

STAT 213 Multiple Regression I

Colin Reimer Dawson

Oberlin College

February 23, 2018

1 / 23

Notes

Outline

The Multiple Regression Model

CHOOSE step

FIT step

2 / 23

Notes

Outline

The Multiple Regression Model

CHOOSE step

FIT step

3 / 23

Notes

Quantitative Vs. Categorical Predictor and Response

		Response	
		Quantitative	Categorical
Predictor(s)	Quantitative	SLR	Logistic Reg.
	Categorical	ANOVA	
	Multiple	Multiple Reg.	

4 / 23

Notes

The Multiple Regression Model

DATA = PATTERN + IDIOSYNCRACIES

The Multiple Regression "Population" Model

$$Y_i = f(X_{i1}, \dots, X_{iK}) + \varepsilon_i$$

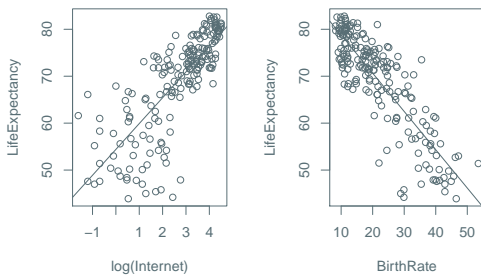
$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK} + \varepsilon_i$$

where i picks out cases, k picks out variables. X_{ik} is the k th predictor for case i . One β_k for each predictor X_k

5 / 23

Notes

Example: Life Expectancy



We could fit separate regression models for each predictor...

6 / 23

Notes

SLR Model Using log(Internet)

```
model1 <- lm(LifeExpectancy ~ log(Internet), data = AllCountries)
summary(model1) %>% coefficients() %>% round(digits = 2)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	54.40	0.95	57.42	0
log(Internet)	5.51	0.32	17.47	0

$$\widehat{\text{LifeExpectancy}}_i = 54.40 + 5.51 \cdot \log(\text{Internet}_i)$$

and

$$\text{LifeExpectancy}_i = 54.40 + 5.51 \cdot \log(\text{Internet}_i) + \varepsilon_i$$

7 / 23

Notes

SLR Model Using BirthRate

```
model2 <- lm(LifeExpectancy ~ BirthRate, data = AllCountries)
summary(model2) %>% coefficients() %>% round(digits = 2)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	86.95	0.88	98.43	0
BirthRate	-0.82	0.04	-22.64	0

$$\widehat{\text{LifeExpectancy}}_i = 86.95 - 0.82 \cdot \text{BirthRate}_i$$

and

$$\text{LifeExpectancy}_i = 86.95 - 0.82 \cdot \text{BirthRate}_i + \varepsilon_i$$

8 / 23

Notes

Or Fit Both at Once

```
model3 <- lm(LifeExpectancy ~ log(Internet) + BirthRate, data = AllCountries)
summary(model3) %>% coefficients() %>% round(digits = 2)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	77.23	2.33	33.15	0
log(Internet)	1.95	0.43	4.59	0
BirthRate	-0.61	0.06	-10.36	0

$$\widehat{\text{LifeExpectancy}}_i = 77.23 + 1.95 \log(\text{Internet}_i) - 0.61 \cdot \text{BirthRate}_i$$

and

$$\text{LifeExpectancy}_i = 77.23 + 1.95 \log(\text{Internet}_i) - 0.61 \cdot \text{BirthRate}_i + \varepsilon_i$$

9 / 23

Notes

Single Multiple Regression vs. Multiple Single Regressions

- What is the added value in using both predictors together, vs. fitting each separately?
- There are disadvantages:
 - Harder to interpret
 - Can't easily plot
- Advantages...
 - If we actually know both predictors, we get a single prediction, instead of two conflicting ones
 - Can "control for" one predictor and test the other

10 / 23

Notes

What is a Case?

- Q: What does a single case consist of for a multiple regression model?
- A: A complete case has a value for Y , and for *each* X .

11 / 23

Notes

The Four-Step Process: Multiple Regression

1. CHOOSE a form of the model
 - Select predictors
 - Choose any transformations of predictors
2. FIT: Estimate
 - coefficients: $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$
 - residual variance $\hat{\sigma}_e^2$
3. ASSESS the fit
 - Examine residuals (may need to return to step 1)
 - Test individual predictors (t -tests)
 - Test/measure overall fit (ANOVA, R^2)
 - Model comparison/selection
4. USE the model
 - Make predictions
 - Construct CIs and PIs

12 / 23

Notes

Outline

The Multiple Regression Model

CHOOSE step

FIT step

13 / 23

Notes

CHOOSE: Active Pulse Rate

```
library(Stat2Data); data("Pulse")
head(Pulse, n = 3)
```

	Active	Rest	Smoke	Gender	Exercise	Hgt	Wgt
1	97	78	0	1	1	63	119
2	82	68	1	0	3	70	225
3	88	62	0	0	3	72	175

$$\text{Active}_i = \beta_0 + \beta_1 \cdot \text{Rest}_i + \beta_2 \cdot \text{Hgt}_i + \beta_3 \cdot \text{Wgt}_i + \varepsilon_i$$

14 / 23

Notes

Outline

The Multiple Regression Model

CHOOSE step

FIT step

15 / 23

Notes

FIT: Estimate Coefficients

The Multiple Regression Population Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK} + \varepsilon_i$$

The Multiple Regression Fitted Model

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_K X_{iK} + \hat{\varepsilon}_i$$

How to choose $\hat{\beta}_k$ s? Minimize SSE! (Requires linear algebra / vector calculus)

16 / 23

Notes

FIT: Estimate Coefficients

```
my.model <- lm(Active ~ Rest + Hgt + Wgt, data = Pulse)
coef(my.model) %>% round(digits = 1)
```

(Intercept)	Rest	Hgt	Wgt
57.3	1.1	-0.9	0.1

$$\text{Active}_i = 57.3 + 1.1 \cdot \text{Rest}_i - 0.9 \cdot \text{Hgt}_i + 0.1 \cdot \text{Wgt}_i + \varepsilon_i$$

17 / 23

Notes

FIT: Estimate Residual Variance

Recall Variance Decomposition for Regression:

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2$$

$$SS_{Total} = SS_{Model} + SS_{Error}$$

Recall ANOVA Table:

$$MS_{Model} = SS_{Model} / df_{Model}$$

$$MS_{Error} = SS_{Error} / df_{Error}$$

where MS_{Error} represents $\hat{\sigma}_\varepsilon^2$. So... what are df_{Model} and df_{Error} ?

18 / 23

Notes

Regression Degrees of Freedom

$df_{Model} = K$ where K is the number of predictors

This is the number of extra “free parameters” (compared to the null model)

$df_{Error} = N - K - 1$ where N is the sample size

This is the number of “pieces of information” we have about the sizes of the residuals. (Can fit any N points exactly with N coefficients including the intercept.)

Notes

FIT: Estimate Residual Variance

$$\hat{\sigma}_\varepsilon^2 = MS_{Error} = \frac{SS_{Error}}{df_{Error}} = \frac{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2}{N - K - 1}$$

Notes

FIT: Estimate Residual Variance

```
summary(my.model)

Call:
lm(formula = Active ~ Rest + Hgt + Wgt, data = Pulse)

Residuals:
    Min       1Q   Median       3Q      Max
-35.245  -9.968  -2.458   6.697  64.716

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.25991    25.01347   2.289  0.0230 **
Rest          1.12647     0.10162  11.086 <2e-16 ***
Hgt          -0.88060     0.40535  -2.172  0.0309 **
Wgt           0.10855     0.04699   2.310  0.0218 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.91 on 228 degrees of freedom
Multiple R-squared:  0.3808, Adjusted R-squared:  0.3726
F-statistic: 46.73 on 3 and 228 DF,  p-value: < 2.2e-16
```

Notes

FIT: The Final Model

$$\text{Active}_i = 57.3 + 1.1 \cdot \text{Rest}_i - 0.9 \cdot \text{Hgt}_i + 0.1 \cdot \text{Wgt} + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0, 14.9)$

22 / 23

Notes

Next

- ASSESSing MLR models
- Binary Predictors

23 / 23

Notes

Notes
