Advanced Multivariate Regression

Jerome Dumortier

ANOVA
Theoretical Concepts
One-Way ANOVA
Two-Way ANOVA

Model Specification
Exclusion of Relevant Variables

Regression Diagnostics

Advanced Multivariate Regression

Jerome Dumortier

Overview of Topics Covered

Analysis of Variance (ANOVA)

- One-way and two-way ANOVA

Model specification

- Exclusion of relevant variables
- Inclusion of irrelevant variables

Regression diagnostics

- Basic diagnostics plots included in R

Overview

Application to public policy

- Difference in average outcomes based on different policy regimes
- Difference in average outcomes based on different regions and policy regimes
- Explanation of outcome variation due to independent variables and residual noise

Decomposition of the variation related to previous concept

- Total variation as the sum of explained and unexplained variation

Regression model with only dummy variables

$$y_i = \beta_0 + \sum_{k=1}^{K-1} \beta_k \cdot D_{ik} + \epsilon_i$$

Introductory Example

Differences in average household electricity bills across utility ownership types

- Dependent variable: *bill* (monthly electricity bill)
- Independent (group) variable: *ownership* (i.e., public, investor-owned, cooperative)

Regression model

$$bill_i = \beta_0 + \beta_1 \cdot D_{i,public} + \beta_2 \cdot D_{i,\text{cooperative}} + \epsilon_i$$

Reference group: Investor-owned utilities

- $D_{i,public}$: Dummy for investor-owned utility
- $D_{i,cooperative}$: Dummy for cooperative

Hypothesis and Interpretation

Hypothesis

$$H_0 : \mu_{public} = \mu_{investor-owned} = \mu_{cooperative}$$

Interpretation

- β_0: Mean bill for investor-owned utilities
- β_1: Mean difference between investor-owned and public utilities
- β_2: mean difference between investor-owned and cooperative utilities

Translation in to coefficient estimates: β_1 and β_0 not statistically significant

One-Way ANOVA for vehicles: Setup

```
df          = subset(vehicles,year==2023,
                     select=c("comb08","drive","vclass"))
df          = na.omit(df)
bhat        = lm(comb08~factor(vclass),data=df)
anova(bhat)
```

```
## Analysis of Variance Table
##
## Response: comb08
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## factor(vclass)   10  60648  6064.8  12.974 < 2.2e-16 ***
## Residuals      1095 511876   467.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One-Way ANOVA for `vehicles`: Interpretation I

Degrees of freedom

- 10 for vehicle class: Comparison across 11 vehicle classes
- 1095 for residuals: Remaining unexplained variation

Between-class variation

- Differences in average fuel economy across vehicle classes
- Variation due to class design (e.g., compact cars versus SUVs)

Within-class noise

- Differences in fuel economy among vehicles within the same class due to engine size, weight, or model-specific features

One-Way ANOVA for `vehicles`: Interpretation II

Mean Square

- Vehicle class: 6,064.8
- Residual: 467.5

Interpretation

- Differences across vehicle-class averages are about 13 times larger than typical within-class variation
- Class membership explains a meaningful share of fuel economy differences

F-statistic (Ratio of between-class variance to within-class variance): 12.97

- Indication of strong signal relative to residual variability
- p-value: $<$ 2.2e-16: Rejection of H_0 of equal mean fuel economy across vehicle classes

Introductory Example

Extension of the one-way ANOVA

- Addition of one more grouping variable
- Factorial ANOVA: Extension to more than two groups

Policy application

- Differences in student test scores across school type and location
- Assessment of whether school-type effects differ by location

Example

- Dependent variable: *score*
- Group 1: *schooltype* (i.e., public, charter)
- Group 2: *location* (i.e., urban, rural)

Model Setup

Regression model

$$score = \beta_0 + \beta_1 \cdot D_{charter} + \beta_2 \cdot D_{urban} + \beta_3 \cdot (D_{charter} \cdot D_{urban}) + \epsilon$$

Coefficients

- β_1: School-type effects
- β_2: Location effects
- β_3: Interaction effects

Disentanglement of the effects by defining a reference group and interpreting everything as deviations from that baseline

Enumeration of possible combinations

School type	Location	Charter	Urban	Charter $\times$ Urban
Public	Rural	0	0	0
Charter	Rural	1	0	0
Public	Urban	0	1	0
Charter	Urban	1	1	1

Interpretation

Baseline

- Mean score for public rural schools (Public and rural): β_0

Scenarios

- Charter–public difference in rural areas (charter and rural): $\beta_0 + \beta_1$
- Urban–rural difference for public schools (public and urban): $\beta_0 + \beta_2$
- Additional charter effect in urban areas (charter and urban): $\beta_0 + \beta_1 + \beta_2 + \beta_3$

Two-Way ANOVA for ToothGrowth

Context

- Tooth growth in 60 guinea pigs based on three levels of vitamin C doses, i.e., 0.5, 1, and 2 mg/day, and two delivery methods, i.e., orange juice (OJ) or ascorbic acid (a form of vitamin C coded as VC)

```
data(ToothGrowth)
ToothGrowth$dose        = factor(ToothGrowth$dose)
bhat1                   = lm(len~supp+dose,data=ToothGrowth)
bhat2                   = lm(len~supp*dose,data=ToothGrowth)
# '*' expands to supp + dose + supp:dose
```

Results Model 1

```
## Analysis of Variance Table
##
## Response: len
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## supp       1  205.35  205.35  14.017 0.0004293 ***
## dose       2 2426.43 1213.22  82.811 < 2.2e-16 ***
## Residuals 56  820.43   14.65
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results Model 2

```
## Analysis of Variance Table
##
## Response: len
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## supp         1  205.35  205.35  15.572 0.0002312 ***
## dose         2 2426.43 1213.22  92.000 < 2.2e-16 ***
## supp:dose    2  108.32   54.16   4.107 0.0218603 *
## Residuals 54  712.11   13.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exclusion of Relevant Variables

Correct model

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \epsilon$$

Estimated model

$$y = \beta_0 + \beta_1 \cdot x_1 + \epsilon$$

Question: Is the estimate of β_1 biased, i.e., incorrect?

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \cdot \frac{Cov(x_1, x_2)}{Var(x_1)}$$

The estimate of β_1 is correct only if x_1 and x_2 are uncorrelated.

- Simulated population parameters: $beta_0 = 50$, $beta_1 = 4$, and $beta_2 = 5$

Estimation Setup

Specification 1: Low correlation between x_1 and x_2

```
df1         = subset(specification,group=="Specification 1")
bhat1       = lm(y~x1+x2,data=df1)
bhat2       = lm(y~x1,data=df1)
covvar1     = c(cov(df1$x1,df1$x2),var(df1$x1))
```

Specification 2: High correlation between x_1 and x_2

```
df2         = subset(specification,group=="Specification 2")
bhat3       = lm(y~x1+x2,data=df2)
bhat4       = lm(y~x1,data=df2)
covvar2     = c(cov(df2$x1,df2$x2),var(df2$x1))
```

Results: Low Correlation

```
##
## ========================================================================
##                                   Dependent variable:
##                     ----------------------------------------------------
##                                              y
##                              (1)                        (2)
## ------------------------------------------------------------------------
## x1                        4.034***                   3.893***
##                            (0.015)                    (0.229)
## x2                        5.020***
##                            (0.015)
## Constant                  47.076***                 298.841***
##                            (1.179)                   (13.290)
## ------------------------------------------------------------------------
## Observations                 499                        499
## R2                          0.997                      0.368
## Adjusted R2                 0.997                      0.367
## Residual Std. Error     9.997 (df = 496)          148.070 (df = 497)
## F Statistic         86,063.530*** (df = 2; 496) 289.879*** (df = 1; 497)
## ========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

Results: High Correlation

```
##
## ===========================================================================
##                                      Dependent variable:
##                         ---------------------------------------------------
##                                                   y
##                                   (1)                        (2)
## ---------------------------------------------------------------------------
## x1                              4.027***                   6.914***
##                                 (0.019)                    (0.188)
## x2                              5.012***
##                                 (0.019)
## Constant                       47.859***                  146.668***
##                                 (0.970)                    (10.932)
## ---------------------------------------------------------------------------
## Observations                      499                        499
## R2                               0.998                      0.731
## Adjusted R2                      0.998                      0.731
## Residual Std. Error        10.011 (df = 496)          121.791 (df = 497)
## F Statistic           136,533.800*** (df = 2; 496) 1,351.303*** (df = 1; 497)
## ===========================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

Bias Calculation

Recall bias calculation

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \cdot \frac{Cov(x_1, x_2)}{Var(x_1)}$$

```
4+5*covvar1[1]/covvar1[2]
```

```
## [1] 3.860031
```

```
4+5*covvar2[1]/covvar2[2]
```

```
## [1] 6.880798
```

Inclusion of Irrelevant Variables

Correct model

$$y = \beta_0 + \beta_1 \cdot x_1 + \epsilon$$

Estimated model

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \epsilon$$

Effects on β_1

- Correct estimation of β_1 but inflated variance

Overview

Assessment of model assumptions and reliability of inference

- Incorrect coefficient estimates leading to potentially errounous policy implementation

Diagnostics

- Linearity and functional form
- Residual analysis
- Normality of errors
- Influential observations and outliers
- Specification errors

Different chapter for more violating of assumptions

Diagnostics Plots in R

Estimating the impact of median household income on student scores in Ohio

```
ohioschools     = merge(ohioscore,ohioincome,by=c("irn"))
bhat1           = lm(score~medianincome,data=ohioschools)
bhat2           = lm(score~medianincome+I(medianincome^2),
                     data=ohioschools)
```

```
plot(bhat1)
plot(bhat2)
```

Residuals vs Fitted

Purpose and desired pattern

- Check linearity and overall fit
- Random scatter around zero
- No clear shape

Problematic pattern

- Curvature: Missing nonlinear term
- Funnel shape: Heteroskedasticity
- Clusters: Omitted variable

Normal Q–Q Plot

Purpose

- Check for normality of residuals
- Comparison between residual quantiles and normal quantiles

Desired pattern

- Points approximately on straight line

Problematic pattern

- Tail deviations: Outliers
- Systematic curvature: Skewness or heavy tails

Scale–Location Plot

Purpose and desired pattern

- Check constant variance (homoskedasticity)
- Horizontal band
- Even spread across fitted values

Problematic pattern

- Upward slope: Increasing variance
- Uneven spread: Heteroskedasticity

Residuals vs Leverage

Purpose: Detection of influential observations

- Inclusion of Cook's distance contours

Problematic patters

- High leverage points
- Large residuals
- Points outside Cook's distance lines

Potential for results driven by a few extreme observations