

Lecture Sixteen

regression: goodness-of-fit, correlation and extensions to multiple independent variables

Fit to the Data?

- 📌 How well does the linear regression fit the data?
- 📌 Equivalently, how dispersed are the data points around the regression line?
- 📌 the linear regression fits *better* if the data are *less dispersed* around the regression line.
- 📌 fit to the data is getting *worse* if data are *more dispersed*.

The quantity σ

- a measure of spread around the regression line
- as fit becomes perfect, $\sigma \rightarrow 0$ (data all on the regression line)

Correlation

- Regression and goodness-of-fit closely tied to concept of *correlation*
- Correlations lie between -1 and 1
- Sign of the correlation indicates *direction* of relationship between y and x (and hence sign of the slope of the regression line).

Correlation

- When correlation is 0, no linear relationship between x and y
- When correlation is -1, perfect negative relationship (all data on downward sloping regression line)
- When correlation is 1, perfect positive relationship (all data on upward sloping regression line)

Correlation and Goodness of Fit

- we use roman letter r or Greek letter ρ (rho) for correlation.
- Hence, when $r = \pm 1$, $\sigma = 0$.
- As $r \rightarrow 0$, regression tends to a horizontal line.

Goodness of fit via r-squared

- r-squared: a measure of how much variation in y is associated with variation in x .
- if $r \in [-1, 1]$, then $r\text{-squared} \in [0, 1]$
- as $r\text{-squared} \rightarrow 1$, $\sigma \rightarrow 0$ (model fits better).
- r-squared is the square of the correlation between x and y .

Examples

1. Example 3.4 in Verzani p84 kids weights
2. Example 3.7, Verzani p97: Florida 2000. Interacting with scatterplots.
3. Example 3.8, Verzani, p99. Outlier deletion.
4. Multiple Regression: Example 10.5, Verzani pp303ff
5. Example 10.6, Galileo & Parabolic Trajectories