Multivariate Regression

POLS 095 | Drake University

Multivariate regression: What is it?

- The properties of linear regression are complex. But we can simplify them.
- Linear regression is a tool for understanding a phenomenon of interest as a linear function of some combination of predictor variables.
- Similar to $y = mx + b$, but we prefer a more malleable form

The regression equation

- Y = $\alpha + \ \beta_1 * X_1 + \beta_2 * X_2 +* \beta_3 * X_3 + \varepsilon$
  - Y = Dependent Variable
  - X = Independent Variable(s)
    - And/or control variables
  - α = Intercept
  - β = Coefficient (Effect)
    - What to look for: Is it positive or negative?
  - ε = Error (sometimes called the “residual”)

An Example

- Y = $\alpha + \beta_1 * X_1 + \beta_2 * X_2 +* \beta_3 * X_3 + \varepsilon$
- $fedspend\_scale$ = $\alpha + \beta_1 * ft\_dem + \beta_2 * polknow_{combined} +* \beta_3 * gender + \varepsilon$
- fedspend_scale = numeric scale of how many programs a respondent wants to increase/decrease spending on
  - Ranges from zero to 16
- ft_dem = feeling thermometer towards Democrats
  - Ranges from 0 to 100
- polknow_combined = numeric scale of how many politicians a respondent could correctly identify
  - Ranges from zero to 8
- gender = Codes for sex (1 = female, zero = male)

The Data (the first 6 lines)

Respondent	fedspend_scale	ft_dem	polknow_combined	sex
1	9	30	5	Male (0)
2	12	95	3	Female (1)
3	5	60	4	Female (1)
4	6	35	6	Female (1)
5	11	70	NA	Male (0)
6	6	85	5	Male (0)

Bivariate Relationship

fedspend_scale = α+ β_1**ft_dem* +ε

```
> mod.a <- svyglm(fedspend_scale ~ ft_dem, design=nesD, na.action="na.omit")
> summary(mod.a)

Call:
svyglm(formula = ..1, design = ..2, na.action = "na.omit")

Survey design:
survey::svydesign(id = ~1, data = nes, weights = ~wt)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.609057   0.122382   54.00   <2e-16 ***
ft_dem      0.063402   0.001944   32.61   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 8.042623)

Number of Fisher Scoring iterations: 2

> fit.svyglm(mod.a)
          R-Squared      Adjusted R-Squared
              0.291                   0.291
```

- Dependent variable:
  - Spending Preferences (0 to 16)
- Independent variable:
  - Feeling thermometer towards Democrats (0 to 100)
- What is the intercept?
  - α =
- What is the effect of feelings towards Democrats on spending preferences?
  - β_1 =
- How much of the variance in spending preferences does the feeling thermometer explain?

Bivariate Relationship

fedspend_scale $= \alpha + \beta_1 * ft\_dem + \varepsilon$

- $\beta_1 = 0.06$, $\alpha = 6.61$

  β = effect α = *intercept*

- *R1:* $6.61 + 0.06 * 30 + \varepsilon = 8.41$
- *R2:* $6.61 + 0.06 * 95 + \varepsilon = 12.31$
- *R3:* $5 = 6.61 + 0.06 * 60 + \varepsilon = 10.21$
- *R4:* $6 = 6.61 + 0.06 * 35 + \varepsilon = 8.71$
- *R5:* $11 = 6.61 + 0.06 * 70 + \varepsilon = 10.81$
- *R6:* $6 = 6.61 + 0.06 * 85 + \varepsilon = 11.71$

DATA		
Respondent	fedspend_scale	ft_dem
1	9	30
2	12	95
3	5	60
4	6	35
5	11	70
6	6	85

Bivariate Relationship

fedspend_scale = $\alpha + \beta_1 * ft\_dem + \varepsilon$

- $\beta_1 = 0.06$, $\alpha = 6.61$

  β = effect *α = intercept*

- *R1: 6.61 + 0.06*30 + ε = 8.41 + ε = 9*
- *R2: 6.61 + 0.06*95 + ε = 12.31 + ε = 12*
- *R3: 5 = 6.61 + 0.06*60 + ε = 10.21 + ε = 5*
- *R4: 6 = 6.61 + 0.06*35 + ε = 8.71 + ε = 6*
- *R5: 11 = 6.61 + 0.06*70 + ε = 10.81 + ε = 11*
- *R6: 6 = 6.61 + 0.06*85 + ε = 11.71 + ε = 6*

DATA		
Respondent	fedspend_scale	ft_dem
1	9	30
2	12	95
3	5	60
4	6	35
5	11	70
6	6	85

Bivariate Relationship

fedspend_scale = α+ β_1**ft_dem* + ε

- $\beta_1 = 0.06$, $\alpha = 6.61$

  β = effect α = *intercept*

- *R1:* $6.61 + 0.06 * 30 + \varepsilon = 8.41 + \varepsilon = 9$
- *R2:* $6.61 + 0.06 * 95 + \varepsilon = 12.31 + \varepsilon = 12$
- *R3:* $5 = 6.61 + 0.06 * 60 + \varepsilon = 10.21 + \varepsilon = 5$
- *R4:* $6 = 6.61 + 0.06 * 35 + \varepsilon = 8.71 + \varepsilon = 6$
- *R5:* $11 = 6.61 + 0.06 * 70 + \varepsilon = 10.81 + \varepsilon = 11$
- *R6:* $6 = 6.61 + 0.06 * 85 + \varepsilon = 11.71 + \varepsilon = 6$

DATA

Respondent	Fedspend _scale	Effect Explained	Error
1	9	8.41	0.59
2	12	12.31	0.31
3	5	10.21	5.21
4	6	8.71	2.71
5	11	10.81	0.19
6	6	11.71	5.71

Is that good?

- The Adj. R^2 tells us we've explained 29% of the variance in federal spending preferences based on feelings towards the Democratic Party
- But we also want to control for other factors that might affect spending preferences
- And reduce our error
- Let's say there's reason to believe that knowledgeable people will have higher spending preferences because they're more aware of government problems
- So we should add a control for political knowledge:
  - polknow_combined
  - Ranges from zero to 8

Multivariate Relationship
$fedspend\_scale = \alpha + \beta_1 * ft\_dem + \beta_2 * polknow\_combined + \varepsilon$

```
> mod.b <- svyglm(fedspend_scale ~ ft_dem + polknow_combined, design=nesD, na.action="na.omit")
> summary(mod.b)

Call:
svyglm(formula = ..1, design = ..2, na.action = "na.omit")

Survey design:
survey::svydesign(id = ~1, data = nes, weights = ~wt)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       7.325183   0.257525  28.445  < 2e-16 ***
ft_dem            0.062086   0.002363  26.269  < 2e-16 ***
polknow_combined -0.180406   0.040241  -4.483 7.57e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 8.100271)

Number of Fisher Scoring iterations: 2

> fit.svyglm(mod.b)
       R-Squared       Adjusted R-Squared
           0.308                    0.308
```

- What is the intercept?
  - α =
- What is the effect of feelings towards Democrats on spending preferences?
  - β_1 =
- What is the effect of political knowledge on spending preferences?
  - β_2 =
- How much of the variance in spending preferences does the feeling thermometer explain?

Multivariate Relationship

$fedspend\_scale = \alpha + \beta_1 * ft\_dem + \beta_2 * polknow\_combined + \beta_3 * gender + \varepsilon$

```
> mod.c <- svyglm(fedspend_scale ~ ft_dem + polknow_combined + gender, design=nesD, na.action="na.omit")
> summary(mod.c)

Call:
svyglm(formula = ..1, design = ..2, na.action = "na.omit")

Survey design:
survey::svydesign(id = ~1, data = nes, weights = ~wt)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       7.256750   0.270424  26.835  < 2e-16 ***
ft_dem            0.061959   0.002371  26.136  < 2e-16 ***
polknow_combined -0.175229   0.040729  -4.302 1.73e-05 ***
genderFemale      0.102194   0.126426   0.808    0.419
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 8.097754)

Number of Fisher Scoring iterations: 2

> fit.svyglm(mod.c)
              R-Squared      Adjusted R-Squared
                  0.309                   0.308
```

- What is the intercept?
  - α =
- What is the effect of feelings towards Democrats on spending preferences?
  - β_1 =
- What is the effect of political knowledge on spending preferences?
  - β_2 =
- What is the effect for females on spending preferences?
  - β_3 =
- How much of the variance in spending preferences does the feeling thermometer explain?

Multivariate Relationship

$$fedspend\_scale = \alpha + \beta_1 * ft\_dem + \beta_2 * polknow\_combined + \beta_3 * gender + \beta_4 * pid\_3 + \varepsilon$$

```
> mod.d <- svyglm(fedspend_scale ~ ft_dem + polknow_combined + gender + as.numeric
(pid_3), design=nesD, na.action="na.omit")
> summary(mod.d)

Call:
svyglm(formula = ..1, design = ..2, na.action = "na.omit")

Survey design:
survey::svydesign(id = ~1, data = nes, weights = ~wt)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)         8.607208   0.420126  20.487  < 2e-16 ***
ft_dem              0.053465   0.003229  16.558  < 2e-16 ***
polknow_combined   -0.179636   0.040809  -4.402 1.10e-05 ***
genderFemale        0.095423   0.125862   0.758    0.448    
as.numeric(pid_3) -0.466212   0.104396  -4.466 8.21e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 8.023211)

Number of Fisher Scoring iterations: 2

> fit.svyglm(mod.d)
         R-squared     Adjusted R-squared
             0.315                  0.315
```

- What is the intercept?
  - α =
- What is the effect of feelings towards Democrats on spending preferences?
  - β_1 =
- What is the effect of political knowledge on spending preferences?
  - β_2 =
- What is the effect for females on spending preferences?
  - β_3 =
- What is the effect of partisanship on spending preferences?
  - β_4 =
- How much of the variance in spending preferences does the feeling thermometer explain?

A nicer R table

```
> mod.d <- svyglm(fedspend_scale ~ ft_dem + polknow_combined + gender + as.numeric
(pid_3), design=nesD, na.action="na.omit")
> summary(mod.d)

Call:
svyglm(formula = ..1, design = ..2, na.action = "na.omit")

Survey design:
survey::svydesign(id = ~1, data = nes, weights = ~wt)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        8.607208   0.420126  20.487  < 2e-16 ***
ft_dem             0.053465   0.003229  16.558  < 2e-16 ***
polknow_combined  -0.179636   0.040809  -4.402 1.10e-05 ***
genderFemale       0.095423   0.125862   0.758    0.448    
as.numeric(pid_3) -0.466212   0.104396  -4.466 8.21e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 8.023211)

Number of Fisher Scoring iterations: 2

> fit.svyglm(mod.d)
              R-Squared      Adjusted R-Squared
                  0.315                   0.315
```

```
> library(broom)
> tidy.mod.d <- tidy(mod.d)
> tidy.mod.d
# A tibble: 5 x 5
  term              estimate std.error statistic  p.value
  <chr>                <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)         8.61     0.420      20.5   1.65e-88
2 ft_dem              0.0535   0.00323    16.6   1.72e-59
3 polknow_combined   -0.180    0.0408     -4.40  1.10e- 5
4 genderFemale        0.0954   0.126       0.758 4.48e- 1
5 as.numeric(pid_3)  -0.466    0.104      -4.47  8.21e- 6
```

A nicer R table

```
> library(broom)
> tidy.mod.d <- tidy(mod.d)
> tidy.mod.d
# A tibble: 5 x 5
  term                estimate std.error statistic  p.value
  <chr>                  <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)           8.61     0.420       20.5   1.65e-88
2 ft_dem                0.0535   0.00323     16.6   1.72e-59
3 polknow_combined     -0.180    0.0408      -4.40  1.10e- 5
4 genderFemale          0.0954   0.126        0.758 4.48e- 1
5 as.numeric(pid_3)    -0.466    0.104       -4.47  8.21e- 6
```

Variable	Coef.	Std. Err.	p-value
Feeling Thermometer: Dems	0.05	(0.00)	0.00
Political Knowledge	-0.18	(0.04)	0.00
Gender	0.10	(0.13)	0.45
Party Identification	-0.47	(0.10)	0.00
Intercept	8.61	(0.42)	0.00
Adjusted R^2	0.32		