Multivariate Regression

The so-called supervised learning problem is the following: we want to approximate the random variable Y with an appropriate function of the random variables $X_1, \ldots, X_p$ with the method of *least squares*. That is,

$$\mathbb{E}(Y - g(X_1, \ldots, X_p))^2$$

is minimized over all p-variate, measurable functions g. From probability theory it is known, that the optimum g is

$$g_{opt}(x_1, \ldots, x_p) = \mathbb{E}(Y|X_1 = x_1, \ldots, X_p = x_p) = \frac{\int_{-\infty}^{\infty} y f(y, x_1, \ldots, x_p) dy}{\int_{-\infty}^{\infty} f(y, x_1, \ldots, x_p) dy},$$

where f is the joint p.d.f. of the above random variables (usually they have an absolutely continuous distribution). g_{opt} is called regression function, and Proposition 5 of Lesson 2 guarantees that it is linear if f is a $(p+1)$-dimensional normal density. Even if our random variables do not have a multivariate normal distribution (which is very usual by the Multivariate Central Limit Theorem), a linear approximation makes sense. 0Frequently, we estimate the regression parameters from a sample, and use so-called linearizing transformations.

1 Linear Regression

Given the expectation vector and covariance matrix of the random vector $(Y, X_1, \ldots, X_p)^T$, we want to approximate Y (target or response variable) with a linear combination of the predictor variable $\mathbf{X} = (X_1, \ldots, X_p)^T$ in such a way that the least squares error is minimized.

The solution is the following. To minimize the function

$$g(a_1, \ldots, a_p, b) = \mathbb{E}(Y - (a_1 X_1 + \cdots + a_p X_p + b))^2$$

let us take its partial derivatives with respect to the parameters $a_1, \ldots, a_p$ and b, then equal them to 0. Under some regularity conditions (which always hold in exponential families, especially in the multivariate Gaussian case), the differentiation with respect to the parameters results in the following system of equations:

$$\frac{\partial g}{\partial b} = -2\mathbb{E}(Y - (a_1 X_1 + \cdots + a_p X_p + b)) = 0$$

which results in

$$b = \mathbb{E}Y - a_1 \mathbb{E}X_1 - \cdots - a_p \mathbb{E}X_p \tag{1}$$

when a_i's are already known.

$$\frac{\partial g}{\partial a_i} = 2\mathbb{E}(Y - (a_1 X_1 + \cdots + a_p X_p + b))(-X_i) = 0 \quad (i = 1, \ldots, p)$$

which, using the solution for b, provides the following system of linear equations:

$$\sum_{j=1}^{p} \mathrm{Cov}(X_i, X_j) a_j = \mathrm{Cov}(Y, X_i), \qquad i = 1, \ldots, p.$$

Denoting by $\mathbf{C}$ the covariance matrix of the random vector $\mathbf{X} = (X_1, \ldots, X_p)$ and $\mathbf{d}$ the p-dimensional vector of cross-covariances between the components of $\mathbf{X}$ and Y, the above system of linear equations has the concise form:

$$\mathbf{Ca} = \mathbf{d}, \tag{2}$$

where $\mathbf{a} = (a_1, \ldots, a_p)^T$ is the vector of the parameters a_j's.

The system of equations (2) has the unique solution

$$\mathbf{a} = \mathbf{C}^{-1}\mathbf{d} \tag{3}$$

whenever $|\mathbf{C}| \neq 0$, which holds if there are no linear relations between $X_1, \ldots, X_p$ (in the multivariate normal case, they do not have a deformed p-variate distribution). Otherwise, we can take generalized inverse of $\mathbf{C}$, and the solution is not unique.

To show that the formulas (1) and (3) indeed give local and global minimum of our objective function, we investigate the Hessian, which is just the covariance matrix of all the $p+1$ variables, and is usually positive definite; hence, we get a unique minimum. If there were linear relations between the variables X_j's and Y, we may eliminate some of them to get a unique solution (see the forthcoming sections about the dimension reduction).

The above minimization is also equivalent to some correlation maximization problem as follows. The above minimization task is equivalent to minimizing $\mathrm{Var}(\varepsilon)$ in the model

$$Y = l(\mathbf{X}) + \varepsilon,$$

where $l(\mathbf{X}) = \sum_{i=1}^{p} a_i X_i + b = \mathbf{a}^T\mathbf{X} + b$ is linear function of the coordinates of $\mathbf{X}$.

In view of (1), $\mathbb{E}(\varepsilon) = 0$, and because the covariance is a bilinear function, not affected by constant shifts of its variables, we get that

$$\begin{aligned} \mathrm{Cov}(l(\mathbf{X}), \varepsilon) &= \mathrm{Cov}(\mathbf{a}^T\mathbf{X}, Y - \mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\mathbf{d} - \mathbf{a}^T\mathbf{Ca} \\ &= \mathbf{d}^T\mathbf{C}^{-1}\mathbf{d} - \mathbf{d}^T\mathbf{C}^{-1}\mathbf{C}\mathbf{C}^{-1}\mathbf{d} = 0, \end{aligned} \tag{4}$$

and consequently,

$$\mathrm{Var}(Y) = \mathrm{Var}(l(\mathbf{X})) + \mathrm{Var}(\varepsilon). \tag{5}$$

Further,

$$\mathrm{Cov}(l(\mathbf{X}), Y) = \mathbf{d}^T\mathbf{C}^{-1}\mathbf{d}$$

which is the first term on the right hand side of (4), and

$$\mathrm{Var}(l(\mathbf{X})) = \mathbf{d}^T\mathbf{C}^{-1}\mathbf{d}$$

which is the second term on the right hand side of (4), are the same.

Definition 1 *The multiple correlation between the target variable Y and the predictor variables $X_1, \ldots, X_p$ is*

$$\mathrm{Corr}(Y, l(\mathbf{X})) = \frac{\mathrm{Cov}(l(\mathbf{X}), Y)}{\sqrt{\mathrm{Var}(Y)\mathrm{Var}(l(\mathbf{X})}} = \frac{\mathbf{d}^T\mathbf{C}^{-1}\mathbf{d}}{\sqrt{\mathbf{d}^T\mathbf{C}^{-1}\mathbf{d}}\sqrt{\mathrm{Var}(Y)}} = \frac{\sqrt{\mathbf{d}^T\mathbf{C}^{-1}\mathbf{d}}}{\sqrt{\mathrm{Var}(Y)}}$$

which is nonnegative and denoted by $r_{Y(X_1,\ldots,X_p)} = r_{Y\mathbf{X}}$.

It is easy to see that in the $p = 1$ case this is the absolute value of the usual correlation coefficient between Y and the only predictor X.

The square of the multiple correlation coefficient can be written in the following form:

$$r^2_{Y\mathbf{X}} = \frac{\mathbf{d}^T\mathbf{C}^{-1}\mathbf{d}}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}(l(\mathbf{X}))}{\mathrm{Var}(Y)}.$$

Therefore, the equation (5) gives rise to the following decomposition of the variance of Y:

$$\mathrm{Var}(Y) = r^2_{Y\mathbf{X}}\mathrm{Var}(Y) + (1 - r^2_{Y\mathbf{X}})\mathrm{Var}(Y). \tag{6}$$

Here the first term is the variance of Y explained by the predictor variables, and the second term is the so-called *residual variance*, that is, the variance of the error term ε. Observe that $r^2_{Y\mathbf{X}} = 1$ is equivalent to $\mathrm{Var}(\varepsilon) = 0$, i.e., there is a linear relation between Y and the components of $\mathbf{X}$ with probability 1. The other extreme case $r^2_{Y\mathbf{X}} = 0$ means that $\mathrm{Var}(l(\mathbf{X})) = 0$, i.e., the best linear approximation is constant with probability 1, consequently $a_1 = \cdots = a_p = 0$, or equivalently, $\mathbf{a} = \mathbf{0}$ and $\mathbf{d} = \mathbf{0}$; in other words, Y is uncorrelated with all the X_j's, and hence, its best linear approximation is its own expectation.

Without proof we state that the above $l(\mathbf{X})$ has the maximal possible correlation with Y among all possible linear combinations of the components of $\mathbf{X}$.

Proposition 1 *For any linear combination $h(\mathbf{X})$ of $X_1, \ldots, X_p$, the following relation holds true:*

$$r_{Y(X_1,\ldots,X_p)} = \mathrm{Corr}(Y, l(\mathbf{X}) \geq |\mathrm{Corr}(Y, h(\mathbf{X})|.$$

Consequently, when subtracting $l(\mathbf{X})$ from Y, ε can be considered as the residual after eliminating the effect of the variables $X_1, \ldots, X_p$ from Y.

Definition 2 *If two target variables Y_1 and Y_2 are expressed as (different) linear combinations of the same predictor $\mathbf{X}$:*

$$Y_1 = l_1(\mathbf{X}) + \varepsilon_1 \quad \textit{and} Y_2 = l_2(\mathbf{X}) + \varepsilon_2,$$

then the partial correlation between Y_1 and Y_2 after eliminating the effect of $\mathbf{X}$ is the usual Pearson correlation coefficient between the error terms ε_1 and ε_2. We use the notation

$$r_{Y_1Y_2|\mathbf{X}} = \mathrm{Corr}(\varepsilon_1, \varepsilon_2).$$

Note that in the $p = 1$ case, when the only predictor is denoted by X, the following formula is used to calculate the partial correlation:

$$r_{Y_1Y_2|X} = \frac{\mathrm{Corr}(Y_1, Y_2) - \mathrm{Corr}(Y_1, X) \cdot \mathrm{Corr}(Y_2, X)}{\sqrt{(1 - \mathrm{Corr}^2(Y_1, X)) \cdot (1 - \mathrm{Corr}^2(Y_2, X))}}.$$

In fact, the partial correlation measures the correlation between two random variables after eliminating the effect of some nuisance variables. Indeed, in multivariate data structures, it can happen that the Pearson correlation coefficient between two variables does not reflect the pure association between them. Because the correlations are highly interlaced through the correlation matrix, we

cannot just pull out two of them. Other variables, which are strongly intercorrelated with both, will disturb their relation; therefore, first we have to eliminate the effect of these nuisance variables. For example, if we consider the correlation between the kiwi consumption and the number of registered cancer cases during the years (in the US), we experience a high correlation between them. However, it does not mean that kiwi causes cancer, it just means that there are other variables (the time and the increasing living standard, which makes rise to buy more tobacco and alcoholic drinks for example, that may cause cancer).

In case of an i.i.d. sample $(Y_1, \mathbf{X}^1), \ldots, (Y_n, \mathbf{X}^n)$ we estimate the parameters by the formulas (1), (3), where we substitute the empirical quantities (ML-estimators) for $\mathbf{C}$ and $\mathbf{d}$. The squared multiple correlation coefficient R^2 is also estimated from the sample, and it gives the proportion of the total variation of Y which is explained by the predictor variables. In the next section we will use hypothesis testing for the significance of R^2.

Note, that with some *linearization* formulas we can use linear regression in the following models:

- *Multiplicative model*:
$$Y \sim b {X_1}^{a_1} \ldots {X_p}^{a_p}.$$
After taking the logarithms, one gets
$$\ln Y \sim \ln b + a_1 \ln X_1 + \cdots + a_p \ln X_p,$$
therefore, we can use linear regression for the log-log data. While, the linear regression performs well for data from a multivariate normal distribution, this model favors lognormally distributed data (for example, chemical concentrations).

- *Polynomial regression*: Now we want to approximate Y with a given degree polynomial of X:
$$Y \sim a_1 X + a_2 X^2 + \cdots + a_p X^p + b.$$
The solution is obtained by applying multivariate linear regression for Y with the predictor variables $X_j = X^j$, $j = 1, \ldots, p$.

2 The linear model (with deterministic predictors)

Now our model is the following.

$$Y_i = \sum_{j=1}^{p} a_j x_{ij} + \varepsilon_i \qquad (i = 1, \ldots, n),$$

where x_{ij} is the prescribed value of the j-th predictor in the i-th measurement. Since the measurement is burdened with the random noise ε_i, the measured value Y_i of the target variable in the i-th measurement is a random variable. For simplicity, the constant term is zero (in fact, it would be $\bar{Y} - \sum_{j=1}^{p} a_j \bar{x}_j$, but we assume that it has already been subtracted from the left hand sides). We also assume that $\mathbb{E}(\varepsilon_i = 0)$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ $(i = 1, \ldots, n)$, and the measurement errors

are uncorrelated. Because of their equal (but unknown) variance they are called *homoscedastic errors*. Therefore, $\mathbb{E}(Y_i) = \sum_{j=1}^{p} a_j x_{ij}$ and $\mathrm{Var}(Y_i) = \sigma^2$ $(i = 1, \ldots, n)$, and the Y_i's are also uncorrelated. Very frequently, the measurement errors are Gaussian, and thus, the random variables $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are also independent, akin to the Y_i's.

With the notation

$$\mathbf{Y} := (Y_1, \ldots, Y_n)^T, \quad \underline{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_n)^T$$

and $\mathbf{X} = (x_{ij})$ $(i = 1, \ldots, n;\ j = 1, \ldots, p)$, our model equation can be put into the following matrix form:

$$\mathbf{Y} = \mathbf{X}\mathbf{a} + \underline{\varepsilon},$$

where the parameter vector $\mathbf{a} = (a_1, \ldots, a_p)^T$ is estimated by the method of least squares, i.e.,

$$\sum_{i=1}^{n} \varepsilon_i^2 = \|\mathbf{Y} - \mathbf{X}\mathbf{a}\|^2$$

is minimized with respect to $\mathbf{a}$.

Here $(\mathbf{Y}, \mathbf{X})^T = (\mathbf{Y}, \mathbf{x}_1, \ldots, \mathbf{x}_p)^T$ is the data matrix, where $\mathbf{x}_j$ denotes the j-th column of the matrix $\mathbf{X}$. If the solution is denoted by $\hat{\mathbf{a}}$, a simple linear algebra guarantees that $\mathbf{X}\hat{\mathbf{a}}$ is the projection of the random vector $\mathbf{Y}$ onto $F = \mathrm{Span}\{\mathbf{x}_1, \ldots, \mathbf{x}_p\} \subset \mathbb{R}^n$. Let us denote the $n \times n$ matrix of this projection by $\mathbf{P}$. Consequently, $\mathbf{X}\hat{\mathbf{a}} = \mathbf{P}\mathbf{Y}$ and $\mathbf{Y} - \mathbf{X}\hat{\mathbf{a}} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$ are orthogonal, and latter vector is also orthogonal to any vector $\mathbf{X}\mathbf{b} \in F$. Therefore,

$$(\mathbf{X}\mathbf{b})^T \cdot (\mathbf{Y} - \mathbf{X}\mathbf{a}) = 0, \quad \forall \mathbf{b} \in \mathbb{R}^p.$$

From this,

$$\mathbf{b}^T\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\mathbf{a}) = 0, \quad \forall \mathbf{b} \in \mathbb{R}^p$$

holds, which implies that

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\mathbf{a}) = \mathbf{0}.$$

In summary, $\hat{\mathbf{a}}$ is the solution of the so-called *Gauss normal equation*

$$\mathbf{X}^T\mathbf{X}\mathbf{a} = \mathbf{X}^T\mathbf{Y}.$$

This equation is always consistent, since $\mathbf{X}^T\mathbf{Y}$ is in F, which is also spanned by the column vectors of $\mathbf{X}^T\mathbf{X}$. In the rank r of F (this is also the rank of $\mathbf{X}$) is equal to $p(\leq n)$, then we have a unique solution:

$$\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

From here,

$$\mathbf{P}\mathbf{Y} = \mathbf{X}\hat{\mathbf{a}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

therefore

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.$$

If the rank of $\mathbf{X}$ is less than p, then there are infinitely many solutions, including the one, obtained by the Moore–Pen rose inverse:

$$\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{+}\mathbf{X}^T\mathbf{Y}.$$

Proposition 2 *If* $\operatorname{rank}(\mathbf{X}) = p \leq n$ *and* $\underline{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$*, then* $\hat{\mathbf{a}} \sim \mathcal{N}_p(\mathbf{a}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.

Therefore $\hat{\mathbf{a}}$ is an unbiased estimator of $\mathbf{a}$. In the Gaussian case, $\hat{\mathbf{a}}$ is also the ML-estimate of $\mathbf{a}$. Further, the ML-estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{SSE}{n}$$

that is biased. The unbiased estimate of σ^2 is $\frac{SSE}{n-p-1}$ or $\frac{SSE}{n-p}$ depending on, whether there is or there is no constant term (intercept) in the model (when the variables are previously transformed to have zero mean, there is no constant term).

The forthcoming Gauss–Markov theorem states that $\hat{\mathbf{a}}$ is also efficient among the linear, unbiased estimators.

Theorem 1 (Gauss –Markov Theorem) *For any other unbiased linear estimator* $\tilde{\mathbf{a}}$ *of* $\mathbf{a}$*:*

$$\operatorname{Var}(\hat{\mathbf{a}}) \leq \operatorname{Var}(\tilde{\mathbf{a}}).$$

This means that the difference of the right-hand and left-hand side $p \times p$ *covariance matrices is positive semidefinite. Shortly,* $\hat{\mathbf{a}}$ *provides a* **BLUE** *(Best, Linear, Unbiased Estimate) for* $\mathbf{a}$.

In view of $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, the minimum of our objective function is

$$SSE := \|\mathbf{Y} - \mathbf{X}\hat{\mathbf{a}}\|^2 = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{a}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{a}}),$$

called *residual variance*. It can also be written as

$$\begin{aligned} SSE &= (\mathbf{Y} - \mathbf{P}\mathbf{Y})^T(\mathbf{Y} - \mathbf{P}\mathbf{Y}) = ((\mathbf{I} - \mathbf{P})\mathbf{Y})^T((\mathbf{I} - \mathbf{P})\mathbf{Y}) = \\ &= \mathbf{Y}^T(\mathbf{I} - \mathbf{P})^2\mathbf{Y} = \mathbf{Y}^T(\mathbf{I} - \mathbf{P})\mathbf{Y}. \end{aligned}$$

Since $\mathbf{I} - \mathbf{P}$ is a projection of rank $n - p$, SSE has $\sigma^2\chi^2(n-p)$-distribution.

To make inference on the significance of the regression, we will intensively use the sample counterpart of the variance decomposition (6):

$$SST = SSR + SSE = R^2 \cdot SST + (1 - R^2) \cdot SST,$$

where $SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$ is the total variation of the measurements (sum of squares total),

$$SSE = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \sum_{j=1}^p \hat{a}_j x_{ij})^2$$

is the *residual sum of squares* (sum of squares due to error), and $SSR = SST - SSE$ is the part of the total variation explained by the regression (sum of squares due to regression). Further, R is the sample estimate of the multiple correlation coefficient.

777

We investigate the alternative

$$H_0 : \mathbf{a} = \mathbf{0} \quad \text{versus} \quad H_1 : \mathbf{a} \neq \mathbf{0}.$$

Under H_0, SSR has $\sigma^2\chi(p)$-distribution, and independent of SSE (see also the ANOVA setup of Lesson 7). Therefore,

$$F = \frac{SSR/p}{SSE/(n-p)} = \frac{R^2}{1-R^2} \cdot \frac{n-p}{p} \sim \mathcal{F}(p, n-p) \tag{7}$$

has Fisher F-distribution with degrees of freedom p and $n-p$ (in fact, $n-p-1$ if there is a constant term as well). If this $F \geq F_\alpha(p, n-p)$ (the upper α-point, or equivalently, the $(1-\alpha)$-quantile value) of this F-distribution, then we reject H_0 with significance α. This means that the regression is significant, and it makes sense to approximate the target variable with the predictors.

When we reject the null-hypothesis, we may further investigate whether the coefficients a_j's significantly differ from zero. For $j = 1, \ldots, p$ we investigate the alternative

$$H_{0j} : a_j = 0 \quad \text{versus} \quad H_{1j} : a_j \neq 0.$$

Under H_{0j}, $\hat{a}_j$ has zero expectation, and standardizing by its standard error

$$s_j = \sqrt{\frac{SSE/(n-p)}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}$$

(which is based on Proposition 2), then under H_{0j}, the test statistic

$$t_j = \frac{\hat{a}_j - 0}{s_j} \sim t(n-p)$$

follows Student's t-distribution with degrees of freedom $n-p$; in fact, $n-p-1$ if there is a constant term (intercept) as well in the model.

If this $t_j \geq t_{\alpha/2}(n-p)$ (the upper $\alpha/2$-point, or equivalently, the $(1-\alpha/2)$-quantile value) of this t-distribution, then we reject H_{0j} with significance α, and conclude that the predictor variable j significantly influences the response.

Estimating parameter functions

Sometimes we want to estimate not directly $\mathbf{a}$, but some linear combintion $\mathbf{b}^T\mathbf{a}$ of its coordinates, where $\mathbf{b} \in \mathbb{R}^p$ is a given vector. (For example, it may contain the unit prices of products, the mass production of which are the coordinates of $\mathbf{a}$.)

Definition 3 *The parameter function $\mathbf{b}^T\mathbf{a}$ is estimable (linearly and on an unbiased way) if there exists a vector $\mathbf{c} \in \mathbb{R}^n$ such that $\mathbb{E}(\mathbf{c}^T\mathbf{Y}) = \mathbf{b}^T\mathbf{a}$.*

Proposition 3 *The parameter function $\mathbf{b}^T\mathbf{a}$ is estimable if and only if $\mathbf{b}$ is in the linear subspace of $\mathbb{R}^p$ spanned by the row vectors of $\mathbf{X}$.*

Proof: The following are equivalent:

$$\begin{aligned}
\mathbf{c}^T\mathbb{E}(\mathbf{Y}) &= \mathbf{b}^T\mathbf{a} \qquad \forall\, \mathbf{a} \in \mathbb{R}^p \\
\mathbf{c}^T\mathbf{X}\mathbf{a} &= \mathbf{b}^T\mathbf{a} \qquad \forall\, \mathbf{a} \in \mathbb{R}^p \\
\mathbf{c}^T\mathbf{X} &= \mathbf{b}^T \\
\mathbf{b} &= \mathbf{X}^T\mathbf{c}
\end{aligned}$$

that means that $\mathbf{b}$ is within the linear subspace spanned by the column vectors of $\mathbf{X}^T$, i.e., the row vectors of $\mathbf{X}$.

If $r = p$, then it is true for any $\mathbf{b} \in \mathbb{R}^p$. If $r < p$, then it is true only for special $\mathbf{b}$'s. The Gauss–Markov theorem implyes the following (sometimes this is called Gauss–Markov theorem).

Theorem 2 *In the $r = p$ case, for any $\mathbf{b} \in \mathbb{R}^p$, $\mathbf{b}^T\hat{\mathbf{a}}$ is a linear and unbiased estimate of $\mathbf{b}^T\mathbf{a}$, and among such estimates it has minimum variance, i.e., BLUE (Best Linear Unbiased Estimate).*

Constructing confidence intervals and testing hypotheses for the parameter function: We saw that in the $r = p$ case, $\mathrm{Var}\hat{\mathbf{a}} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$, and so,

$$\mathrm{Var}(\mathbf{b}^T\hat{\mathbf{a}}) = \mathbf{b}^T\mathrm{Var}(\hat{\mathbf{a}})\mathbf{b} = \sigma^2\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1})\mathbf{b}.$$

Let us consider the estimable parameter function $\theta = \mathbf{b}^T\mathbf{a}$, and its least square estimate $\hat{\theta} = \mathbf{b}^T\hat{\mathbf{a}}$. To make inferences on θ, we assume multivariate normality of $\underline{\varepsilon}$, consequently, that of $\mathbf{Y}$. By Proposition 2, $\hat{\mathbf{a}}$ is also multivariate normally distributed, and so, $\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1})\mathbf{b})$ and $SSE \sim \sigma^2\chi^2(n-p)$ are independent. Therefore, by the standardized $\hat{\theta}$ and SSE a Student t-statistic and confidence interval for θ can be constructed.

Likewise, we can also test the hypothesis

$$H_0 : \theta = \theta_0$$

by means of the above t-statistic.

If we consider k independent estimable parameter functions $\theta_1 = \mathbf{b}_1^T\mathbf{a}, \ldots, \theta_k = \mathbf{b}_k^T\mathbf{a}$ and investigate

$$H_0 : \theta_1 = \theta_{10}, \ldots, \theta_k = \theta_{k0}$$

then from two independent χ^2 variables an F-statistic can be constructed.

Examples for Multivariate Regression

The following examples are from the book
Gary Koop: Analysis of economic data, Wiley, 2005.

1. Example: Consider the following Multivariate Regression problem emerging in Electric Power Industry in the USA.

$$\begin{aligned} Y &= \text{cost of production (million dollar/year)} \\ X_1 &= \text{yield (kWh/year)} \\ X_2 &= \text{cost of labour (dollar/year/worker)} \\ X_3 &= \text{cost of capital (dollar/unit)} \\ X_4 &= \text{cost of fuel (dollar/million BTU)} \end{aligned}$$

Regression results:

	Coefficient	Standard error	t-value	p-value	Lower 95%	Upper 95%
Intercept	-70.49511	12.69501	-5.55298	1.76E-07	-95.6347	-45.3556
X_1	0.00474	0.00011	43.22597	3.41E-74	0.004514	0.004948
X_2	0.00363	0.00106	3.43660	0.000814	0.001537	0.005717
X_3	0.28008	0.12949	2.16301	0.032557	0.023663	0.536503
X_4	0.78346	0.16759	4.72566	6.391E-06	0.455154	1.11177

$R^2 = 0.94$, the p-value of the $H_0 : R^2 = 0$ is $9.73E-73$.

This shows that the regression is significant, and all the predictors are significant, except X_3. We can see that X_3 has relatively 'small' correlation with the other variables, where the correlation matrix of the predictors is

$$\begin{pmatrix} 1 & & & \\ 0.056399 & 1 & & \\ 0.021481 & -0.078686 & 1 & \\ 0.053507 & 0.318349 & 0.155224 & 1 \end{pmatrix}.$$

When we leave out X_3 from the regression, the results do not change

significantly:

	Coefficient	Standard error	t-value	p-value	Lower 95%	Upper 95%
Intercept	-49.75804	8.44931	-5.88900	3.68E-08	-71.8765	-27.6396
X_1	0.00473	0.00011	42.6218	6.4E-74	0.004445	0.005027
X_2	0.00331	0.00006	3.12145	0.002259	0.000535	0.006091
X_4	0.851586	0.165266	5.15282	1.03E-06	0.418956	1.284216

$R^2 = 0.94$, the p-value of the $H_0 : R^2 = 0$ is $3.5E - 73$.

If there were *collinearities* between the variables, we could not simply leave out one.

2. Example: the following Multivariate Regression problem is to predict the apartment prices in the USA.

$$\begin{aligned} Y &= \text{selling price} \\ X_1 &= \text{site area} \\ X_2 &= \text{number of bedrooms} \\ X_3 &= \text{number of bathrooms} \\ X_4 &= \text{number of levels} \end{aligned}$$

Regression results:

	Coefficient	Standard error	t-value	p-value	Lower 95%	Upper 95%
Intercept	-4009.5500	3603.109	-1.1128	0.266287	-11087.3	3068.248
X_1	5.42917	0.36925	14.70325	2.05E-41	4.703835	6.154513
X_2	2824.61379	1214.808	2.325153	0.020433	438.2961	5210.931
X_3	17105.1745	1734.434	9.862107	3.29E-21	13698.12	20512.22
X_4	7634.897	1007.974	7.574494	1.57E-13	5654.874	9614.92

$R^2 = 0.54$, the p-value of the $H_0 : R^2 = 0$ is $1.18E - 88$.

Therefore, the predicted price is

$$\hat{Y} = -4009.55 + 5.43X_1 + 2824.6X_2 + 17105.17X_3 + 7634.90X_4.$$

This means that in the apartments with the same number of bedrooms, bathrooms, and levels, the increase of the site with 1 $foot^2$ will result in

the increase of the price with 5.43 dollar. Likewise, keeping the site are, number of bathrooms, and levels fixed, a 3-bedroom apartment costs with 2824.6 dollar more than a 2-bedroom one, on average, 'ceteris paribus'.

In fact, here 1-, 2-, or 3-way ANOVA can also be used with the discrete variables X_2, X_3, X_4, as well as an Analysis of Covariance with the site area as concomitant variable.