Multivariate regression notes

Daniel J. Eck

Contents

The Multivariate Regression Model and estimation

The multivariate (multiresponse) regression model is an extension of the classical linear regression model to settings involving multiple possible correlated responses. The model is of the form

$$Y_i = \beta X_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \Sigma), \tag{1}$$

where $Y_i \in \mathbb{R}^r$ is the response vector, $X_i \in \mathbb{R}^p$ is a vector of predictors which we assume are fixed, $\beta \in \mathbb{R}^{r\times p}$ is a coefficient matrix, and $\Sigma \in \mathbb{R}^{r\times r}$ is a positive definite covariance matrix. This model is very similar to the classical linear regression in its construction and motivation.

Model parameters β and Σ are estimated using the least squares approach and maximum likelihood estimation. Estimation of model parameters is easiest and more elegant using a multivariate representation of model (1). Let $\mathbb{Y} \in \mathbb{R}^{n\times r}$ be a matrix with rows Y_i' and $\mathbb{X} \in \mathbb{R}^{n\times p}$ be a matrix with rows X_i'. We can then rewrite model (1) as

$$\text{vec}(\mathbb{Y}') \sim N(\text{vec}(\beta\mathbb{X}'), I_n \otimes \Sigma), \tag{2}$$

where vec denotes the vec operator that stacks the columns of a matrix, and $\otimes$ is the Kronecker product operator which we define with respect to matrices $A \in \mathbb{R}^{mn}$ and $B \in \mathbb{R}^{pq}$

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix} \in \mathbb{R}^{mp\times nq}.$$

The representation (2) allows for a straightforward procedure for obtaining estimates of β. Maximizing the likelihood (2) is equivalent to minimizing the sum of squares $\|\text{vec}(\mathbb{Y}') - \text{vec}(\beta\mathbb{X}')\|^2$. After a bit of algebra we arrive at an aesthetically similar ordinary least squares problem

$$\|\text{vec}(\mathbb{Y}') - \text{vec}(\beta\mathbb{X}')\|^2 = \|\text{vec}(\mathbb{Y}) - \text{vec}(\mathbb{X}\beta')\|^2$$

$$= \|\mathbb{Y} - \mathbb{X}\beta'\|_F^2,$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix. Continuing on,

$$\begin{aligned}\|\mathbb{Y} - \mathbb{X}\beta'\|_F^2 &= \text{tr}\left((\mathbb{Y} - \mathbb{X}\beta')(\mathbb{Y} - \mathbb{X}\beta')'\right) \\ &= \text{tr}\left(\mathbb{Y}\mathbb{Y}'\right) - 2\text{tr}\left(\mathbb{Y}'\mathbb{X}\beta'\right) + \text{tr}\left(\beta\mathbb{X}'\mathbb{X}\beta'\right).\end{aligned}$$

Taking derivatives with respect to a matrix yields

$$\frac{\partial\|\mathbb{Y} - \mathbb{X}\beta'\|_F^2}{\partial\beta} = -2\mathbb{X}'\mathbb{Y} + 2\mathbb{X}'\mathbb{X}\beta'.$$

The OLS solution then takes a somewhat familiar form

$$\widehat{\beta} = \mathbb{Y}'\mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}.$$

We now estimate Σ. Estimation of Σ involves tensor algebra and tensor calculus that we will go into. First, the log likelihood of (2) as a function of Σ is written as

$$l(\Sigma) \propto -\frac{1}{2}\log\left(|I_n \otimes \Sigma|\right) - \frac{1}{2}\text{vec}\left(\mathbb{Y}' - \beta\mathbb{X}'\right)'\left(I_n \otimes \Sigma\right)^{-1}\text{vec}\left(\mathbb{Y}' - \beta\mathbb{X}'\right). \tag{3}$$

The term $|I_n \otimes \Sigma| = \|\Sigma\|^n$ and the term $(I_n \otimes \Sigma)^{-1} = I_n \otimes \Sigma^{-1}$ so we can rewrite the above as

$$-\frac{n}{2}\log\left(|\Sigma|\right) - \frac{1}{2}\text{vec}\left(\mathbb{Y}' - \beta\mathbb{X}'\right)'\left(I_n \otimes \Sigma^{-1}\right)\text{vec}\left(\mathbb{Y}' - \beta\mathbb{X}'\right).$$

Using Kronecker product, trace, and vec operator algebra and the property $\text{vec}(AXB) = (B' \otimes A)\text{vec}(X)$ we can write

$$\begin{aligned}\text{vec}\left(\mathbb{Y}' - \beta\mathbb{X}'\right)'\left(I_n \otimes \Sigma^{-1}\right)\text{vec}\left(\mathbb{Y}' - \beta\mathbb{X}'\right) &= \text{vec}\left(\mathbb{Y}' - \beta\mathbb{X}'\right)'\left(I_n \otimes \Sigma^{-1/2}\right)\left(I_n \otimes \Sigma^{-1/2}\right)\text{vec}\left(\mathbb{Y}' - \beta\mathbb{X}'\right) \\ &= \text{vec}\left(\Sigma^{-1/2}(\mathbb{Y}' - \beta\mathbb{X}')\right)'\text{vec}\left(\Sigma^{-1/2}(\mathbb{Y}' - \beta\mathbb{X}')\right) \\ &= \text{tr}\left((\mathbb{Y}' - \beta\mathbb{X}')'\Sigma^{-1}(\mathbb{Y}' - \beta\mathbb{X}')\right)\end{aligned}$$

Using matrix derivative properties [Lutkepohl, 1996], we have

$$\begin{aligned}\frac{\partial\log\left(|\Sigma|\right)}{\partial\Sigma} &= \Sigma^{-1}, \\ \frac{\partial\text{tr}\left((\mathbb{Y}' - \beta\mathbb{X}')'\Sigma^{-1}(\mathbb{Y}' - \beta\mathbb{X}')\right)}{\partial\Sigma} &= -\Sigma^{-1}(\mathbb{Y}' - \beta\mathbb{X}')(\mathbb{Y}' - \beta\mathbb{X}')'\Sigma^{-1}\end{aligned}$$

We use the the derivatives above to obtain the MLE for Σ,

$$\frac{\partial l(\Sigma)}{\partial\Sigma} = -\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(\mathbb{Y}' - \beta\mathbb{X}')(\mathbb{Y}' - \beta\mathbb{X}')'\Sigma^{-1}.$$

Setting the above equal to zero and solving for Σ yields

$$\Sigma = \frac{1}{n}(\mathbb{Y}' - \beta\mathbb{X}')(\mathbb{Y}' - \beta\mathbb{X}')'.$$

Plugging in the solution for β that was previously derived above yields our estimator

$$\widehat{\Sigma} = \frac{1}{n}(\mathbb{Y}' - \widehat{\beta}\mathbb{X}')(\mathbb{Y}' - \widehat{\beta}\mathbb{X}')'.$$

Properties of estimators

We now investigate some properties of these estimators. First of all

$$\mathrm{E}(\widehat{\beta}) = \mathrm{E}\left(\mathbb{Y}'\mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\right) = \mathrm{E}(\mathbb{Y}')\mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1} = \beta\mathbb{X}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} = \beta,$$

so that $\widehat{\beta}$ is an unbiased estimator of β. We now investigate the variance of $\mathrm{vec}(\widehat{\beta}')$, a transformation of the matrix $\widehat{\beta}$ that allows us to represent the variance of $\widehat{\beta}$ in matrix form,

$$\begin{aligned}
\mathrm{Var}(\mathrm{vec}(\widehat{\beta})) &= \mathrm{Var}(\mathrm{vec}(\mathbb{Y}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1})) \\
&= \mathrm{Var}([(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}' \otimes I_r]\mathrm{vec}(\mathbb{Y}')) \\
&= [(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}' \otimes I_r](I_n \otimes \Sigma)[\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} \otimes I_r] \\
&= (\mathbb{X}'\mathbb{X})^{-1} \otimes \Sigma.
\end{aligned}$$

Similarly, we have $\mathrm{Var}(\mathrm{vec}(\widehat{\beta}')) = \Sigma \otimes (\mathbb{X}'\mathbb{X})^{-1}$. These results yield

$$\begin{aligned}
\mathrm{vec}(\widehat{\beta}) &\sim N\left(\mathrm{vec}(\beta), (\mathbb{X}'\mathbb{X})^{-1} \otimes \Sigma\right), \\
\mathrm{vec}(\widehat{\beta}') &\sim N\left(\mathrm{vec}(\beta'), \Sigma \otimes (\mathbb{X}'\mathbb{X})^{-1}\right).
\end{aligned}$$

As in classical regression we can perform inferences with respect elements of $\mathrm{vec}(\beta)$ as

$$\frac{\mathrm{vec}(\widehat{\beta})_j}{\sqrt{((\mathbb{X}'\mathbb{X})^{-1} \otimes \widehat{\Sigma})_{jj}}} \sim t_{n-p}$$

where n is the number of observations and p is the number of predictors. We can also compute the covariance between two columns of $\widehat{\beta}$ which has the form

$$\mathrm{Cov}(\hat{\beta}_k, \hat{\beta}_l) = \sigma_{kl}(\mathbb{X}'\mathbb{X})^{-1},$$

and the covariance between two rows of $\widehat{\beta}$ has the form

$$\mathrm{Cov}(\hat{\beta}_g, \hat{\beta}_j) = (\mathbb{X}'\mathbb{X})^{-1}_{gj}\Sigma,$$

where $(\mathbb{X}'\mathbb{X})^{-1}_{gj}$ denotes the (g, j)-th element of $(\mathbb{X}'\mathbb{X})^{-1}$. For the expectation of $\widehat{\Sigma}$ we re-express

$$\begin{aligned}
\widehat{\Sigma} &= \frac{1}{n}(\mathbb{Y}' - \widehat{\beta}\mathbb{X}')(\mathbb{Y}' - \widehat{\beta}\mathbb{X}')' \\
&= \frac{1}{n}(\mathbb{Y}' - \mathbb{Y}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}')(\mathbb{Y}' - \mathbb{Y}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}')' \\
&= \frac{1}{n}(\mathbb{Y}' - \mathbb{Y}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}')(\mathbb{Y} - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}) \\
&= \frac{1}{n}\mathbb{Y}'(I_n - H)\mathbb{Y}
\end{aligned}$$

where $H = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$ is an idempotent matrix and $I_n - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$ is also an idempotent matrix which is the projection into the orthogonal complement of the space spanned by the columns of $\mathbb{X}$, $(I_n - H)\mathbb{X} = 0$. With this in mind

$$\mathrm{E}(\widehat{\Sigma}) = \frac{1}{n}\mathrm{E}\left(\mathbb{Y}'(I_n - H)\mathbb{Y}\right) = \frac{n-p-1}{n}\Sigma,$$

and an unbiased estimator of Σ is given as

$$\widetilde{\Sigma} = \frac{\mathbb{Y}'(I_n - H)\mathbb{Y}}{n-p-1}.$$

As in classical regression we see that the expectations of the residuals satisfies

$$\mathrm{E}(\mathbb{Y} - \widehat{\mathbb{Y}}) = (I_n - H)\mathrm{E}(\mathbb{Y}) = (I_n - H)\mathbb{X}\beta' = 0,$$

and we have

$$\begin{aligned}
\mathrm{Var}(\mathrm{vec}(\mathbb{Y} - \widehat{\mathbb{Y}})') &= \mathrm{Var}[\mathrm{vec}(\mathbb{Y}'(I_n - H))] \\
&= \mathrm{Var}[((I_n - H) \otimes I_r)\mathrm{vec}(\mathbb{Y}')] \\
&= ((I_n - H) \otimes I_r)\mathrm{Var}[\mathrm{vec}(\mathbb{Y}')]((I_n - H) \otimes I_r) \\
&= ((I_n - H) \otimes I_r)(I_n \otimes \Sigma)((I_n - H) \otimes I_r) \\
&= (I_n - H) \otimes \Sigma,
\end{aligned}$$

and we could similarly write $\mathrm{Var}(\mathrm{vec}(\mathbb{Y} - \widehat{\mathbb{Y}})) = \Sigma \otimes (I_n - H)$.

Motor Trend Cars example

We first load in the dataset. The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). The variables which we consider in this analysis are

```
mpg:    Miles/(US) gallon
disp:   Displacement (cu.in.)
hp: Gross horsepower
wt: Weight (1000 lbs)
cyl:    Number of cylinders
am: Transmission (0 = automatic, 1 = manual)
carb:   Number of carburetors
```

The first four variables are response variables corresponding to engine performance and size. The next three variables are engine design variables.

```
data(mtcars)
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

The standard `lm` function in R can fit multivariate linear regression models.

```
mtcars$cyl = factor(mtcars$cyl)
Y = as.matrix(mtcars[,c("mpg","disp","hp","wt")])
m = lm(Y ~ cyl + am + carb, data=mtcars, x = TRUE)

# estimate of beta'
betahat = coef(m)
betahat
```

```
##                  mpg      disp         hp         wt
## (Intercept) 25.320303 134.32487 46.5201421  2.7612069
## cyl6        -3.549419  61.84324  0.9116288  0.1957229
## cyl8        -6.904637 218.99063 87.5910956  0.7723077
## am           4.226774 -43.80256  4.4472569 -1.0254749
## carb        -1.119855   1.72629 21.2764930  0.1749132
```

```
# estimates of Sigma
SSE = crossprod(Y - m$fitted.values)
n = nrow(Y)
p = nrow(coef(m))
SigmaMLE = SSE / n
SigmaMLE
```

```
##             mpg       disp          hp         wt
## mpg    6.638633  -44.94796 -16.6232233 -0.5548030
## disp -44.947964 2113.48487 358.7058833 15.2773945
## hp   -16.623223  358.70588 487.0718440  0.3933977
## wt    -0.554803   15.27739   0.3933977  0.2171394
```

```
Sigmahat = SSE / (n - p)
Sigmahat
```

```
##              mpg       disp          hp         wt
## mpg    7.8680094  -53.27166 -19.7015979 -0.6575443
## disp -53.2716607 2504.87095 425.1328988 18.1065416
## hp   -19.7015979  425.13290 577.2703337  0.4662491
## wt    -0.6575443   18.10654   0.4662491  0.2573503
```

We can see that the R's `vcov` function provides an estimate of $\text{Var}(\text{vec}(\hat{\beta}'))$ as its default

```
X = m$x
unique(round(vcov(m) - kronecker(Sigmahat, solve(crossprod(X))), 10))
```

```
##                 mpg:(Intercept) mpg:cyl6 mpg:cyl8 mpg:am mpg:carb
## mpg:(Intercept)               0        0        0      0        0
##                 disp:(Intercept) disp:cyl6 disp:cyl8 disp:am disp:carb
## mpg:(Intercept)                0         0         0       0         0
##                 hp:(Intercept) hp:cyl6 hp:cyl8 hp:am hp:carb wt:(Intercept)
## mpg:(Intercept)              0       0       0     0       0              0
##                 wt:cyl6 wt:cyl8 wt:am wt:carb
## mpg:(Intercept)       0       0     0       0
```

We obtain inferences for regression coefficients corresponding to the first response variable `mpg` and compare those to what is obtained using theory.

```
# summary table from lm
msum = summary(m)
msum[[1]]
```

```
##
## Call:
## lm(formula = mpg ~ cyl + am + carb, data = mtcars, x = TRUE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9074 -1.1723  0.2538  1.4851  5.4728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.3203     1.2238  20.690  < 2e-16 ***
## cyl6         -3.5494     1.7296  -2.052 0.049959 *
## cyl8         -6.9046     1.8078  -3.819 0.000712 ***
## am            4.2268     1.3499   3.131 0.004156 **
```

```
## carb         -1.1199     0.4354  -2.572 0.015923 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.805 on 27 degrees of freedom
## Multiple R-squared:  0.8113, Adjusted R-squared:  0.7834
## F-statistic: 29.03 on 4 and 27 DF,  p-value: 1.991e-09
```

```
# summary table from theory (they are the same)
msum2 = cbind(coef(m)[, 1], sqrt(diag( kronecker(Sigmahat, solve(crossprod(X))) ))[1:5])
msum2 = cbind(msum2, msum2[, 1] / msum2[, 2])
msum2 = cbind(msum2, sapply(msum2[, 3], function(x) pt(abs(x), df = n - p, lower = FALSE)*2 ))
msum2
```

```
##                  [,1]      [,2]      [,3]         [,4]
## (Intercept) 25.320303 1.2237903 20.690067 4.303483e-18
## cyl6        -3.549419 1.7295506 -2.052221 4.995947e-02
## cyl8        -6.904637 1.8078219 -3.819313 7.124146e-04
## am           4.226774 1.3499249  3.131118 4.156214e-03
## carb        -1.119855 0.4353558 -2.572274 1.592280e-02
```

Inference and nested models

We can test for the presence of multiple regression coefficients simultaneously. Assume that $q < p$ and we want to test if a reduced model with q terms is sufficient:

$$H_0 : \beta_2 = 0_{(p-q)\times r}$$
$$H_1 : \beta_2 \neq 0_{(p-q)\times r}$$

where $\beta = (\beta_1 \ \beta_2)$. To test the above hypothesis we could use the familiar likelihood ratio test statistic:

$$\Lambda = \frac{\max_{\beta_1,\Sigma} L(\beta_1, \Sigma)}{\max_{\beta,\Sigma} L(\beta, \Sigma)}.$$

For large n, we can use the modified test statistics

$$-\nu \log(\Lambda) \sim \chi^2_{r(p-q)}$$

where $\nu = n - p - 1 - (1/2)(m - p + q + 1)$. There are a plethora of additional tests that we could perform in this setting (we will not stress each test's origins in this course). These include the Wilk's lambda, Pillai's trace, Hotelling-Lawley trace, and Roy's greatest root. The corresponding tests statistics follow below. First, let $E = n\tilde{\Sigma}$ where $\tilde{\Sigma}$ is the MLE for the full model with β unconstrained and let $\tilde{H} = n(\tilde{\Sigma}_1 - \tilde{\Sigma})$ where $\tilde{\Sigma}_1$ is the MLE of Σ in the reduced model constrained by $\beta_2 = 0$. The test statistics now follow:

- Wilk's lambda $= \prod_{i=1}^{s} \frac{1}{1+\eta_i} = \frac{|\tilde{E}|}{|\tilde{E}+\tilde{H}|}$
- Pillai's trace $= \sum_{i=1}^{s} \frac{\eta_i}{1+\eta_i} = \text{tr}[\tilde{H}(\tilde{E}+\tilde{H})^{-1}]$
- Hotelling-Lawley trace $= \sum_{i=1}^{s} \eta_i = \text{tr}(\tilde{H}\tilde{E}^{-1})$
- Roy's greatest root $= \frac{\eta_1}{1+\eta_1}$

where $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_s$ denote the nonzero eigenvalues of $\tilde{H}\tilde{E}^{-1}$. Let's demonstrate this on the motor trends cars example:

```
## anova implements these methods
m0 = lm(Y ~ am + carb, data=mtcars)
anova(m0, m, test="Wilks")
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ am + carb
## Model 2: Y ~ cyl + am + carb
##   Res.Df Df Gen.var.   Wilks approx F num Df den Df    Pr(>F)
## 1     29      43.692
## 2     27 -2   29.862 0.16395   8.8181      8     48 2.525e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m0, m, test="Pillai")
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ am + carb
## Model 2: Y ~ cyl + am + carb
##   Res.Df Df Gen.var. Pillai approx F num Df den Df    Pr(>F)
## 1     29      43.692
## 2     27 -2   29.862 1.0323   6.6672      8     50 6.593e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Etilde = n * SigmaMLE
SigmaTilde1 = crossprod(Y - m0$fitted.values) / n
Htilde = n * (SigmaTilde1 - SigmaMLE)
HEi = Htilde %*% solve(Etilde)
HEi.values = eigen(HEi)$values
c(Wilks = prod(1 / (1 + HEi.values)), Pillai = sum(HEi.values / (1 + HEi.values)))
```

```
##     Wilks    Pillai
## 0.1639527 1.0322975
```

Interval estimation

Suppose we want to estimate $\hat{E}(Y|X = x_h)$ and the variability about $\hat{E}(Y|X = x_h)$. Given predictor value x_h, we have $\hat{E}(Y|X = x_h) = \hat{\beta}x_h$. Note that $\hat{Y}_h \sim N(\beta x_h, x_h'(\mathbb{X}'\mathbb{X})^{-1}x_h\Sigma)$. We can test

$$H_0 : E(Y_h|X = x_h) = Y_h^*$$
$$H_1 : E(Y_h|X = x_h) \neq Y_h^*$$

We have

$$T^2 = \left(\frac{\hat{\beta}x_h - \beta x_h}{\sqrt{x_h'(\mathbb{X}'\mathbb{X})x_h}}\right)' \hat{\Sigma}^{-1} \left(\frac{\hat{\beta}x_h - \beta x_h}{\sqrt{x_h'(\mathbb{X}'\mathbb{X})x_h}}\right) \sim \frac{r(n-p-1)}{n-p-r} F_{r,(n-p-r)}.$$

Then a $100(1-\alpha)\%$ confidence interval for component k of $\mathrm{E}(Y_{hk}|X = x_h)$ is

$$\hat{Y}_{hk} \pm \sqrt{\frac{r(n-p-1)}{n-p-r} F_{r,(n-p-r)}(\alpha)} \sqrt{x_h'(\mathbb{X}'\mathbb{X})^{-1}x_h\hat{\sigma}_{kk}}$$

Now suppose that we want to estimate an observed value of the response for a given predictor, here we want estimates and inferences for $\widehat{Y}_h$ at x_h instead of $\hat{E}(\widehat{Y}_h|X = x_h)$. Given x_h, the fitted value $\widehat{Y}_h = \widehat{\beta}x_h$ remains the proper point estimator. We can test

$$H_0 : E(Y_h|X = x_h) = Y_h^*$$
$$H_1 : E(Y_h|X = x_h) \neq Y_h^*$$

We have

$$T^2 = \left(\frac{\hat{\beta}x_h - \beta x_h}{\sqrt{1 + x_h'(\mathbb{X}'\mathbb{X})x_h}} \right)' \widehat{\Sigma}^{-1} \left(\frac{\hat{\beta}x_h - \beta x_h}{\sqrt{1 + x_h'(\mathbb{X}'\mathbb{X})x_h}} \right) \sim \frac{r(n-p-1)}{n-p-r} F_{r,(n-p-r)}.$$

Then a $100(1-\alpha)\%$ confidence interval for component k of Y_{hk} is

$$\hat{Y}_{hk} \pm \sqrt{\frac{r(n-p-1)}{n-p-r} F_{r,(n-p-r)}(\alpha)} \sqrt{1 + x_h'(\mathbb{X}'\mathbb{X})^{-1} x_h \hat{\sigma}_{kk}}$$

We will now demonstrate confidence and prediction intervals in the motor trends cars example. Note that we have to code our own functions because R does not have the capability to provide these quantities. R does provide point predictions.

```
# confidence interval
newdata = data.frame(cyl=factor(6, levels=c(4,6,8)), am=1, carb=4)
predict(m, newdata, interval="confidence")
```

```
##        mpg     disp      hp       wt
## 1 21.51824 159.2707 136.985 2.631108
```

```
# prediction interval
newdata = data.frame(cyl=factor(6, levels=c(4,6,8)), am=1, carb=4)
predict(m, newdata, interval="prediction")
```

```
##        mpg     disp      hp       wt
## 1 21.51824 159.2707 136.985 2.631108
```

Here is the function which produces confidence and prediction intervals (credit to Nathaniel Helwig)

```
pred.mlm = function(object, newdata, level=0.95,
                    interval = c("confidence", "prediction")){
  form = as.formula(paste("~",as.character(formula(object))[3]))
  xnew = model.matrix(form, newdata)
  fit = predict(object, newdata)
  Y = model.frame(object)[,1]
  X = model.matrix(object)
  n = nrow(Y)
  r = ncol(Y)
  p = ncol(X) - 1
  sigmas = colSums((Y - object$fitted.values)^2) / (n - p - 1)
  fit.var = diag(xnew %*% tcrossprod(solve(crossprod(X)), xnew))
  if(interval[1]=="prediction") fit.var = fit.var + 1
  const = qf(level, df1=r, df2=n-p-r) * r * (n - p - 1) / (n - p - r)
  vmat = (n/(n-p-1)) * outer(fit.var, sigmas)
  lwr = fit - sqrt(const) * sqrt(vmat)
  upr = fit + sqrt(const) * sqrt(vmat)
  if(nrow(xnew)==1L){
    ci = rbind(fit, lwr, upr)
    rownames(ci) = c("fit", "lwr", "upr")
  } else {
    ci = array(0, dim=c(nrow(xnew), r, 3))
    dimnames(ci) = list(1:nrow(xnew), colnames(Y), c("fit", "lwr", "upr") )
    ci[,,1] = fit
    ci[,,2] = lwr
    ci[,,3] = upr
  }
```

```
  ci
}
```

```
# confidence interval
newdata = data.frame(cyl=factor(6, levels=c(4,6,8)), am=1, carb=4)
pred.mlm(m, newdata)
```

```
##          mpg     disp        hp       wt
## fit 21.51824 159.2707 136.98500 2.631108
## lwr 16.65593  72.5141  95.33649 1.751736
## upr 26.38055 246.0273 178.63351 3.510479
```

```
# prediction interval
newdata = data.frame(cyl=factor(6, levels=c(4,6,8)), am=1, carb=4)
pred.mlm(m, newdata, interval="prediction")
```

```
##           mpg      disp        hp        wt
## fit 21.518240 159.27070 136.98500 2.6311076
## lwr  9.680053 -51.95435  35.58397 0.4901152
## upr 33.356426 370.49576 238.38603 4.7720999
```

The Envelope Model

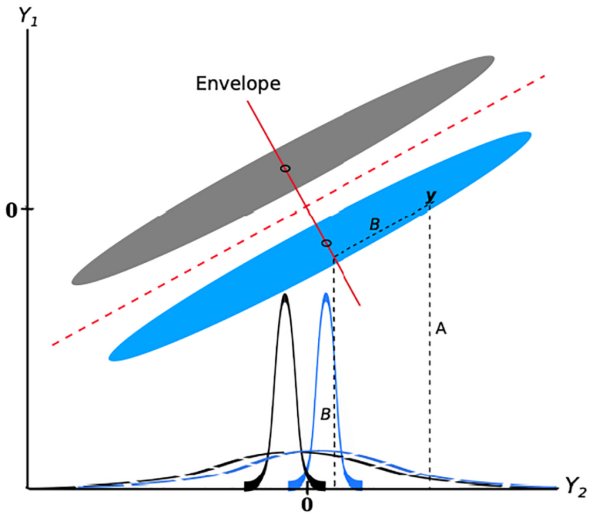
Envelope methodology [Cook, 2018] is a class of multivariate statistical techniques which aim to reduce variation in estimation without altering estimation objectives. Envelope methodology has its origins in providing variance reductions for the estimation of β in the multivariate linear regression model [Cook et al., 2010]. Before we begin, we are going to reparameterize the multivariate linear regression model (1) in a form more conducive for envelope methodology. Let α be the intercept vector so that $\tilde{\beta}[1, X] = \alpha + \beta X$ where $\tilde{\beta}$ is β in model (1). Then the multivariate linear model can be written as

$$Y = \alpha + \beta X + \varepsilon, \tag{4}$$

where $\beta \in \mathbb{R}^{r \times p}$ and $X \in \mathbb{R}^p$. The terms α and β in model (4) would be combined to form β in model (1). Similarly, X in model (4) would be $[1, X]$ in model (1).

We first provide envelope heuristics before we jump into the formal details. The following picture is taken from Cook [2020]:

FIGURE 1 Schematic illustration of response envelopes for the regression of a bivariate response $\mathbf{Y} = (Y_1, Y_2)^T$ on a binary predictor. The distributions of $\mathbf{Y} \mid X$ are represented by the gray and blue ellipses. The red solid and dashed lines represent the envelope and its orthogonal complement



The motivation for response envelopes in multivariate linear regression comes from allowing for the possibility that there are linear combinations of the response vector whose distribution is invariant to changes in the fixed predictor vector. Such linear combinations are referred to as X-invariants. If X-invariants exist, then allowing for them in model (4) can result in substantial reduction in estimated variance. The linear transformation G^TY, where $G \in \mathbb{R}^{r\times q}$ with $q \leq r$, is X-invariant if and only if A^TG^TY for any fixed full-rank matrix $A \in \mathbb{R}^{q\times q}$. Thus, a particular G is not identifiable, but span(G) is identifiable. This leads us to consider subspaces rather than specific coordinates.

The envelope model arises by parameterizing the multivariate linear regression model (4) in terms of the smallest subspace $\mathcal{E} \subseteq \mathbb{R}^r$ with the property that, for all relevant x_1 and x_2,

$$(i)\ Q_{\mathcal{E}}Y|(X = x_1) \sim Q_{\mathcal{E}}Y|(X = x_2) \qquad \text{and} \qquad (ii)\ P_{\mathcal{E}}Y \perp\!\!\!\perp Q_{\mathcal{E}}Y|X,$$

where $P_{\mathcal{E}}$ is the projection into $\mathcal{E}$ and $Q_{\mathcal{E}} = I_r - P_{\mathcal{E}}$. These conditions have the following meaning:

- Condition (i) stipulates that the marginal distribution of $Q_{\mathcal{E}}Y$ must be unaffected by changes in X. It holds if and only if span$(\beta) \subseteq \mathcal{E}$, because then

$$Q_{\mathcal{E}}Y = Q_{\mathcal{E}}\alpha + Q_{\mathcal{E}}\beta X + Q_{\mathcal{E}}\mathcal{E} = Q_{\mathcal{E}}\alpha + Q_{\mathcal{E}}\mathcal{E}.$$

- Condition (ii) requires that $Q_{\mathcal{E}}Y$ be unaffected by changes in X through an association with $P_{\mathcal{E}}Y$. This condition holds if and only if

$$\text{cov}(P_{\mathcal{E}}Y, Q_{\mathcal{E}}Y|X) = P_{\mathcal{E}}\Sigma Q_{\mathcal{E}} = 0.$$

Taken together, conditions (i) and (ii) imply that any dependence of Y on X must be concentrated in $P_{\mathcal{E}}Y$, the X-variant part of Y that is material to the regression. The X-invariant part of Y, $Q_{\mathcal{E}}Y$, represents variation in the response variable that is immaterial. The next rwo definitions formalize the construction of an envelope in general:

Definition 1: A subspace $\mathcal{R} \subseteq \mathbb{R}^r$ is said to be a reducing subspace of a symmetric matrix $M \in \mathbb{R}^{r\times r}$ if $\mathcal{R}$ decomposes M as $M = P_{\mathcal{R}}MP_{\mathcal{R}} + Q_{\mathcal{R}}MQ_{\mathcal{R}}$. If $\mathcal{R}$ is a reducing subspace of M, we say $\mathcal{R}$ reduces M.

Definition 2: Let $\mathcal{B} = \text{span}(\beta)$ and let $M \in \mathbb{R}^{r\times r}$ be a symmetric matrix and suppose $\mathcal{B} \subseteq \text{span}(M)$. Then the M-envelope of $\mathcal{B}$, denoted by $\mathcal{E}_M(\mathcal{B})$, is the intersection of all reducing subspaces of M that contain $\mathcal{B}$.

Returning to model (4), let $\mathcal{B} = \text{span}(\beta)$. The projection $P_{\mathcal{E}}$ is then the projection into $\mathcal{E}_\Sigma(\mathcal{B})$. We can then parameterize the envelope structure into model (4). Let $u = \dim(\mathcal{E}_\Sigma(\mathcal{B}))$ and let $(\Gamma, \Gamma_0) \in \mathbb{R}^{r\times r}$ be an orthogonal matrix with $\Gamma \in \mathbb{R}^{r\times u}$ and $\text{span}(\Gamma) = \mathcal{E}_\Sigma(\mathcal{B})$. Then the envelope model can be written as

$$Y = \alpha + \Gamma\eta X + \varepsilon, \qquad \text{with} \qquad \Sigma = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T, \tag{5}$$

where:

- The coefficient vector $\beta = \Gamma\eta$, where $\eta \in \mathbb{R}^{u\times p}$ carries the coordinates of β relative to the basis matrix Γ.
- The matrices $\Omega \in \mathbb{R}^{r\times r}$ and $\Omega_0 \in \mathbb{R}^{(r-u)\times(r-u)}$ are positive definite representing, respectively, material and immaterial variation.

To see how this model reflects the X-invariant part of Y, consider Γ_0^TY,

$$\begin{aligned}\Gamma_0^TY &= \Gamma_0^T(\alpha + \Gamma\eta X + \varepsilon)\\ &= \Gamma_0^T\alpha + \Gamma_0^T\varepsilon,\end{aligned}$$

where $\text{Var}(\Gamma_0^T \varepsilon) = \Omega_0$. In the above, we see that the marginal distribution of $\Gamma_0^T Y$ does not contain X, so condition (i) is satisfied. Furthermore,

$$\begin{aligned}\text{cov}(\Gamma^T Y, \Gamma_0^T Y | X) &= \Gamma^T \Sigma \Gamma_0 \\ &= \Gamma^T \left(\Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T \right) \Gamma_0 \\ &= 0.\end{aligned}$$

Thus condition (ii) is also satisfied.

Note: The envelope model decomposes variation Σ into a material portion $\Gamma\Omega\Gamma^T$ and an immaterial portion $\Gamma_0\Omega_0\Gamma_0^T$. The material part represents the variability in Y necessary for estimation of $\beta = \Gamma\eta$. When the errors in model (5) are normally distributed then $\beta X = \Gamma\eta X$ is the mean of a normal distribution (centered predictors and responses) and the envelope model yields variance reduction by exploiting a connection between the mean and the variance. This is conflicts from lessons learned from Basu's Theorem which tells us that estimation of the mean and variance are independent in some normal models.

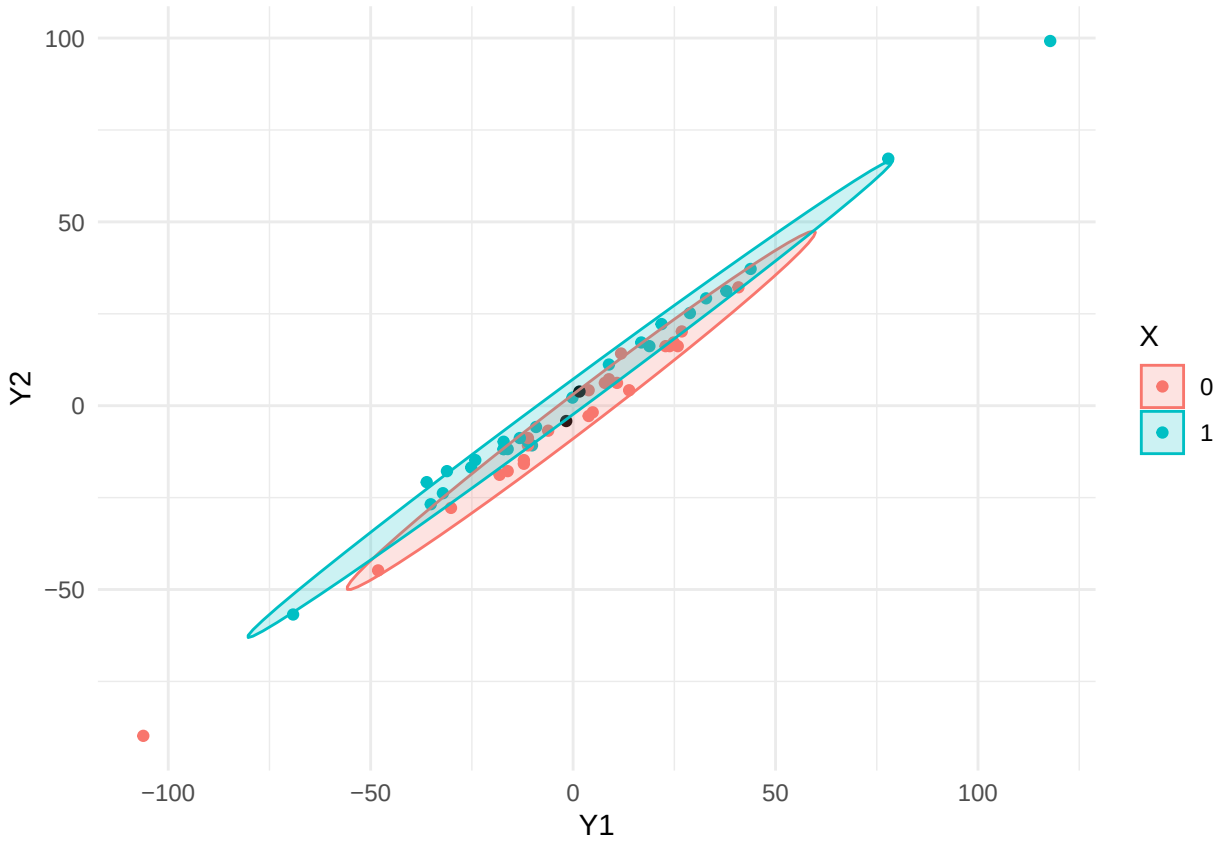
Example: wheat protein

This data contains measurements on protein content and the logarithms of near-infrared reflectance at six wavelengths across the range 1680-2310 nm measured on each of $n = 50$ samples of ground wheat. We will consider an analysis of the first two responses (Y_1, Y_2) and convert the continuous measure of protein content into a categorical variable indicating low and high levels of protein. Here, the mean difference, $\mu_2 - \mu_1$ corresponds to β in model (4) where $X = 0$ indicates a high level of protein and $X = 1$ indicates a low level of protein. This dataset is in the `Renvlp` package..

First, let's plot the data with the means of both responses at each protein level plotted.

```
library(Renvlp)
library(tidyverse)
library(ggplot2)
library(reshape2)
data(wheatprotein)

dat = data.frame(Y1 = wheatprotein[, 1] - mean( wheatprotein[, 1]),
                 Y2 = wheatprotein[, 2] - mean( wheatprotein[, 2]),
                 X  = wheatprotein[, 8])
dat$X = as.factor(dat$X)
foo = unlist(lapply(split(dat, f = dat$X), function(xx) colMeans(xx[, 1:2])))
dat_means = data.frame(Y1 = foo[c(1,3)], Y2 = foo[c(2,4)])
ggplot(dat) + aes(x = Y1, y = Y2, color = X) +
  geom_point() +
  theme_minimal() +
  geom_point(data=dat_means,  mapping=aes(x = Y1, y = Y2), col="black") +
  stat_ellipse(geom = "polygon", aes(fill = X), alpha = 0.20)
```



We will now consider an envelope model with $\hat{u} = 1$. Model selection criteria support such a choice.

```
# which dimension?
u.env(X = as.numeric(dat$X), Y = dat[, 1:2])
```

```
## $u.aic
## [1] 1
##
## $u.bic
## [1] 1
##
## $u.lrt
## [1] 1
##
## $loglik.seq
## [1] -383.5512 -364.3534 -364.1719
##
## $aic.seq
## [1] 777.1024 740.7067 742.3438
##
## $bic.seq
## [1] 786.6625 752.1788 755.7279
```

The ratios of the standard multivariate linear regression asymptotic standard errors over those of the envelope estimator are displayed below (for each element in β). THe envelope model yields substantial variance reduction.

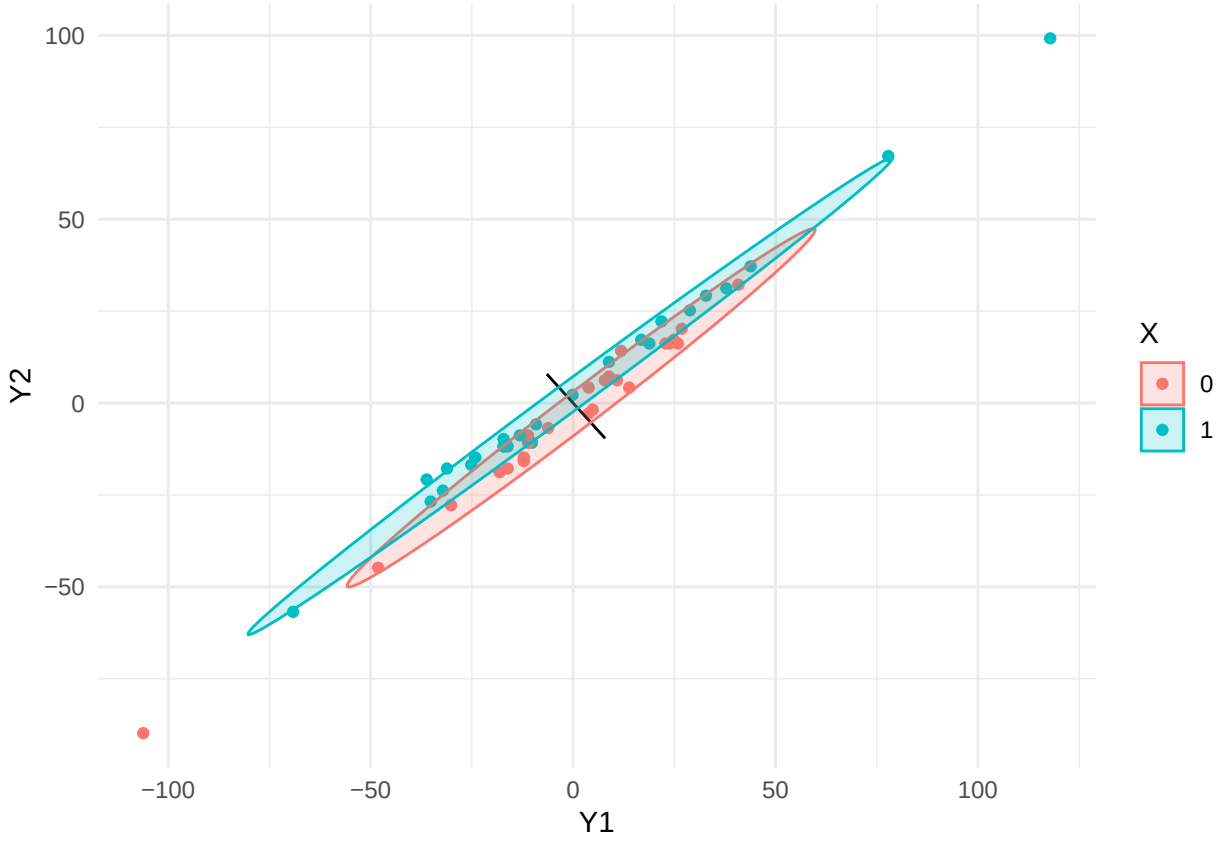
```
# ratios at u = 1
env_mod = env(X = as.numeric(dat$X), Y = dat[, 1:2], u = 1)
```

```
env_mod$ratio
```

```
##          [,1]
## [1,] 28.40504
## [2,] 19.23553
```

We depict the envelope subspace added to the previous plot that only displayed the distribution of responses. We can see how the envelope model works by thinking through this visualization. The separate marginal distributions of the response have considerable overlap. Projecting into the envelope subspace reveals a separation across the categories of X.

```
dat_means2 = data.frame(Y1 = c(env_mod$mu[1] + 4*env_mod$beta[1], env_mod$mu[1] - 1.5*env_mod$beta[1]),
                        Y2 = c(env_mod$mu[2] + 4*env_mod$beta[2], env_mod$mu[2] - 1.5*env_mod$beta[2]))
ggplot(dat) + aes(x = Y1, y = Y2, color = X) +
  geom_point() +
  theme_minimal() +
  geom_line(data=dat_means2,  mapping=aes(x = Y1, y = Y2), col="black") +
  stat_ellipse(geom = "polygon", aes(fill = X), alpha = 0.20)
```



Maximum likelihood estimation

In this section we discuss MLEs of the parameters in an envelope model assuming centered predictors when the dimension $u = \dim(\mathcal{E}_\Sigma(\mathcal{B}))$ is known. The log likelihood $l_u(\alpha, \eta, \mathcal{E}_\Sigma(\mathcal{B}), \Omega, \Omega_0)$ with known u can be expressed as

$$l_u = -\frac{nr}{2}\log(2\pi) - \frac{n}{2}\log|\Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T|$$

$$-\frac{1}{2}\sum_{i=1}^{n}(Y_i-\alpha-\Gamma\eta X_i)^T(\Gamma\Omega\Gamma^T+\Gamma_0\Omega_0\Gamma_0^T)^{-1}(Y_i-\alpha-\Gamma\eta X_i).$$

From Corollary A.1 in Cook [2018] we have that

$$\log|\Gamma\Omega\Gamma^T+\Gamma_0\Omega_0\Gamma_0^T|=\log|\Omega|+\log|\Omega_0|,\qquad\text{and}\qquad(\Gamma\Omega\Gamma^T+\Gamma_0\Omega_0\Gamma_0^T)^{-1}=\Gamma\Omega^{-1}\Gamma^T+\Gamma_0\Omega_0^{-1}\Gamma_0^T.$$

Thus,

$$\begin{aligned}l_u=&-\frac{nr}{2}\log(2\pi)-\frac{n}{2}\log|\Omega|-\frac{n}{2}\log|\Omega_0|\\&-\frac{1}{2}\sum_{i=1}^{n}(Y_i-\alpha-\Gamma\eta X_i)^T(\Gamma\Omega^{-1}\Gamma^T+\Gamma_0\Omega_0^{-1}\Gamma_0^T)(Y_i-\alpha-\Gamma\eta X_i).\end{aligned}$$

The MLE of α is $\hat{\alpha}=\bar{Y}$. Substituting $\hat{\alpha}$ for α and decomposing $Y_i-\bar{Y}=P_\Gamma(Y_i-\bar{Y})+Q_\Gamma(Y_i-\bar{Y})$ and simplifying we arrive at the first partially maximized log likelihood,

$$l_u^{(1)}(\eta,\mathcal{E}_\Sigma(\mathcal{B}),\Omega,\Omega_0)=-\frac{nr}{2}\log(2\pi)+l_u^{(11)}(\eta,\mathcal{E}_\Sigma(\mathcal{B}),\Omega)+l_u^{(12)}(\mathcal{E}_\Sigma(\mathcal{B}),\Omega_0)$$

where

$$\begin{aligned}l_u^{(11)}&=-\frac{n}{2}\log|\Omega|-\frac{1}{2}\sum_{i=1}^{n}[\Gamma^T(Y_i-\bar{Y})-\eta X_i]^T\Omega^{-1}[\Gamma^T(Y_i-\bar{Y})-\eta X_i],\\l_u^{(12)}&=-\frac{n}{2}\log|\Omega_0|-\frac{1}{2}\sum_{i=1}^{n}(Y_i-\bar{Y})^T\Gamma_0\Omega_0^{-1}\Gamma_0^T(Y_i-\bar{Y}).\end{aligned}$$

Holding Γ fixed, $l_u^{(11)}$ can be seen as the log likelihood for the multivariate regression of $\Gamma^T(Y_i-\bar{Y})$ on X_i and thus $l_u^{(11)}$ is maximized over η at $\eta=\Gamma^T\hat{\beta}_{\text{OLS}}$ (with respect to centered responses). Substituting this into $l_u^{(11)}$ we obtained the second partially maximized log likelihood,

$$l_u^{(21)}(\mathcal{E}_\Sigma(\mathcal{B}),\Omega)=-\frac{n}{2}\log|\Omega|-\frac{1}{2}\sum_{i=1}^{n}(\Gamma^T r_i)^T\Omega^{-1}\Gamma^T r_i,$$

where r_i is the ith residual vector corresponding to the standard model (4).

Now, with Γ fixed, $l_u^{(21)}$ is maximized over Ω at $\Omega=\Gamma^T S_{Y|X}\Gamma$ where $S_{Y|X}=\frac{1}{n}\sum_{i=1}^{n}r_i r_i^T=\frac{1}{n}\mathbb{Y}^T Q_{\mathbb{X}}\mathbb{Y}$. Consequently, we arrive at the third partially maximized log-likelihood,

$$l_u^{(31)}(\mathcal{E}_\Sigma(\mathcal{B}))=-\frac{n}{2}\log|\Gamma^T S_{Y|X}\Gamma|-\frac{nu}{2}.$$

By similar reasoning, the value of Ω_0 that maximizes $l_u^{(22)}(\mathcal{E}_\Sigma(\mathcal{B}),\Omega_0)$ over Ω_0 is $\Omega_0=\Gamma_0^T S_Y\Gamma_0$ where $S_Y=\frac{1}{n}\mathbb{Y}^T\mathbb{Y}$. This leads to

$$l_u^{(32)}(\mathcal{E}_\Sigma(\mathcal{B}))=-\frac{n}{2}\log|\Gamma_0^T S_Y\Gamma_0|-\frac{n(r-u)}{2}.$$

Combining the above steps, we arrive at the partially maximized log likelihood

$$l_u^{(3)}(\mathcal{E}_\Sigma(\mathcal{B}))\propto-\frac{n}{2}\log|\Gamma^T S_{Y|X}\Gamma|-\frac{n}{2}\log|\Gamma_0^T S_Y\Gamma_0|.$$

Lemma A.13 in Cook [2018] states that $\log|\Gamma_0^T S_Y\Gamma_0|=\log|S_Y|+\log|\Gamma^T S_Y^{-1}\Gamma|$ which means that the above can be written as a function of Γ alone. We now have

$$l_u^{(3)}(\mathcal{E}_\Sigma(\mathcal{B}))\propto-\frac{n}{2}\log|\Gamma^T S_{Y|X}\Gamma|-\frac{n}{2}\log|\Gamma^T S_Y^{-1}\Gamma|.$$

The MLEs are now all functions of Γ, and we obtain these MLEs as

$$\begin{aligned}
\hat{\mathcal{E}}_\Sigma(\mathcal{B}) &= \text{span}\{\text{argmin}_G(\log|G^T S_{Y|X} G| + \log|G^T S_Y^{-1} G|)\},\\
\hat{\eta} &= \widehat{\Gamma}^T \hat{\beta}_{\text{OLS}},\\
\hat{\beta}_{\text{env}} &= \widehat{\Gamma}\eta = \widehat{\Gamma}\widehat{\Gamma}^T\hat{\beta}_{\text{OLS}} = P_{\hat{\mathcal{E}}_\Sigma(\mathcal{B})}\hat{\beta}_{\text{OLS}},\\
\widehat{\Omega} &= \widehat{\Gamma}^T S_{Y|X}\widehat{\Gamma},\\
\widehat{\Omega}_0 &= \widehat{\Gamma}_0^T S_Y \widehat{\Gamma}_0,\\
\widehat{\Sigma} &= \widehat{\Gamma}\widehat{\Omega}\widehat{\Gamma}^T + \widehat{\Gamma}_0\widehat{\Omega}_0\widehat{\Gamma}_0^T,
\end{aligned}$$

where $\min_G$ is over all semi-orthogonal matrices $G \in \mathbb{R}^{r\times u}$, and $\widehat{\Gamma}$ is any semi-orthogonal basis matrix for $\hat{\mathcal{E}}_\Sigma(\mathcal{B})$.

Weighted envelope estimation

Estimation for all of the above quantities is conducted via the `env` function in the `Renvlp` package which we saw earlier with the wheat protein data. We also plugged in an estimated $\hat{u}$ when fitting the envelope model using `env`. The `u.env` function showed that model selection criteria such as AIC and BIC and a LRT at the $\alpha = 0.05$ testing level supported the choice $\hat{u} = 1$. All of these model comparisons are based on maximizing the likelihood, details of which are presented above.

Note that further inferences are conditional on $\hat{u} = 1$. Also note that the LRT is nonstandard in the envelope context. Envelope models fit at different u values are not necessarily nested, but an envelope model is always nested within the original model ($u = r$). Also note the critical role of estimating u. Overestimation of the correct dimension implies that extraneous variation will be included. However, underestimation of the correct dimension implies that estimation will be inconsistent and is perhaps the more serious error. We can use model averaging across bootstrap samples to estimate envelope estimation variability where such a procedure accounts for the variability in the estimated envelope dimension. Consider the weighted envelope estimator

$$\hat{\beta}_w = \sum_{u=1}^{r} w_u \hat{\beta}_u,$$

where weights are computed as

$$w_u = \frac{\exp(-b_u)}{\sum_{k=1}^{r}\exp(-b_k)},$$

and b_k is the BIC criterion evaluated at the envelope estimator $\hat{\beta}_k$. See Eck and Cook [2017] for more details.

Example: wheat protein (again)

We can use the `weighted.env` function to estimation the variability of the weighted envelope estimator in the wheat protein example. We see that accounting for model selection variability drastically lowers the efficiency gains produced by an analysis that conditions on $\hat{u} = 1$, but meaningful variance reduction is still observed.

```
set.seed(13)
wtenv = weighted.env(X = as.numeric(dat$X), Y = dat[, 1:2], bstrpNum = 1e3)

## ratios wrt to weighted envelope estimator after bootstrapping
wtenv$ratios
```

```
##          [,1]
## [1,] 2.334444
## [2,] 2.333241
```

```
## ratios conditional on u = 1
env_mod$ratio
```

```
##          [,1]
## [1,] 28.40504
## [2,] 19.23553
```

```
## number of times each dimension is selected
wtenv$bic_select
```

```
##   1   2
## 953  47
```

Acknowledgments

These notes borrow materials from Nathaniel E. Helwig's Multivariate Linear Regression notes and Cook [2018].

References

R Dennis Cook. *An introduction to envelopes: dimension reduction for efficient estimation in multivariate statistics*. John Wiley & Sons, 2018.

R Dennis Cook. Envelope methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(2):e1484, 2020.

R Dennis Cook, Bing Li, and Francesca Chiaromonte. Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, pages 927–960, 2010.

Daniel J Eck and R Dennis Cook. Weighted envelope estimation to handle variability in model selection. *Biometrika*, 104(3):743–749, 2017.

Helmut Lutkepohl. *Handbook of matrices*. Wiley, 1996.