Cross-Validation

Example: \((X_1, y_1), (X_n, y_n)\). Polynomial regression of degree \(r\). Which \(r\) to use?

Let \(\hat{Y}(r) = Z^{(r)} \hat{B}(r)\)
Features \(1, x, x^2, \ldots, x^r\)

Divide data at random into \(K\) roughly equal sized parts

\[
\begin{array}{cccccc}
1 & 2 & 3 & 4 & 5 \\
\end{array}
\]
\(K=5\)

For \(k=1, 2, \ldots, K\)
Fit \(\hat{Y}^{(r)}(-k) = Z^{(r)}(-k) \hat{B}^{(r)}(-k)\)
Predict \(\hat{Y}^{(r)}(k) = Z^{(r)}(k) \hat{B}^{(r)}(k)\) for all \(r\)

Then CV estimate is
\(CV(r) = \frac{1}{N} \sum (\hat{Y}^{(r)}_i - y_i)^2\)

Choose \(\hat{r} = \text{argmin}_r CV(r)\)

Final Fit
\[
\hat{Y}(\hat{r}) = Z^{(\hat{r})} \hat{B}(\hat{r}) \quad \leftarrow \text{Fit from full data}\
\]