Linear Regression with One Regressor

Michael Ash

Lecture 12

Goodness of Fit

What fraction of the variation in Y is explained by X?
Reminder (by definition)

$$Y_i = \hat{Y}_i + \hat{u}_i$$

Total Sum of Squares (TSS) expresses the total variation in Y_i (ignoring X) around the mean of Y:

$$TSS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

Explained Sum of Squares (ESS) expresses the variation in $\hat{Y}_i$, the prediction of Y_i using X, around the mean of Y:

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

The $\mathbf{R}^2$ (R-squared) I

If the variation of the prediction of Y_i using X captures a lot of the overall variation in variation in Y_i, then the regression has high explanatory value.
In a perfect regression, because $\hat{Y}_i = Y_i$, the variation of the prediction of Y_i using X would capture all of the overall variation in variation in Y_i.

$$\begin{aligned} R^2 &= \frac{ESS}{TSS} \\ 0 \leq R^2 &\leq 1 \end{aligned}$$

The $\mathbf{R}^2$ (R-squared) II

The Sum of Squared Residuals (SSR) expresses the variation in Y_i around the mean of Y **not** predicted by $\hat{Y}_i$.

$$\begin{aligned} SSR &= \sum_{i=1}^{n} \hat{u}_i^2 \\ TSS &= ESS + SSR \end{aligned}$$

All of the variation can be decomposed into the explained and unexplained variation. (This is not self-evident and depends on the absence of correlation between the explained and unexplained portions).

In the worst possible regression, $\hat{Y}_i = \overline{Y}$ the variation of the prediction of Y_i using X would capture none of the overall variation in variation in Y_i.

$$\begin{aligned} R^2 &= 1 - \frac{SSR}{TSS} \\ 0 \leq & R^2 \leq 1 \end{aligned}$$

The $\mathbf{R}^2$ (R-squared) III

In bivariate regression, $R^2 = r^2$, R-squared is the square of the correlation between X and Y, a direct measure of how well a **line** fits the data.

The $\mathbf{R}^2$ (R-squared) IV

Three competing regressions

1. No-information regression: ignore X; always predict same Y.

$$\begin{aligned} Y_i &= \mu_Y + v_i \\ \hat{Y}_i &= \overline{Y} \\ \hat{v}_i &= Y_i - \overline{Y} \end{aligned}$$

2. OLS regression: does X add any explanatory value?

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ \hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_i \\ \hat{u}_i &= Y_i - \hat{Y}_i \end{aligned}$$

3. Magical regression: know Y_i perfectly

$$\begin{aligned} \hat{Y}_i &= Y_i \\ \hat{w}_i &= 0 \end{aligned}$$

Method for ranking regressions

$$
\begin{aligned}
SSR_1 &= \sum_{i=1}^{n} \hat{v}_i^2, \;\; R^2 = 0 \\
SSR_2 &= \sum_{i=1}^{n} \hat{u}_i^2 \\
SSR_3 &= \sum_{i=1}^{n} \hat{w}_i^2 = 0, \;\; R^2 = 1
\end{aligned}
$$

R^2 expresses how OLS (method 2) fares between method 1 (guessing the mean every time) and method 3 (predicting all of the Y_i perfectly).

What's a "good" $\mathbf{R}^2$?

- Completely context dependent
  - Time-series macroeconomics: typical $R^2 \approx 0.9$
  - Models of individual wages: typical $R^2 \approx 0.3$
- β large and significant but R^2 low
  - Lots of individual randomness (u_i) in the data
  - Regression results useful for average (budgeting, etc.) but not individual prediction

Standard Error of the Regression

- Estimator of the standard deviation of the regression error u_i.
- How much spread in Y_i due to "other factors" remains after the portion explainable by the regression line has been removed? On average, how much of the spread remains after we use knowledge of X to explain spread.
  - Variation in Y_i: some is explainable by X; some is explainable by other factors u
- Measures actual underlying variation in the world (not the sampling variance of an estimator).

$$SER = s_{\hat{u}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2}$$

SER expresses residual variation

- Measured in same units as Y
- Role of "other factors"
- Example: Increasing Standard Error of the Regression in wage regressions since 1979.

Heteroskedasticity and Homoskedasticity

Figure 4.7: All three regression assumptions are maintained. Most importantly, the regression line goes through the middle of the data and observations are evenly spread both above and below the regression line.

- (Non-)Consequences of Heteroskedasticity
  - OLS estimators remain unbiased and consistent
  - Standard errors of OLS estimators are wrong, which can interfere with inference and hypothesis testing.
- Use Heteroskedasticity-Robust Standard Errors (also called Heteroskedasticity-Consistent Standard Errors, Huber-White Standard Errors, Robust Standard Errors, Asymptotic Standard Errors, and Sandwich Estimator)

  ```
  regress testscr str, robust
  ```

Review

- Linear Regression means estimating an intercept and a slope to best fit (X_i, Y_i) data to the equation $Y_i = \beta_0 + \beta_1 X_i + u_i$.
- The prediction equation $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ can be used for policy and simulation.
- The Least Squares formulas are used to determine best fit in an actual sample of data.
- Because the estimates are based on sampled, they are subject to sampling error. We use standard errors to test hypotheses and construct confidence intervals.
- $$\begin{aligned} \widehat{\text{TestScore}} = \quad & 698.9 \qquad - \qquad 2.28\ \text{STR} \\ & (10.4) \qquad\qquad (0.52) \end{aligned}$$

Toward Multiple Regression

- Causal Interpretation and Threats to Causal Interpretation
- Other Factors and Omitted Variables
- Is it possible to “hold other factors constant” while examining a key explanatory variable?