Summary of Week 7

- Categorical predictors
- Interactions
- residual plots

this week:

- other nonlinear models
- overfitting bias variance (trade-off)
- a bit of statistical learning theory

Statistical learning (Ch 2 ISLP)

Consider general problem of predicting Y based on X w/ model

$$E[Y|X] = f(X), \quad Y|X \sim \text{Dist depending on } X$$

Prediction Vs. inference

Prediction: want to know $Y|X$ for X which are not in our data set

Example: predict CO_2 in the future

Inference: want to understand relationships between variables

Example: what % of CO_2 variation is driven by seasonal trend.

inform each other

Goal: Understand issues that emerge when adding more complexity to models

(focus on prediction for now)

Easier to work w/ 1 predictor and complex Y vs. X relationship than multiple predictors w/ simple relationship between each X and Y (as we have before)

Polynomial Regression

Suppose $X \in \mathbb{R}$

$$Y = \beta_0 + \sum_{j=1}^{K} \beta_j X^j + \varepsilon \iff Y|X \sim N\left(\sum_{j=0}^{K} \beta_j X^j, \sigma^2\right)$$

Can define predictors $X_j = X^j$ and estimate β_j using least squares

Lets review how this works:

$$Cov(Y, X^m) = Cov\left(\sum_{j=0}^{K} \beta_j X^j + \varepsilon, X^m\right) = \sum_{j=0}^{K} \beta_j Cov(X^j, X^m)$$

Given data $D = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ can approximate

$$Cov(X^j X^m) \approx \frac{1}{N} \sum_{i=1}^{N} (X_i^j - \overline{X^j})(X_i^m - \overline{X^m})$$

$$Cov(Y, X^m) \approx \frac{1}{N} \sum_{i=1}^{N} (Y_i - \overline{Y})(X_i^m - \overline{X^m})$$

and solve for β. this gives same result as

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_K \end{bmatrix} = \underset{\hat{\beta}}{\operatorname{argmin}} \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{K} \beta_j X_i^j\right)^2$$

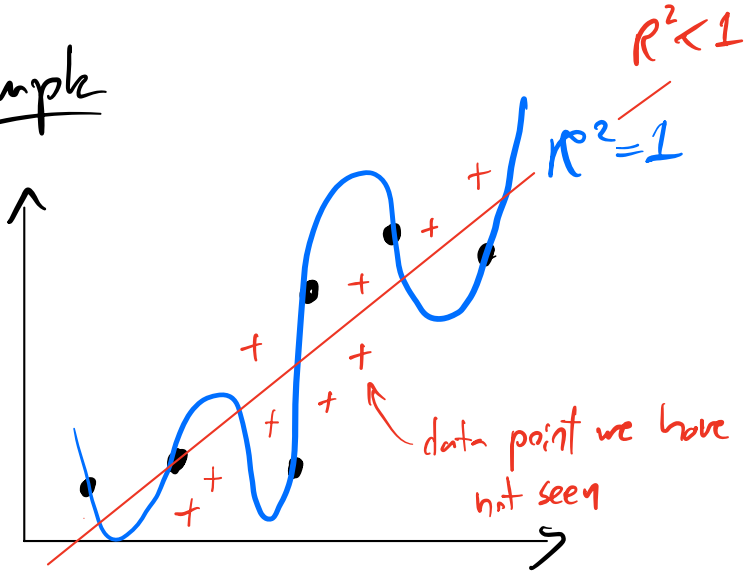
Notation: Let $\hat{y}(x, D) = \sum_{j=0}^{K} \hat{\beta}_j x^j$ when coefficients are computed from data set D

and $RSS = \sum_{i=1}^{N} (Y_i - \hat{y}(x_i, D))^2$, recall $\frac{RSS}{N} \to \hat{\sigma}_\varepsilon^2$

How big should we make K?

previously we used $R^2 = 1 - \frac{RSS}{\hat{\sigma}_y^2}$, but w/ $K = N-1$ we can always make $R^2 = 1$

Example



R^2 is computed based on the data we used to fit the coefficients

$\Rightarrow$ doesn't tell us about ability to generalize!

Cross Validation

Break up into two sets

$$D = D^{train} \cup D^{test}$$

D^{train}: treat as seen data

D^{test}: treat as unseen data

② fit the model using only data in D^{train}

③ for each X_i^{test} in D^{test} compute

$$\hat{y}(X_i^{test}, D^{train}) = \sum_{j=1}^{K} \hat{\beta}_j X_{j,i}^{test}$$

($X_{j,i}^{test}$ — $(X_i^{test})^j$ for polynomial regression)

④ look at $\overline{\varepsilon^2}_{test} = \frac{1}{N_{test}} RSS_{test} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} (Y_i - \hat{y}(X_i^{test}, D^{train}))^2$

[eq. 2.6 in ISLP]

Now we can use $\overline{\varepsilon^2}_{test}$ as measure of how well model predicts new data points!

Also note that we can define

$$\overline{\mathcal{E}^2_{train}} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \left(Y_i - \hat{y}(X_i^{train}, D^{train})\right)^2$$

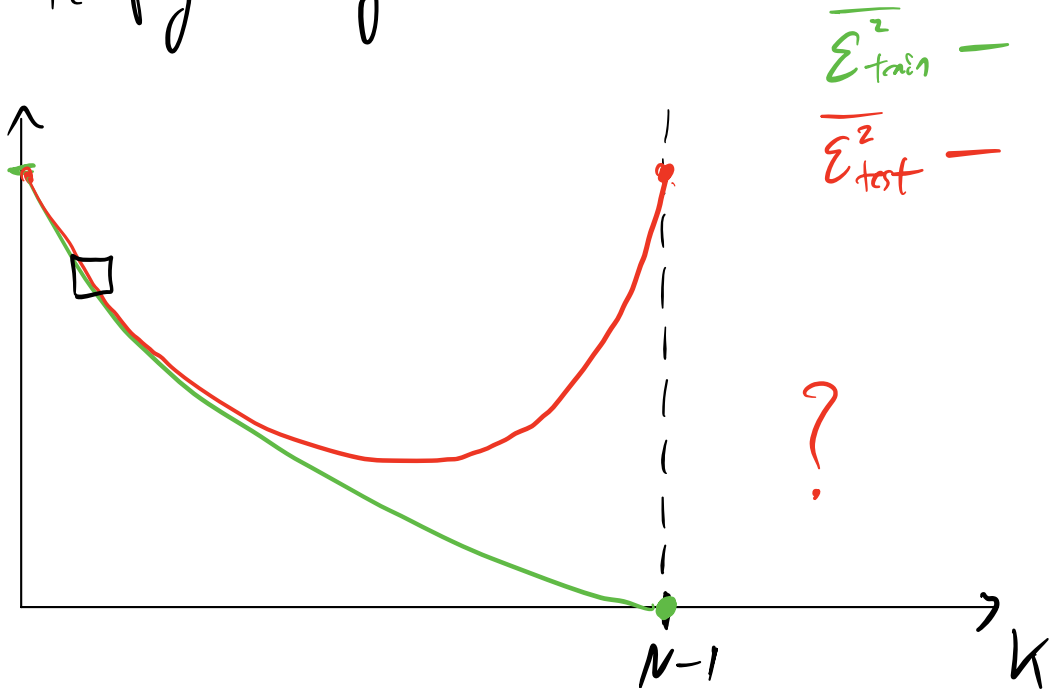
[eq. 2.5 in ISLP]

and notice

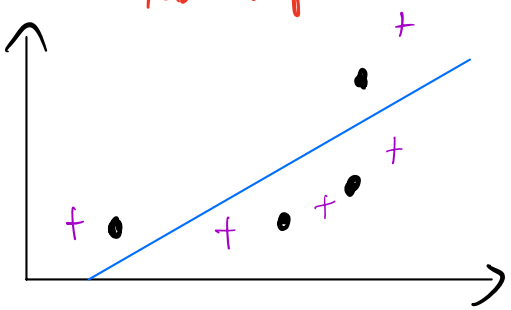
$$R^2 = R^2_{train} = 1 - \frac{\overline{\mathcal{E}^2_{train}}}{\hat{\sigma}^2_{Y^{train}}}$$

we will assume $\hat{\sigma}^2_Y$ is close for test and training points so we can just work w/ $\mathcal{E}$s

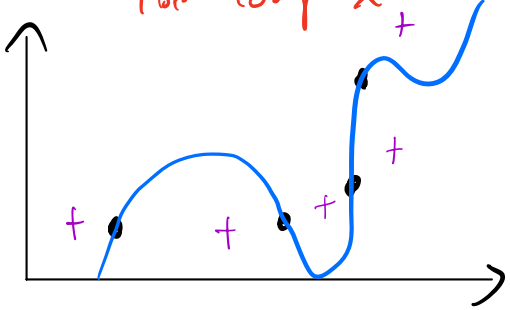
Now lets look at behavior of $\overline{\mathcal{E}^2_{train}}$, $\overline{\mathcal{E}^2_{test}}$ as function of K for the polynomial regression model



too simple



too complex



Bias Variance Decomposition

Let θ be any quantity we would like to estimate and $\hat{\theta}$ be an estimator of it

in discussion above $\hat{\theta} = \hat{y}$, $\theta = \mathbb{E}[Y|X]$

let $MSE_{\hat{\theta}} = \mathbb{E}[(\theta - \hat{\theta})^2]$

Because adding a constant does not change the variance

$$Var(\hat{\theta}) = Var(\hat{\theta} - \theta) = \mathbb{E}[(\hat{\theta} - \theta)^2] - \mathbb{E}[\hat{\theta} - \theta]^2$$

$$= MSE_{\hat{\theta}} - \mathbb{E}[\hat{\theta} - \theta]^2$$

$\mathbb{E}[\hat{\theta} - \theta]^2 = Bias^2_{\hat{\theta}}$

Bias-Variance decomposition

$$MSE_{\hat{\theta}} = Var(\hat{\theta}) + Bias^2_{\hat{\theta}}$$

large Bias means result will differ systematically from truth

large Variance means less confident in result / will depend more on specific data set

Example Laplace Rule of Succession (see HW5)

$$\hat{q} = \frac{S+1}{N+2} \quad \left(\text{compare to } \hat{q} = \frac{S}{N}\right)$$

Now to relate back to $\overline{\mathcal{E}^2_{test}}$ vs $\overline{\mathcal{E}^2_{train}}$ picture:

Suppose $y = f(x) = \mathbb{E}[Y|X=x]$ is true relationship between x and avg. of Y

then $\hat{\theta} = \hat{y}(x, D)$ is estimator of $\theta = f(x)$

$$\overline{\mathcal{E}^2_{test}} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \left(Y_i - \hat{y}(X_i^{test}, D^{train})\right)^2$$

$$\approx \mathbb{E}[(Y - \hat{y})^2] = \mathbb{E}[(f(X) + \varepsilon - \hat{y}(X, D))^2]$$

$$= \mathbb{E}[(f(X) - \hat{y}(X, D))^2 + 2f(X)\varepsilon - 2\varepsilon\hat{y} + \varepsilon^2]$$

$$= MSE_{\hat{y}} + \mathbb{E}[\varepsilon^2] = MSE_{\hat{y}} + \sigma_\varepsilon^2$$

↑ assumes N_{test} large

Hence $\overline{\mathcal{E}^2_{test}}$ almost approximates MSE of $\hat{y}$, but since we don't know the true relationship there is an extra σ_ε^2

this extra term is called irreducible error

Note when $K \approx N-1$ this relationship breaks down, but the curve looks the way it still gives us an idea of why it does