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 Regression Equation

Correlation tells us the strength of a relationship on a $[-1, 1]$ *scale. Regression gives us something more actionable: a line through the data with a slope in the units of the actual variables.*

$$\hat{y} = \hat{\alpha} + \hat{\beta} x_i$$

- $\hat{\alpha}$: intercept (predicted y when $x = 0$)
- $\hat{\beta}$: slope (change in y for a 1-unit change in x)

R function: `lm(y ~ x, data = df)`.

How R Picks the Line — Ordinary Least Squares

For each observation, the residual *is the vertical distance from the point to the line:*

$$\hat{u}_i = y_i - \hat{y}_i$$

OLS chooses $\hat{\alpha}$ *and* $\hat{\beta}$ *to* minimize the sum of squared residuals:

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

Some calculus (setting partial derivatives to zero) gives the closed-form solutions:

$$\hat{\beta} = \frac{\hat{\sigma}_{x,y}}{\hat{\sigma}_x^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Neat result: the slope coefficient is just the covariance of x and y divided by the variance of x. *And $\hat{\alpha}$ ensures the line passes through* $(\bar{x}, \bar{y})$.

Interpreting Regression Output

Example: California school data — student/teacher ratio (STR) predicting 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
```

Reading the output

Intercept ($\hat{\alpha}$): predicted y when $x = 0$. Here that's test scores when the student/teacher ratio is zero — meaningless. Interpret only when

$x = 0$ is a sensible value.

Slope ($\hat{\beta}$): "for every 1-unit increase in x, y changes by $\hat{\beta}$ on average." Here, one additional student per teacher is associated with a 2.3-point decrease in test scores.

Scale sensitivity: if x is rescaled (e.g., "students per 10 teachers"), $\hat{\beta}$ rescales proportionally. Same relationship, different-looking number.

t-value, p-value: hypothesis test that the true coefficient is 0. A high $|t|$ / low p means "we can reject the null of a flat line."

Standard error: the estimated standard deviation of $\hat{\beta}$ across hypothetical repeated samples.

Regression with a Binary Independent Variable

If x is 0/1 (an indicator), a "1-unit change" means moving from one group to the other. This is why indicator variables are so tidy in regression.

Example: schools with above-median vs below-median ESL enrollment.

```
(Intercept)    high.esl
    664.354     -20.395
```

- $\hat{\alpha}$: mean of y among the reference group (low-ESL schools)
- $\hat{\alpha} + \hat{\beta}$: mean of y among the other group (high-ESL schools)
- $\hat{\beta}$: difference between the two group means

This is identical to a difference-in-means computed via `t.test()`.

Regression with a Binary Dependent Variable

If y is 0/1, regression becomes a linear probability model *— $\hat{y}$ is the predicted probability of being 1.*

Caution: OLS can predict values outside $[0, 1]$*, which is nonsensical as a probability. Fine for* explanation*, bad for* prediction*. For prediction with binary y, use logit or probit (not in this course).*

Goodness of Fit: R^2

Two sums of squares:

$$\text{SST} = \sum (y_i - \bar{y})^2 \qquad \text{(total variation in } y\text{)}$$

$$\text{SSR} = \sum (\hat{y}_i - \bar{y})^2 \qquad \text{(variation explained by the model)}$$

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

Interpretation: proportion of variation in y explained by the model. Ranges $[0, 1]$.

But — don't fetishize R^2. Whether a low R^2 matters depends on your goal:

- **Prediction**: R^2 matters a lot. A low R^2 means your predictions will be poor.
- **Explanation**: R^2 matters much less. You're not trying to explain every unit of variation, just to identify whether x has a non-zero effect. A low R^2 with a strong, well-identified coefficient is a perfectly fine result.

Age has a low R^2 for predicting BLM support, but the coefficient is negative, large, and highly significant. That's a real result — of course lots of things beyond age influence BLM support.