^2\,\|\,\mathbf{d} - \mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{gT}\boldsymbol{\lambda}\,\|^2 = 0.$$

So equality occurs (and an estimator with equal variance) iff $\mathbf{d} = \mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{gT}\boldsymbol{\lambda}$ or $\mathbf{d}^T\mathbf{y} = \boldsymbol{\lambda}^T\hat{\mathbf{b}}$. In other words, the best linear unbiased estimator is unique. □

Note that in the proof above, the crucial step is showing

$$\mathrm{Cov}(\boldsymbol{\lambda}^T\hat{\mathbf{b}}, \mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) = 0, \text{ where } \mathbf{a}^T\mathbf{X} = \boldsymbol{\lambda}^T.$$

Now what is $\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}$ estimating? Notice that

$$E(\,\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}) = \mathbf{a}^T\mathbf{X}\mathbf{b} - \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}\mathbf{b} = \boldsymbol{\lambda}^T\mathbf{b} - \boldsymbol{\lambda}^T\mathbf{b} = 0$$

so that $\mathbf{d}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}$ is an *unbiased estimator of zero*, and the Best Linear Unbiased Estimator $\lambda^T\hat{\mathbf{b}}$ is uncorrelated with it.

Result 4.1. The BLUE $\boldsymbol{\lambda}^T\hat{\mathbf{b}}$ of estimable $\boldsymbol{\lambda}^T\mathbf{b}$ is uncorrelated with all unbiased estimators of zero.

Proof: First, characterize unbiased estimators of zero as $c + \mathbf{a}^T\mathbf{y}$ such that

$$E(c + \mathbf{a}^T\mathbf{y}) = c + \mathbf{a}^T\mathbf{X}\mathbf{b} = 0 \text{ for all } \mathbf{b},$$

or $c = 0$ and $\mathbf{X}^T\mathbf{a} = \mathbf{0}$, or $\mathbf{a} \in \mathcal{N}(\mathbf{X}^T)$. Computing the covariance between $\boldsymbol{\lambda}^T\hat{\mathbf{b}}$ and $\mathbf{a}^T\mathbf{y}$ we have

$$\text{Cov}(\boldsymbol{\lambda}^T\hat{\mathbf{b}}, \mathbf{a}^T\mathbf{y}) = \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T \text{Cov}(\mathbf{y})\, \mathbf{a} = \sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\, \mathbf{X}^T\mathbf{a} = 0. \square$$

The Gauss-Markov Theorem can be extended to the vector case. Let the columns $\boldsymbol{\lambda}^{(j)}$ of the matrix $\boldsymbol{\Lambda}$ be linear independent such that $\boldsymbol{\lambda}^{(j)\,T}\mathbf{b}$ are linearly independent estimable functions, then

$$\text{Cov}(\boldsymbol{\Lambda}^T\hat{\mathbf{b}}) = \text{Cov}(\boldsymbol{\Lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y}) = \sigma^2\boldsymbol{\Lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{gT}\boldsymbol{\Lambda} = \sigma^2\boldsymbol{\Lambda}^T(\mathbf{X}^T\mathbf{X})^g\boldsymbol{\Lambda}.$$

If we have any set of unbiased estimator $\boldsymbol{\Lambda}^T\mathbf{b}$, say $\mathbf{C}^T\mathbf{y}$, that is, $E(\mathbf{C}^T\mathbf{y}) = \boldsymbol{\Lambda}^T\mathbf{b}$ for all $\mathbf{b}$, then the difference in their covariance matrices

$$\text{Cov}(\mathbf{C}^T\mathbf{y}) - \text{Cov}(\boldsymbol{\Lambda}^T\hat{\mathbf{b}})$$

is nonnegative definite. If $\mathbf{X}$ has full column rank and applying $\boldsymbol{\Lambda} = \mathbf{I}$, we have $\text{Cov}(\hat{\mathbf{b}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Moreover, in the full column rank model, the Gauss-Markov Theorem says that if $\tilde{\mathbf{b}}$ is any other unbiased estimator of $\mathbf{b}$, then

$$\text{Cov}(\tilde{\mathbf{b}}) - \text{Cov}(\hat{\mathbf{b}})$$

is nonnegative definite.

4.3. Variance Estimation

Throughout this discussion, our focus has been on estimating linear functions of the coefficients $\mathbf{b}$ and no attention has been paid to the other unknown parameter σ^2. So far, we've used $\mathbf{P}_\mathbf{X}\mathbf{y}$ to estimate $\mathbf{b}$, as another version of the normal equations is $\mathbf{X}\mathbf{b} = \mathbf{P}_\mathbf{X}\mathbf{y}$. As the reader might guess, we will use $(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{y}$, or, more specifically its sum of squares $\text{SSE} = \|(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{y}\|^2$ to estimate σ^2. To construct an unbiased estimator for σ^2, we will use the following lemma.

Lemma 4.1. Let $\mathbf{Z}$ be a vector random variable with $E(\mathbf{Z}) = \boldsymbol{\mu}$ and $\text{Cov}(\mathbf{Z}) = \boldsymbol{\Sigma}$. Then $E(\mathbf{Z}^T\mathbf{A}\mathbf{Z}) = \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} + \text{tr}(\mathbf{A}\boldsymbol{\Sigma})$.

Proof: Note that

$$E(Z_i - \mu_i)(Z_j - \mu_j) = \Sigma_{ij} = E(Z_iZ_j) - \mu_iE(Z_j) - \mu_jE(Z_i) + \mu_i\mu_j$$

so that

$$[\, E(\mathbf{Z}\,\mathbf{Z}^T)\,]_{ij} = E(Z_i - \mu_i)(Z_j - \mu_j) + \mu_i\mu_j = \Sigma_{ij} + \mu_i\mu_j,$$

or $E(\mathbf{Z}\,\mathbf{Z}^T) = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}$. Now using the linearity of trace and expectation operations, we have

$$E(\mathbf{Z}^T\mathbf{A}\,\mathbf{Z}) = E(\text{tr}(\mathbf{Z}^T\mathbf{A}\,\mathbf{Z})) = E(\text{tr}(\mathbf{A}\,\mathbf{Z}\,\mathbf{Z}^T)) = \text{tr}(\mathbf{A}\,E(\mathbf{Z}\,\mathbf{Z}^T)) = \text{tr}(\mathbf{A}\,(\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}))$$
$$= \text{tr}(\mathbf{A}\boldsymbol{\mu}\boldsymbol{\mu}^T) + \text{tr}(\mathbf{A}\boldsymbol{\Sigma}) = \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} + \text{tr}(\mathbf{A}\boldsymbol{\Sigma}). \quad \square$$

Result 4.2. Consider the Gauss-Markov Model given in (4.1) and (4.2). An unbiased estimator of σ^2 is $\hat{\sigma}^2 = \text{SSE}/(N - r)$, where $\text{SSE} = \hat{\mathbf{e}}^T\hat{\mathbf{e}} = \mathbf{y}^T(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{y}$ and $r = \text{rank}(\mathbf{X})$.

Proof: Since $E(\mathbf{y}) = \mathbf{X}\mathbf{b}$, $\text{Cov}(\mathbf{y}) = \sigma^2\mathbf{I}_N$, applying Lemma 4.1 above yields

$E(\mathbf{y}^T(\mathbf{I}-\mathbf{P_X})\mathbf{y}) = (\mathbf{Xb})^T(\mathbf{I}-\mathbf{P_X})\mathbf{Xb} + \text{tr}((\mathbf{I}-\mathbf{P_X})(\sigma^2\mathbf{I}_N)) = \sigma^2(N-r)$.
Dividing by the degrees of freedom parameter $\text{tr}(\mathbf{I}-\mathbf{P_X}) = (N-r)$ associated with $SSE = \mathbf{y}^T(\mathbf{I}-\mathbf{P_X})\mathbf{y}$ gives the desired result.□

If we examine the regression sum of squares, $SSR = \hat{\mathbf{y}}^T\hat{\mathbf{y}} = \mathbf{y}^T\mathbf{P_X}\,\mathbf{y}$, we find
$E(SSR) = E(\mathbf{y}^T\mathbf{P_X}\,\mathbf{y}) = \mathbf{b}^T\mathbf{X}^T\mathbf{P_X}\,\mathbf{Xb} + \text{tr}(\mathbf{P_X}\,\sigma^2\mathbf{I}_N) = \|\mathbf{Xb}\|^2 + r\,\sigma^2$.

Little else can be said about $\hat{\sigma}^2$ without further distributional assumptions, even when e_i are iid (independent, identically distributed) since its variance depends on the third and fourth moments of the underlying distribution. The algebra for computing $\text{Var}(\hat{\sigma}^2)$ in terms of the third and fourth moments is given in an addendum to this Chapter.

4.4. Implications of Model Selection

In the application of linear models, sometimes the appropriate statistical model is not obvious. Even in designed experiments, researchers will debate whether to include certain interactions into the model. In the case of observational studies with continuous covariates the problem of model selection becomes quite difficult and leads to many diagnostics and proposed methodologies (see, e.g. Rawlings, Pantula, & Dickey, 1998). In the context of this book, however, we are concerned with the theoretical aspects that drive the practical model selection. At this point an important distinction must be made between the underlying true model for the data, and the model that the researcher is employing. In practice, of course, we cannot know the true model.

The issue of model selection is often described as steering between two unpleasant situations: *overfitting*, that is, including explanatory variables which are not needed, and *underfitting*, not including an important explanatory variable. We will focus on the implications of these two situations.

In the case of underfitting, also referred to as *misspecification*, we can write the model for the truth as

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{\eta} + \mathbf{e} \tag{4.3}$$

where $\mathbf{X}$ is the design matrix for the model that the researcher is using, and $\boldsymbol{\eta}$ includes the omitted variables and their coefficients. Assume, as usual, $E(\mathbf{e}) = \mathbf{0}$ and $\text{Cov}(\mathbf{e}) = \sigma^2\mathbf{I}_N$, and notice that if $E(\mathbf{e})$ were not zero, $\boldsymbol{\eta}$ would capture its effect. Consider first the least squares estimators:

$$E(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) = \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^TE(\mathbf{y}) = \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T(\mathbf{Xb}+\boldsymbol{\eta}) = \boldsymbol{\lambda}^T\mathbf{b} + \mathbf{a}^T\mathbf{P_X}\boldsymbol{\eta}$$

where $\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a}$. When the model is misspecified, the least squares estimators are *biased*:

$$E(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) - \boldsymbol{\lambda}^T\mathbf{b} = \mathbf{a}^T\mathbf{P_X}\boldsymbol{\eta} = \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\boldsymbol{\eta}$$

where the bias above depends on:

- the magnitude of the misspecified effect $\boldsymbol{\eta}$
- how much of that effect lies in $\mathcal{C}(\mathbf{X})$, $\mathbf{P_X}\boldsymbol{\eta}$, and
- how much does it relate to the function at hand $\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\boldsymbol{\eta}$.

If the missing signal $\boldsymbol{\eta}$ is orthogonal to $\mathcal{C}(\mathbf{X})$, then $\mathbf{P_X}\boldsymbol{\eta} = \mathbf{0}$ and the estimation of coefficients will be unaffected. If the misspecification is due to omitted explanatory variables that are uncorrelated with those included in $\mathbf{X}$, that is, $\mathbf{X}^T\boldsymbol{\eta} = \mathbf{0}$, the estimates are not biased.

The estimation of the variance is also affected by misspecification:

$$E(\mathbf{y}^T(\mathbf{I}-\mathbf{P_X})\mathbf{y}) = (\mathbf{Xb}+\boldsymbol{\eta})^T(\mathbf{I}-\mathbf{P_X})(\mathbf{Xb}+\boldsymbol{\eta}) + \mathrm{tr}((\mathbf{I}-\mathbf{P_X})(\sigma^2\mathbf{I}_N))$$
$$= \boldsymbol{\eta}^T(\mathbf{I}-\mathbf{P_X})\boldsymbol{\eta} + \sigma^2(N-r). \quad (4.4)$$

Therefore, the bias in the variance estimate disappears only when the misspecification disappears. Note that the bias in $\hat{\sigma}^2$ is zero if and only if

$\boldsymbol{\eta}^T(\mathbf{I}-\mathbf{P_X})\boldsymbol{\eta} = 0$, or $\boldsymbol{\eta} \in \mathcal{C}(\mathbf{X})$, or $\boldsymbol{\eta} = \mathbf{Xd}$ for some $\mathbf{d}$.

In that case, the model is then

$\mathbf{y} = \mathbf{Xb} + \mathbf{Xd} + \mathbf{e} = \mathbf{X}(\mathbf{b}+\mathbf{d}) + \mathbf{e}$.

One view of misspecification is that the signal that is not accounted for $\boldsymbol{\eta}$ is partitioned into two pieces:

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{\eta} + \mathbf{e} = \mathbf{P_X y} + (\mathbf{I}-\mathbf{P_X})\mathbf{y} = (\mathbf{Xb} + \mathbf{P_X}\boldsymbol{\eta} + \mathbf{P_X e}) + (\mathbf{I}-\mathbf{P_X})(\boldsymbol{\eta}+\mathbf{e})$$

One piece $\mathbf{P_X}\boldsymbol{\eta}$ in $\mathcal{C}(\mathbf{X})$ affects the estimation of $\boldsymbol{\lambda}^T\mathbf{b}$; the other part $(\mathbf{I}-\mathbf{P_X})\boldsymbol{\eta}$ affects the estimation of the variance.

Example 4.4. Simple misspecification. Suppose $y_i = \beta_0 + \beta_1 x_i + e_i$, but the covariate x_i is ignored and we just estimate the mean and variance. Here we have $\eta_i = \beta_1 x_i$ and $\mathbf{X} = \mathbf{1}$ so that $\hat{\beta}_0 = \bar{y}$ and $E(\bar{y}) = \beta_0 + \beta_1 \bar{x}$. As for the variance, our usual variance estimate is $\hat{\sigma}^2 = \sum_i (y_i - \bar{y})^2/(N-1)$. Applying (4.4), the size of the bias,

$$E(\hat{\sigma}^2) - \sigma^2 = \beta_1^2 \sum_i (x_i - \bar{x})^2/(N-1)$$

depends on the size of the departure of the mean response from the model, whose mean is constant.

Example 4.5. Electricity problem. Consider the analysis of electricity consumption, using the following model as the true model for households:

$$\text{bill}_i = \beta_0 + \beta_1 \text{income}_i + \beta_2 \text{persons}_i + \beta_3 \text{area}_i + e_i \quad (4.5)$$

where bill_i = monthly electric bill for household i, income_i = monthly disposable income, persons_i = number in household, area_i = heating living area of home or apartment. In such analyses, often income is may not be available, so consider the consequences in estimating the regression coefficients β_2 and β_3 when income is dropped from the model:

$$E(\text{bill}_i) = \beta_0 + \beta_2 \text{persons}_i + \beta_3 \text{area}_i. \quad (4.6)$$

In this situation, rows of $\mathbf{X}$ contain $[\,1 \quad \text{persons}_i \quad \text{area}_i\,]$, $\mathbf{b}^T = [\,\beta_0 \quad \beta_2 \quad \beta_3\,]$, and $\eta_i = \beta_1 \text{income}_i$. Another approach is to construct a regression model for the missing variable income, using the remaining variables as explanatory variables:

$$\text{income}_i = \gamma_0 + \gamma_1 \text{persons}_i + \gamma_2 \text{area}_i + f_i. \quad (4.7)$$

Combining the true model (4.5) and the expression above for the missing variable (4.7), the misspecified model we are fitting (4.6) really now becomes

$$\text{bill}_i = (\beta_0 + \beta_1\gamma_0) + (\beta_2 + \beta_1\gamma_1)\text{persons}_i + (\beta_3 + \beta_1\gamma_2)\text{area}_i + (e_i + \beta_1 f_i). \quad (4.8)$$

If income is not related to persons, then $\gamma_1 = 0$, and the estimate of β_2 remains unbiased; however, if income is related to area, then $E(\hat{\beta}_3) = \beta_3 + \beta_1\gamma_2$, with the size of the bias depending on the importance of the missing variable β_1 and the strength of the relationship with the variable of interest γ_2. Also note the expression for the error in the misspecified model (4.8), $e_i + \beta_1 f_i$, contains both the original error e_i, and the effect of the misspecification.

The case of *overfitting* can be viewed in the following context, by partitioning the explanatory variables into two groups or blocks:

$$\mathbf{y} = \mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2\mathbf{b}_2 + \mathbf{e}$$

where the second group of explanatory variables is not needed, since $\mathbf{b}_2 = \mathbf{0}$. To make comparisons, let us simplify matters and assume the Gauss-Markov model and that both $\mathbf{X}_1$ and $\mathbf{X} = (\mathbf{X}_1\ \mathbf{X}_2)$ have full column rank. Using the smallest model with a design matrix of $\mathbf{X}_1$ leads to the least squares estimator we will denote as $\tilde{\mathbf{b}}_1$, constructed as

$$\tilde{\mathbf{b}}_1 = (\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{y}\ .$$

Using previous results, we can show that this estimator is unbiased, $\mathrm{E}(\tilde{\mathbf{b}}_1) = \mathbf{b}_1$, and its covariance matrix is

$$\mathrm{Cov}(\tilde{\mathbf{b}}_1) = \sigma^2(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}.$$

Including the second block of explanatory variables into the model that the researcher fits leads to the least squares estimators

$$\begin{bmatrix} \hat{\mathbf{b}}_1 \\ \hat{\mathbf{b}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1 & \mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2 \\ \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1 & \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^{\mathrm{T}}\mathbf{y} \\ \mathbf{X}_2^{\mathrm{T}}\mathbf{y} \end{bmatrix}.$$

Again, it is easy to show similar results for this estimator:

$$\mathrm{E}\left[\begin{bmatrix} \hat{\mathbf{b}}_1 \\ \hat{\mathbf{b}}_2 \end{bmatrix}\right] = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{0} \end{bmatrix} \text{ and } \mathrm{Cov}\left(\begin{bmatrix} \hat{\mathbf{b}}_1 \\ \hat{\mathbf{b}}_2 \end{bmatrix}\right) = \sigma^2 \begin{bmatrix} \mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1 & \mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2 \\ \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1 & \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_2 \end{bmatrix}^{-1}.$$

Using Exercise A.74 (partitioned inverse) we can show

$$\mathrm{Cov}(\hat{\mathbf{b}}_1) = \sigma^2(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1} + \sigma^2(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2[\mathbf{X}_2^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2]^{-1}\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1} \qquad (4.9)$$

and so that the penalty for including the second block of variables $\mathbf{X}_2$ is increased variance in the coefficient estimators.

Denoting $\mathrm{rank}(\mathbf{X}_1) = \mathrm{r}_1$ and $\mathrm{rank}(\mathbf{X}) = \mathrm{r}$, then the variance estimators are both unbiased

$$\mathrm{E}[\mathbf{y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{y}]/(\mathrm{N} - \mathrm{r}_1) = \sigma^2$$
$$\mathrm{E}[\mathbf{y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}]/(\mathrm{N} - \mathrm{r}) = \sigma^2$$

since there is no misspecification, and the only difference betweeen the two estimators is in the degrees of freedom. Except for some applications where error degrees of freedom are scarce, there is little lost in variance estimation by overfitting.

Example 4.6. No intercept. Suppose we have the simple linear regression problem, with $\mathrm{y_i} = \beta_0 + \beta_1\mathrm{x_i} + \mathrm{e_i}$, but $\beta_0 = 0$. The variance of the least squares estimate of the slope, where we include the intercept is $\sigma^2/\sum_{\mathrm{i}}(\mathrm{x_i} - \bar{\mathrm{x}})^2$. If we drop the intercept, the slope estimator is simply $\sum_{\mathrm{i}}\mathrm{x_i}\mathrm{y_i}/\sum_{\mathrm{i}}\mathrm{x_i^2}$ and its variance is $\sigma^2/\sum_{\mathrm{i}}\mathrm{x_i^2}$, which is smaller, since $\sum_{\mathrm{i}}\mathrm{x_i^2} > \sum_{\mathrm{i}}(\mathrm{x_i} - \bar{\mathrm{x}})^2$. See also Exercise 4.7.

Returning to the coefficient estimation, examination of the difference of the two covariance matrices shows the effect of including the second block of variables:

$$\mathrm{Cov}(\hat{\mathbf{b}}_1) - \mathrm{Cov}(\tilde{\mathbf{b}}_1) = \sigma^2(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2[\mathbf{X}_2^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2]^{-1}\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}$$

If the second block of explanatory variables is orthogonal to the first, that is, $\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2 = \mathbf{0}$, then the estimators not only have the same variance, they are the same estimators, as $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ becomes block diagonal . As $\mathbf{X}_2$ gets closer to $\mathcal{C}(\mathbf{X}_1)$, then $\mathbf{X}_2^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2$ gets smaller, and, when inverted, causes the $\mathrm{Cov}(\hat{\mathbf{b}}_1)$ to explode. This condition, known as *multicollinearity*, is the other feared consequence in model selection. One signal for severe multicollinearity is the

appearance of unexpected signs of coefficient estimators due to their wild inaccuracy. Without multicollinearity, and with enough degrees of freedom to estimate the variance, there is little to be lost in overfitting. With multicollinearity, overfitting can be catastrophic.

One measure of the effect of multicollinearity is called the *variance inflation factor* or VIF. If the explanatory variables were mutually orthogonal, then, following, say, Example 4.3, then $\text{Var}(\hat{b}_j) = \sigma^2/S_{xx}$ where $S_{xx} = \sum_i (X_{ij} - \bar{X}_{.j})^2$. In practice, when the explanatory variables are not orthogonal, then the defining expression is:

$$\text{Var}(\hat{b}_j) = (\text{VIF}) \times \sigma^2/S_{xx} \tag{4.10}$$

This relationship arises from the other form of the partitioned inverse result (Exercise A.74). Without loss of generality, consider $j = 1$, and partition the design matrix $\mathbf{X}$ as above, with just the first column in $\mathbf{X}_1$, and all of the other columns in $\mathbf{X}_2$. The employing the other form of the partitioned inverse result, we have

$$\text{Cov}(\hat{\mathbf{b}}_1) = \sigma^2[\mathbf{X}_1^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_2})\mathbf{X}_1]^{-1}. \tag{4.11}$$

Putting (4.10) and (4.11) together, we find that

$$\text{VIF} = \frac{S_{xx}}{\mathbf{X}_1^T(\mathbf{I}-\mathbf{P}_{\mathbf{X}_2})\mathbf{X}_1} = \frac{1}{1-R_j^2} \tag{4.12}$$

where R_j^2 is what we would have for R^2 if we took column j for the response vector, and employed the remaining explanatory variables. Clearly, when an explanatory variable can itself be closely approximated by a linear combination of other variables, then R_j^2 will be close to one and VIF very large, indicating a serious multicollinearity problem.

The quantity $\mathbf{X}_2^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2$ is employed by some algorithms, e.g. the sweep operator in SAS (see, e.g. Goodnight(1979) or Monahan(2001)), to determine rank of $\mathbf{X}$ or $\mathbf{X}^T\mathbf{X}$ in regression problems. In practice, this works quite well, although the finite precision of floating point arithmetic limits its effectiveness. When most of the columns of the design matrix $\mathbf{X}$ are composed of 0, 1, or some small integer, the computed elements of $\mathbf{X}_2^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2$ can be zero or nearly so. However, in the case of continuous covariates, $\mathbf{X}_2^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2$ will rarely be zero even in the case of perfect collinearity due to the effects of rounding error. In practice, most computer software for regression determine rank by testing whether $\mathbf{X}_2^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2$, or something similar, is close to zero. While the mathematics presented here appears to proceed smoothly in the case of dependence in the columns of $\mathbf{X}$ -- by using the generalized inverse when the inverse doesn't exist -- the reality is that the effect is dramatic: certain components or functions of $\mathbf{b}$ are no longer estimable. The practical matter is that the rank of the design matrix should be known in advance by the researcher. If the computer software determines a smaller rank than expected, then either an unexpected dependency or catastrophic multicollinearity problem exists. Finding a larger rank than expected indicates the inability of the software to detect dependence. (See Exercises 4.15, 4.16)

4.5 The Aitken Model and Generalized Least Squares

The Aitken model is a slight extension of the Gauss-Markov Model in that only different moment assumptions are made on the errors. The Aitken Model takes the form

$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, where $E(\mathbf{e}) = \mathbf{0}$, but $\text{Cov}(\mathbf{e}) = \sigma^2\mathbf{V}$

where the matrix $\mathbf{V}$ is a KNOWN positive definite matrix. In this way, it is similar to the Gauss-Markov Model in that the covariance is known up to a scalar σ^2. If $\mathbf{V} = \mathbf{I}$, then we have the Gauss-Markov model. In practice, however, usually $\mathbf{V}$ is unknown, or has just a few parameters; this case will be addressed later. In an Aitken model, the least squares estimator $\boldsymbol{\lambda}^T\hat{\mathbf{b}}$ of an estimable function $\boldsymbol{\lambda}^T\mathbf{b}$ may no longer be the BLUE for$\boldsymbol{\lambda}^T\mathbf{b}$. We will now construct a generalized least squares (GLS) estimator of $\boldsymbol{\lambda}^T\mathbf{b}$ and show that it is the BLUE for $\boldsymbol{\lambda}^T\mathbf{b}$.

The crucial step in GLS is the construction of a square root of the unscaled covariance matrix, that is, find $\mathbf{R}$ such that $\mathbf{RVR}^T = \mathbf{I}_N$. As discussed in Appendix A, there are two approaches for constructing a square root of a positive definite matrix, *Cholesky factorization* and *Spectral decomposition*. The Cholesky factorization writes $\mathbf{V}$ as the product of a lower triangular matrix $\mathbf{L}$ and its transpose, $\mathbf{V} = \mathbf{LL}^T$. Taking this route, use $\mathbf{R} = \mathbf{L}^{-1}$. The spectral decomposition uses the eigenvector-eigenvalue decomposition $\mathbf{V} = \mathbf{Q\Lambda Q}^T$ where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues, and $\mathbf{Q}$ is the (orthogonal) matrix of eigenvectors stacked as columns. This route suggests taking $\mathbf{R} = \mathbf{Q\Lambda}^{-1/2}\mathbf{Q}^T$, so that the 'square root' matrix is symmetric in this case. Note that we will insist that $\mathbf{V}$ be positive definite, so that $\mathbf{L}$ or $\boldsymbol{\Lambda}$ are nonsingular. If $\mathbf{V}$ were singular, then there would be a linear combination of observations with zero variance, and this case would more properly be treated as a linear contraint.

Using the matrix $\mathbf{R}$, we can reformulate the Aitken Model using a tranformed response variable

$\mathbf{z} = \mathbf{Ry} = \mathbf{RXb} + \mathbf{Re}$, or

$\mathbf{z} = \mathbf{Ub} + \mathbf{f}$, where $E(\mathbf{f}) = \mathbf{0}$ and $Cov(\mathbf{f}) = \sigma^2\mathbf{I}_N$, (4.13)

which looks just like the Gauss-Markov Model. Now we can tackle all of the same issues as before, and then transform back to the Aitken model:

a) Estimability. The linear function $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable if $\boldsymbol{\lambda}$ is in the column space of the transpose of the design matrix. Here this means $\boldsymbol{\lambda} \in \mathcal{C}(\mathbf{U}) = \mathcal{C}(\mathbf{X}^T\mathbf{R}^T) = \mathcal{C}(\mathbf{X}^T)$ since $\mathbf{R}$ is nonsingular. From another viewpoint, estimability did not involve the second moment anyway, so that estimability should not be affected by the fact that $\mathbf{V}$ is not a constant diagonal matrix.

b) Linear estimator. Note that any linear estimator $\mathbf{g} + \mathbf{h}^T\mathbf{y}$ that is linear in $\mathbf{z}$ is also a linear estimator $\mathbf{g} + \mathbf{h}^T\mathbf{Ry} = \mathbf{g} + \mathbf{a}^T\mathbf{y}$ and vice versa. In other words, the class of linear estimators in $\mathbf{z}$ is the same as the class of linear estimators in $\mathbf{y}$.

c) Generalized Least Squares Estimators. In the transformed model, the normal equations are

$\mathbf{U}^T\mathbf{U}\,\mathbf{b} = \mathbf{U}^T\mathbf{z}$ (4.14)

and so the least squares estimator from (4.13) solves (4.14) above. However, these normal equations can be easily rewritten as

$(\mathbf{RX})^T(\mathbf{RX})\,\mathbf{b} = (\mathbf{RX})^T(\mathbf{Ry})$, or

$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\,\mathbf{b} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ (4.15)

which are known as the *Aitken equations*, and the solution to (4.8) will be denoted as $\hat{\mathbf{b}}_{GLS}$, a generalized least squares estimator of $\mathbf{b}$. When needed for clarity, the solution to the usual normal equations $\mathbf{X}^T\mathbf{Xb} = \mathbf{X}^T\mathbf{y}$ will be denoted by $\hat{\mathbf{b}}_{OLS}$, for *Ordinary Least Squares*. From Section 4.1, we should expect that $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_{GLS}$ is BLUE for $\boldsymbol{\lambda}^T\mathbf{b}$, but this will be examined further.

Theorem 4.2. (Aitken's Theorem) Consider the Aitken Model given by $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, where $E(\mathbf{e}) = \mathbf{0}$, and $Cov(\mathbf{e}) = \sigma^2\mathbf{V}$, where $\mathbf{V}$ is a known positive definite matrix. If $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable, then $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_{GLS}$ is the BLUE for $\boldsymbol{\lambda}^T\mathbf{b}$.
Proof: Follows from (a), (b), (c), and the Gauss-Markov Theorem.□

From Chapter 2, we know that $\hat{\mathbf{b}}_{GLS}$ minimizes a sum of squares, in particular,

$$\| \mathbf{z} - \mathbf{Ub} \|^2 = \|\mathbf{R}(\mathbf{y} - \mathbf{Xb}) \|^2 = (\mathbf{y} - \mathbf{Xb})\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb}),$$

so that this sum of squares is often called *weighted least squares* or *generalized least squares*. In the simplest case, say, simple linear regression and $\mathbf{V}$ diagonal, then we have

$$(\mathbf{y} - \mathbf{Xb})\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb}) = \sum_i \tfrac{1}{V_{ii}} (y_i - \beta_0 - \beta_1 x_i)^2$$

and the name should be apparent.

c) Estimation of σ^2. Since the transformed model follows the Gauss-Markov assumptions, the estimator constructed in Section 4.3 is the natural unbiased estimator for σ^2:

$$\sigma^2_{GLS} = (\mathbf{z} - \mathbf{U}\hat{\mathbf{b}}_{GLS})^T(\mathbf{z} - \mathbf{U}\hat{\mathbf{b}}_{GLS}) / (N - r) = (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_{GLS})\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_{GLS}) / (N - r)$$

Example 4.7. Heteroskedasticity.
Consider the simple linear regression model through the origin with heteroskedastic (different variances) errors, with the variances proportional to the squares of the explanatory variables:

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

where $E(e_i) = 0$, $Var(e_i) = \sigma^2 x_i^2$, e_i uncorrelated and $x_i \neq 0$. Notice that

$$Var(e_i/x_i) = \sigma^2$$

so that the obvious step is to transform by dividing by x_i:

$$z_i = y_i/x_i = \beta + e_i/x_i.$$

The BLUE of β, then is $\hat{\beta}_{GLS} = \bar{z} = \frac{1}{N}\sum_i (y_i/x_i)$, and $Var(\hat{\beta}_{GLS}) = \sigma^2/N$. For comparison, $\hat{\beta}_{OLS} = \sum_i x_i y_i / \sum_i x_i^2$, and see Exercise 4.4.

Example 4.8. Autoregressive errors.
Suppose we have the usual multiple regression model

$$y_i = \mathbf{x}_i^T\mathbf{b} + e_i,$$

where the errors have the usual zero mean $E(e_i) = 0$, but the covariance structure induced by the model

$$e_i = \rho e_{i-1} + a_i$$

where the a_i's are uncorrelated with zero mean and variance σ^2. Then it can be shown that $Var(e_i) = \sigma^2/(1 - \rho^2)$ and the covariance matrix of the original errors e_i is given by

$$Cov(\mathbf{y}) = Cov(\mathbf{e}) = \sigma^2\mathbf{V} = \tfrac{\sigma^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{N-1} \\ \rho & 1 & \rho & \dots & \rho^{N-2} \\ \rho^2 & \rho & 1 & \dots & \\ \dots & & & \dots & \rho \\ \rho^{N-1} & \dots & & \rho & 1 \end{bmatrix}$$

so that $V_{ij} = \rho^{|i-j|}/(1 - \rho^2)$. This error structure is known as a first-order *autoregressive* model. The following transformation, known as *Cochrane-Orcutt transformation*, restores the usual Gauss-Markov assumptions:

$z_1 = \sqrt{1-\rho^2}\ y_1 = \sqrt{1-\rho^2}\ \mathbf{x}_1^T\mathbf{b} + \sqrt{1-\rho^2}\ e_1$

$z_i = y_i - \rho\, y_{i-1} = \mathbf{x}_i^T\mathbf{b} - \rho\, \mathbf{x}_{i-1}^T\mathbf{b} + e_i - \rho\, e_{i-1}$
$= (\mathbf{x}_i - \rho\, \mathbf{x}_{i-1})^T\mathbf{b} + a_i = \mathbf{u}_i^T\mathbf{b} + a_i$ for i = 2, ..., N.

The single parameter ρ of this model is not usually known, and the usual approach is to begin with ordinary least squares to estimate **b**, estimate ρ from the residuals, do GLS with estimated ρ to reestimate **b**, reestimate ρ, and iterate until convergence. This procedure is known as estimated generalized least squares (EGLS) since ρ is estimated.

Finding the BLUE estimator under the Aitken model really is not difficult, as you've seen; all that was necessary was to transform **y** and **X** to **z** and **U** so that the Gauss-Markov assumptions held. A more interesting pursuit, however, is the set of conditions that make the usual $\hat{\mathbf{b}}_{OLS}$ to be BLUE under the Aitken model assumptions.

Result 4.3. (Generalization of Result 4.1) The estimator $\mathbf{t}^T\mathbf{y}$ is the BLUE for $E(\mathbf{t}^T\mathbf{y})$ iff $\mathbf{t}^T\mathbf{y}$ is uncorrelated with all unbiased estimators of zero.

Proof: (If) Let $\mathbf{a}^T\mathbf{y}$ be another unbiased estimator of $E(\mathbf{t}^T\mathbf{y})$, that is, $E(\mathbf{a}^T\mathbf{y}) = E(\mathbf{t}^T\mathbf{y})$. Then

$\text{Var}(\mathbf{a}^T\mathbf{y}) = \text{Var}(\mathbf{t}^T\mathbf{y} + \mathbf{a}^T\mathbf{y} - \mathbf{t}^T\mathbf{y})$
$= \text{Var}(\mathbf{t}^T\mathbf{y}) + \text{Var}(\mathbf{a}^T\mathbf{y} - \mathbf{t}^T\mathbf{y}) + 2\ \text{Cov}(\mathbf{t}^T\mathbf{y}, \mathbf{a}^T\mathbf{y} - \mathbf{t}^T\mathbf{y})$

Since $\mathbf{a}^T\mathbf{y} - \mathbf{t}^T\mathbf{y}$ is an unbiased estimator of zero, the covariance terms drops out of the equation above, leading to

$\text{Var}(\mathbf{a}^T\mathbf{y}) = \text{Var}(\mathbf{t}^T\mathbf{y}) + \text{Var}(\mathbf{a}^T\mathbf{y} - \mathbf{t}^T\mathbf{y}) \geq \text{Var}(\mathbf{t}^T\mathbf{y})$.

(Only if) Suppose there is another unbiased estimator of zero $\mathbf{h}^T\mathbf{y}$, so that $E(\mathbf{h}^T\mathbf{y}) = 0$ where $\mathbf{h} \neq \mathbf{0}$, and let $\text{Cov}(\mathbf{t}^T\mathbf{y}, \mathbf{h}^T\mathbf{y}) = c$ and $\text{Var}(\mathbf{h}^T\mathbf{y}) = d$. Then consider the estimator of $E(\mathbf{t}^T\mathbf{y})$ given by

$\mathbf{a}^T\mathbf{y} = \mathbf{t}^T\mathbf{y} - (c/d)\ \mathbf{h}^T\mathbf{y}$.

This estimator is also unbiased for $E(\mathbf{t}^T\mathbf{y})$. Its variance is

$\text{Var}(\mathbf{a}^T\mathbf{y}) = \text{Var}(\mathbf{t}^T\mathbf{y}) + (c/d)^2\text{Var}(\mathbf{h}^T\mathbf{y}) - 2\ (c/d)\ \text{Cov}(\mathbf{t}^T\mathbf{y}, \mathbf{h}^T\mathbf{y})$
$= \text{Var}(\mathbf{t}^T\mathbf{y}) - c^2/d \leq \text{Var}(\mathbf{t}^T\mathbf{y})$.

So that if $\mathbf{t}^T\mathbf{y}$ is the BLUE, then $c = 0$, otherwise the estimator $\mathbf{a}^T\mathbf{y}$ constructed above will also be unbiased and have smaller variance. □

Corollary 4.1. Under the Aitken Model, the estimator $\mathbf{t}^T\mathbf{y}$ is the BLUE for $E(\mathbf{t}^T\mathbf{y})$ iff $\mathbf{Vt} \in \mathcal{C}(\mathbf{X})$.

Proof: From the preceding Result 4.3, $\mathbf{t}^T\mathbf{y}$ is the BLUE for $E(\mathbf{t}^T\mathbf{y})$ iff $\text{Cov}(\mathbf{t}^T\mathbf{y}, \mathbf{h}^T\mathbf{y}) = 0$ for all **h** such that $E(\mathbf{h}^T\mathbf{y}) = 0$. Note that if $\mathbf{h}^T\mathbf{y}$ is an unbiased estimator of zero, then we have

$E(\mathbf{h}^T\mathbf{y}) = \mathbf{h}^T\mathbf{X}\mathbf{b} = 0$ for all **b**.

This means $\mathbf{h}^T\mathbf{X} = \mathbf{0}$ or $\mathbf{h} \in \mathcal{N}(\mathbf{X}^T)$. Now $\text{Cov}(\mathbf{t}^T\mathbf{y}, \mathbf{h}^T\mathbf{y}) = \sigma^2\mathbf{t}^T\mathbf{V}\mathbf{h}$, and this is zero iff **h** is orthogonal to **Vt**, or, since $\mathbf{h} \in \mathcal{N}(\mathbf{X}^T)$, $\mathbf{Vt} \in \mathcal{C}(\mathbf{X})$. □

Corollary 4.2. Under the Aitken Model, $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_{OLS} = \mathbf{a}^T\mathbf{P}_\mathbf{X}\mathbf{y}$ (that is, $\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{a}$) is the BLUE for an estimable $\boldsymbol{\lambda}^T\mathbf{b}$ iff $\mathbf{V}\mathbf{P}_\mathbf{X}\mathbf{a} = \mathbf{X}\mathbf{q}$ for some **q**.

Proof: Let $\mathbf{t} = \mathbf{P}_\mathbf{X}\mathbf{a}$ in Corollary 4.1.

Result 4.4. Under the Aitken Model, all OLS estimators are BLUE (that is, each $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_{OLS}$ is the BLUE for the corresponding estimable $\boldsymbol{\lambda}^T\mathbf{b}$) iff there exists a matrix **Q** such that $\mathbf{VX} = \mathbf{XQ}$.

Proof: (If $\mathbf{VX} = \mathbf{XQ}$) First write $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_{OLS} = \boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y} = \mathbf{t}^T\mathbf{y}$ for $\mathbf{t} = \mathbf{X}\,(\mathbf{X}^T\mathbf{X})^g\boldsymbol{\lambda}$ so that $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_{OLS}$ is the BLUE for $\boldsymbol{\lambda}^T\mathbf{b}$ iff $\mathbf{Vt} \in \mathcal{C}(\mathbf{X})$. Now if $\mathbf{VX} = \mathbf{XQ}$, then

$$\mathbf{Vt} = \mathbf{VX}(\mathbf{X}^T\mathbf{X})^g = \mathbf{XQ}(\mathbf{X}^T\mathbf{X})^g\boldsymbol{\lambda} \in \mathcal{C}(\mathbf{X})$$

and employ Corollary 4.2.

(Only if) Now if $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_{OLS}$ is the BLUE for $\boldsymbol{\lambda}^T\mathbf{b}$, take $\boldsymbol{\lambda}^{(j)}$ as column j of $\mathbf{X}^T\mathbf{X}$, so that $\mathbf{t}^{(j)} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^g\boldsymbol{\lambda}^{(j)}$, then from Corollary 4.2, $\mathbf{Vt}^{(j)} = \mathbf{Xq}^{(j)}$. Stacking these columns side by side to form matrices $\mathbf{T}$ and $\mathbf{Q}$, we have

$$\mathbf{VT} = \mathbf{VX}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X} = \mathbf{VX} = \mathbf{XQ}. \square$$

Example 4.9. Consider the regression problem $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ with an equicorrelated covariance structure, that is $\text{Cov}(\mathbf{e}) = \mathbf{V} = \sigma^2\mathbf{I}_N + \tau^2\mathbf{1}_N\mathbf{1}_N^T$, and with an intercept and the following partitioning:

$$\mathbf{X} = [\,\mathbf{1} \quad \mathbf{X}^*\,] \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ \mathbf{b}_2 \end{bmatrix} \begin{matrix} 1 \\ p-1 \end{matrix}$$

In this case, is $\boldsymbol{\lambda}^T\hat{\mathbf{b}}_{OLS}$ the BLUE for estimable $\boldsymbol{\lambda}^T\mathbf{b}$? It will be if we can find $\mathbf{Q}$ such that $\mathbf{VX} = \mathbf{XQ}$.

$$\mathbf{VX} = (\sigma^2\mathbf{I}_N + \tau^2\mathbf{1}_N\mathbf{1}_N^T)\;[\,\mathbf{1} \quad \mathbf{X}^*\,] = [\,\sigma^2\mathbf{1}_N + \tau^2\mathbf{1}_N N, \quad \sigma^2\mathbf{X}^* + \tau^2\mathbf{1}_N\mathbf{1}_N^T\mathbf{X}^*\,]$$

$$\mathbf{XQ} = [\,\mathbf{1} \quad \mathbf{X}^*\,]\begin{bmatrix} Q_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix} = [\,\mathbf{1}Q_{11} + \mathbf{X}^*\mathbf{Q}_{21} \quad \mathbf{1}\mathbf{Q}_{12} + \mathbf{X}^*\mathbf{Q}_{22}\,]$$

Matching the first entries suggests choosing $Q_{11} = \sigma^2 + N\tau^2$ and $\mathbf{Q}_{21} = \mathbf{0}$, and matching the second yields $\mathbf{Q}_{12} = \tau^2\mathbf{1}^T\mathbf{X}^*$ and $\mathbf{Q}_{22} = \sigma^2\mathbf{I}_N$. See also Exercise 4.8.

Example 4.10. Seemingly Unrelated Regression. Suppose we have m individuals, each with n responses following regression models:

$$\mathbf{y}^{(i)} = \mathbf{X}^{(i)}\mathbf{b}^{(i)} + \mathbf{e}^{(i)},\ i = 1, ..., m$$

where $\mathbf{y}^{(i)}$ and $\mathbf{e}^{(i)}$ are $n \times 1$, $\mathbf{X}^{(i)}$ is $n \times p$, and $\mathbf{b}^{(i)}$ is $p \times 1$. The covariances in the errors $e^{(i)}$ ties these regressions together:

$$\text{Cov}(\,\mathbf{e}^{(i)}, \mathbf{e}^{(j)}) = \sigma_{ij}\mathbf{I}_n.$$

For example, the individuals may be companies, and the responses are quarterly sales which would be contemporaneously correlated. We can write this as one large linear model by combining these pieces:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \\ ... \\ \mathbf{y}^{(m)} \end{bmatrix},\ \mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \mathbf{X}^{(2)} & & \mathbf{0} \\ ... & & ... & \\ \mathbf{0} & \mathbf{0} & & \mathbf{X}^{(m)} \end{bmatrix},\ \mathbf{b} = \begin{bmatrix} \mathbf{b}^{(1)} \\ \mathbf{b}^{(2)} \\ ... \\ \mathbf{b}^{(m)} \end{bmatrix},\ \sigma^2\mathbf{V} = \begin{bmatrix} \sigma_{11}\mathbf{I}_n & \sigma_{12}\mathbf{I}_n & ... & \sigma_{1m}\mathbf{I}_n \\ \sigma_{21}\mathbf{I}_n & \sigma_{22}\mathbf{I}_n & & \sigma_{2m}\mathbf{I}_n \\ & ... & & ... \\ \sigma_{m1}\mathbf{I}_n & \sigma_{m2}\mathbf{I}_n & ... & \sigma_{mm}\mathbf{I}_n \end{bmatrix}$$

What are the best estimators in this case? Or, more specifically, when are the least squares estimators BLUE? In general, the least squares estimators are not always the best. However, some specific cases are interesting.

If $\sigma_{ij} = 0$ for $i \neq j$, then the problem completely decouples into m individual least squares problems and, not surprisingly, the least squares estimators are BLUE. The other interesting case has $\mathbf{X}^{(i)} = \mathbf{X}^{(1)}$, that is, the design matrices are the same for each company. In this case, we can show $\mathbf{VX} = \mathbf{XV}$, so that taking $\mathbf{Q} = \mathbf{V}$, the least squares estimators are BLUE. This latter situation is known as *multivariate regression*, although it is usually written differently.

4.6. Summary

1) The Gauss-Markov assumptions on the errors in a linear model are introduced. They specify that the errors have zero mean, are uncorrelated and have constant variance.

2) The Gauss-Markov Theorem says that the least squares estimator $\boldsymbol{\lambda}^T\hat{\mathbf{b}}$ has the smallest variance of all linear unbiased estimators of an estimable function $\boldsymbol{\lambda}^T\mathbf{b}$ when the Gauss-Markov assumptions hold.

3) The estimator $\hat{\sigma}^2 = \text{SSE}/(N - r) = \mathbf{y}^T(\mathbf{I} - \mathbf{P_X})\mathbf{y}/(N - r)$ is an unbiased estimator of the variance parameter σ^2.

4) The consequences of underfitting or misspecification and overfitting are evaluated.

5) Generalized least squares estimators are introduced for cases where the Gauss-Markov assumptions on the errors may not hold.

4.7. Exercises

1) Suppose the random variable Y_i represents the number of votes that a candidate receives in county i. A reasonable model would be that Y_i would be independent binomial random variables with parameters n_i = number of voters in county i and p = Pr(voter correctly votes for candidate).
a) What are $E(Y_i)$ and $\text{Var}(Y_i)$?
b) Write this as a linear model.
c) Find the BLUE of p in this situation.

2) Under the Gauss-Markov Model, show that for $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}^T\mathbf{b}$ is estimable, $\text{Var}(\boldsymbol{\lambda}^T\hat{\mathbf{b}}) = \sigma^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^g\boldsymbol{\lambda}$ does not depend on the choice of generalized inverse $(\mathbf{X}^T\mathbf{X})^g$.

3) In Example 4.6 (heteroskedasticity) find $\text{Var}(\hat{\mathbf{b}}_{\text{OLS}})$ and compare it to $\text{Var}(\hat{\mathbf{b}}_{\text{GLS}})$.

4) Consider the heteroskedasticity situation in Example 4.6, but suppose $\text{Var}(e_i) = \sigma^2 x_i$ where $x_i > 0$. Find $\text{Var}(\hat{\mathbf{b}}_{\text{OLS}})$ and compare it to $\text{Var}(\hat{\mathbf{b}}_{\text{GLS}})$.

5) Find the Cholesky factor of $\mathbf{V}$ in Example 4.6.

6) Suppose we have the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, ..., N$$

where e_i are uncorrelated and $\text{Var}(e_i) = \sigma^2$. Consider the instrumental variables estimator of the slope

$$\tilde{b}_1 = \sum_{i=1}^{N} (z_i - \bar{z})(y_i - \bar{y}) \,/\, \sum_{i=1}^{N} (z_i - \bar{z})(x_i - \bar{x})$$

where z_1, ..., z_N are known constants.

a) Is $\tilde{b}_1$ an unbiased estimator of the slope parameter β_1?

b) Find the variance of $\tilde{b}_1$.

c) We know that taking $z_i = x_i$ gives our familiar least squares estimator $\hat{b}_1$. Find the ratio of the two variances: $\text{Var}(\hat{b}_1) / \text{Var}(\tilde{b}_1)$ and show that it is less than or equal to 1 ($\hat{b}_1$ is BLUE). You may find the following results useful: $\sum_{i=1} (x_i - \bar{x}) = \sum_{i=1} (z_i - \bar{z}) = 0$

$$\sum_{i=1} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1} (x_i - \bar{x})y_i = \sum_{i=1} x_i(y_i - \bar{y})$$

7) In a reversal of Example 4.6, suppose we have the simple linear regression problem, with $y_i = \beta_0 + \beta_1 x_i + e_i$, and the usual Gauss-Markov assumptions. Compute the bias in the slope estimator $\sum_i x_i y_i / \sum_i x_i^2$ when $\beta_0 \neq 0$.

8) (Compare with Example 4.9 (equicorrelation)) Consider the linear model $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ where $\mathbf{e} = \mathbf{u} + Z\mathbf{1}$ (Z is a scalar random variable) where $\text{Cov}(\mathbf{u}) = \sigma^2 \mathbf{I}_N$, $\text{Var}(Z) = \tau^2$, and Z and $\mathbf{u}$ are uncorrelated. Find $\mathbf{V}$ and derive conditions under which the OLS estimator of every estimable function is BLUE.

9) Suppose $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$, where $E(e_i) = 0$, $\text{Var}(e_i) = \sigma^2$ and the e_i are independent. But suppose we fit a simple linear regression model: $E(y_i) = \beta_0 + \beta_1 x_i$. To illustrate, consider the following simple situation: $x_i = i$, $i = 1, 2, ..., n = 8$, for $\beta_0 = 2$, $\beta_1 = 3$. Consider also various values of β_2, say, -2 to 4 by ones. (Hint: I suggest using PROC REG for computations.)

a) Compute the bias in the least squares estimators, $\hat{\beta}_0$, $\hat{\beta}_1$.

b) Compute the bias in our usual variance estimate $\hat{\sigma}^2$.

c) Would your results change if the values of β_0 and β_1 were changed? Explain.

10) Consider the simple linear regression problem, $y_i = \beta_0 + \beta_1 x_i + e_i$, for $i = 1, ..., 4 = n$ with $x_i = i$ and the Gauss-Markov assumptions on e_i. A direct route for getting the BLUE would be to construct ALL linear unbiased estimators and then directly minimize the variance. Write the unbiased estimators as $\sum_{i=1}^{n} a_i y_i$ and focus on estimating the slope β_1.

a) Write the two linear equations in the a_i's that express constraints so that $\sum a_i y_i$ is an unbiased estimator of β_1.

b) Construct the family of solutions to the equations in (a). (Hint: you'll need two z's) This will parameterize all unbiased estimators with just two parameters.

c) Compute the variance of the estimators in (b) and minimize the variance. You should get a familiar solution.

11) Prove the Gauss-Markov Theorem directly, that is, by constructing all linear estimators $\mathbf{a}^T\mathbf{y}$ which are unbiased for $\boldsymbol{\lambda}^T\mathbf{b}$ (find a family of solutions $\mathbf{a}(\mathbf{z})$), and then minimizing the variance $\sigma^2\mathbf{a}^T\mathbf{a}$.

12) Show that if $\mathbf{R}$ is square and nonsingular and $\mathbf{R}^T\mathbf{V}\mathbf{R} = \mathbf{I}$ then $\mathbf{V}^{-1} = \mathbf{R}^T\mathbf{R}$. Do you need the assumption that $\mathbf{R}$ is square? nonsingular?

13) Let $\mathbf{Q} = \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^g\mathbf{X}^T\mathbf{V}^{-1}$. Show that $\mathbf{Q}$ is a projection onto $\mathcal{C}(\mathbf{X})$. (This is difficult without the Hint: first factor $\mathbf{V} = \mathbf{L}\mathbf{L}^T$ with Cholesky and work with a symmetric version $\mathbf{Q}^* = \mathbf{L}^{-1}\mathbf{Q}\mathbf{L}$)

14) Let $\sigma^2\mathbf{V}$ be the N $\times$ N covariance matrix for a first order moving average process:

$V_{ij} = 1 + \alpha^2$ if i = j

$V_{ij} = \alpha$ if | i − j | = 1

$V_{ij} = 0$ if | i − j | > 1

Notice that V is banded with zeros outside of three bands, on the diagonal and above and below the diagonal.

a) Show that the Cholesky factor of $\mathbf{V}$ is also banded (lower triangular, with nonzeros on the diagonal, and below the diagonal).

b) Find the limit of the two nonzero elements in row N as N$\rightarrow\infty$.

15) Consider the multiple regression problem including an intercept with the following list of explanatory variables:

c1 = cos(2πi/7)	s1 = sin(2πi/7)
c2 = cos(2π2i/7)	s2 = sin(2π2i/7)
c3 = cos(2π3i/7)	s3 = sin(2π3i/7)
c4 = cos(2π4i/7)	s4 = sin(2π4i/7)
c5 = cos(2π5i/7)	s5 = sin(2π5i/7)
c6 = cos(2π6i/7)	s6 = sin(2π6i/7)

for i = 1, ..., N.

a) Show thtat the last six variables (c4, s4, ..., s6) are linearly dependent on the first six (c1, s1, ..., s3) and an intercept.

b) Test whether the regression software that you commonly use can detect dependencies among the explanatory variables, using 3.1416 as your approximation for π, and various values of N.

c) Repeat this exercise with a cruder approximation 3.14 for π.

d) Repeat this exercise with 4 in place of 7 (that is, 2πi/4, 4πi/4, etc.).

16) Using the variables (1, c1, s1, c2, s2, c3, s3) from Exercise 15 above, is there any multicollinearity problem?

17) Consider the usual simple linear regression situation

$E(y_i) = \alpha + \beta\, x_i$ for i = 1, ..., 5 with x_i = i

$Var(y_i) = \sigma^2$ and y_i are independent

Note the simple form of x_i and that we have only 5 observations.

a) Find the least squares estimator $\hat{\beta}$ and express it in terms of $\mathbf{t}^T\mathbf{y}$ by explicitly giving $\mathbf{t}$.

b) Show that $\hat{\beta}$ is unbiased and find its variance.

c) Show that $\hat{\gamma} = (y_4 - y_2)/2$ and $\hat{\eta} = (y_5 - y_1)/4$ are also unbiased for β and also find the variance of each.

Consider now another estimator of the slope parameter β the estimator $\hat{\delta} = c\,\hat{\gamma} + (1\text{-}c)\,\hat{\eta}$.

d) Find the variance of $\hat{\delta}$ in terms of c and σ^2.

e) Find the value of c that minimizes the variance found in part d).

f) Suppose we knew that $\alpha = 0$, would $\hat{\beta}$ still be BLUE?

18) Under Gauss-Markov assumptions, show that if cov($\mathbf{a}^T\mathbf{y}$, $\mathbf{d}^T\hat{\mathbf{e}}$) $= 0$ for all $\mathbf{d}$, then $\mathbf{a}^T\mathbf{y}$ is the BLUE for its expectation.

19) Under the Aitken Model, with Cov($\mathbf{e}$) $= \sigma^2\mathbf{V}$, if cov($\mathbf{a}^T\mathbf{y}$, $\mathbf{d}^T\hat{\mathbf{e}}$) $= 0$ for all $\mathbf{d}$, then is $\mathbf{a}^T\mathbf{y}$ still the BLUE for its expectation?

20) Prove Aitken's Theorem (Theorem 4.2) directly. Consider the Aitken Model, and let $\boldsymbol{\lambda}^T\mathbf{b}$ be estimable. Let $\mathbf{a}^T\mathbf{y}$ be a linear unbiased estimator of $\boldsymbol{\lambda}^T\mathbf{b}$. Show that

$$\text{Var}(\mathbf{a}^T\mathbf{y}) = \text{Var}(\boldsymbol{\lambda}^T\hat{\mathbf{b}}_{GLS}) + \text{Var}(\mathbf{a}^T\mathbf{y} - \boldsymbol{\lambda}^T\hat{\mathbf{b}}_{GLS}).$$

21) Let Y_i, i = 1, .., N be iid exponential(λ), that is, each has density $f(y) = \lambda^{-1}e^{-y/\lambda}$ for $y > 0$.

a) Find E Y_i^k for k = 1, 2, 3, 4.

b) Find the variance of the usual variance estimator, Var($\sum_i (Y_i - \bar{Y})^2/(N-1)$).

22) Recall the analysis of the US population in Exercise 3.23. Find the VIF's for each of the coefficients in the uncentered model, using t = 1790 through 2000. If you center using c = 1890, what happens to the VIF's?

23) Prove that if $\mathbf{V}$ is positive definite, then $\mathcal{C}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}) = \mathcal{C}(\mathbf{X}^T)$.

24) Ridge regression is a technique that has been recommended by some statisticians to address multicollinearity problems arising in multiple regression. In our usual linear models framework with E($\mathbf{y}$) $= \mathbf{X}\mathbf{b}$, and Cov($\mathbf{y}$) $= \sigma^2\mathbf{I}_N$, the ridge regression estimator takes the form

$$\tilde{\mathbf{b}} = (\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}$$

where $k > 0$. Assume here that the $\mathbf{X}$ has full column rank, that is, rank($\mathbf{X}$) $= p$.

a) Find E($\tilde{\mathbf{b}}$).

b) Is $\boldsymbol{\lambda}^T\tilde{\mathbf{b}}$ an unbiased estimator of $\boldsymbol{\lambda}^T\mathbf{b}$?

c) Find Cov($\tilde{\mathbf{b}}$).

d) Mean squared error is commonly used to assess the quality of an estimator. For the ridge regression estimator $\tilde{\mathbf{b}}$, find its mean squared error

$$\text{E}(\|\tilde{\mathbf{b}} - \mathbf{b}\|^2) = \text{E}((\tilde{\mathbf{b}} - \mathbf{b})^T(\tilde{\mathbf{b}} - \mathbf{b})).$$

Consider applying ridge regression to a multivariate regression problem with two centered covariates ($\sum x_i = \sum z_i = 0$), taking the form

$$\mathbf{Xb} = \begin{bmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \dots & \dots & \dots \\ 1 & x_N & z_N \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

yielding the inner product matrix

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} N & 0 & 0 \\ 0 & \sum x_i^2 & \sum x_i z_i \\ 0 & \sum x_i z_i & \sum z_i^2 \end{bmatrix}.$$

For simplicity, suppose $N = 10$, $\sum x_i^2 = \sum z_i^2 = 5$ and $\sum x_i z_i = 4$.

e) Find the covariance matrix of our usual least squares estimator $\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ in this situation.
f) Find a value of k such that one component of $\tilde{\mathbf{b}}$ (your choice of component) has smaller variance than the corresponding least squares estimator.
g) Part (f) looks like it violates the Gauss-Markov Theorem. Does it?

4.10. References

D. Cochrane and G. H. Orcutt (1949) "Applications of least squares regression to relationships containing autocorrelated error terms,' Journal of the American Statistical Association, 44:32-61.

J. H. Goodnight (1979) 'A Tutorial on the Sweep Operator,' The American Statistician 33:149-158.

J. F. Monahan (2001) Numerical Methods of Statistics, Cambridge University Press.

J. O. Rawlings, S. G. Pantula, and D. A. Dickey (1998) Applied Regression Analysis: A Research Tool, Springer-Verlag.

4.11. Addendum -- Variance of variance estimator

Computing the $\text{Var}(\hat{\sigma}^2)$ requires considerable detail, athough the result is none too deep. The main result will be given with some generality.

Result 4.9. Let **P** be a symmetric matrix, and e be a random vector. The components e_i are iid with the following four moments: $E(e_i) = 0$, $\text{Var}(e_i) = E(e_i^2) = \sigma^2$, $E(e_i^3) = \gamma_3$, $E(e_i^4) = \gamma_4$, then

$$\text{Var}(\ (\boldsymbol{\mu} + \mathbf{e})^T\mathbf{P}(\boldsymbol{\mu} + \mathbf{e})\ ) = 4\sigma^2(\boldsymbol{\mu}^T\mathbf{P}^2\boldsymbol{\mu}) + 4\,\gamma_3 \sum_i \mu_i P_{ii} \sum_j P_{ji}$$

$$+ 2\sigma^4 \sum_{i \neq j} P_{ij}^2 + \sum_i (\gamma_4 - \sigma^4) P_{ii}^2$$

Proof: Begin with

Var($(\boldsymbol{\mu} + \mathbf{e})^T \mathbf{P} (\boldsymbol{\mu} + \mathbf{e})$) =

$\quad E[(\boldsymbol{\mu} + \mathbf{e})^T \mathbf{P} (\boldsymbol{\mu} + \mathbf{e})(\boldsymbol{\mu} + \mathbf{e})^T \mathbf{P} (\boldsymbol{\mu} + \mathbf{e})] - E[(\boldsymbol{\mu} + \mathbf{e})^T \mathbf{P} (\boldsymbol{\mu} + \mathbf{e})]^2$

Note that $E[(\boldsymbol{\mu} + \mathbf{e})^T \mathbf{P} (\boldsymbol{\mu} + \mathbf{e})] = \boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu} + \sigma^2 \text{tr}(\mathbf{P})$ from Lemma 4.3. Now write out all 16 terms of $E[(\boldsymbol{\mu} + \mathbf{e})^T \mathbf{P} (\boldsymbol{\mu} + \mathbf{e})(\boldsymbol{\mu} + \mathbf{e})^T \mathbf{P} (\boldsymbol{\mu} + \mathbf{e})]$ as

$E[\boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu} \boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu}] = (\boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu})^2$

$E[\boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu} \mathbf{e}^T \mathbf{P} \boldsymbol{\mu}] = 0$

$E[\boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu} \boldsymbol{\mu}^T \mathbf{P} \mathbf{e}] = 0$

$E[\boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu} \mathbf{e}^T \mathbf{P} \mathbf{e}] = (\boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu}) E[\mathbf{e}^T \mathbf{P} \mathbf{e}] = (\boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu}) \sigma^2 \text{trace}(\mathbf{P})$

$E[\mathbf{e}^T \mathbf{P} \boldsymbol{\mu} \boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu}] = 0$

$E[\mathbf{e}^T \mathbf{P} \boldsymbol{\mu} \mathbf{e}^T \mathbf{P} \boldsymbol{\mu}] = \text{Var}(\boldsymbol{\mu}^T \mathbf{P} \mathbf{e}) = E[\mathbf{e}^T \mathbf{P} \boldsymbol{\mu} \boldsymbol{\mu}^T \mathbf{P} \mathbf{e}] = \sigma^2 (\boldsymbol{\mu}^T \mathbf{P}^2 \boldsymbol{\mu})$

$E[\mathbf{e}^T \mathbf{P} \boldsymbol{\mu} \boldsymbol{\mu}^T \mathbf{P} \mathbf{e}] = \sigma^2 (\boldsymbol{\mu}^T \mathbf{P}^2 \boldsymbol{\mu})$

$E[\mathbf{e}^T \mathbf{P} \boldsymbol{\mu} \mathbf{e}^T \mathbf{P} \mathbf{e}] = E[\boldsymbol{\mu}^T \mathbf{P} \mathbf{e} \mathbf{e}^T \mathbf{P} \mathbf{e}]$ (see below)

$E[\boldsymbol{\mu}^T \mathbf{P} \mathbf{e} \boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu}] = 0$

$E[\boldsymbol{\mu}^T \mathbf{P} \mathbf{e} \mathbf{e}^T \mathbf{P} \boldsymbol{\mu}] = \sigma^2 (\boldsymbol{\mu}^T \mathbf{P}^2 \boldsymbol{\mu})$

$E[\boldsymbol{\mu}^T \mathbf{P} \mathbf{e} \boldsymbol{\mu}^T \mathbf{P} \mathbf{e}] = \sigma^2 (\boldsymbol{\mu}^T \mathbf{P}^2 \boldsymbol{\mu})$

$E[\boldsymbol{\mu}^T \mathbf{P} \mathbf{e} \mathbf{e}^T \mathbf{P} \mathbf{e}] =$ (see below)

$E[\mathbf{e}^T \mathbf{P} \mathbf{e} \boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu}] = (\boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu}) E[\mathbf{e}^T \mathbf{P} \mathbf{e}] = (\boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu}) \sigma^2 \text{trace}(\mathbf{P})$

$E[\mathbf{e}^T \mathbf{P} \mathbf{e} \mathbf{e}^T \mathbf{P} \boldsymbol{\mu}] = E[\boldsymbol{\mu}^T \mathbf{P} \mathbf{e} \mathbf{e}^T \mathbf{P} \mathbf{e}]$ (see below)

$E[\mathbf{e}^T \mathbf{P} \mathbf{e} \boldsymbol{\mu}^T \mathbf{P} \mathbf{e}] = E[\boldsymbol{\mu}^T \mathbf{P} \mathbf{e} \mathbf{e}^T \mathbf{P} \mathbf{e}]$ (see below)

$E[\ \mathbf{e}^T \mathbf{P} \mathbf{e} \mathbf{e}^T \mathbf{P} \mathbf{e}] =$ (see below)

Only two difficult expressions remain: $E[\boldsymbol{\mu}^T \mathbf{P} \mathbf{e} \mathbf{e}^T \mathbf{P} \mathbf{e}]$ and $E[\mathbf{e}^T \mathbf{P} \mathbf{e} \mathbf{e}^T \mathbf{P} \mathbf{e}]$. For both we need the following algebra:

$$E[\mathbf{a}^T \mathbf{P} \mathbf{b} \mathbf{c}^T \mathbf{P} \mathbf{d}] = \sum_i \sum_j \sum_k \sum_l E[\ a_i b_j c_k d_l P_{ij} P_{kl}\ ]$$

We're interested in all cases where the indices are the same:

iiii	(all same)	n
iiij	(one differs)	n(n-1) of each of (iiij, iiji, ijii, jiii)
iijj	(two pair)	n(n-1) of each of (iijj, ijij, ijji)
iijk	(one pair)	n(n-1)(n-2) of each of (iijk, ijik, ijki, jiik, jiki, jkii)
ijkl	(all different)	n(n-1)(n-2)(n-3)

For $E[\boldsymbol{\mu}^T \mathbf{P} \mathbf{e} \mathbf{e}^T \mathbf{P} \mathbf{e}] = \sum_i \sum_j \sum_k \sum_l E[\ \mu_i e_j e_k e_l P_{ij} P_{kl}\ ]$ we have

iiii $\quad \sum_i E[\ \mu_i e_i e_i e_i P_{ii} P_{ii}\ ] = \sum_i \mu_i \gamma_3 P_{ii}^2$

iiij $\quad \sum_{i \neq j} E[\ \mu_j e_i e_i e_i P_{ji} P_{ii}\ ] = \sum_{i \neq j} \mu_i \gamma_3 P_{ji} P_{ii}$

Note that the other three cases are zero, for example, $E[\ \mu_i e_j e_i e_i P_{ij} P_{ii}\ ] = 0$. This extends to the other three possible combinations (two pair, one pair, all differ). Adding these two pieces produces

$$E[\boldsymbol{\mu}^T \mathbf{P} \mathbf{e} \mathbf{e}^T \mathbf{P} \mathbf{e}] = \sum_i \mu_i \gamma_3 P_{ii}^2 \ + \ \sum_{i \neq j} \mu_i \gamma_3 P_{ji} P_{ii} \ = \gamma_3 \sum_i \mu_i P_{ii} \sum_j P_{ji}$$

which does not appear to simplify further.

For $E[\mathbf{e}^T \mathbf{P} \mathbf{e} \mathbf{e}^T \mathbf{P} \mathbf{e}] = \sum_i \sum_j \sum_k \sum_l E[\ e_i e_j e_k e_l P_{ij} P_{kl}\ ]$ we have

iiii $\quad \sum_i E[\ e_i e_i e_i e_i P_{ii} P_{ii}\ ] = \sum_i \gamma_4 P_{ii}^2$

iiij $\quad \sum_{i \neq j} E[\ e_j e_i e_i e_i P_{ji} P_{ii}\ ] = \ 0$ (all four cases: jiii, ijii, iiji, iiij)

iijj $\quad \sum_{i \neq j} E[\ e_i e_i e_j e_j P_{ii} P_{jj}\ ] = \sigma^4 \sum_{i \neq j} P_{ii} P_{jj}$

ijij $\quad \sum_{i \neq j} E[\ e_i e_j e_i e_j P_{ij} P_{ij}\ ] = \sigma^4 \sum_{i \neq j} P_{ij} P_{ij}$

ijji $\quad \sum_{i \neq j} E[\ e_i e_j e_j e_i P_{ij} P_{ji}\ ] = \sigma^4 \sum_{i \neq j} P_{ij} P_{ji}$

and the other cases are all zero. Gathering up the pieces produces the following:

$$\begin{aligned}
&\text{Var}(\ (\boldsymbol{\mu} + \mathbf{e})^T \mathbf{P} (\boldsymbol{\mu} + \mathbf{e})\ ) = \\
&\quad E[(\boldsymbol{\mu} + \mathbf{e})^T \mathbf{P} (\boldsymbol{\mu} + \mathbf{e}) (\boldsymbol{\mu} + \mathbf{e})^T \mathbf{P} (\boldsymbol{\mu} + \mathbf{e})] - E[(\boldsymbol{\mu} + \mathbf{e})^T \mathbf{P} (\boldsymbol{\mu} + \mathbf{e})]^2 \\
&\quad = (\boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu})^2 + 2(\boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu}) \sigma^2 \text{trace}(\mathbf{P}) + 4\sigma^2 (\mu^T P^2 \mu) + 4\ \gamma_3 \sum_i \mu_i P_{ii} \sum_j P_{ji} \\
&\qquad\qquad + \sigma^4 \sum_{i \neq j} (P_{ii} P_{jj} + 2P_{ij}^2) + \sum_i \gamma_4 P_{ii}^2 \ - (\boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu} + \sigma^2 \text{tr}(\mathbf{P}))(\boldsymbol{\mu}^T \mathbf{P} \boldsymbol{\mu} + \sigma^2 \text{tr}(\mathbf{P})) \\
&\quad = 4\sigma^2 (\boldsymbol{\mu}^T \mathbf{P}^2 \boldsymbol{\mu}) + 4\ \gamma_3 \sum_i \mu_i P_{ii} \sum_j P_{ji} \ + \sigma^4 \sum_{i \neq j} (P_{ii} P_{jj} + 2P_{ij}^2) - \sigma^4 \text{tr}(\mathbf{P})^2 + \sum_i \gamma_4 P_{ii}^2 \\
&\quad = 4\sigma^2 (\boldsymbol{\mu}^T \mathbf{P}^2 \boldsymbol{\mu}) + 4\ \gamma_3 \sum_i \mu_i P_{ii} \sum_j P_{ji} \ + 2\sigma^4 \sum_{i \neq j} P_{ij}^2 + \sum_i (\gamma_4 - \sigma^4) P_{ii}^2
\end{aligned}$$

The last step employs

$\text{tr}(\mathbf{P})^2 = (\sum_i P_{ii})^2 = \sum_i P_{ii}^2 + \sum_{i \neq j} P_{ii} P_{jj}$. □

Corollary 4.10. Let **P** be a symmetric and idempotent, and the components of the random vector **e** iid with four moments existing, then

$$\text{Var}(\mathbf{e}^T \mathbf{P} \mathbf{e}) = \ 2\sigma^4 \sum_{i \neq j} P_{ij}^2 + \sum_i (\gamma_4 - \sigma^4) P_{ii}^2$$

Corollary 4.11. Let **P** be symmetric and $\gamma_3 = 0$, $\gamma_4 = 3\sigma^4$ (which hold if e_i are iid Normal(0, σ^2)), then

$$\text{Var}(\mathbf{e}^T \mathbf{P} \mathbf{e}) = 4\sigma^2 (\boldsymbol{\mu}^T \mathbf{P}^2 \boldsymbol{\mu}) + 2\sigma^4 \text{tr}(\mathbf{P}^2)$$

Proof: Here $\gamma_3 = 0$, $\gamma_4 = 3\sigma^4$ and starting with

$$\mathrm{Var}(\mathbf{e}^{\mathrm{T}}\mathbf{P}\mathbf{e}) = 4\sigma^2(\boldsymbol{\mu}^{\mathrm{T}}\mathbf{P}^2\boldsymbol{\mu}) + 2\sigma^4\sum_{i\neq j} P_{ij}^2 + \sum_i (3\sigma^4 - \sigma^4)P_{ii}^2$$

and now use

$$\mathrm{tr}(\mathbf{P}^2) = \sum_i \sum_j P_{ij}^2 = \sum_i P_{ii}^2 + \sum_{i\neq j} P_{ij}^2 . \square$$

JFM -- 05 January 2006