Multivariate Linear Regression

Max Turgeon

STAT 7200–Multivariate Statistics

Objectives

- Introduce the linear regression model for a multivariate outcome
- Discuss inference for the regression parameters
- Discuss model selection
- Discuss influence measures

Multivariate Linear Regression model

- We are interested in the relationship between p outcomes $Y_1, \ldots, Y_p$ and q covariates $X_1, \ldots, X_q$.
    - We will write $\mathbf{Y} = (Y_1, \ldots, Y_p)$ and $\mathbf{X} = (1, X_1, \ldots, X_q)$.
- We will assume a **linear relationship**:
    - $E(\mathbf{Y} \mid \mathbf{X}) = B^T\mathbf{X}$, where B is a $(q+1) \times p$ matrix of *regression coefficients*.
- We will also assume **homoscedasticity**:
    - $\mathrm{Cov}(\mathbf{Y} \mid \mathbf{X}) = \Sigma$, where Σ is positive-definite.
    - In other words, the (conditional) covariance of $\mathbf{Y}$ does not depend on $\mathbf{X}$.

Relationship with Univariate regression i

- Let σ_i^2 be the i-th diagonal element of Σ.
- Let β_i be the i-th column of B.
- From the model above, we get p univariate regressions:
    - $E(Y_i \mid \mathbf{X}) = \mathbf{X}^T \beta_i$;
    - $\mathrm{Var}(Y_i \mid \mathbf{X}) = \sigma_i^2$.
- However, we will use the correlation between outcomes for hypothesis testing
- This follows from the assumption that each component Y_i is linearly associated with the *same* covariates $\mathbf{X}$.

Relationship with Univariate regression ii

- If we assumed a different set of covariates $\mathbf{X}_i$ for each outcome Y_i and still wanted to use the correlation between the outcomes, we would get the Seemingly Unrelated Regressions (SUR) model.
    - This model is sometimes used by econometricians.

Least-Squares Estimation i

- Let $\mathbf{Y}_1 \ldots, \mathbf{Y}_n$ be a random sample of size n, and let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be the corresponding sample of covariates.
- We will write $\mathbb{Y}$ and $\mathbb{X}$ for the matrices whose i-th row is $\mathbf{Y}_i$ and $\mathbf{X}_i$, respectively.
    - We can then write $E(\mathbb{Y} \mid \mathbb{X}) = \mathbb{X}B$.
- For Least-Squares Estimation, we will be looking for the estimator $\hat{B}$ of B that minimises a least-squares criterion:
    - $LS(B) = \mathrm{tr}\left[(\mathbb{Y} - \mathbb{X}B)^T(\mathbb{Y} - \mathbb{X}B)\right]$
    - **Note**: This criterion is also known as the (squared) *Frobenius norm*; i.e. $LS(B) = \|\mathbb{Y} - \mathbb{X}B\|_F^2$.

Least-Squares Estimation ii

- **Note 2**: If you expand the matrix product and look at the diagonal, you can see that the Frobenius norm is equivalent to the sum of the squared entries.

- To minimise $LS(B)$, we could use matrix derivatives...
- Or, we can expand the matrix product along the diagonal and compute the trace.
- Let $\mathbf{Y}_{(j)}$ be the j-th column of $\mathbb{Y}$.

Least-Squares Estimation iii

- In other words, $\mathbf{Y}_{(j)} = (Y_{1j}, \ldots, Y_{nj})$ contains the n values for the outcome Y_j. We then have

$$\begin{aligned} LS(B) &= \mathrm{tr}\left[(\mathbb{Y} - \mathbb{X}B)^T(\mathbb{Y} - \mathbb{X}B)\right] \\ &= \sum_{j=1}^{p}(\mathbf{Y}_{(j)} - \mathbb{X}\beta_j)^T(\mathbf{Y}_{(j)} - \mathbb{X}\beta_j) \\ &= \sum_{j=1}^{p}\sum_{i=1}^{n}(Y_{ij} - \beta_j^T\mathbf{X}_i)^2. \end{aligned}$$

- For each j, the sum $\sum_{i=1}^{n}(Y_{ij} - \beta_j^T\mathbf{X}_i)^2$ is simply the least-squares criterion for the corresponding univariate linear regression.

Least-Squares Estimation iv

  - $\hat{\beta}_j = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}_{(j)}$
- But since $LS(B)$ is a sum of p positive terms, each minimised at $\hat{\beta}_j$, the whole is sum is minimised at

$$\hat{B} = \begin{pmatrix} \hat{\beta}_1 & \cdots & \hat{\beta}_p \end{pmatrix}.$$

- Or put another way:

$$\hat{B} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}.$$

Comments i

- We still have not made any distributional assumptions on $\mathbf{Y}$.
    - We do not need to assume normality to derive the least-squares estimator.
- The least-squares estimator is *unbiased*:

$$\begin{aligned} E(\hat{B} \mid \mathbb{X}) &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}E(\mathbb{Y} \mid \mathbb{X}) \\ &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}B \\ &= B. \end{aligned}$$

Comments ii

- We did not use the covariance matrix Σ anywhere in the estimation process. But note that:

$$\begin{aligned}
\mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) &= \mathrm{Cov}\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}_{(i)}, (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}_{(j)}\right) \\
&= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathrm{Cov}\left(\mathbf{Y}_{(i)}, \mathbf{Y}_{(j)}\right)\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\right)^T \\
&= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\left(\sigma_{ij}I_n\right)\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} \\
&= \sigma_{ij}(\mathbb{X}^T\mathbb{X})^{-1},
\end{aligned}$$

where σ_{ij} is the (i, j)-th entry of Σ.

Example i

```r
# Let's revisit the plastic film data
library(heplots)
library(tidyverse)

Y <- Plastic %>%
  select(tear, gloss, opacity) %>%
  as.matrix

X <- model.matrix(~ rate, data = Plastic)
head(X)
```

Example ii

```
##   (Intercept) rateHigh
## 1           1        0
## 2           1        0
## 3           1        0
## 4           1        0
## 5           1        0
## 6           1        0
```

```r
(B_hat <- solve(crossprod(X)) %*% t(X) %*% Y)
```

Example iii

```
##               tear gloss opacity
## (Intercept) 6.49  9.57    3.79
## rateHigh    0.59 -0.51    0.29
```

```r
# Compare with lm output
fit <- lm(cbind(tear, gloss, opacity) ~ rate,
          data = Plastic)
coef(fit)
```

```
##               tear gloss opacity
## (Intercept) 6.49  9.57    3.79
## rateHigh    0.59 -0.51    0.29
```

Geometry of LS i

- Let $P = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$.
- P is symmetric and *idempotent*:

$$P^2 = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T = P.$$

- Let $\hat{\mathbb{Y}} = \mathbb{X}\hat{B}$ be the fitted values, and $\hat{\mathbb{E}} = \mathbb{Y} - \hat{\mathbb{Y}}$, the residuals.
    - We have $\hat{\mathbb{Y}} = P\mathbb{Y}$.
    - We also have $\hat{\mathbb{E}} = (I - P)\mathbb{Y}$.

Geometry of LS ii

- Putting all this together, we get

$$\begin{aligned}\hat{\mathbb{Y}}^T\hat{\mathbb{E}} &= (P\mathbb{Y})^T(I-P)\mathbb{Y}\\ &= \mathbb{Y}^T P(I-P)\mathbb{Y}\\ &= \mathbb{Y}^T(P-P^2)\mathbb{Y}\\ &= 0.\end{aligned}$$

- In other words, the fitted values and the residuals are **orthogonal**.
- Similarly, we can see that $\mathbb{X}^T\hat{\mathbb{E}} = 0$ and $P\mathbb{X} = \mathbb{X}$.
- **Interpretation**: $\hat{\mathbb{Y}}$ is the orthogonal projection of $\mathbb{Y}$ onto the column space of $\mathbb{X}$.

Example (cont'd) i

```r
Y_hat <- fitted(fit)
residuals <- residuals(fit)

crossprod(Y_hat, residuals)
```

```
##                  tear         gloss       opacity
## tear     1.776357e-15 -1.998401e-15  1.776357e-15
## gloss   -8.881784e-16 -1.998401e-15 -1.065814e-14
## opacity -4.440892e-16 -1.887379e-15  1.776357e-15
```

```r
crossprod(X, residuals)
```

Example (cont'd) ii

```
##                     tear         gloss       opacity
## (Intercept) 1.110223e-16 -3.330669e-16 -4.440892e-16
## rateHigh    3.330669e-16 -3.330669e-16 -4.440892e-16
```

```
# Is this really zero?
isZero <- function(mat) {
  all.equal(mat, matrix(0, ncol = ncol(mat),
                           nrow = nrow(mat)),
            check.attributes = FALSE)
}

isZero(crossprod(Y_hat, residuals))
```

Example (cont'd) iii

```
## [1] TRUE
```

```r
isZero(crossprod(X, residuals))
```

```
## [1] TRUE
```

Maximum Likelihood Estimation i

- We now introduce distributional assumptions on $\mathbf{Y}$:

$$\mathbf{Y} \mid \mathbf{X} \sim N_p(B^T\mathbf{X}, \Sigma).$$

- This is the same conditions on the mean and covariance as above. The only difference is that we now assume the residuals are normally distributed.
- **Note**: The distribution above is conditional on $\mathbf{X}$. It could happen that the marginal distribution of $\mathbf{Y}$ is not normal.

- **Theorem**: Suppose $\mathbb{X}$ has full rank $q+1$, and assume that $n \geq q+p+1$. Then the least-squares estimator $\hat{B} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$ of B is also the *maximum likelihood estimator*. Moreover, we have
    1. $\hat{B}$ is normally distributed.
    2. The maximum likelihood estimator for Σ is $\hat{\Sigma} = \frac{1}{n}\hat{\mathbb{E}}^T\hat{\mathbb{E}}$.
    3. $n\hat{\Sigma}$ follows a Wishart distribution $W_p(n-q-1, \Sigma)$ on $n-q-1$ degrees of freedom.
    4. The maximised likelihood is $L(\hat{B}, \hat{\Sigma}) = (2\pi)^{-np/2}|\hat{\Sigma}|^{-n/2}\exp(-pn/2)$.

Maximum Likelihood Estimation iii

- **Note**: Looking at the degrees of freedom of the Wishart distribution, we can infer that $\hat{\Sigma}$ is a biased estimator of Σ. An *unbiased* estimator is

$$S = \frac{1}{n-q-1}\hat{\mathbb{E}}^T\hat{\mathbb{E}}.$$

Example i

```r
library(heplots)

head(NLSY)
```

```
##    math  read antisoc hyperact income educ
## 1 50.00 45.24       4        3 52.518   14
## 2 28.57 28.57       0        0 42.600   12
## 3 50.00 53.57       2        2 50.000   12
## 4 32.14 34.52       0        2  6.082   12
## 5 21.43 22.62       0        2  7.410   14
## 6 15.48 40.48       1        0 12.988   12
```

Example ii

```r
# Fit model and look at coefficients
fit <- lm(cbind(math, read) ~ income + educ,
          data = NLSY)

coef(fit)
```

```
##                  math        read
## (Intercept) 8.7828704 15.88479888
## income      0.0893217  0.01366238
## educ        1.2755492  0.94949980
```

Example iii

```r
range(NLSY$income)
```

```
## [1]   0.000 146.942
```

```r
range(NLSY$educ)
```

```
## [1]  6 20
```

- Suppose we have a new observation $\mathbf{X}_0$. We are interested in making predictions and inference about the corresponding outcome vector $\mathbf{Y}_0$.
- First, since $\hat{B}$ is an unbiased estimator of B, we see that

$$E(\mathbf{X}_0^T \hat{B}) = \mathbf{X}_0^T E(\hat{B}) = \mathbf{X}_0^T B = E(\mathbf{Y}_0).$$

  Therefore, it makes sense to estimate $\mathbf{Y}_0$ using $\mathbf{X}_0^T \hat{B}$.
- *What is the estimation error?* Let's look at the covariance of $\mathbf{X}_0^T \hat{\beta}_i$ and $\mathbf{X}_0^T \hat{\beta}_j$

$$\begin{aligned} \mathrm{Cov}\left(\mathbf{X}_0^T \hat{\beta}_i, \mathbf{X}_0^T \hat{\beta}_j\right) &= \mathbf{X}_0^T \mathrm{Cov}\left(\hat{\beta}_i, \hat{\beta}_j\right) \mathbf{X}_0 \\ &= \sigma_{ij} \mathbf{X}_0^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbf{X}_0. \end{aligned}$$

Confidence and Prediction Regions ii

- *What is the forecasting error?* In that case, we also need to take into account the extra variation coming from the residuals.
- In other words, we also need to sample a new "error" term $\mathbf{E}_0 = (E_{01}, \ldots, E_{0p})$ independently of $\mathbf{X}_0$.
    - Let $\tilde{\mathbf{Y}}_0 = \mathbf{X}_0^T B + \mathbf{E}_0$ be the new value.
- The **forecast error** is given by

$$\tilde{\mathbf{Y}}_0 - \mathbf{X}_0^T \hat{B} = \mathbf{E}_0 - \mathbf{X}_0^T (\hat{B} - B).$$

- Since $E(\tilde{\mathbf{Y}}_0 - \mathbf{X}_0^T \hat{B}) = 0$, we can still deduce that $\mathbf{X}_0^T \hat{B}$ is an unbiased predictor of $\mathbf{Y}_0$.

Confidence and Prediction Regions iii

- Now let's look at the covariance of the forecast errors in each component:

$$
\begin{aligned}
&E\left[\left(\tilde{Y}_{0i} - \mathbf{X}_0^T\hat{\beta}_i\right)\left(\tilde{Y}_{0j} - \mathbf{X}_0^T\hat{\beta}_j\right)\right] \\
&= E\left[\left(E_{0i} - \mathbf{X}_0^T(\hat{\beta}_i - \beta_i)\right)\left(E_{0j} - \mathbf{X}_0^T(\hat{\beta}_j - \beta_j)\right)\right] \\
&= E(E_{0i}E_{0j}) + \mathbf{X}_0^T E\left[(\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j)\right]\mathbf{X}_0 \\
&= \sigma_{ij} + \sigma_{ij}\mathbf{X}_0^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{X}_0 \\
&= \sigma_{ij}\left(1 + \mathbf{X}_0^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{X}_0\right).
\end{aligned}
$$

- Therefore, we can see that the difference between the *estimation* error and the *forecasting* error is σ_{ij}.

Example i

```r
# Recall our model for Plastic
fit <- lm(cbind(tear, gloss, opacity) ~ rate,
          data = Plastic)

new_x <- data.frame(rate = factor("High",
                                  levels = c("Low",
                                             "High")))
(prediction <- predict(fit, newdata = new_x))
```

```
##   tear gloss opacity
## 1 7.08  9.06    4.08
```

Example ii

```
X <- model.matrix(fit)
S <- crossprod(resid(fit))/(nrow(Plastic) - ncol(X))
new_x <- model.matrix(~rate, new_x)

quad_form <- drop(new_x %*% solve(crossprod(X)) %*%
                    t(new_x))

# Estimation covariance
(est_cov <- S * quad_form)
```

Example iii

```
##                   tear        gloss      opacity
## tear      0.014027778 0.003994444 -0.006083333
## gloss     0.003994444 0.021027778  0.014716667
## opacity  -0.006083333 0.014716667  0.409916667
```

```
# Forecasting covariance
(fct_cov <- S *(1 + quad_form))
```

```
##                 tear      gloss     opacity
## tear      0.15430556 0.04393889 -0.06691667
## gloss     0.04393889 0.23130556  0.16188333
## opacity  -0.06691667 0.16188333  4.50908333
```

Example iv

```
# Estimation CIs
cbind(drop(prediction) - 1.96*sqrt(diag(est_cov)),
      drop(prediction) + 1.96*sqrt(diag(est_cov)))
```

```
##               [,1]     [,2]
## tear    6.847860 7.312140
## gloss   8.775781 9.344219
## opacity 2.825115 5.334885
```

```
# Forecasting CIs
cbind(drop(prediction) - 1.96*sqrt(diag(fct_cov)),
      drop(prediction) + 1.96*sqrt(diag(fct_cov)))
```

Example v

```
##                  [,1]      [,2]
## tear       6.31007778  7.849922
## gloss      8.11735297 10.002647
## opacity   -0.08198204  8.241982
```

Likelihood Ratio Tests i

- We can use a Likelihood Ratio test to assess the evidence in support of two nested models.
- Write
$$B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \qquad \mathbb{X} = \begin{pmatrix} \mathbb{X}_1 & \mathbb{X}_2 \end{pmatrix},$$
where B_1 is $(r+1) \times p$, B_2 is $(q-r) \times p$, $\mathbb{X}_1$ is $n \times (r+1)$, $\mathbb{X}_2$ is $n \times (q-r)$, and $r \geq 0$ is a non-negative integer.

Likelihood Ratio Tests ii

- We want to compare the following models:

$$\text{Full model} : E(\mathbf{Y} \mid \mathbf{X}) = B^T\mathbf{X}$$
$$\text{Nested model} : E(\mathbf{Y} \mid \mathbf{X}_1) = B_1^T\mathbf{X}_1$$

- According to our previous theorem, the corresponding maximised likelihoods are

$$\text{Full model} : L(\hat{B}, \hat{\Sigma}) = (2\pi)^{-np/2}|\hat{\Sigma}|^{-n/2}\exp(-pn/2)$$
$$\text{Nested model} : L(\hat{B}_1, \hat{\Sigma}_1) = (2\pi)^{-np/2}|\hat{\Sigma}_1|^{-n/2}\exp(-pn/2)$$

Likelihood Ratio Tests iii

- Therefore, taking the ratio of the likelihoods of the nested model to the full model, we get

$$\Lambda = \frac{L(\hat{B}_1, \hat{\Sigma}_1)}{L(\hat{B}, \hat{\Sigma})} = \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|}\right)^{n/2}.$$

- Or equivalently, we get *Wilks' lambda statistic*:

$$\Lambda^{2/n} = \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|}.$$

- As discussed in the lecture on MANOVA, there is no closed-form solution for the distribution of this statistic under the null hypothesis $H_0 : B_2 = 0$, but there are many approximations.

Likelihood Ratio Tests iv

- Two important special cases:
  - When $r = 0$, we are testing the full model against the empty model (i.e. only the intercept).
  - When $\mathbb{X}_2$ only contains one covariate, we are testing the full model against a simpler model without that covariate. In other words, we are testing for the *significance* of that covariate.

- The Wilks' lambda statistic can actually be expressed in terms of the (generalized) eigenvalues of a pair of matrices (H, E):
    - $E = n\hat{\Sigma}$ is the **error** matrix.
    - $H = n(\hat{\Sigma}_1 - \hat{\Sigma})$ is the **hypothesis** matrix.
- Under our assumptions about the rank of $\mathbb{X}$ and the sample size, E is (almost surely) invertible, and therefore we can look at the nonzero eigenvalues of HE^{-1}:
    - Let $\eta_1 \geq \cdots \geq \eta_s$ be those nonzero eigenvalues, where $s = \min(p, q - r)$.
    - Equivalently, these eigenvalues are the nonzero roots of the determinantal equation $\det\left((\hat{\Sigma}_1 - \hat{\Sigma}) - \eta\hat{\Sigma}\right) = 0$.

Other Multivariate Test Statistics ii

- Recall the four classical multivariate test statistics:

$$\text{Wilks' lambda}: \prod_{i=1}^{s} \frac{1}{1+\eta_i} = \frac{|E|}{|E+H|}$$

$$\text{Pillai's trace}: \sum_{i=1}^{s} \frac{\eta_i}{1+\eta_i} = \mathrm{tr}\left(H(H+E)^{-1}\right)$$

$$\text{Hotelling-Lawley trace}: \sum_{i=1}^{s} \eta_i = \mathrm{tr}\left(HE^{-1}\right)$$

$$\text{Roy's largest root}: \frac{\eta_1}{1+\eta_1}$$

- Under the null hypothesis $H_0 : B_2 = 0$, all four statistics can be well-approximated using the F distribution.

Other Multivariate Test Statistics iii

- **Note**: When $r = q - 1$, all four tests are equivalent.
- In general, as the sample size increases, all four tests give similar results. For finite sample size, Roy's largest root has good power only if there the leading eigenvalue η_1 is significantly larger than the other ones.

Example i

```r
# Going back to our NLSY example
full_model <- lm(cbind(math, read) ~ income + educ +
                     antisoc + hyperact,
                   data = NLSY)

library(pander)
pander(anova(full_model, test = "Wilks"))
```

Example ii

	Df	Wilks	approx F	num Df	den Df	Pr(>F)
(Intercept)	1	0.09	1243.04	2	237	0.00
income	1	0.93	9.19	2	237	0.00
educ	1	0.95	6.57	2	237	0.00
antisoc	1	0.99	1.16	2	237	0.31
hyperact	1	0.99	1.74	2	237	0.18
Residuals	238	NA	NA	NA	NA	NA

Example iii

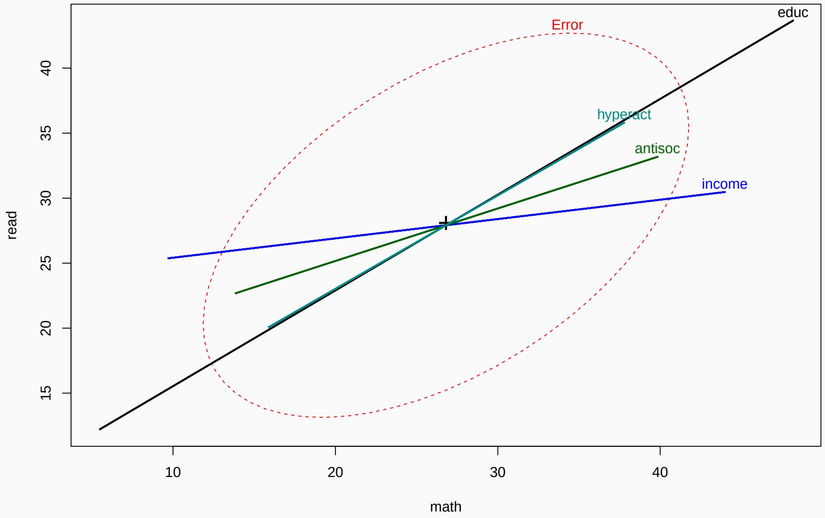
```
pander(anova(full_model, test = "Roy"))
```

	Df	Roy	approx F	num Df	den Df	Pr(>F)
(Intercept)	1	10.49	1243.04	2	237	0.00
income	1	0.08	9.19	2	237	0.00
educ	1	0.06	6.57	2	237	0.00
antisoc	1	0.01	1.16	2	237	0.31
hyperact	1	0.01	1.74	2	237	0.18
Residuals	238	NA	NA	NA	NA	NA

Example iv

```r
# Visualize the error and hypothesis ellipses
heplot(full_model)
```

Example v



Error
educ
hyperact
antisoc
income
read
math
10
20
30
40
15
20
25
30
35
40

Example vi

```r
# Fit a model with only income and educ
rest_model <- lm(cbind(math, read) ~ income + educ,
                 data = NLSY)

pander(anova(full_model, rest_model,
             test = "Wilks"))
```

Res.Df	Df	Gen.var.	Wilks	approx F	num Df	den Df	Pr(>F)
238	NA	82.87	NA	NA	NA	NA	NA
240	2	83.18	0.98	1.44	4	474	0.22

Example vii

```r
pander(anova(full_model, rest_model,
             test = "Roy"))
```

Res.Df	Df	Gen.var.	Roy	approx F	num Df	den Df	Pr(>F)
238	NA	82.87	NA	NA	NA	NA	NA
240	2	83.18	0.02	2.64	2	238	0.07

Example viii

```r
# Let's look at the eigenvalues
E <- crossprod(residuals(full_model))
H <- crossprod(residuals(rest_model)) - E

result <- eigen(H %*% solve(E),
                only.values = TRUE)
result$values[seq_len(2)]
```

```
## [1] 0.022196515 0.002277582
```

Information Criteria i

- We can use hypothesis testing for model building:
  - Add covariates that significantly improve the model (*forward selection*);
  - Remove non-significant covariates (*backward elimination*).
- Another approach is to use *Information Criteria*.
- The general form of Akaike's information criterion:

$$-2\log L(\hat{B}, \hat{\Sigma}) + 2d,$$

  where d is the number of parameters to estimate.
  - In multivariate regression, this would be $d = (q+1)p + p(p+1)/2$.

Information Criteria ii

- Therefore, we get (up to a constant):

$$AIC = n\log|\hat{\Sigma}| + 2(q+1)p + p(p+1).$$

- The intuition behind AIC is that it estimates the Kullback-Leibler divergence between the posited model and the true data-generating mechanism.
    - So smaller is better.
- Model selection using information criteria proceeds as follows:
    1. Select models of interest $\{M_1, \ldots, M_K\}$. They do not need to be nested, and they do not need to involve the same variables.
    2. Compute the AIC for each model.
    3. Select the model with the smallest AIC.

Information Criteria iii

- The set of interesting models should be selected using domain-specific knowledge when possible.
    - If it is not feasible, you can look at all possible models between the empty model and the full model.
- There are many variants of AIC, each with their own trade-offs.
    - For more details, see Timm (2002) Section 4.2.d.

Example (cont'd) i

```
## AIC(full_model)
# Error in logLik.lm(full_model) :
#   'logLik.lm' does not support multiple responses
class(full_model)
```

```
## [1] "mlm" "lm"
```

Example (cont'd) ii

```
logLik.mlm <- function(object, ...) {
  resids <- residuals(object)
  Sigma_ML <- crossprod(resids)/nrow(resids)
  ans <- sum(mvtnorm::dmvnorm(resids, log = TRUE,
                                sigma = Sigma_ML))
  df <- prod(dim(coef(object))) +
    choose(ncol(Sigma_ML) + 1, 2)
  attr(ans, "df") <- df
  class(ans) <- "logLik"
  return(ans)
}
```

Example (cont'd) iii

```r
logLik(full_model)
```

```
## 'log Lik.' -1757.947 (df=13)
```

```r
AIC(full_model)
```

```
## [1] 3541.894
```

```r
AIC(rest_model)
```

```
## [1] 3539.781
```

Example of model selection i

```r
# Model selection for Plastic data
lhs <- "cbind(tear, gloss, opacity) ~"
rhs_form <- c("1", "rate", "additive",
              "rate+additive", "rate*additive")

purrr::map_df(rhs_form, function(rhs) {
  form <- formula(paste(lhs, rhs))
  fit <- lm(form, data = Plastic)
  return(data.frame(model = rhs, aic = AIC(fit),
                    stringsAsFactors = FALSE))
})
```

Example of model selection ii

```
##              model      aic
## 1                1 155.4330
## 2             rate 143.7768
## 3         additive 150.9542
## 4 rate+additive 137.9592
## 5 rate*additive 138.9157
```

Multivariate Influence Measures i

- Earlier we introduced the projection matrix

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

  and we noted that $\hat{\mathbb{Y}} = P\mathbb{Y}$.
- Looking at one row at a time, we can see that

$$\begin{aligned}\hat{\mathbf{Y}}_i &= \sum_{j=1}^n P_{ij}\mathbf{Y}_j \\ &= P_{ii}\mathbf{Y}_i + \sum_{j\neq i} P_{ij}\mathbf{Y}_i,\end{aligned}$$

  where P_{ij} is the (i, j)-th entry of P.

Multivariate Influence Measures ii

- In other words, the diagonal element P_{ii} represents the *leverage* (or influence) of observation $\mathbf{Y}_i$ on the fitted value $\hat{\mathbf{Y}}_i$.
  - Observation $\mathbf{Y}_i$ is said to have a **high leverage** if P_{ii} is large compared to the other element on the diagonal.
- Let $S = \frac{1}{n-q-1}\hat{\mathbb{E}}^T\hat{\mathbb{E}}$ be the unbiased estimator of Σ, and let $\hat{\mathbf{E}}_i$ be the i-th row of $\hat{\mathbb{E}}$.
- We define the multivariate **internally Studentized residuals** as follows:

$$r_i = \frac{\hat{\mathbf{E}}_i^T S^{-1}\hat{\mathbf{E}}_i}{1 - P_{ii}}.$$

- If we let $S_{(i)}$ be the estimator of Σ where we have removed row i from the residual matrix $\hat{\mathbb{E}}$, we define the multivariate **externally Studentized residuals** as follows:

$$T_i^2 = \frac{\hat{\mathbf{E}}_i^T S_{(i)}^{-1} \hat{\mathbf{E}}_i}{1 - P_{ii}}.$$

- An observation $\mathbf{Y}_i$ may be considered a potential outlier if

$$\left(\frac{n - q - p - 1}{p(n - q - 2)}\right) T_i^2 > F_\alpha(p, n - q - 2).$$

- Yet another measure of influence is the multivariate **Cook's distance**.

$$C_i = \frac{P_{ii}}{(1 - P_{ii})^2} \hat{\mathbf{E}}_i^T S^{-1} \hat{\mathbf{E}}_i / (q + 1).$$

- An observation $\mathbf{Y}_i$ may be considered a potential outlier if C_i is larger than the median of a chi square distribution with $\nu = p(n - q - 1)$ degrees of freedom.

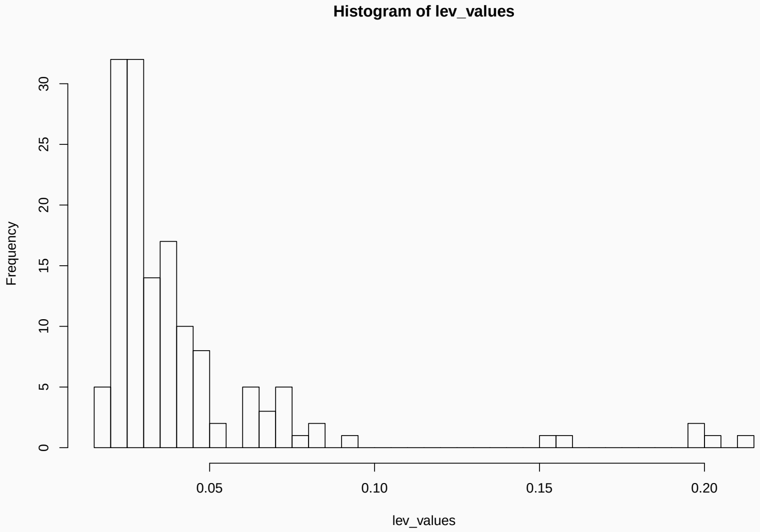
Example i

```r
library(openintro)
model <- lm(cbind(startPr, totalPr) ~
              nBids + cond + sellerRate +
              wheels + stockPhoto,
            data = marioKart)

X <- model.matrix(model)
P <- X %*% solve(crossprod(X)) %*% t(X)
lev_values <- diag(P)

hist(lev_values, 50)
```

Example ii


Histogram of lev_values
Frequency
0
5
10
15
20
25
30
0.05
0.10
0.15
0.20
lev_values

Example iii

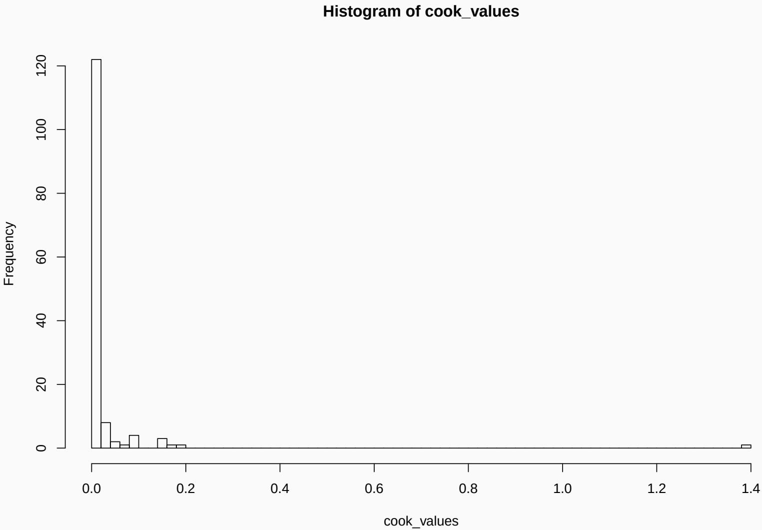
```r
n <- nrow(marioKart)
resids <- residuals(model)
S <- crossprod(resids)/(n - ncol(X))

S_inv <- solve(S)

const <- lev_values/((1 - lev_values)^2*ncol(X))
cook_values <- const * diag(resids %*% S_inv
                                  %*% t(resids))

hist(cook_values, 50)
```

Example iv



Histogram of cook_values
Frequency
0
20
40
60
80
100
120
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
cook_values

Example v

```r
# Cut-off value
(cutoff <- qchisq(0.5, ncol(S)*(n - ncol(X))))
```

```
## [1] 273.3336
```

```r
which(cook_values > cutoff)
```

```
## named integer(0)
```

Strategy for Multivariate Model Building

1. Try to identify outliers.
    - This should be done graphically at first.
    - Once the model is fitted, you can also look at influence measures.
2. Perform a multivariate test of hypothesis.
3. If there is evidence of a multivariate difference, calculate Bonferroni confidence intervals and investigate component-wise differences.
    - The projection of the confidence region onto each variable generally leads to confidence intervals that are too large.

- Recall from our lecture on MANOVA: assume the data comes from g populations:

$$\begin{matrix} \mathbf{Y}_{11}, & \dots, & \mathbf{Y}_{1n_1} \\ \vdots & \ddots & \vdots \\ \mathbf{Y}_{g1}, & \dots, & \mathbf{Y}_{gn_g} \end{matrix},$$

where $\mathbf{Y}_{\ell 1}, \dots, \mathbf{Y}_{\ell n_\ell} \sim N_p(\mu_\ell, \Sigma)$.

- We obtain an equivalent model if we set

$$\mathbb{Y} = \begin{pmatrix} \mathbf{Y}_{11} \\ \vdots \\ \mathbf{Y}_{1n_1} \\ \vdots \\ \mathbf{Y}_{g1} \\ \vdots \\ \mathbf{Y}_{gn_g} \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

- Here, $\mathbb{Y}$ is $n \times p$ and $\mathbb{X}$ is $n \times g$.
  - The first column of $\mathbb{X}$ is all ones.
  - The $(i, \ell + 1)$ entry of $\mathbb{X}$ is 1 iff the i-th row belongs to the ℓ-th group.
  - **Note**: It is common to have a different constraint on the parameters τ_ℓ in regression; here, we assume that $\tau_g = 0$.
- In other words, we model group membership using a single categorial covariate and therefore $g - 1$ dummy variables.
- More complicated designs for MANOVA can also be expressed in terms of linear regression:

Multivariate Regression and MANOVA iv

- For example, for two-way MANOVA, we would have two categorical variables. We would also need to include an interaction term to get all combinations of the two treatments.
- In general, fractional factorial designs can be expressed as a linear regression with a carefully selected series of dummy variables.