Multivariate Regression

PSCI 1801 · Class Handout

Dr. Marc Trussler

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

The Standard Error of a Regression Coefficient

Just like the sample mean, a regression coefficient is a random variable with a sampling distribution — normal, centered on the true population slope, with a standard error we can estimate.

Every regression line passes through $(\bar{x}, \bar{y})$. *That's why re-sampled regression lines look like a "double fan" — they converge at the sample means and spread as we move away.*

What R reports in a regression summary:

- **Estimate**: the sample $\hat{\beta}$
- **Std. Error**: the estimated SE of $\hat{\beta}$
- **t value**: $\hat{\beta}/SE$ — how many SEs the estimate is from zero
- **Pr(>|t|)**: two-sided p-value for $H_0 : \beta = 0$

Everything you learned about SEs, CIs, and hypothesis tests transfers directly. A regression coefficient is just another statistic with a sampling distribution.

Regression with More Than One Independent Variable

Add more predictors on the right-hand side of ~. The model becomes:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}$$

Example: California schools — student/teacher ratio (STR) and test scores.

```
Call:
lm(formula = score ~ STR, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-47.727 -14.251   0.483  12.822  48.540

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 698.9329     9.4675  73.825  < 2e-16 ***
STR          -2.2798     0.4798  -4.751 2.78e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.58 on 418 degrees of freedom
Multiple R-squared:  0.05124,   Adjusted R-squared:  0.04897
F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

Univariate result: one extra student per teacher associated with a 2.3-point drop in test scores.

Interpreting Multivariate Coefficients

The magic phrase: "holding the other variables constant."

$$\hat{\beta}_j = \text{change in } y \text{ for a 1-unit change in } x_j, \text{ holding all other } x\text{s constant}$$

Same intercept rule, generalized: $\hat{\alpha}$ is the predicted y when all *predictors equal zero.*

Adding an ESL indicator

```
Call:
lm(formula = score ~ STR + high.esl, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-37.857 -10.853  -0.626  10.198  43.834

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 691.3267     8.1315  85.019  < 2e-16 ***
STR          -1.3963     0.4171  -3.348 0.000889 ***
high.esl    -19.4911     1.5763 -12.365  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.91 on 417 degrees of freedom
Multiple R-squared:  0.3058,    Adjusted R-squared:  0.3025
F-statistic: 91.84 on 2 and 417 DF,  p-value: < 2.2e-16
```

The STR coefficient shrinks from -2.3 *to* -1.4*. Some of the "effect" of STR was really about ESL composition: schools with more ESL students happen to have both higher STR and lower test scores. Once we control for that, the STR effect is smaller (but still significant).*

Omitted Variable Bias

A variable is an omitted variable *if it (a) affects the outcome* y *and (b) is correlated with an included predictor* x*. Omitting it biases* $\hat{\beta}$*.*

Classic example: number of firefighters at a fire correlates with property damage. Both are driven by fire severity. Firefighters don't cause damage.

Not everything correlated is confounded. But you have to think carefully about what you might be missing.

Multiple Continuous Variables

Same math, harder to visualize. With two predictors, the "line" becomes a plane in 3D. With three or more predictors, no visualization is possible — you have to reason directly from the coefficients.

```
Call:
lm(formula = score ~ STR + english, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.845 -10.240  -0.308   9.815  43.461 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 686.03224    7.41131  92.566  < 2e-16 ***
STR          -1.10130    0.38028  -2.896  0.00398 ** 
english      -0.64978    0.03934 -16.516  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.46 on 417 degrees of freedom
Multiple R-squared:  0.4264,	Adjusted R-squared:  0.4237 
F-statistic:   155 on 2 and 417 DF,  p-value: < 2.2e-16
```

Holding % ESL constant, +1 student per teacher → −1.1 test score points. Holding STR constant, +1 percentage point ESL → −0.65 test score points.

Many Variables at Once

Just keep adding to the right-hand side.

```
Call:
lm(formula = score ~ STR + english + lunch + income, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
```

```
-30.7085  -4.9977  -0.2014   4.8411  28.7000

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 675.60821    5.30886 127.261  < 2e-16 ***
STR          -0.56039    0.22861  -2.451   0.0146 *
english      -0.19433    0.03138  -6.193 1.42e-09 ***
lunch        -0.39637    0.02741 -14.461  < 2e-16 ***
income        0.67498    0.08333   8.100 6.19e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.448 on 415 degrees of freedom
Multiple R-squared:  0.8053,    Adjusted R-squared:  0.8034
F-statistic: 429.1 on 4 and 415 DF,  p-value: < 2.2e-16
```

Interpretation stays the same — each coefficient is a partial effect holding all others constant.

Important caution: *more variables is not always better.* You'll see in coming weeks that adding the wrong variables can *worsen* your estimate (post-treatment bias, collider bias, multicollinearity). Specifying a regression is theoretical work, not a "throw everything in" exercise.