The General Linear Model

Peter Hoff

October 21, 2020

Contents

1 The linear model with correlated outcomes 2

2 GLS for multivariate regression 6

3 Covariance estimation for the GLM 9

4 Testing the GLH 13

5 Example: NHANES analysis 21

Abstract

We study the linear model with correlated outcomes, and relate
it to multivariate linear regression, also known as the general linear model.

1 The linear model with correlated outcomes

Recall that the linear model is

\[ y = X\beta + \Sigma^{1/2}z \]

\[ \mathbb{E}[z] = 0 \]

\[ \text{Var}[z] = I, \]

where \( X \in \mathbb{R}^{n \times p}, \beta \in \mathbb{R}^p \) and \( \Sigma \in S^n_+ \). The term “linear” here typically refers to the mean parameters: The expectation of any element \( y_i \) of \( y \) is linear in the parameters. If the parameter space for \( \Sigma \) is unrestricted, then the covariance model is “linear” as well. If the covariance is restricted, say \( \Sigma = \Sigma(\theta) \) for \( \theta \in \Theta \), then typically the covariance model is nonlinear in its parameters (e.g. autoregressive covariance).

A wide variety of models for random matrices can be viewed as submodels of the linear model:

1. Multivariate random sample: Let \( y_1, \ldots, y_n \) be uncorrelated, each with mean \( \mu \) and variance \( \Sigma \). Then

\[ Y = 1\mu^\top + Z\Sigma^{1/2} \]

\[ y = (I \otimes 1)\mu + (\Sigma^{1/2} \otimes I)z. \]

2. Row and column effects models: Let \( \{y_{i,j} : i = 1, \ldots, n; j = 1, \ldots, p\} \)
be uncorrelated, with $E[y_{i,j}] = \mu + a_i + b_j$ and $\text{Var}[y_{i,j}] = \psi_i^2 \sigma_j^2$

$$Y = \mu 1 1^T + a 1^T + 1 b^T + \Psi^{1/2} Z \Sigma^{1/2}$$

$$y = 1 \mu + (1 \otimes I) a + (I \otimes 1) b + (\Sigma \otimes \Psi)^{1/2} z$$

$$= [1 (1 \otimes I) (I \otimes 1)] \begin{pmatrix} \mu \\ a \\ b \end{pmatrix} + (\Sigma \otimes \Psi)^{1/2} z.$$

3. General linear model: Let $y_1, \ldots, y_n$ be uncorrelated, with $E[y_i] = B^\top x_i$ and $\text{Var}[y_i] = \Sigma$. Then

$$Y = XB + Z \Sigma^{1/2}$$

$$y = (I \otimes X) b + (\Sigma^{1/2} \otimes I) z.$$  

4. Correlated rows: Suppose $\text{Var}[Y_{[i,]}] \propto \Sigma$, $\text{Var}[Y_{[j,]}] \propto \Psi$. Then

$$Y = XB + \Psi^{1/2} Z \Sigma^{1/2}$$

$$y = (I \otimes X) b + (\Sigma^{1/2} \otimes \Psi^{1/2}) z.$$  

5. Row and column covariates: Suppose $E[y_{i,j}] = x_i^\top B w_j$, for example,

$$Y = X BW^T + \Psi^{1/2} Z \Sigma^{1/2}$$

$$y = (W \otimes X) b + (\Sigma^{1/2} \otimes \Psi^{1/2}) z.$$  

6. Random effects models: Integrate any of the above over the possible values of some subset of parameters, with respect to a probability distribution.

So we begin by studying the deceptively simple general form of the linear model, $y = X \beta + \Sigma^{1/2} z$.  

3
**OLS:** For a given value of $\beta$, the residual sum of squares (RSS) is given by

$$\text{RSS}(\beta) = (y - X\beta)^\top(y - X\beta).$$

Differentiation or completing the square shows that the minimizer in $\beta$ is

$$\hat{\beta} = (X^\top X)^{-1}X^\top y.$$ 

By writing $\hat{\beta} = \beta + (X^\top X)^{-1}X^\top \Sigma^{1/2} z$, it is straightforward to show that the OLS estimator is unbiased,

$$E[\hat{\beta}] = \beta,$$

and that the variance of this estimator is

$$\text{Var}[\hat{\beta}] = (X^\top X)^{-1}(X^\top \Sigma X^\top)(X^\top X)^{-1}.$$ 

If $\Sigma = \sigma^2 I$ then the OLS estimator is optimal in terms of variance:

**Theorem 1** (Gauss-Markov). Suppose $\text{Var}[y] = \sigma^2 I$ and let $\tilde{\beta} = Ay$ be an unbiased estimator for $\beta$. Then

$$\text{Var}[\tilde{\beta}] = \text{Var}[\hat{\beta}] + E[(\tilde{\beta} - \hat{\beta})(\tilde{\beta} - \hat{\beta})^\top]$$

**Exercise 1.** Characterize the class of unbiased estimators of $\beta$.

**Exercise 2.** Prove Theorem 1 and show that $\text{Var}[a^\top \tilde{\beta}] \geq \text{Var}[a^\top \hat{\beta}]$ for any $a \in \mathbb{R}^p$.

If $\Sigma \neq \sigma^2 I$ then the OLS estimator does not have minimum variance among unbiased estimators.

**Exercise 3.** Obtain the form of $\beta$, $\hat{y} = X\beta$, and $\text{Var}[\hat{\beta}]$ in terms of components of the SVD of $X$. Comment on the following intuitive statement: “If $y \approx X\beta$ then $X^{-1}y \approx \beta$.”
**GLS:** If we knew $\Sigma$ we could obtain a decorrelated vector $\tilde{y} = \Sigma^{-1/2}y$. The model for $\tilde{y}$ is

$$\tilde{y} = \Sigma^{-1/2}y = \Sigma^{-1/2}X\beta + \Sigma^{-1/2}\Sigma^{1/2}z$$

$$\tilde{y} = \tilde{X}\beta + z.$$  

The OLS estimator based on the decorrelated data and corresponding regressors is

$$\hat{\beta}_{OLS} = (\tilde{X}^\top\tilde{X})^{-1}\tilde{X}^\top\tilde{y}.$$  

By the Gauss-Markov theorem, this estimator is optimal among all linear unbiased estimators, that is unbiased estimators of the form $A\tilde{y}$. Obviously, this includes unbiased estimators of the form $Ay$, since $\tilde{y}$ is a linear transformation of $y$.

What is this estimator in terms of the original data and covariates?

$$\hat{\beta}_{GLS} = \hat{\beta}_{OLS} = (X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}y.$$  

The variance of this estimator is easily shown to be

$$\text{Var}[\hat{\beta}_{GLS}] = (X^\top\Sigma^{-1}X)^{-1}.$$  

This is the “smallest” variance attainable among unbiased linear estimators, since the GLS estimate is the BLUE - best linear unbiased estimator.

Furthermore, suppose we assume normal errors, $z \sim N_p(0, I)$. For known $\Sigma$, the MLE of $\beta$ is the minimizer of

$$-2 \log p(y|X, \beta) = c + (y - X\beta)^\top\Sigma^{-1}(y - X\beta).$$
Taking derivatives, the MLE is the GLS estimator.

**Big picture:** If \( y \sim N_n(X\beta, \Sigma) \) and \( \Sigma \) is known, \((X^\top \Sigma^{-1}X)^{-1}X^\top \Sigma^{-1}y\) is

- the OLS estimator based on the decorrelated data \( \Sigma^{-1/2}y \);
- the GLS estimator;
- the BLUE;
- the ML estimator.

Now the problem is that \( \Sigma \) is rarely known. In some cases, particularly if \( \Sigma = \Sigma(\theta) \) for some low dimensional parameter \( \theta \), we can jointly estimate \( \Sigma \) and \( \beta \). This is known as "feasible GLS."

## 2 GLS for multivariate regression

Perhaps the most commonly used model for multivariate data analysis is multivariate regression, a.k.a. the general linear model:

\[
Y = XB + Z\Sigma^{1/2} \\
y = (I \otimes X)b + (\Sigma^{1/2} \otimes I)z
\]

**GLS:** Using the vector form for the model, the GLS estimator of \( b = \text{vec}(B) \) is

\[
\hat{b}_{GLS} = ((I \otimes X)^\top (\Sigma \otimes I)^{-1}(I \otimes X))^{-1} ((I \otimes X)^\top (\Sigma^{-1} \otimes I)y) \\
= (\Sigma^{-1} \otimes X^\top X)^{-1}(\Sigma^{-1} \otimes X^\top)y \\
= (I \otimes (X^\top X)^{-1}X^\top)y
\]
Now “unvectorize”:

\[
\text{vec}(\hat{B}_{GLS}) = (I \otimes (X^\top X)^{-1}X^\top)y
\]

\[
\hat{B}_{GLS} = (X^\top X)^{-1}X^\top Y.
\]

Notice that \( \Sigma \) does not appear in formula for the GLS estimator. Think about why this is:

- The multivariate regression model is just a regression model for each variable:
  \[
y_j = Xb_j + e_j
\]
  So the \( j \)th column of \( B \in \mathbb{R}^{q \times p} \) is \( b_j \), the vector of regression coefficients for the \( j \)th variable.

- In this model, there is no correlation among the elements of \( y_j \)/rows of \( Y \).

- We already knew that in this case the BLUE for \( b_j \) based on \( y_j \) is the OLS estimator: \( \hat{\beta}_j = (X^\top X)^{-1}X^\top y_j \).

- Now is this BLUE among estimators based on \( Y \)? The question comes down to whether or not observing some variables \( Y_{[-j]} \) correlated with \( y_j \) helps with estimating \( b_j \). The answer given by the above calculation is no.

- Binding the OLS estimators columnwise gives the multivariate OLS/GLS estimator: \( \hat{B} = [\hat{\beta}_1 \cdots \hat{\beta}_p] \)

To summarize, the estimator \((X^\top X)^{-1}X^\top Y\) is

- the column-bound univariate OLS estimator;
• the GLS and OLS estimator;
• the BLUE;
• the MLE under normal errors.

What is the covariance of $\hat{B}$? How do we even write this out, since $\hat{B}$ is a matrix? As usual, we vectorize. Recall the vector form for the GLM:

\[
Y = XB + Z\Sigma^{1/2} \\
y = (I \otimes X)b + (\Sigma^{1/2} \otimes I)z \\
= \tilde{X}b + \tilde{\Sigma}^{1/2}z
\]

Now we already calculated the variance of a GLS estimator:

\[
\text{Var}[\hat{B}] \equiv \text{Var}[\hat{b}] = (\tilde{X}^T \tilde{\Sigma}^{-1} \tilde{X})^{-1} \\
= ((I \otimes X^T)(\Sigma^{-1} \otimes I)(I \otimes X))^{-1} \\
= (\Sigma^{-1} \otimes X^T X)^{-1} \\
= \Sigma \otimes (X^T X)^{-1}.
\]

This means that

\[
\text{Cov}[\hat{B}_{[i,j]}, \hat{B}_{[j',k]}] = \sigma_{j,j'}(X^T X)^{-1} \\
\text{Cov}[\hat{B}_{[k,j]}, \hat{B}_{[k',j']} = s_{k,k'}\Sigma \\
\text{Cov}[\hat{b}_{k,j}, \hat{b}_{k',j'}] = \sigma_{j,j'}s_{k,k'}
\]

where $s_{k,k'} = (X^T X)^{-1}_{k,k'}$. In particular, $\text{Cov}[\hat{b}_{k,j}, \hat{b}_{k',j'}] = \sigma_{j,j'}s_{k,k'}$. So in a sense, $\Sigma$ is the column covariance of $\hat{B}$ (corresponding to variables) and $(X^T X)^{-1}$ is the row covariance (corresponding to predictors).

**Exercise 4.** Obtain the posterior distribution of $B$ based on data from the model $Y \sim N_{n \times p}(XB, \Sigma \otimes I)$ and the following prior distributions:
1. \( \text{vec}(B) = b \sim N_{qp}(0, \Psi) \).

2. \( B \sim N_{q \times p}(0, \Sigma \otimes \Psi) \)

3. \( B \sim N_{q \times p}(0, \Psi \otimes (X^T X)^{-1}) \)

4. \( B \sim N_{q \times p}(0, a \times (\Sigma \otimes (X^T X)^{-1})) \)

3 Covariance estimation for the GLM

Since the OLS/GLS estimator of \( B \) is \( (X^T X)^{-1} X^T Y \), the fitted values for \( Y \) are

\[
\hat{Y} = X \hat{B}
= X (X^T X)^{-1} X^T Y
= P_X Y.
\]

So the fitted values of the matrix \( Y \) are just the column-wise projections of \( Y \) on to the space spanned by the columns of \( X \).

Obvious but important is that \( \hat{B} \) itself is a function of \( Y \) through \( \hat{Y} \):

\[
\hat{B} = (X^T X)^{-1} X^T Y
= (X^T X)^{-1} (X^T X) (X^T X)^{-1} X^T Y
= (X^T X)^{-1} X^T (X (X^T X)^{-1} X^T) Y
= (X^T X)^{-1} X^T P_X Y
= (X^T X)^{-1} X^T \hat{Y}
\]
The residuals are the projection onto the null space of $X$

$$
\hat{E} = Y - \hat{Y}
= (I - P_x)Y
= (I - X(X^T X)^{-1}X)Y \equiv P_0 Y.
$$

This gives the decomposition

$$
Y = \hat{Y} + \hat{E}.
$$

Is $\hat{Y}$ correlated with $\hat{E}$? Let’s find out. Let $\hat{y}$ and $\hat{e}$ be the vectorizations of $\hat{Y}$ and $\hat{E}$ respectively. The covariance between these two matrices is the covariance between their vectorizations.

$$
\text{Cov}[\hat{y}, \hat{e}] = E[(\hat{y} - E[\hat{y}])\hat{e}^T]
$$

Now what are these items?

$$
\hat{Y} - E[\hat{Y}] = P_x Y - E[P_x Y]
= P_x (Y - E[Y])
= P_x E,
$$

$$
\hat{y} - E[\hat{y}] = (I \otimes P_x)e.
$$

$$
\hat{E} = P_0 Y
= P_0 (XB + E)
= P_0 E,
$$

$$
\hat{e} = (I \otimes P_0)e.
$$
Therefore,

\[
\text{Cov}[\hat{y}, \hat{e}] = E[(\hat{y} - E[\hat{y}])\hat{e}^\top] \\
= E[(I \otimes P_X)e e^\top (I \otimes P_0)] \\
= 0.
\]

So as with ordinary regression, the residuals are uncorrelated with the fitted values, and hence also uncorrelated with \( \hat{B} \). This is useful, as it allows us to obtain an estimate of \( \Sigma \) that is independent of our estimate of \( B \) if the errors are normal.

Now let’s consider estimation of \( \Sigma \) from \( \hat{E} = P_0E \). Under the model we have \( E = Z\Sigma^{1/2} \), so \( \hat{E} = P_0Z\Sigma^{1/2} \), and

\[
S = \hat{E}^\top \hat{E} = \Sigma^{1/2}Z^\top P_0P_0Z\Sigma^{1/2} \\
E[S] = \Sigma^{1/2}E[Z^\top P_0Z]\Sigma^{1/2} \\
= \Sigma^{1/2}(\text{tr}(P_0)I)\Sigma^{1/2} \\
= \Sigma \times (n - q)
\]

Therefore, \( \hat{\Sigma} = \hat{E}^\top \hat{E} / (n - q) \) is an unbiased estimator of \( \Sigma \).

If the data are Gaussian then we can say more. Suppose

\[
Y = XB + Z\Sigma^{1/2} \\
Z \sim N_{n \times p}(0, I \otimes I),
\]

or equivalently, \( Y \sim N_{n \times p}(XB, \Sigma \otimes I) \). Then

\[
\hat{E} = P_0Z\Sigma^{1/2}.
\]

Now recall from several chapters ago the following:
• $P_0$ is a rank-$(n - q)$ idempotent projection matrix;

• $P_0 = GG^T$ for some $G \in \mathcal{V}_{n-q,n} \subset \mathbb{R}^{n \times (n-q)}$;

• If $Z \sim N_{n \times p}(0, I \otimes I)$ then $G^\top Z \sim N_{(n-q) \times p}(0, I \otimes I)$.

So we have

$$S = \hat{E}^\top \hat{E} = \Sigma^{1/2} Z^\top P_0 P_0 Z \Sigma^{1/2} = \Sigma^{1/2} Z^\top P_0 Z \Sigma^{1/2} = \Sigma^{1/2} Z^\top G G^\top Z \Sigma^{1/2} = W^\top W,$$

where

$$W = G^\top Z \Sigma^{1/2} \sim N_{(n-q) \times p}(0, \Sigma).$$

Therefore $S = \hat{E}^\top \hat{E} = W^\top W \sim \text{Wishart}_{n-q}(\Sigma)$.

We’ve seen a simpler version of this before: Suppose $Y \sim N_{n \times p}(1\mu^\top, \Sigma \otimes I)$. Although it doesn’t look like it, this is a multivariate linear regression model where the regression model for each column $j$ is $y_j = 1\mu_j + e_j$. That is, there is $q = 1$ “predictor” and it is constant for all rows:

• $X = 1$ and $B = \mu^\top$;

• $P_X = 1(1^\top 1)^{-1}1^\top = 1^\top/n$;

• $P_0 = I - 11^\top/n = C$;

• $S = \hat{E}^\top \hat{E} = Y^\top CY \sim \text{Wishart}_{n-1}(\Sigma)$.

**Exercise 5.** *Show that the MLE of $(B, \Sigma)$ is $(\hat{B}, S/n)$.*
4 Testing the GLH

We now derive some classical testing and confidence procedures for the regression matrix $B$ of the Gaussian GLM:

$$Y \sim N_{n \times p}(XB, \Sigma \otimes I).$$

Some results we’ve obtained so far:

- $\hat{B} = (X^\top X)^{-1}X^\top Y \sim N_{q \times p}(B, \Sigma \otimes (X^\top X)^{-1})$;
- $S = \hat{E}^\top \hat{E} \sim \text{Wishart}(n-q)(\Sigma)$;
- $\hat{B}$ and $\hat{E}$ are uncorrelated and so independent, as are $\hat{B}$ and $S$.

Consider first inference for $B_{[k,]} = b_k^\top$, the (transposed) $k$th row of the matrix $B$. This $p \times 1$ vector represents the effects of the $k$th predictor $X_{[k]}$ on the $p$ column variables of $Y$: $Y$:

$$Y = XB + E = \sum_{k=1}^{q} X_{[q]} b_k^\top + E.$$

In the context of making inference about $b_k$, our results above tell us that

- $\hat{b}_k \sim N_p(b_k, h_{kk} \times \Sigma)$;
- $S \sim \text{Wishart}(\Sigma, n - q)$.
- $\hat{b}_k$ is independent of $S$.

where $h_{kk} = (X^\top X)^{-1}_{kk}$. Our objective is to make a confidence region or test a hypothesis about $b_k$ based on $\hat{b}_k$. 

13
Recall that a particularly simple case of this is when \( q = 1 \) and \( X = 1 \). This is just the multivariate normal model \( Y \sim N_{n \times p}(1\mu^\top, \Sigma \otimes I) \). Recall our results, both old and new, imply that

- \( \bar{y} \sim N(\mu, \Sigma/n) \)
- \( S \sim \text{Wishart}(\Sigma, n - 1) \).
- \( \bar{y} \) is independent of \( S \).

The point is that a statistical method developed to infer \( \mu \) from \((\bar{y}, S)\) can be used to infer \( b_k \) from \((\hat{b}_k, S)\). So we proceed with the former since the notation is simpler.

Suppose \( \Sigma \) were known and we wanted to test \( H : E[y] = \mu \). We then might construct a test statistic as follows:

\[
\bar{y} \sim N(\mu, \Sigma/n)
\]
\[
\sqrt{n}(\bar{y} - \mu) \sim N(0, \Sigma)
\]
\[
z(\mu) = \sqrt{n}\Sigma^{-1/2}(\bar{y} - \mu) \sim N(0, I)
\]
\[
z^2(\mu) = n(\bar{y} - \mu)^\top\Sigma^{-1}(\bar{y} - \mu) \sim \chi_p^2.
\]

So in this case the null distribution is \( \chi_p^2 \), and of course does not depend on the parameters. In this case that \( \Sigma \) is known, this statistic is a pivot statistic, that is, has a distribution that doesn’t depend on any unknown parameters. We use such statistics to test and infer values of \( \mu \).

Since \( \Sigma \) is unknown we might try replacing it with the unbiased estimator \( \hat{\Sigma} \):

\[
t^2(\mu) = n(\bar{y} - \mu)^\top\hat{\Sigma}^{-1}(\bar{y} - \mu).
\]

Can we use this for testing and confidence region construction? To do so requires the following:
• show that $t^2$ is pivotal, that is, doesn’t depend on unknown parameters;
• derive the distribution of $t^2$.

**Pivotal quantity:** In general, an unbiased estimate of $\Sigma$ based on $S \sim W(\Sigma, \nu)$ is $\hat{\Sigma} = S/\nu$. Recall that the Wishart matrix $S$ has a representation as

$$S = E^T E$$

$$E \sim N_{p, p}(0, \Sigma \otimes I).$$

Now of course $E \overset{d}{=} Z\Sigma^{1/2}$ where the elements of $Z$ are i.i.d. standard normal, so

$$S \overset{d}{=} \Sigma^{1/2}Z^T Z\Sigma^{1/2}$$

$$\hat{\Sigma} \overset{d}{=} \Sigma^{1/2}(W/\nu)\Sigma^{1/2}.$$

Now recall

$$\sqrt{n}(\bar{y} - \mu) \sim N(0, \Sigma)$$

$$\sqrt{n}\Sigma^{-1/2}(\bar{y} - \mu) \sim N(0, I).$$

Therefore

$$t^2 = n(\bar{y} - \mu)^T \hat{\Sigma}^{-1}(\bar{y} - \mu)$$

$$\overset{d}{=} \sqrt{n}(\bar{y} - \mu)^T \Sigma^{-1/2}(W/\nu)^{-1}\Sigma^{-1/2}(\bar{y} - \mu)\sqrt{n}$$

$$\overset{d}{=} z^T (W/\nu)^{-1} z$$

where

• $z \sim N_p(0, I)$
• \( \mathbf{W} \sim W(I, \nu) \)

• \( \mathbf{z} \) and \( \mathbf{W} \) are independent.

Thus \( t^2 \) is indeed pivotal: It has a distribution that does not depend on any unknown parameters (although the form does depend on the unknown parameter \( \mu \)).

**Definition 1** (Hotelling \( T^2 \) distribution). If \( \mathbf{z} \sim N(0, I_p) \) and \( \mathbf{W} \sim W(I, \nu) \), then the distribution of \( t^2 = \mathbf{z}^\top (\mathbf{W}/\nu)^{-1} \mathbf{z} \) is said to be a Hotelling \( T^2 \) distribution, denoted \( t^2 \sim T^2(p, \nu) \).

More useful for our purposes is the following equivalent definition:

**Definition 2** (Hotelling \( T^2 \) distribution). If \( \mathbf{z} \sim N(0, \Sigma) \) and \( \mathbf{W} \sim W(\Sigma, \nu) \), then the distribution of \( t^2 = \mathbf{z}^\top (\mathbf{W}/\nu)^{-1} \mathbf{z} \) is said to be a Hotelling \( T^2 \) distribution, denoted \( t^2 \sim T^2(p, \nu) \), which does not depend on \( \Sigma \).

That \( t^2 \) is pivotal means that we can use it for frequentist testing and confidence region construction.

**Test of a multivariate mean:** Consider testing \( H : \mu = \mu_0 \) versus \( K : \mu \neq \mu_0 \) based on data \( \mathbf{Y} \sim N_{n \times p}(\mu, \Sigma \otimes I) \). From these data we construct the statistics

• \( \bar{\mathbf{y}} = \mathbf{Y}^\top \mathbf{1}/n \)

• \( \hat{\Sigma} = \mathbf{Y}^\top \mathbf{C} \mathbf{Y}/(n - 1) \).

Based on the above discussion, if the null hypothesis is true then

\[
t^2(\mu_0) = n(\bar{\mathbf{y}} - \mu_0)^\top \Sigma^{-1}(\bar{\mathbf{y}} - \mu_0)
\]
has a $T^2(p, n - 1)$ distribution. If we reject $H$ when $t^2 > T_{1-\alpha}^2(p, n - 1)$ then our type I error is $\alpha$.

**Confidence region construction:**

Let $C(\hat{y}, \hat{\Sigma}) = \{\mu : t^2(\hat{y}, \hat{\Sigma}, \mu_0) < T_{1-\alpha}^2(p, n - 1)\}$. Then

$$\Pr(\mu \in C(\hat{y}, \hat{\Sigma})|\mu) = 1 - \alpha \quad \forall \mu \in \mathbb{R}^p.$$

**Inference for multivariate regression coefficients:**

Now let’s return to the multivariate regression model

$$Y = XB + Z\Sigma^{1/2}$$

$$Z \sim N_{n \times p}(0, I \otimes I).$$

For this model,

- the number of observations is $n$;
- the number of variables is $p$
- the number of predictors is $q$

So $X \in \mathbb{R}^{n \times q}$ and $B \in \mathbb{R}^{q \times p}$. We interpret the elements of $B$ as follows:

- column $j$ of $B$ gives the effects of all predictors on variable $j$.
- row $k$ of $B$ gives the effects of predictor $k$ on all variables.

Recall that if you have taken a linear regression class, you are already capable of making inference for a single element $b_{k,j}$ of $B$, or even an entire column
\textbf{b}_j: \text{ Just consider the regression model for the relevant variable } j. \text{ Our new results allow us to make inference for } \textit{individual predictors} \text{ on all columns. For example we may be interested in testing whether or not}

- an Alzheimer’s treatment is associated with any of several cognitive measures;
- a genotype is associated with any of several metabolomic variables.

Let \( \textbf{b}_k \) be a row of \( \textbf{B} \), and consider testing \( H : \textbf{b}_k = \textbf{b}_k^0 \). Our results are the following:

- \( \hat{\textbf{b}}_k \sim \mathcal{N}_p(\textbf{b}_k, h_{kk} \times \Sigma) \);
- \( \textbf{S} \sim \text{Wishart}(\Sigma, n - q) \).
- \( \hat{\textbf{b}}_k \) is independent of \( \textbf{S} \).

Under the null hypothesis,

- \( (\hat{\textbf{b}}_k - \textbf{b}_k^0)/\sqrt{h_{kk}} \sim \mathcal{N}_p(0, \Sigma) \);
- \( (n - q)\hat{\Sigma} \sim \mathcal{W}(\Sigma, n - q) \);
- \( \hat{\textbf{b}}_k \) and \( \hat{\Sigma} \) are independent.

So all the pieces are in place: Under the null hypothesis,

\[
(\hat{\textbf{b}}_k - \textbf{b}_k^0)^\top \hat{\Sigma}^{-1}(\hat{\textbf{b}}_k - \textbf{b}_k^0)/h_{kk,k} \sim T^2_{p,n-q}.
\]

**Calculating the \( T^2 \) distribution:**

Monte Carlo approximation to the \( T^2 \) distribution is straightforward:
1. simulate $z \sim N(0, I_p)$

2. simulate $S \sim W(I, \nu)$

3. compute $t^2 = z^\top (W/\nu)^{-1} z$

Alternatively, you can show the following result:

**Exercise 6.** *Show that if* $f \sim F_{p,\nu-p+1}$ *then*

$$\frac{\nu p}{\nu - p + 1} f \sim T_{p,\nu}^2.$$

*It is good to check results like this numerically when you can:*

```r
n<-50 ; p<-3 ; q<-6
X<-matrix(rnorm(n*q),n,q)
B<-matrix(rnorm(q*p),q,p)
hSigma<-matrix(rnorm(p*p),p,p)

XtXitX<-solve(crossprod(X))%*%t(X)
k<-3
hkk<-solve(crossprod(X))[k,k]

T2vals<-NULL
for(s in 1:50000)
{
  ## data from a regression model
  Y<-X%*%B+matrix(rnorm(n*p),n,p)%*%hSigma
  Bhat<-XtXitX%*%Y
  bk<-Bhat[k,]

  z_dat<- (bk - B[k,] )/sqrt(hkk)
  W_dat<-crossprod( Y-X%*%Bhat )

  # other code...
}
```

19
t2_dat<-t(z_dat)%*% solve(W_dat/(n-q)) %*% z_dat

## simulation based on pivot result
z_sim<-rnorm(p)
W_sim<-crossprod( matrix(rnorm((n-q)*p),n-q,p) )
t2_sim<-t(z_sim)%*% solve(W_sim/(n-q)) %*% z_sim

## simulation from F distribution
t2_f<-rf(1,p,n-q-p+1)*(n-q)*p/(n-q-p+1)

T2vals<-rbind(T2vals, c(t2_dat, t2_sim, t2_f) )

apply(T2vals,2,mean)

[1] 3.30 3.32 3.32

apply(T2vals,2,sd)

[1] 2.88 2.88 2.88
5 Example: NHANES analysis

As a numerical example, let’s use the NHANES data to model several health-related variables (BMI, blood pressure) as a function of some dietary variables, and control for some demographic effects such as age and sex.

First we obtain the complete cases, select out the variables and predictors, remove a small number of cases with anomalous blood pressures, and log transform the calorie data:

```r
## data processing
nhanes<-nhanes[ !is.na(apply(nhanes,1,mean)) ,]
X<-nhanes[,c(1:8,12)] ; X[,2]<-X[,2]-1 ; colnames(X)[2]<"FEMALE"
Y<-nhanes[,c(15,17,18)]

## people with zero BP
X<-X[ Y[,3]>20 ,]
Y<-Y[ Y[,3]>20 ,]
```
## transform diet variables

```
```

## what are the variables and predictors?

```
colnames(Y)
```

```
[1] "BMXBMI" "BPXSY1" "BPXDI1"
```

```
colnames(X)
```

```
[1] "RIDAGEYR" "FEMALE" "DR1TKCAL" "DR1TPROT" "DR1TCARB" "DR1TSUGR"
[7] "DR1TFIBE" "DR1TTFAT" "DR1TCHOL"
```
We probably want to add a quadratic term for age:
## add intercept and quadratic age term
\[ X <- \text{cbind}(1, X[,1], X[,1]^2, X[,-1]) \]
\[
\text{colnames}(X)[1:3] <- c("intercept", "RIDAGEYR", "RIDAGEYR2")
\]
\[ n <- \text{nrow}(Y) ; p <- \text{ncol}(Y) ; q <- \text{ncol}(X) \]

Here are the OLS/GLS/MLEs from R:

\[ \text{lm}(Y \sim -1 + X) \]

Call:
\[ \text{lm}(\text{formula} = Y \sim -1 + X) \]

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>BMXBMI</th>
<th>BPXSY1</th>
<th>BPXDI1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xintercept</td>
<td>25.45274</td>
<td>80.71650</td>
<td>36.39620</td>
</tr>
<tr>
<td>XRIDAGEYR</td>
<td>0.530558</td>
<td>0.459628</td>
<td>1.232645</td>
</tr>
<tr>
<td>XRIDAGEYR2</td>
<td>-0.005067</td>
<td>0.000298</td>
<td>-0.012197</td>
</tr>
<tr>
<td>XFEMALE</td>
<td>0.733693</td>
<td>-3.232005</td>
<td>-1.456368</td>
</tr>
<tr>
<td>XDR1TKCAL</td>
<td>-2.822738</td>
<td>4.777430</td>
<td>1.599476</td>
</tr>
<tr>
<td>XDR1TPROT</td>
<td>0.478016</td>
<td>-0.354521</td>
<td>0.212984</td>
</tr>
<tr>
<td>XDR1TCARB</td>
<td>0.656858</td>
<td>0.582261</td>
<td>0.511036</td>
</tr>
<tr>
<td>XDR1TSUGR</td>
<td>0.258376</td>
<td>-1.104426</td>
<td>-0.173892</td>
</tr>
<tr>
<td>XDR1TFIBE</td>
<td>-0.887301</td>
<td>-1.837439</td>
<td>-1.398397</td>
</tr>
<tr>
<td>XDR1TTFAT</td>
<td>1.604038</td>
<td>-2.017769</td>
<td>-0.795543</td>
</tr>
<tr>
<td>XDR1TCHOL</td>
<td>0.160630</td>
<td>0.280249</td>
<td>-0.104081</td>
</tr>
</tbody>
</table>

Here they are “by hand”, along with the estimate of \( \Sigma \):
B <- solve(t(X) %*% X) %*% t(X) %*% Y

B

<table>
<thead>
<tr>
<th></th>
<th>BMXBMI</th>
<th>BPXSY1</th>
<th>BPXDI1</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>25.45273</td>
<td>80.716503</td>
<td>36.3962</td>
</tr>
<tr>
<td>RIDAGEYR</td>
<td>0.53056</td>
<td>0.459628</td>
<td>1.2326</td>
</tr>
<tr>
<td>RIDAGEYR2</td>
<td>-0.00507</td>
<td>0.000298</td>
<td>-0.0122</td>
</tr>
<tr>
<td>FEMALE</td>
<td>0.73369</td>
<td>-3.232005</td>
<td>-1.4564</td>
</tr>
<tr>
<td>DR1TKCAL</td>
<td>-2.82274</td>
<td>4.777430</td>
<td>1.5995</td>
</tr>
<tr>
<td>DR1TPROT</td>
<td>0.47802</td>
<td>-0.354521</td>
<td>0.2130</td>
</tr>
<tr>
<td>DR1TCARB</td>
<td>0.65686</td>
<td>0.582261</td>
<td>0.5110</td>
</tr>
<tr>
<td>DR1TSUGR</td>
<td>0.25838</td>
<td>-1.104426</td>
<td>-0.1739</td>
</tr>
<tr>
<td>DR1TFIBE</td>
<td>-0.88730</td>
<td>-1.837439</td>
<td>-1.3984</td>
</tr>
<tr>
<td>DR1TTFAT</td>
<td>1.60404</td>
<td>-2.017769</td>
<td>-0.7955</td>
</tr>
<tr>
<td>DR1TCHOL</td>
<td>0.16063</td>
<td>0.280249</td>
<td>-0.1041</td>
</tr>
</tbody>
</table>

SigmaHat <- crossprod(Y - X %*% B) / (n - q)

SigmaHat

<table>
<thead>
<tr>
<th></th>
<th>BMXBMI</th>
<th>BPXSY1</th>
<th>BPXDI1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMXBMI</td>
<td>41.88</td>
<td>14.5</td>
<td>5.92</td>
</tr>
<tr>
<td>BPXSY1</td>
<td>14.54</td>
<td>202.8</td>
<td>59.57</td>
</tr>
<tr>
<td>BPXDI1</td>
<td>5.92</td>
<td>59.6</td>
<td>118.84</td>
</tr>
</tbody>
</table>

cov2cor(SigmaHat)

<table>
<thead>
<tr>
<th></th>
<th>BMXBMI</th>
<th>BPXSY1</th>
<th>BPXDI1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMXBMI</td>
<td>1.0000</td>
<td>0.158</td>
<td>0.0839</td>
</tr>
<tr>
<td>BPXSY1</td>
<td>0.1577</td>
<td>1.000</td>
<td>0.3838</td>
</tr>
<tr>
<td>BPXDI1</td>
<td>0.0839</td>
<td>0.384</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Compare to marginal variance

\[ \text{var}(Y) \]

<table>
<thead>
<tr>
<th></th>
<th>BMXBMI</th>
<th>BPXSY1</th>
<th>BPXDI1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMXBMI</td>
<td>51.8</td>
<td>37.5</td>
<td>26.4</td>
</tr>
<tr>
<td>BPXSY1</td>
<td>37.5</td>
<td>311.8</td>
<td>108.6</td>
</tr>
<tr>
<td>BPXDI1</td>
<td>26.4</td>
<td>108.6</td>
<td>165.6</td>
</tr>
</tbody>
</table>

\[ \text{cor}(Y) \]

<table>
<thead>
<tr>
<th></th>
<th>BMXBMI</th>
<th>BPXSY1</th>
<th>BPXDI1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMXBMI</td>
<td>1.000</td>
<td>0.295</td>
<td>0.285</td>
</tr>
<tr>
<td>BPXSY1</td>
<td>0.295</td>
<td>1.000</td>
<td>0.478</td>
</tr>
<tr>
<td>BPXDI1</td>
<td>0.285</td>
<td>0.478</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Now we will test \( B_{[k,]} = 0 \) for \( k = 1, \ldots, q \):

\[
h2 \leftarrow \text{diag} \left( \text{solve}(t(X)X) \right)
\]

\[
t2stats \leftarrow \text{NULL}
\]

\[
\text{for}(k \text{ in } 1:q)
\]

\[
\{ t2stats \leftarrow c( t2stats, B[k,] \times \text{solve}(\SigmaHat) \times c(B[k,]) / h2[k] ) \}
\]

\[
\text{names}(t2stats) \leftarrow \text{colnames}(X)
\]

\[
t2stats
\]

<table>
<thead>
<tr>
<th></th>
<th>intercept</th>
<th>RIDAGEYR</th>
<th>RIDAGEYR2</th>
<th>FEMALE</th>
<th>DR1TKCAL</th>
<th>DR1TPROT</th>
<th>DR1TCARB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>232.22</td>
<td>2480.67</td>
<td>2015.05</td>
<td>108.20</td>
<td>17.12</td>
<td>2.59</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>DR1TSUGR</td>
<td>DR1TFIBE</td>
<td>DR1TTFAT</td>
<td>DR1TCHOL</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\[ F_{\text{stats}} = \frac{t^2_{\text{stats}} \cdot (n-q-p+1)}{(n-q) \cdot p} \]

<table>
<thead>
<tr>
<th>Intercept</th>
<th>RIDAGEYR</th>
<th>RIDAGEYR2</th>
<th>FEMALE</th>
<th>DR1TKCAL</th>
<th>DR1TPROT</th>
<th>DR1TCARB</th>
<th>DR1TSUGR</th>
<th>DR1TFIBE</th>
<th>DR1TTFAT</th>
<th>DR1TCHOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>77.381</td>
<td>826.626</td>
<td>671.468</td>
<td>36.055</td>
<td>5.706</td>
<td>0.862</td>
<td>0.399</td>
<td>2.621</td>
<td>12.077</td>
<td>9.358</td>
<td>0.772</td>
</tr>
</tbody>
</table>

\[1 - \text{pf}(F_{\text{stats}}, p, n-q)\]

<table>
<thead>
<tr>
<th>Intercept</th>
<th>RIDAGEYR</th>
<th>RIDAGEYR2</th>
<th>FEMALE</th>
<th>DR1TKCAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0000000000</td>
<td>0.0000000000</td>
<td>0.0000000000</td>
<td>0.0000000000</td>
<td>0.0006752172</td>
</tr>
</tbody>
</table>

Now let's do the same tests using the `anova` command in R:

#### testing total calories

```r
anova( lm(Y ~ -1 + X), lm(Y ~ -1 + X[-5]) )
```

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>Df</th>
<th>Gen.var.</th>
<th>Pillai approx F</th>
<th>num Df</th>
<th>den Df</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6258</td>
<td>94.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6259</td>
<td>1</td>
<td>94.4</td>
<td>0.00273</td>
<td>5.71</td>
<td>6256</td>
</tr>
</tbody>
</table>
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### testing total carbs

```r
anova(  lm(Y ~ -1 + X), lm(Y ~ -1 + X[-7] ))
```

Analysis of Variance Table

<table>
<thead>
<tr>
<th></th>
<th>Res.Df</th>
<th>Df</th>
<th>Gen.var.</th>
<th>Pillai</th>
<th>approx F</th>
<th>num Df</th>
<th>den Df</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6258</td>
<td>94.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6259</td>
<td>1</td>
<td>94.3</td>
<td>0.00191</td>
<td>0.399</td>
<td>3</td>
<td>6256</td>
<td>0.75</td>
</tr>
</tbody>
</table>

### testing total sugars

```r
anova(  lm(Y ~ -1 + X), lm(Y ~ -1 + X[-8] ))
```

Analysis of Variance Table

<table>
<thead>
<tr>
<th></th>
<th>Res.Df</th>
<th>Df</th>
<th>Gen.var.</th>
<th>Pillai</th>
<th>approx F</th>
<th>num Df</th>
<th>den Df</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6258</td>
<td>94.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6259</td>
<td>1</td>
<td>94.3</td>
<td>0.00126</td>
<td>2.62</td>
<td>3</td>
<td>6256</td>
<td>0.049</td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1