Multivariate Linear Regression

Nathaniel E. Helwig

Associate Professor of Psychology and Statistics
University of Minnesota

March 2, 2024

Table of Contents

Content adapted from:
Johnson, R. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis (6th ed).

Table of Contents

MLR Model: Scalar Form

The multiple linear regression model has the form

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

for $i \in \{1, \ldots, n\}$ where

- $y_i \in \mathbb{R}$ is the real-valued response for the i-th observation
- $\beta_0 \in \mathbb{R}$ is the regression intercept
- $\beta_j \in \mathbb{R}$ is the j-th predictor's regression slope
- $x_{ij} \in \mathbb{R}$ is the j-th predictor for the i-th observation
- $\epsilon_i \stackrel{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is a Gaussian error term

MLR Model: Nomenclature

The model is multiple because we have $p > 1$ predictors.

- If $p = 1$, we have a simple linear regression model

The model is linear because y_i is a linear function of the parameters (β_0, β_1, ..., β_p are the parameters).

The model is a regression model because we are modeling a response variable (Y) as a function of predictor variables ($X_1, \ldots, X_p$).

MLR Model: Assumptions

The fundamental assumptions of the MLR model are:

1. Relationship between X_j and Y is linear (given other predictors)
2. x_{ij} and y_i are observed random variables (known constants)
3. $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is an unobserved random variable
4. $\beta_0, \beta_1, \ldots, \beta_p$ are unknown constants
5. $(y_i | x_{i1}, \ldots, x_{ip}) \overset{\text{ind}}{\sim} \mathrm{N}(\beta_0 + \sum_{j=1}^p \beta_j x_{ij}, \sigma^2)$
   note: homogeneity of variance

Note: β_j is expected increase in Y for 1-unit increase in X_j with all other predictor variables held constant

MLR Model: Matrix Form

The multiple linear regression model has the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

- $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ is the $n \times 1$ response vector
- $\mathbf{X} = [\mathbf{1}_n, \mathbf{x}_1, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n \times (p+1)}$ is the $n \times (p+1)$ design matrix
    - $\mathbf{1}_n$ is an $n \times 1$ vector of ones
    - $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^\top \in \mathbb{R}^n$ is j-th predictor vector $(n \times 1)$
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^{p+1}$ is $(p+1) \times 1$ vector of coefficients
- $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top \in \mathbb{R}^n$ is the $n \times 1$ error vector

MLR Model: Matrix Form (another look)

Matrix form writes MLR model for all n points simultaneously

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

MLR Model: Assumptions (revisited)

In matrix terms, the error vector is multivariate normal:

$$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$$

In matrix terms, the response vector is multivariate normal given $\mathbf{X}$:

$$(\mathbf{y}|\mathbf{X}) \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$$

Ordinary Least Squares

The ordinary least squares (OLS) problem is

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \textstyle\sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

where $\| \cdot \|$ denotes the Frobenius norm.

The OLS solution has the form

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{y}$$

Fitted Values and Residuals

SCALAR FORM:

Fitted values are given by

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$$

and residuals are given by

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

MATRIX FORM:

Fitted values are given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

and residuals are given by

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}}$$

Hat Matrix

Note that we can write the fitted values as

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} \\ &= \mathbf{H}\mathbf{y}\end{aligned}$$

where $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ is the hat matrix.

$\mathbf{H}$ is a symmetric and idempotent matrix: $\quad \mathbf{H}\mathbf{H} = \mathbf{H}$

$\mathbf{H}$ projects $\mathbf{y}$ onto the column space of $\mathbf{X}$.

Multiple Regression Example in R

```
> 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
> mtcars$cyl <- factor(mtcars$cyl)
> mod <- lm(mpg ~ cyl + am + carb, data=mtcars)
> coef(mod)
(Intercept)        cyl6        cyl8          am        carb
  25.320303   -3.549419   -6.904637    4.226774   -1.119855
```

Regression Sums-of-Squares: Scalar Form

In MLR models, the relevant sums-of-squares are

- Sum-of-Squares Total: $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Sum-of-Squares Regression: $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- Sum-of-Squares Error: $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

The corresponding degrees of freedom are

- SST: $df_T = n - 1$
- SSR: $df_R = p$
- SSE: $df_E = n - p - 1$

Regression Sums-of-Squares: Matrix Form

In MLR models, the relevant sums-of-squares are

$$
\begin{aligned}
SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= \mathbf{y}^\top \left[\mathbf{I}_n - (1/n)\mathbf{J}\right] \mathbf{y} \\
SSR &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\
&= \mathbf{y}^\top \left[\mathbf{H} - (1/n)\mathbf{J}\right] \mathbf{y} \\
SSE &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \mathbf{y}^\top \left[\mathbf{I}_n - \mathbf{H}\right] \mathbf{y}
\end{aligned}
$$

Note: $\mathbf{J}$ is an $n \times n$ matrix of ones

Partitioning the Variance

We can partition the total variation in y_i as

$$
\begin{aligned}
SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n} (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) \\
&= SSR + SSE + 2\sum_{i=1}^{n} (\hat{y}_i - \bar{y})\hat{\epsilon}_i \\
&= SSR + SSE
\end{aligned}
$$

Regression Sums-of-Squares in R

```
> anova(mod)
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value   Pr(>F)
cyl        2 824.78  412.39 52.4138 5.05e-10 ***
am         1  36.77   36.77  4.6730  0.03967 *
carb       1  52.06   52.06  6.6166  0.01592 *
Residuals 27 212.44    7.87
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Anova(mod, type=3)
Anova Table (Type III tests)

Response: mpg
            Sum Sq Df  F value    Pr(>F)
(Intercept) 3368.1  1 428.0789 < 2.2e-16 ***
cyl          121.2  2   7.7048  0.002252 **
am            77.1  1   9.8039  0.004156 **
carb          52.1  1   6.6166  0.015923 *
Residuals    212.4 27
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Coefficient of Multiple Determination

The coefficient of multiple determination is defined as

$$\begin{aligned} R^2 &= \frac{SSR}{SST} \\ &= 1 - \frac{SSE}{SST} \end{aligned}$$

and gives the amount of variation in y_i that is explained by the linear relationships with $x_{i1}, \ldots, x_{ip}$.

When interpreting R^2 values, note that…

- $0 \leq R^2 \leq 1$
- Large R^2 values do not necessarily imply a good model

Adjusted Coefficient of Multiple Determination (R_{a}^2)

Including more predictors in a MLR model can artificially inflate R^2:

- Capitalizing on spurious effects present in noisy data
- Phenomenon of over-fitting the data

The adjusted R^2 is a relative measure of fit:

$$
\begin{aligned}
R_{\text{a}}^2 &= 1 - \frac{SSE/df_E}{SST/df_T} \\
&= 1 - \frac{\hat{\sigma}^2}{s_Y^2}
\end{aligned}
$$

where $s_Y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$ is the sample estimate of the variance of Y.

Note: R^2 and R_{a}^2 have different interpretations!

Regression Sums-of-Squares in R

```
> smod <- summary(mod)
> names(smod)
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
> summary(mod)$r.squared
[1] 0.8113434
> summary(mod)$adj.r.squared
[1] 0.7833943
```

Relation to ML Solution

Remember that $(\mathbf{y}|\mathbf{X}) \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, which implies that $\mathbf{y}$ has pdf

$$f(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}$$

As a result, the log-likelihood of $\boldsymbol{\beta}$ given $(\mathbf{y}, \mathbf{X}, \sigma^2)$ is

$$\ln\{L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2)\} = -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + c$$

where c is a constant that does not depend on $\boldsymbol{\beta}$.

Relation to ML Solution (continued)

The maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ is the estimate satisfying

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Now, note that…

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \max_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} -(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} -(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

The OLS and ML estimate of $\boldsymbol{\beta}$ is the same: $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$

Estimated Error Variance (Mean Squared Error)

The estimated error variance is

$$\begin{aligned}\hat{\sigma}^2 &= SSE/(n-p-1) \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p-1) \\ &= \|(\mathbf{I}_n - \mathbf{H})\mathbf{y}\|^2/(n-p-1)\end{aligned}$$

which is an unbiased estimate of error variance σ^2.

The estimate $\hat{\sigma}^2$ is the mean squared error (MSE) of the model.

Maximum Likelihood Estimate of Error Variance

$\tilde{\sigma}^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n$ is the MLE of σ^2.

From our previous results using $\hat{\sigma}^2$, we have that

$$\mathrm{E}(\tilde{\sigma}^2) = \frac{n-p-1}{n}\sigma^2$$

Consequently, the bias of the estimator $\tilde{\sigma}^2$ is given by

$$\frac{n-p-1}{n}\sigma^2 - \sigma^2 = -\frac{(p+1)}{n}\sigma^2$$

and note that $-\frac{(p+1)}{n}\sigma^2 \to 0$ as $n \to \infty$.

Comparing $\hat{\sigma}^2$ and $\tilde{\sigma}^2$

Reminder: the MSE and MLE of σ^2 are given by

$$\hat{\sigma}^2 = \|(\mathbf{I}_n - \mathbf{H})\mathbf{y}\|^2/(n-p-1)$$
$$\tilde{\sigma}^2 = \|(\mathbf{I}_n - \mathbf{H})\mathbf{y}\|^2/n$$

From the definitions of $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ we have that

$$\tilde{\sigma}^2 < \hat{\sigma}^2$$

so the MLE produces a smaller estimate of the error variance.

Estimated Error Variance in R

```
# get mean-squared error in 3 ways
> n <- length(mtcars$mpg)
> p <- length(coef(mod)) - 1
> smod$sigma^2
[1] 7.868009
> sum((mod$residuals)^2) / (n - p - 1)
[1] 7.868009
> sum((mtcars$mpg - mod$fitted.values)^2) / (n - p - 1)
[1] 7.868009

# get MLE of error variance
> smod$sigma^2 * (n - p - 1) / n
[1] 6.638633
```

Summary of Results

Given the model assumptions, we have

$$\hat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$$

$$\hat{\mathbf{y}} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{H})$$

$$\hat{\boldsymbol{\epsilon}} \sim \mathrm{N}(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H}))$$

Typically σ^2 is unknown, so we use the MSE $\hat{\sigma}^2$ in practice.

ANOVA Table and Regression F Test

We typically organize the SS information into an ANOVA table:

Source	SS	df	MS	F	p-value
SSR	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	p	MSR	F^*	p^*
SSE	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-p-1$	MSE		
SST	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$			

$MSR = \frac{SSR}{p}$ and $MSE = \frac{SSE}{n-p-1}$

$F^* = \frac{MSR}{MSE} \sim F_{p,n-p-1}$ and $p^* = P(F_{p,n-p-1} > F^*)$

F^*-statistic and p^*-value are testing $H_0 : \beta_1 = \cdots = \beta_p = 0$ versus $H_1 : \beta_k \neq 0$ for some $k \in \{1, \ldots, p\}$

Inferences about $\hat{\beta}_j$ with σ^2 Known

If σ^2 is known, form $100(1-\alpha)\%$ CIs using

$$\hat{\beta}_0 \pm Z_{\alpha/2}\sigma_{\beta_0} \qquad\qquad \hat{\beta}_j \pm Z_{\alpha/2}\sigma_{\beta_j}$$

where

- $Z_{\alpha/2}$ is normal quantile such that $P(X > Z_{\alpha/2}) = \alpha/2$
- σ_{β_0} and σ_{β_j} are square-roots of diagonals of $\mathrm{V}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$

To test $H_0 : \beta_j = \beta_j^*$ vs. $H_1 : \beta_j \neq \beta_j^*$ (for some $j \in \{0, 1, \ldots, p\}$) use

$$Z = (\hat{\beta}_j - \beta_j^*)/\sigma_{\beta_j}$$

which follows a standard normal distribution under H_0.

Inferences about $\hat{\beta}_j$ with σ^2 Unknown

If σ^2 is unknown, form $100(1-\alpha)\%$ CIs using

$$\hat{\beta}_0 \pm t_{n-p-1}^{(\alpha/2)} \hat{\sigma}_{\beta_0} \qquad\qquad \hat{\beta}_j \pm t_{n-p-1}^{(\alpha/2)} \hat{\sigma}_{\beta_j}$$

where

- $t_{n-p-1}^{(\alpha/2)}$ is t_{n-p-1} quantile with $P(X > t_{n-p-1}^{(\alpha/2)}) = \alpha/2$
- $\hat{\sigma}_{\beta_0}$ and $\hat{\sigma}_{\beta_j}$ are square-roots of diagonals of $\hat{\mathrm{V}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}$

To test $H_0 : \beta_j = \beta_j^*$ vs. $H_1 : \beta_j \neq \beta_j^*$ (for some $j \in \{0, 1, \ldots, p\}$) use

$$T = (\hat{\beta}_j - \beta_j^*)/\hat{\sigma}_{\beta_j}$$

which follows a t_{n-p-1} distribution under H_0.

Coefficient Inference in R

```
> summary(mod)

Call:
lm(formula = mpg ~ cyl + am + carb, data = mtcars)

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

> confint(mod)
                 2.5 %        97.5 %
(Intercept)  22.809293 27.8313132711
cyl6         -7.098164 -0.0006745487
cyl8        -10.613981 -3.1952927942
am            1.456957  6.9965913486
carb         -2.013131 -0.2265781401
```

Inferences about Multiple $\hat{\beta}_j$

Assume that $q < p$ and want to test if a reduced model is sufficient:

$$H_0 : \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = \beta^*$$
$$H_1 : \text{at least one } \beta_k \neq \beta^*$$

Compare the SSE for full and reduced (constrained) models:
(a) Full Model: $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$
(b) Reduced Model: $y_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_{ij} + \beta^* \sum_{k=q+1}^{p} x_{ik} + \epsilon_i$

Note: set $\beta^* = 0$ to remove $X_{q+1}, \ldots, X_p$ from model.

Inferences about Multiple $\hat{\beta}_j$ (continued)

Test Statistic:

$$\begin{aligned} F^* &= \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_F} \\ &= \frac{SSE_R - SSE_F}{(n-q-1)-(n-p-1)} \div \frac{SSE_F}{n-p-1} \\ &\sim F_{(p-q,n-p-1)} \end{aligned}$$

where

- SSE_R is sum-of-squares error for reduced model
- SSE_F is sum-of-squares error for full model
- df_R is error degrees of freedom for reduced model
- df_F is error degrees of freedom for full model

Inferences about Linear Combinations of $\hat{\beta}_j$

Assume that $\mathbf{c} = (c_1, \ldots, c_{p+1})^\top$ and want to test:

$$H_0 : \mathbf{c}^\top \boldsymbol{\beta} = \beta^*$$
$$H_1 : \mathbf{c}^\top \boldsymbol{\beta} \neq \beta^*$$

Test statistic:

$$t^* = \frac{\mathbf{c}^\top \hat{\boldsymbol{\beta}} - \beta^*}{\hat{\sigma}\sqrt{\mathbf{c}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{c}}}$$
$$\sim t_{n-p-1}$$

Confidence Interval for σ^2

Note that $\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} = \frac{SSE}{\sigma^2} = \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sigma^2} \sim \chi^2_{n-p-1}$

This implies that

$$\chi^2_{(n-p-1;1-\alpha/2)} < \frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} < \chi^2_{(n-p-1;\alpha/2)}$$

where $P(Q > \chi^2_{(n-p-1;\alpha/2)}) = \alpha/2$ so a $100(1-\alpha)\%$ CI is given by

$$\frac{(n-p-1)\hat{\sigma}^2}{\chi^2_{(n-p-1;\alpha/2)}} < \sigma^2 < \frac{(n-p-1)\hat{\sigma}^2}{\chi^2_{(n-p-1;1-\alpha/2)}}$$

Interval Estimation

Idea: estimate expected value of response for a given predictor score.

Given $\mathbf{x}_h = (1, x_{h1}, \ldots, x_{hp})$, the fitted value is $\hat{y}_h = \mathbf{x}_h \hat{\boldsymbol{\beta}}$.

Variance of $\hat{y}_h$ is $\sigma^2_{\bar{y}_h} = \mathrm{V}(\mathbf{x}_h \hat{\boldsymbol{\beta}}) = \mathbf{x}_h \mathrm{V}(\hat{\boldsymbol{\beta}}) \mathbf{x}_h^\top = \sigma^2 \mathbf{x}_h (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_h^\top$

- Use $\hat{\sigma}^2_{\bar{y}_h} = \hat{\sigma}^2 \mathbf{x}_h (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_h^\top$ if σ^2 is unknown

We can test $H_0 : \mathrm{E}(y_h) = y_h^*$ vs. $H_1 : \mathrm{E}(y_h) \neq y_h^*$

- Test statistic: $T = (\hat{y}_h - y_h^*)/\hat{\sigma}_{\bar{y}_h}$, which follows t_{n-p-1} distribution
- $100(1-\alpha)\%$ CI for $E(y_h)$: $\quad \hat{y}_h \pm t_{n-p-1}^{(\alpha/2)} \hat{\sigma}_{\bar{y}_h}$

Predicting New Observations

Idea: estimate observed value of response for a given predictor score.

- Note: interested in actual $\hat{y}_h$ value instead of $\mathrm{E}(\hat{y}_h)$

Given $\mathbf{x}_h = (1, x_{h1}, \ldots, x_{hp})$, the fitted value is $\hat{y}_h = \mathbf{x}_h \hat{\boldsymbol{\beta}}$.

- Note: same as interval estimation

When predicting a new observation, there are two uncertainties:

- location of the distribution of Y for $X_1, \ldots, X_p$ (captured by $\sigma^2_{\bar{y}_h}$)
- variability within the distribution of Y (captured by σ^2)

Predicting New Observations (continued)

Two sources of variance are independent so $\sigma^2_{y_h} = \sigma^2_{\bar{y}_h} + \sigma^2$

- Use $\hat{\sigma}^2_{y_h} = \hat{\sigma}^2_{\bar{y}_h} + \hat{\sigma}^2$ if σ^2 is unknown

We can test $H_0 : y_h = y_h^*$ vs. $H_1 : y_h \neq y_h^*$

- Test statistic: $T = (\hat{y}_h - y_h^*)/\hat{\sigma}_{y_h}$, which follows t_{n-p-1} distribution
- $100(1-\alpha)\%$ Prediction Interval (PI) for y_h: $\quad \hat{y}_h \pm t_{n-p-1}^{(\alpha/2)} \hat{\sigma}_{y_h}$

Confidence and Prediction Intervals in R

```
# confidence interval
> newdata <- data.frame(cyl=factor(6, levels=c(4,6,8)), am=1, carb=4)
> predict(mod, newdata, interval="confidence")
       fit      lwr      upr
1 21.51824 18.92554 24.11094
```

```
# prediction interval
> newdata <- data.frame(cyl=factor(6, levels=c(4,6,8)), am=1, carb=4)
> predict(mod, newdata, interval="prediction")
       fit      lwr      upr
1 21.51824 15.20583 27.83065
```

Simultaneous Confidence Regions

Given the distribution of $\hat{\boldsymbol{\beta}}$ (and some probability theory), we have that

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{X}^\top \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} \sim \chi^2_{p+1} \quad \text{and} \quad \frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$$

which implies that

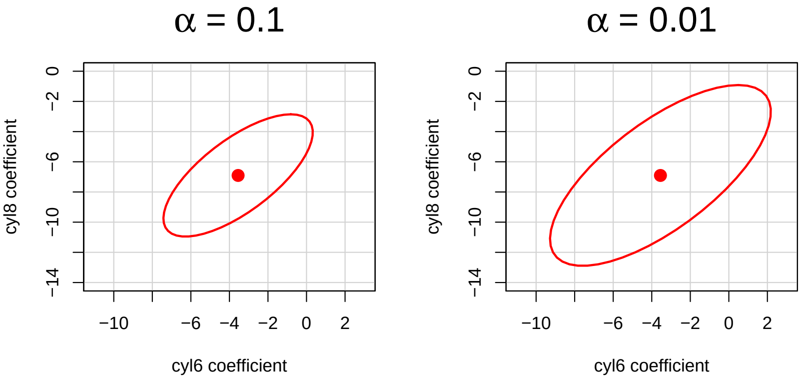
$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{X}^\top \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{(p+1)\hat{\sigma}^2} \sim \frac{\chi^2_{p+1}/(p+1)}{\chi^2_{n-p-1}/(n-p-1)} \equiv F_{(p+1,n-p-1)}$$

To form a $100(1-\alpha)\%$ confidence region (CR) use limits such that

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{X}^\top \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq (p+1)\hat{\sigma}^2 F^{(\alpha)}_{(p+1,n-p-1)}$$

where $F^{(\alpha)}_{(p+1,n-p-1)}$ is the critical value for significance level α.

Simultaneous Confidence Regions in R




```
dev.new(height=4,width=8,noRStudioGD=TRUE)
par(mfrow=c(1,2))
confidenceEllipse(mod,c(2,3),levels=.9,xlim=c(-11,3),ylim=c(-14,0),
                  main=expression(alpha*" = "*.1),cex.main=2)
confidenceEllipse(mod,c(2,3),levels=.99,xlim=c(-11,3),ylim=c(-14,0),
                  main=expression(alpha*" = "*.01),cex.main=2)
```

Table of Contents

MvLR Model: Scalar Form

The multivariate (multiple) linear regression model has the form

$$y_{ik} = b_{0k} + \sum_{j=1}^{p} b_{jk} x_{ij} + e_{ik}$$

for $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, m\}$ where

- $y_{ik} \in \mathbb{R}$ is the k-th real-valued response for the i-th observation
- $b_{0k} \in \mathbb{R}$ is the regression intercept for k-th response
- $b_{jk} \in \mathbb{R}$ is the j-th predictor's regression slope for k-th response
- $x_{ij} \in \mathbb{R}$ is the j-th predictor for the i-th observation
- $(e_{i1}, \ldots, e_{im})^\top \stackrel{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}_m, \boldsymbol{\Sigma})$ is a multivariate Gaussian error vector

MvLR Model: Nomenclature

The model is multivariate because we have $m > 1$ response variables.

The model is multiple because we have $p > 1$ predictors.

- If $p = 1$, we have a multivariate simple linear regression model

The model is linear because y_{ik} is a linear function of the parameters (b_{jk} are the parameters for $j \in \{1, \ldots, p+1\}$ and $k \in \{1, \ldots, m\}$).

The model is a regression model because we are modeling response variables $(Y_1, \ldots, Y_m)$ as a function of predictor variables $(X_1, \ldots, X_p)$.

MvLR Model: Assumptions

The fundamental assumptions of the MLR model are:

1. Relationship between X_j and Y_k is linear (given other predictors)
2. x_{ij} and y_{ik} are observed random variables (known constants)
3. $(e_{i1}, \ldots, e_{im})^\top \stackrel{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}_m, \boldsymbol{\Sigma})$ is an unobserved random vector
4. $\mathbf{b}_k = (b_{0k}, b_{1k}, \ldots, b_{pk})^\top$ for $k \in \{1, \ldots, m\}$ are unknown constants
5. $(y_{ik} | x_{i1}, \ldots, x_{ip}) \sim \mathrm{N}(b_{0k} + \sum_{j=1}^p b_{jk} x_{ij}, \sigma_{kk})$ for all $k \in \{1, \ldots, m\}$
   note: homogeneity of variance for each response

Note: b_{jk} is expected increase in Y_k for 1-unit increase in X_j with all other predictor variables held constant

MvLR Model: Matrix Form

The multivariate multiple linear regression model has the form

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

where

- $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_m] \in \mathbb{R}^{n \times m}$ is the $n \times m$ response matrix
  - $\mathbf{y}_k = (y_{1k}, \ldots, y_{nk})^\top \in \mathbb{R}^n$ is k-th response vector ($n \times 1$)
- $\mathbf{X} = [\mathbf{1}_n, \mathbf{x}_1, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n \times (p+1)}$ is the $n \times (p+1)$ design matrix
  - $\mathbf{1}_n$ is an $n \times 1$ vector of ones
  - $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^\top \in \mathbb{R}^n$ is j-th predictor vector ($n \times 1$)
- $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_m] \in \mathbb{R}^{(p+1) \times m}$ is $(p+1) \times m$ matrix of coefficients
  - $\mathbf{b}_k = (b_{0k}, b_{1k}, \ldots, b_{pk})^\top \in \mathbb{R}^{p+1}$ is k-th coefficient vector ($p + 1 \times 1$)
- $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_m] \in \mathbb{R}^{n \times m}$ is the $n \times m$ error matrix
  - $\mathbf{e}_k = (e_{1k}, \ldots, e_{nk})^\top \in \mathbb{R}^n$ is k-th error vector ($n \times 1$)

MvLR Model: Matrix Form (another look)

Matrix form writes MLR model for all nm points simultaneously

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

$$\begin{pmatrix} y_{11} & \cdots & y_{1m} \\ y_{21} & \cdots & y_{2m} \\ y_{31} & \cdots & y_{3m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} b_{01} & \cdots & b_{0m} \\ b_{11} & \cdots & b_{1m} \\ b_{21} & \cdots & b_{2m} \\ \vdots & \ddots & \vdots \\ b_{p1} & \cdots & b_{pm} \end{pmatrix} + \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ e_{21} & \cdots & e_{2m} \\ e_{31} & \cdots & e_{3m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix}$$

MvLR Model: Assumptions (revisited)

Assuming that the n subjects are independent, we have that

- $\mathbf{e}_k \sim \mathrm{N}(\mathbf{0}_n, \sigma_{kk}\mathbf{I}_n)$ where $\mathbf{e}_k$ is k-th column of $\mathbf{E}$
- $\mathbf{e}_i \stackrel{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}_m, \boldsymbol{\Sigma})$ where $\mathbf{e}_i$ is i-th row of $\mathbf{E}$
- $\mathrm{vec}(\mathbf{E}) \sim \mathrm{N}(\mathbf{0}_{nm}, \boldsymbol{\Sigma} \otimes \mathbf{I}_n)$ where $\otimes$ denotes the Kronecker product
- $\mathrm{vec}(\mathbf{E}^\top) \sim \mathrm{N}(\mathbf{0}_{nm}, \mathbf{I}_n \otimes \boldsymbol{\Sigma})$ where $\otimes$ denotes the Kronecker product

The response matrix is multivariate normal given $\mathbf{X}$

$$(\mathrm{vec}(\mathbf{Y})|\mathbf{X}) \sim \mathrm{N}([\mathbf{B}^\top \otimes \mathbf{I}_n]\mathrm{vec}(\mathbf{X}), \boldsymbol{\Sigma} \otimes \mathbf{I}_n)$$
$$(\mathrm{vec}(\mathbf{Y}^\top)|\mathbf{X}) \sim \mathrm{N}([\mathbf{I}_n \otimes \mathbf{B}^\top]\mathrm{vec}(\mathbf{X}^\top), \mathbf{I}_n \otimes \boldsymbol{\Sigma})$$

where $[\mathbf{B}^\top \otimes \mathbf{I}_n]\mathrm{vec}(\mathbf{X}) = \mathrm{vec}(\mathbf{XB})$ and
$[\mathbf{I}_n \otimes \mathbf{B}^\top]\mathrm{vec}(\mathbf{X}^\top) = \mathrm{vec}(\mathbf{B}^\top\mathbf{X}^\top)$

MvLR Model: Mean and Covariance

Note that the assumed mean vector for $\mathrm{vec}(\mathbf{Y}^\top)$ is

$$[\mathbf{I}_n \otimes \mathbf{B}^\top]\mathrm{vec}(\mathbf{X}^\top) = \mathrm{vec}(\mathbf{B}^\top\mathbf{X}^\top) = \begin{pmatrix} \mathbf{B}^\top\mathbf{x}_1 \\ \vdots \\ \mathbf{B}^\top\mathbf{x}_n \end{pmatrix}$$

where $\mathbf{x}_i$ is the i-th row of $\mathbf{X}$

The assumed covariance matrix for $\mathrm{vec}(\mathbf{Y}^\top)$ is block diagonal

$$\mathbf{I}_n \otimes \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0}_{m\times m} & \cdots & \mathbf{0}_{m\times m} \\ \mathbf{0}_{m\times m} & \boldsymbol{\Sigma} & \cdots & \mathbf{0}_{m\times m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{m\times m} & \mathbf{0}_{m\times m} & \cdots & \boldsymbol{\Sigma} \end{pmatrix}$$

Ordinary Least Squares

The ordinary least squares (OLS) problem is

$$\min_{\mathbf{B}\in\mathbb{R}^{(p+1)\times m}} \|\mathbf{Y}-\mathbf{X}\mathbf{B}\|^2 = \min_{\mathbf{B}\in\mathbb{R}^{(p+1)\times m}} \sum_{i=1}^{n}\sum_{k=1}^{m}\left(y_{ik} - b_{0k} - \textstyle\sum_{j=1}^{p} b_{jk}x_{ij}\right)^2$$

where $\|\cdot\|$ denotes the Frobenius norm.

- $\mathrm{OLS}(\mathbf{B}) = \|\mathbf{Y}-\mathbf{X}\mathbf{B}\|^2 = \mathrm{tr}(\mathbf{Y}^\top\mathbf{Y}) - 2\mathrm{tr}(\mathbf{Y}^\top\mathbf{X}\mathbf{B}) + \mathrm{tr}(\mathbf{B}^\top\mathbf{X}^\top\mathbf{X}\mathbf{B})$
- $\frac{\partial \mathrm{OLS}(\mathbf{B})}{\partial \mathbf{B}} = -2\mathbf{X}^\top\mathbf{Y} + 2\mathbf{X}^\top\mathbf{X}\mathbf{B}$

The OLS solution has the form

$$\hat{\mathbf{B}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \quad\Longleftrightarrow\quad \hat{\mathbf{b}}_k = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}_k$$

where $\mathbf{b}_k$ and $\mathbf{y}_k$ denote the k-th columns of $\mathbf{B}$ and $\mathbf{Y}$, respectively.

Fitted Values and Residuals

SCALAR FORM:

Fitted values are given by

$$\hat{y}_{ik} = \hat{b}_{0k} + \textstyle\sum_{j=1}^{p} \hat{b}_{jk} x_{ij}$$

and residuals are given by

$$\hat{e}_{ik} = y_{ik} - \hat{y}_{ik}$$

MATRIX FORM:

Fitted values are given by

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$$

and residuals are given by

$$\hat{\mathbf{E}} = \mathbf{Y} - \hat{\mathbf{Y}}$$

Hat Matrix

Note that we can write the fitted values as

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\hat{\mathbf{B}} \\ &= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \\ &= \mathbf{H}\mathbf{Y}\end{aligned}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is the hat matrix.

$\mathbf{H}$ is a symmetric and idempotent matrix: $\quad \mathbf{H}\mathbf{H} = \mathbf{H}$

$\mathbf{H}$ projects $\mathbf{y}_k$ onto the column space of $\mathbf{X}$ for $k \in \{1, \ldots, m\}$.

Multivariate Regression Example in R

```
> 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
> mtcars$cyl <- factor(mtcars$cyl)
> Y <- as.matrix(mtcars[,c("mpg","disp","hp","wt")])
> mvmod <- lm(Y ~ cyl + am + carb, data=mtcars)
> coef(mvmod)
                  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
```

Sums-of-Squares and Crossproducts: Vector Form

In MvLR models, the relevant sums-of-squares and crossproducts are

- Total: $\mathrm{SSCP}_T = \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top$
- Regression: $\mathrm{SSCP}_R = \sum_{i=1}^n (\hat{\mathbf{y}}_i - \bar{\mathbf{y}})(\hat{\mathbf{y}}_i - \bar{\mathbf{y}})^\top$
- Error: $\mathrm{SSCP}_E = \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)(\mathbf{y}_i - \hat{\mathbf{y}}_i)^\top$

where $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ denote the i-th rows of $\mathbf{Y}$ and $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$, respectively.

The corresponding degrees of freedom are

- SSCP_T: $df_T = m(n-1)$
- SSCP_R: $df_R = mp$
- SSCP_E: $df_E = m(n-p-1)$

Sums-of-Squares and Crossproducts: Matrix Form

In MvLR models, the relevant sums-of-squares are

$$\begin{aligned}
\mathrm{SSCP}_T &= \sum_{i=1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top \\
&= \mathbf{Y}^\top \left[\mathbf{I}_n - (1/n)\mathbf{J}\right] \mathbf{Y} \\
\mathrm{SSCP}_R &= \sum_{i=1}^{n} (\hat{\mathbf{y}}_i - \bar{\mathbf{y}})(\hat{\mathbf{y}}_i - \bar{\mathbf{y}})^\top \\
&= \mathbf{Y}^\top \left[\mathbf{H} - (1/n)\mathbf{J}\right] \mathbf{Y} \\
\mathrm{SSCP}_E &= \sum_{i=1}^{n} (\mathbf{y}_i - \hat{\mathbf{y}}_i)(\mathbf{y}_i - \hat{\mathbf{y}}_i)^\top \\
&= \mathbf{Y}^\top \left[\mathbf{I}_n - \mathbf{H}\right] \mathbf{Y}
\end{aligned}$$

Note: $\mathbf{J}$ is an $n \times n$ matrix of ones

Partitioning the SSCP Total Matrix

We can partition the total covariation in $\mathbf{y}_i$ as

$$\begin{aligned}
\mathrm{SSCP}_T &= \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top \\
&= \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i + \hat{\mathbf{y}}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \hat{\mathbf{y}}_i + \hat{\mathbf{y}}_i - \bar{\mathbf{y}})^\top \\
&= \sum_{i=1}^n (\hat{\mathbf{y}}_i - \bar{\mathbf{y}})(\hat{\mathbf{y}}_i - \bar{\mathbf{y}})^\top + \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)(\mathbf{y}_i - \hat{\mathbf{y}}_i)^\top \\
&\quad\quad + 2\sum_{i=1}^n (\hat{\mathbf{y}}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \hat{\mathbf{y}}_i)^\top \\
&= \mathrm{SSCP}_R + \mathrm{SSCP}_E + 2\sum_{i=1}^n (\hat{\mathbf{y}}_i - \bar{\mathbf{y}})\hat{\mathbf{e}}_i^\top \\
&= \mathrm{SSCP}_R + \mathrm{SSCP}_E
\end{aligned}$$

Multivariate Regression SSCP in R

```
> ybar <- colMeans(Y)
> n <- nrow(Y)
> m <- ncol(Y)
> Ybar <-  matrix(ybar, n, m, byrow=TRUE)
> SSCP.T <- crossprod(Y - Ybar)
> SSCP.R <- crossprod(mvmod$fitted.values - Ybar)
> SSCP.E <- crossprod(Y - mvmod$fitted.values)
> SSCP.T
            mpg       disp          hp          wt
mpg    1126.0472 -19626.01   -9942.694  -158.61723
disp -19626.0134 476184.79  208355.919  3338.21032
hp    -9942.6938 208355.92  145726.875  1369.97250
wt     -158.6172   3338.21    1369.972    29.67875
> SSCP.R + SSCP.E
            mpg       disp          hp          wt
mpg    1126.0472 -19626.01   -9942.694  -158.61723
disp -19626.0134 476184.79  208355.919  3338.21033
hp    -9942.6938 208355.92  145726.875  1369.97250
wt     -158.6172   3338.21    1369.973    29.67875
```

Relation to ML Solution

Remember that $(\mathbf{y}_i|\mathbf{x}_i) \sim \mathrm{N}(\mathbf{B}^\top \mathbf{x}_i, \mathbf{\Sigma})$, which implies that $\mathbf{y_i}$ has pdf

$$f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{B}, \mathbf{\Sigma}) = (2\pi)^{-m/2}|\mathbf{\Sigma}|^{-1/2} \exp\{-(1/2)(\mathbf{y}_i - \mathbf{B}^\top \mathbf{x}_i)^\top \mathbf{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{B}^\top \mathbf{x}_i)\}$$

where $\mathbf{y}_i$ and $\mathbf{x}_i$ denote the i-th rows of $\mathbf{Y}$ and $\mathbf{X}$, respectively.

As a result, the log-likelihood of $\mathbf{B}$ given $(\mathbf{Y}, \mathbf{X}, \mathbf{\Sigma})$ is

$$\ln\{L(\mathbf{B}|\mathbf{Y}, \mathbf{X}, \mathbf{\Sigma})\} = -\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{B}^\top \mathbf{x}_i)^\top \mathbf{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{B}^\top \mathbf{x}_i) + c$$

where c is a constant that does not depend on $\mathbf{B}$.

Relation to ML Solution (continued)

The maximum likelihood estimate (MLE) of $\mathbf{B}$ satisfy

$$\max_{\mathbf{B}\in\mathbb{R}^{(p+1)\times m}} \mathrm{MLE}(\mathbf{B}) = \max_{\mathbf{B}\in\mathbb{R}^{(p+1)\times m}} -\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{B}^\top\mathbf{x}_i)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{B}^\top\mathbf{x}_i)$$

and note that
$(\mathbf{y}_i - \mathbf{B}^\top\mathbf{x}_i)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{B}^\top\mathbf{x}_i) = \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{B}^\top\mathbf{x}_i)(\mathbf{y}_i - \mathbf{B}^\top\mathbf{x}_i)^\top\}$

Taking the derivative with respect to $\mathbf{B}$ we see that

$$\begin{aligned}\frac{\partial \mathrm{MLE}(\mathbf{B})}{\partial \mathbf{B}} &= -2\sum_{i=1}^{n}\mathbf{x}_i\mathbf{y}_i^\top\boldsymbol{\Sigma}^{-1} + 2\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{B}\boldsymbol{\Sigma}^{-1} \\ &= -2\mathbf{X}^\top\mathbf{Y}\boldsymbol{\Sigma}^{-1} + 2\mathbf{X}^\top\mathbf{X}\mathbf{B}\boldsymbol{\Sigma}^{-1}\end{aligned}$$

The OLS and ML estimate of $\mathbf{B}$ is the same: $\hat{\mathbf{B}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$

Estimated Error Covariance

The estimated error variance is

$$\begin{aligned}\hat{\boldsymbol{\Sigma}} &= \frac{\mathrm{SSCP}_E}{n-p-1} \\ &= \frac{\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)(\mathbf{y}_i - \hat{\mathbf{y}}_i)^\top}{n-p-1} \\ &= \frac{\mathbf{Y}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{Y}}{n-p-1}\end{aligned}$$

which is an unbiased estimate of error covariance matrix $\boldsymbol{\Sigma}$.

The estimate $\hat{\boldsymbol{\Sigma}}$ is the mean SSCP error of the model.

Maximum Likelihood Estimate of Error Covariance

$\tilde{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{Y}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{Y}$ is the MLE of $\boldsymbol{\Sigma}$.

From our previous results using $\hat{\boldsymbol{\Sigma}}$, we have that

$$\mathrm{E}(\tilde{\boldsymbol{\Sigma}}) = \frac{n-p-1}{n}\boldsymbol{\Sigma}$$

Consequently, the bias of the estimator $\tilde{\boldsymbol{\Sigma}}$ is given by

$$\frac{n-p-1}{n}\boldsymbol{\Sigma} - \boldsymbol{\Sigma} = -\frac{(p+1)}{n}\boldsymbol{\Sigma}$$

and note that $-\frac{(p+1)}{n}\boldsymbol{\Sigma} \to \mathbf{0}_{m\times m}$ as $n \to \infty$.

Comparing $\hat{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\Sigma}}$

Reminder: the MSSCPE and MLE of $\boldsymbol{\Sigma}$ are given by

$$\hat{\boldsymbol{\Sigma}} = \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{H}) \mathbf{Y}/(n-p-1)$$
$$\tilde{\boldsymbol{\Sigma}} = \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{H}) \mathbf{Y}/n$$

From the definitions of $\hat{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\Sigma}}$ we have that

$$\tilde{\sigma}_{kk} < \hat{\sigma}_{kk} \quad \text{for all } k$$

where $\hat{\sigma}_{kk}$ and $\tilde{\sigma}_{kk}$ denote the k-th diagonals of $\hat{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\Sigma}}$, respectively.

- MLE produces smaller estimates of the error variances

Estimated Error Covariance Matrix in R

```
> n <- nrow(Y)
> p <- nrow(coef(mvmod)) - 1
> SSCP.E <- crossprod(Y - mvmod$fitted.values)
> SigmaHat <- SSCP.E / (n - p - 1)
> SigmaTilde <- SSCP.E / n
> 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
> SigmaTilde
           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
```

Expected Value of Least Squares Coefficients

The expected value of the estimated coefficients is given by

$$\begin{aligned} E(\hat{\mathbf{B}}) &= E[(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}] \\ &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top E(\mathbf{Y}) \\ &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{X}\mathbf{B} \\ &= \mathbf{B} \end{aligned}$$

so $\hat{\mathbf{B}}$ is an unbiased estimator of $\mathbf{B}$.

Covariance Matrix of Least Squares Coefficients

The covariance matrix of the estimated coefficients is given by

$$
\begin{aligned}
V\{\mathrm{vec}(\hat{\mathbf{B}}^\top)\} &= V\{\mathrm{vec}(\mathbf{Y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1})\} \\
&= V\{[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \otimes \mathbf{I}_m]\mathrm{vec}(\mathbf{Y}^\top)\} \\
&= [(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \otimes \mathbf{I}_m] V\{\mathrm{vec}(\mathbf{Y}^\top)\}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \otimes \mathbf{I}_m]^\top \\
&= [(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \otimes \mathbf{I}_m][\mathbf{I}_n \otimes \boldsymbol{\Sigma}][\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \otimes \mathbf{I}_m] \\
&= [(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \otimes \mathbf{I}_m][\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \otimes \boldsymbol{\Sigma}] \\
&= (\mathbf{X}^\top\mathbf{X})^{-1} \otimes \boldsymbol{\Sigma}
\end{aligned}
$$

Note: we could also write $V\{\mathrm{vec}(\hat{\mathbf{B}})\} = \boldsymbol{\Sigma} \otimes (\mathbf{X}^\top\mathbf{X})^{-1}$

Distribution of Coefficients

The estimated regression coefficients are a linear function of $\mathbf{Y}$ so we know that $\hat{\mathbf{B}}$ follows a multivariate normal distribution.

- $\mathrm{vec}(\hat{\mathbf{B}}) \sim \mathrm{N}[\mathrm{vec}(\mathbf{B}), \boldsymbol{\Sigma} \otimes (\mathbf{X}^\top \mathbf{X})^{-1}]$
- $\mathrm{vec}(\hat{\mathbf{B}}^\top) \sim \mathrm{N}[\mathrm{vec}(\mathbf{B}^\top), (\mathbf{X}^\top \mathbf{X})^{-1} \otimes \boldsymbol{\Sigma}]$

The covariance between two columns of $\hat{\mathbf{B}}$ has the form

$$\mathrm{Cov}(\hat{\mathbf{b}}_k, \hat{\mathbf{b}}_\ell) = \sigma_{k\ell}(\mathbf{X}^\top \mathbf{X})^{-1}$$

and the covariance between two rows of $\hat{\mathbf{B}}$ has the form

$$\mathrm{Cov}(\hat{\mathbf{b}}_g, \hat{\mathbf{b}}_j) = (\mathbf{X}^\top \mathbf{X})^{-1}_{gj} \boldsymbol{\Sigma}$$

where $(\mathbf{X}^\top \mathbf{X})^{-1}_{gj}$ denotes the (g, j)-th element of $(\mathbf{X}^\top \mathbf{X})^{-1}$.

Expectation and Covariance of Fitted Values

The expected value of the fitted values is given by

$$E(\hat{\mathbf{Y}}) = E(\mathbf{X}\hat{\mathbf{B}}) = \mathbf{X}E(\hat{\mathbf{B}}) = \mathbf{X}\mathbf{B}$$

and the covariance matrix has the form

$$\begin{aligned}
V\{\mathrm{vec}(\hat{\mathbf{Y}}^\top)\} &= V\{\mathrm{vec}(\hat{\mathbf{B}}^\top\mathbf{X}^\top)\} \\
&= V\{(\mathbf{X}\otimes\mathbf{I}_m)\mathrm{vec}(\hat{\mathbf{B}}^\top)\} \\
&= (\mathbf{X}\otimes\mathbf{I}_m)V\{\mathrm{vec}(\hat{\mathbf{B}}^\top)\}(\mathbf{X}\otimes\mathbf{I}_m)^\top \\
&= (\mathbf{X}\otimes\mathbf{I}_m)[(\mathbf{X}^\top\mathbf{X})^{-1}\otimes\boldsymbol{\Sigma}](\mathbf{X}\otimes\mathbf{I}_m)^\top \\
&= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\otimes\boldsymbol{\Sigma}
\end{aligned}$$

Note: we could also write $V\{\mathrm{vec}(\hat{\mathbf{Y}})\} = \boldsymbol{\Sigma}\otimes\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$

Distribution of Fitted Values

The fitted values are a linear function of $\mathbf{Y}$ so we know that $\hat{\mathbf{Y}}$ follows a multivariate normal distribution.

- $\mathrm{vec}(\hat{\mathbf{Y}}) \sim \mathrm{N}[(\mathbf{B}^\top \otimes \mathbf{I}_n)\mathrm{vec}(\mathbf{X}), \boldsymbol{\Sigma} \otimes \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]$
- $\mathrm{vec}(\hat{\mathbf{Y}}^\top) \sim \mathrm{N}[(\mathbf{I}_n \otimes \mathbf{B}^\top)\mathrm{vec}(\mathbf{X}^\top), \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \otimes \boldsymbol{\Sigma}]$

where $(\mathbf{B}^\top \otimes \mathbf{I}_n)\mathrm{vec}(\mathbf{X}) = \mathrm{vec}(\mathbf{XB})$ and
$(\mathbf{I}_n \otimes \mathbf{B}^\top)\mathrm{vec}(\mathbf{X}^\top) = \mathrm{vec}(\mathbf{B}^\top\mathbf{X}^\top)$.

The covariance between two columns of $\hat{\mathbf{Y}}$ has the form

$$\mathrm{Cov}(\hat{\mathbf{y}}_k, \hat{\mathbf{y}}_\ell) = \sigma_{k\ell}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$$

and the covariance between two rows of $\hat{\mathbf{Y}}$ has the form

$$\mathrm{Cov}(\hat{\mathbf{y}}_g, \hat{\mathbf{y}}_j) = h_{gj}\boldsymbol{\Sigma}$$

where h_{gj} denotes the (g, j)-th element of $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

Expectation and Covariance of Residuals

The expected value of the residuals is given by

$$E(\mathbf{Y} - \hat{\mathbf{Y}}) = E([\mathbf{I}_n - \mathbf{H}]\mathbf{Y}) = (\mathbf{I}_n - \mathbf{H})E(\mathbf{Y}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\mathbf{B} = \mathbf{0}_{n \times m}$$

and the covariance matrix has the form

$$\begin{aligned} V\{\mathrm{vec}(\hat{\mathbf{E}}^\top)\} &= V\{\mathrm{vec}(\mathbf{Y}^\top[\mathbf{I}_n - \mathbf{H}])\} \\ &= V\{([\mathbf{I}_n - \mathbf{H}] \otimes \mathbf{I}_m)\mathrm{vec}(\mathbf{Y}^\top)\} \\ &= ([\mathbf{I}_n - \mathbf{H}] \otimes \mathbf{I}_m)V\{\mathrm{vec}(\mathbf{Y}^\top)\}([\mathbf{I}_n - \mathbf{H}] \otimes \mathbf{I}_m) \\ &= ([\mathbf{I}_n - \mathbf{H}] \otimes \mathbf{I}_m)[\mathbf{I}_n \otimes \mathbf{\Sigma}]([\mathbf{I}_n - \mathbf{H}] \otimes \mathbf{I}_m) \\ &= (\mathbf{I}_n - \mathbf{H}) \otimes \mathbf{\Sigma} \end{aligned}$$

Note: we could also write $V\{\mathrm{vec}(\hat{\mathbf{E}})\} = \mathbf{\Sigma} \otimes (\mathbf{I}_n - \mathbf{H})$

Distribution of Residuals

The residuals are a linear function of $\mathbf{Y}$ so we know that $\hat{\mathbf{E}}$ follows a multivariate normal distribution.

- $\mathrm{vec}(\hat{\mathbf{E}}) \sim \mathrm{N}[\mathbf{0}_{mn}, \boldsymbol{\Sigma} \otimes (\mathbf{I}_n - \mathbf{H})]$
- $\mathrm{vec}(\hat{\mathbf{E}}^\top) \sim \mathrm{N}[\mathbf{0}_{mn}, (\mathbf{I}_n - \mathbf{H}) \otimes \boldsymbol{\Sigma}]$

The covariance between two columns of $\hat{\mathbf{E}}$ has the form

$$\mathrm{Cov}(\hat{\mathbf{e}}_k, \hat{\mathbf{e}}_\ell) = \sigma_{k\ell}(\mathbf{I}_n - \mathbf{H})$$

and the covariance between two rows of $\hat{\mathbf{E}}$ has the form

$$\mathrm{Cov}(\hat{\mathbf{e}}_g, \hat{\mathbf{e}}_j) = (\delta_{gj} - h_{gj})\boldsymbol{\Sigma}$$

where δ_{gj} is a Kronecker's δ and h_{gj} denotes the (g, j)-th element of $\mathbf{H}$.

Summary of Results

Given the model assumptions, we have

$$\mathrm{vec}(\hat{\mathbf{B}}) \sim \mathrm{N}[\mathrm{vec}(\mathbf{B}), \boldsymbol{\Sigma} \otimes (\mathbf{X}^\top \mathbf{X})^{-1}]$$

$$\mathrm{vec}(\hat{\mathbf{Y}}) \sim \mathrm{N}[\mathrm{vec}(\mathbf{XB}), \boldsymbol{\Sigma} \otimes \mathbf{H}]$$

$$\mathrm{vec}(\hat{\mathbf{E}}) \sim \mathrm{N}[\mathbf{0}_{mn}, \boldsymbol{\Sigma} \otimes (\mathbf{I}_n - \mathbf{H})]$$

where $\mathrm{vec}(\mathbf{XB}) = (\mathbf{B}^\top \otimes \mathbf{I}_n)\mathrm{vec}(\mathbf{X})$.

Typically $\boldsymbol{\Sigma}$ is unknown, so we use the mean SSCP error matrix $\hat{\boldsymbol{\Sigma}}$.

Coefficient Inference in R

```
> mvsum <- summary(mvmod)
> mvsum[[1]]

Call:
lm(formula = mpg ~ cyl + am + carb, data = mtcars)

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
```

Coefficient Inference in R (continued)

```
> mvsum <- summary(mvmod)
> mvsum[[3]]

Call:
lm(formula = hp ~ cyl + am + carb, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-41.520 -17.941  -4.378  19.799  41.292

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  46.5201    10.4825   4.438 0.000138 ***
cyl6          0.9116    14.8146   0.062 0.951386
cyl8         87.5911    15.4851   5.656 5.25e-06 ***
am            4.4473    11.5629   0.385 0.703536
carb         21.2765     3.7291   5.706 4.61e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.03 on 27 degrees of freedom
Multiple R-squared:  0.893,Adjusted R-squared:  0.8772
F-statistic: 56.36 on 4 and 27 DF,  p-value: 1.023e-12
```

Inferences about Multiple $\hat{b}_{jk}$

Assume that $q < p$ and want to test if a reduced model is sufficient:

$$H_0 : \mathbf{B}_2 = \mathbf{0}_{(p-q)\times m}$$
$$H_1 : \mathbf{B}_2 \neq \mathbf{0}_{(p-q)\times m}$$

where

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix}$$

is the partitioned coefficient vector.

Compare the SSCP-Error for full and reduced (constrained) models:
(a) Full Model: $y_{ik} = b_{0k} + \sum_{j=1}^{p} b_{jk}x_{ij} + e_{ik}$
(b) Reduced Model: $y_{ik} = b_{0k} + \sum_{j=1}^{q} b_{jk}x_{ij} + e_{ik}$

Inferences about Multiple $\hat{b}_{jk}$ (continued)

Likelihood Ratio Test Statistic:

$$\begin{aligned}\Lambda &= \frac{\max_{\mathbf{B}_1, \boldsymbol{\Sigma}} L(\mathbf{B}_1, \boldsymbol{\Sigma})}{\max_{\mathbf{B}, \boldsymbol{\Sigma}} L(\mathbf{B}, \boldsymbol{\Sigma})} \\ &= \left(\frac{|\tilde{\boldsymbol{\Sigma}}|}{|\tilde{\boldsymbol{\Sigma}}_1|} \right)^{n/2}\end{aligned}$$

where

- $\tilde{\boldsymbol{\Sigma}}$ is the MLE of $\boldsymbol{\Sigma}$ with $\mathbf{B}$ unconstrained
- $\tilde{\boldsymbol{\Sigma}}_1$ is the MLE of $\boldsymbol{\Sigma}$ with $\mathbf{B}_2 = \mathbf{0}_{(p-1)\times m}$

For large n, we can use the modified test statistic

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

where $\nu = n - p - 1 - (1/2)(m - p + q + 1)$

Some Other Test Statistics

Let $\tilde{\mathbf{E}} = n\tilde{\mathbf{\Sigma}}$ denote the SSCP error matrix from the full model, and let $\tilde{\mathbf{H}} = n(\tilde{\mathbf{\Sigma}}_1 - \tilde{\mathbf{\Sigma}})$ denote the hypothesis (or extra) SSCP error matrix.

Test statistics for $H_0 : \mathbf{B}_2 = \mathbf{0}_{(p-1)\times m}$ versus $H_1 : \mathbf{B}_2 \neq \mathbf{0}_{(p-1)\times m}$

- Wilks' lambda $= \prod_{i=1}^{s} \frac{1}{1+\eta_i} = \frac{|\tilde{\mathbf{E}}|}{|\tilde{\mathbf{E}}+\tilde{\mathbf{H}}|}$
- Pillai's trace $= \sum_{i=1}^{s} \frac{\eta_i}{1+\eta_i} = \mathrm{tr}[\tilde{\mathbf{H}}(\tilde{\mathbf{E}} + \tilde{\mathbf{H}})^{-1}]$
- Hotelling-Lawley trace $= \sum_{i=1}^{s} \eta_s = \mathrm{tr}(\tilde{\mathbf{H}}\tilde{\mathbf{E}}^{-1})$
- 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 $\tilde{\mathbf{H}}\tilde{\mathbf{E}}^{-1}$

Testing a Reduced Multivariate Linear Model in R

```
> mvmod0 <- lm(Y ~ am + carb, data=mtcars)
> anova(mvmod, mvmod0, test="Wilks")
Analysis of Variance Table

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

> anova(mvmod, mvmod0, test="Pillai")
Analysis of Variance Table

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

> Etilde <- n * SigmaTilde
> SigmaTilde1 <- crossprod(Y - mvmod0$fitted.values) / n
> Htilde <- n * (SigmaTilde1 - SigmaTilde)
> HEi <- Htilde %*% solve(Etilde)
> HEi.values <- eigen(HEi)$values
> c(Wilks = prod(1 / (1 + HEi.values)), Pillai = sum(HEi.values / (1 + HEi.values)))
    Wilks    Pillai
0.1639527 1.0322975
```

Interval Estimation

Idea: estimate expected value of response for a given predictor score.

Given $\mathbf{x}_h = (1, x_{h1}, \ldots, x_{hp})$, we have $\hat{\mathbf{y}}_h = (\hat{y}_{h1}, \ldots, \hat{y}_{hk})^\top = \hat{\mathbf{B}}^\top \mathbf{x}_h$.

Note that $\hat{\mathbf{y}}_h \sim \mathrm{N}(\mathbf{B}^\top \mathbf{x}_h, \mathbf{x}_h^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_h \boldsymbol{\Sigma})$ from our previous results.

We can test $H_0 : \mathrm{E}(\mathbf{y}_h) = \mathbf{y}_h^*$ versus $H_1 : \mathrm{E}(\mathbf{y}_h) \neq \mathbf{y}_h^*$

- $T^2 = \left(\frac{\hat{\mathbf{B}}^\top \mathbf{x}_h - \mathbf{B}^\top \mathbf{x}_h}{\sqrt{\mathbf{x}_h^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_h}} \right)^\top \hat{\boldsymbol{\Sigma}}^{-1} \left(\frac{\hat{\mathbf{B}}^\top \mathbf{x}_h - \mathbf{B}^\top \mathbf{x}_h}{\sqrt{\mathbf{x}_h^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_h}} \right) \sim \frac{m(n-p-1)}{n-p-m} F_{m,(n-p-m)}$
- $100(1-\alpha)\%$ simultaneous CI for $E(y_{hk})$:
  $\hat{y}_{hk} \pm \sqrt{\frac{m(n-p-1)}{n-p-m} F_{m,(n-p-m)}} \sqrt{\mathbf{x}_h^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_h \hat{\sigma}_{kk}}$

Predicting New Observations

Idea: estimate observed value of response for a given predictor score.

- Note: interested in actual $\hat{\mathbf{y}}_h$ value instead of $\mathrm{E}(\hat{\mathbf{y}}_h)$
- Given $\mathbf{x}_h = (1, x_{h1}, \ldots, x_{hp})$, the fitted value is still $\hat{\mathbf{y}}_h = \hat{\mathbf{B}}^\top \mathbf{x}_h$.

When predicting a new observation, there are two uncertainties:

- location of distribution of $Y_1, \ldots, Y_m$ for $X_1, \ldots, X_p$, i.e., $V(\hat{\mathbf{y}}_h)$
- variability within the distribution of $Y_1, \ldots, Y_m$, i.e., $\boldsymbol{\Sigma}$

We can test $H_0 : \mathbf{y}_h = \mathbf{y}_h^*$ versus $H_1 : \mathbf{y}_h \neq \mathbf{y}_h^*$

- $T^2 = \left(\frac{\hat{\mathbf{B}}^\top \mathbf{x}_h - \mathbf{B}^\top \mathbf{x}_h}{\sqrt{1 + \mathbf{x}_h^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_h}} \right)^\top \hat{\boldsymbol{\Sigma}}^{-1} \left(\frac{\hat{\mathbf{B}}^\top \mathbf{x}_h - \mathbf{B}^\top \mathbf{x}_h}{\sqrt{1 + \mathbf{x}_h^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_h}} \right) \sim \frac{m(n-p-1)}{n-p-m} F_{m,(n-p-m)}$
- $100(1-\alpha)\%$ simultaneous PI for $E(y_{hk})$: $\hat{y}_{hk} \pm \sqrt{\frac{m(n-p-1)}{n-p-m} F_{m,(n-p-m)}(\alpha)} \sqrt{(1 + \mathbf{x}_h^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_h) \hat{\sigma}_{kk}}$

Confidence and Prediction Intervals in R

Note: R does not yet have this capability!

```
> # confidence interval
> newdata <- data.frame(cyl=factor(6, levels=c(4,6,8)), am=1, carb=4)
> predict(mvmod, 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(mvmod, newdata, interval="prediction")
       mpg     disp      hp       wt
1 21.51824 159.2707 136.985 2.631108
```

R Function for Multivariate Regression CIs and PIs

```
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)
  m <- 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=m, df2=n-p-m) * m * (n - p - 1) / (n - p - m)
  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), m, 3))
    dimnames(ci) <- list(1:nrow(xnew), colnames(Y), c("fit", "lwr", "upr") )
    ci[,,1] <- fit
    ci[,,2] <- lwr
    ci[,,3] <- upr
  \}
  ci
\}
```

Confidence and Prediction Intervals in R (revisited)

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

Confidence and Prediction Intervals in R (revisited 2)

```
# confidence interval (multiple new observations)
> newdata <- data.frame(cyl=factor(c(4,6,8), levels=c(4,6,8)), am=c(0,1,1), carb=c(2,4,6))
> pred.mlm(mvmod, newdata)
, , fit

       mpg     disp        hp       wt
1 23.08059 137.7774  89.07313 3.111033
2 21.51824 159.2707 136.98500 2.631108
3 15.92331 319.8707 266.21745 3.557519

, , lwr

       mpg     disp        hp       wt
1 17.76982  43.0190  43.58324 2.150555
2 16.65593  72.5141  95.33649 1.751736
3 10.65231 225.8219 221.06824 2.604233

, , upr

       mpg     disp       hp       wt
1 28.39137 232.5359 134.5630 4.071512
2 26.38055 246.0273 178.6335 3.510479
3 21.19431 413.9195 311.3667 4.510804
```