Evaluating Regression Models

Recap

1. Linear Regression

- A supervised learning method.
- Fits a straight line (or hyperplane in higher dimensions) between predictors and response.
- Useful but limited to linear relationships.

2. Polynomial Regression

- Extends linear regression by adding higher-order terms ($x^2, x^3, \ldots$).
- Captures non-linear patterns while remaining linear in parameters.

3. Underfitting vs Overfitting

- **Underfitting:** Model too simple, fails to capture data trends (high bias, low variance).
- **Overfitting:** Model too complex, fits noise instead of trend (low bias, high variance).
- Goal = find a balance (bias-variance tradeoff).

Practical Considerations

1. Is Linear Regression a Machine Learning method?

- **Yes.** Because it can **generalize** to out-of-sample (unseen) data.
- ML is not just about using complex models → even simple regression qualifies if it generalizes.

---

2. Is E_out always greater than E_in?

- **Definitions:**
  - E_{in}: Error on training data (in-sample error).
  - E_{out}: Error on unseen/test data (out-of-sample error).

- **Reasoning:**
  - E_{in} is minimized during training → so it's usually optimistic (smaller).
  - E_{out} reflects true generalization ability → typically larger.
- **Important nuance:**
  - For individual data points, $E_{out}(x_k)$ could be smaller than E_{in}.
  - But **on average**, across all points:

    $E_{out} \geq E_{in}$

Hypothesis Testing & Variable Importance

1. Is at least one predictor related to the response?

- Use the **F-statistic**:
  - Null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_d = 0$ (no predictors matter).
  - Large F-statistic → reject H_0, meaning predictors are useful.
  - p-value from F-stat can guide acceptance/rejection.
- **Comparison with t-test:**
  - **t-test:** Checks individual predictor significance.
  - **F-test:** Checks overall significance (all predictors together).
  - When many predictors exist, t-tests may falsely suggest many are important → F-test adjusts for this.

The relationship between t and F

q is the number of restrictions (coefficients tested)

- **When q=1 (only one coefficient tested)**:

  $F = t^2$

  with the same degrees of freedom.

  This means:
  - A **t-test** for one coefficient and an **F-test** for that same single restriction are exactly equivalent.

- Same p-value, same decision.

- **When q>1**:
  - You cannot replace an F-test with a single t-test, because you're testing a *joint hypothesis* (multiple coefficients simultaneously).
  - Example: "Are both β_1 and β_2 zero?" — F-statistic can handle this, t cannot.

Practical connections

- **t-statistic**: use when you care about *one predictor* at a time.
- **F-statistic**: use when you care about *overall model fit* or *joint significance of a group of predictors*.
- Both are based on the same idea: compare "signal explained by predictors" vs "unexplained noise."

---

2. Deciding on Important Variables → Variable Selection Methods

- **Forward Selection:**
  - Start with no predictors.
  - Add features one by one (choose the one that reduces RSS the most).
- **Backward Selection:**
  - Start with all features.
  - Remove the least significant one (largest p-value based on the F-statistic).
  - Repeat until the optimal set is left.
- **Mixed Selection:**
  - Combines forward and backward.
  - Add predictors while also dropping those that become insignificant.

Model Fit Assessment

1. Residual Standard Error (RSE)

- **Definition:** Average amount by which the observed responses deviate from the true regression line.

- Formula (conceptual):

$$RSE = \sqrt{\frac{RSS}{n-d-1}}$$

  where $RSS = \sum (y_i - \hat{y}_i)^2$.
- Interpretation: Smaller RSE → better fit.
- Caveat: Depends on the scale of y (so interpretability can be weak).

---

2. R^2 Statistic (Coefficient of Determination)

- Measures the proportion of variance in yy explained by the predictors.
- Formula:

  $R^2 = 1 - \frac{RSS}{TSS}$

  where TSS = total sum of squares.
- Range: 0 $\leq R^2 \leq$ 1.
- Higher = better fit. But "good" value depends on context.

---

3. Adjusted R^2

- Problem with plain R^2: It always increases when more features are added, even if they are irrelevant.
- Adjusted R^2: Penalizes inclusion of unnecessary predictors.
- Useful for:
  - **Multiple regression** (more than 1 predictor).
  - **Model comparison** (different number of predictors).
  - **Preventing overfitting**.

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}$$

- Denominator adjustment:
  - (n−p−1) is the degrees of freedom for error (residuals).
  - (n−1) is the degrees of freedom for total variation.

- This penalizes the inclusion of irrelevant predictors.

Key Properties

- $R^2_{\text{adj}} \leq R^2$ always.
- Adding a predictor:
  - If it improves the model enough → adjusted R^2 goes **up**.
  - If it's mostly noise → adjusted R^2 goes **down**.

👉 So adjusted R^2 balances **goodness of fit** with **model parsimony**.

Interpretation

- Adjusted R^2 still measures "proportion of variance explained," but with a correction for the number of predictors.
- It is more trustworthy than plain R^2 when comparing models of different sizes.

Accuracy of Predictions

1. Two Sources of Error

- **Reducible Error (Bias + Variance):**
  - Comes from the model itself.
  - It can be reduced by choosing better models, more data, or proper regularization.
- **Irreducible Error (Noise):**
  - Comes from randomness or factors outside the model (measurement error, unknown variables).
  - It cannot be eliminated, no matter how good the model is.

---

2. Bias-Variance Decomposition (High-Level)

$E_{out} = \text{Bias}^2 + \text{Variance} + \text{Noise}$

- **Bias:** Error from oversimplifying (underfitting).
- **Variance:** Error from model being too sensitive to training data (overfitting).
- **Noise:** Unavoidable error due to randomness.

Bias-Variance Tradeoff

1. Key Idea

- Model complexity influences **bias** and **variance** in opposite ways.
  - Simple models → high bias, low variance (underfit).
  - Complex models → low bias, high variance (overfit).
- The **optimal model** balances bias and variance → lowest E_{out}.

---

2. Practical Implications

- Choosing the number of parameters (or polynomial degree d) = a balancing act.
- **Too few parameters** → model misses important patterns (underfitting).
- **Too many parameters** → model chases noise (overfitting).

📘 Bias–Variance Decomposition of Out-of-Sample Error (explicit)

Setting / assumptions

- Data generated by

$y = f(x) + \varepsilon$,

where ε is noise with $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$, independent of x.

- A learning algorithm produces a predictor $\hat{f}(x)$ that depends on the **training set** $\mathcal{D}$. Expectations below are over the randomness of the training set (and of ε when applicable).

We examine the expected squared prediction error at a fixed test point x:

$\mathbb{E}_{\mathcal{D},\varepsilon}\big[(y - \hat{f}(x))^2\big]$.

Decomposition at a fixed x

Start with

$$\mathbb{E}\big[(y - \hat{f}(x))^2\big] = \mathbb{E}\big[(f(x) + \varepsilon - \hat{f}(x))^2\big].$$

Expand and use $\mathbb{E}[\varepsilon] = 0$ and independence:

$$\begin{aligned}\mathbb{E}\big[(f(x) + \varepsilon - \hat{f}(x))^2\big] &= \mathbb{E}\big[(f(x) - \hat{f}(x))^2\big] + \mathbb{E}[\varepsilon^2] + 2\,\mathbb{E}[\varepsilon(f(x) - \hat{f}(x))] \\ &= \mathbb{E}\big[(f(x) - \hat{f}(x))^2\big] + \sigma^2.\end{aligned}$$

So the noise (irreducible error) separates out. Now decompose $\mathbb{E}\big[(f(x) - \hat{f}(x))^2\big]$. Define the expected predictor

$$\bar{f}(x) \equiv \mathbb{E}_{\mathcal{D}}[\hat{f}(x)].$$

Add and subtract $\bar{f}(x)$:

$$\mathbb{E}\big[(f(x) - \hat{f}(x))^2\big] = \mathbb{E}\big[(f(x) - \bar{f}(x) + \bar{f}(x) - \hat{f}(x))^2\big].$$

Expand the square and use linearity and $\mathbb{E}[\hat{f}(x) - \bar{f}(x)] = 0$:

$$\begin{aligned}\mathbb{E}\big[(f - \hat{f})^2\big] &= (f(x) - \bar{f}(x))^2 + \mathbb{E}\big[(\hat{f}(x) - \bar{f}(x))^2\big] + 2(f(x) - \bar{f}(x))\underbrace{\mathbb{E}[\bar{f}(x) - \hat{f}(x)]}_{0} \\ &= \underbrace{(\bar{f}(x) - f(x))^2}_{\mathrm{Bias}^2(x)} + \underbrace{\mathrm{Var}_{\mathcal{D}}(\hat{f}(x))}_{\mathrm{Variance}(x)}.\end{aligned}$$

Combine with the noise term σ^2 to get the full decomposition at x:

$$\boxed{\mathbb{E}_{\mathcal{D},\varepsilon}\big[(y - \hat{f}(x))^2\big] = \sigma^2 \;+\; \big(\mathrm{Bias}(\hat{f}, x)\big)^2 \;+\; \mathrm{Var}(\hat{f}, x)}$$

where

$$\mathrm{Bias}(\hat{f}, x) \equiv \bar{f}(x) - f(x), \qquad \mathrm{Var}(\hat{f}, x) \equiv \mathbb{E}_{\mathcal{D}}\big[(\hat{f}(x) - \bar{f}(x))^2\big].$$

Integrated (Out-of-Sample / Expected Test MSE)

If X is random with distribution $p_X(x)$, the expected out-of-sample MSE is the expectation over x:

$$\mathbb{E}_{X,\mathcal{D},\varepsilon}\big[(Y - \hat{f}(X))^2\big] = \sigma^2 \;+\; \mathbb{E}_X\big[(\bar{f}(X) - f(X))^2\big] \;+\; \mathbb{E}_X\big[\,\mathrm{Var}_{\mathcal{D}}(\hat{f}(X))\big].$$

Often written compactly as

$$\text{Test MSE} = \text{Irreducible noise} + \text{Bias}^2 + \text{Variance},$$

with the $bias^2$ and variance interpreted as averages over the input distribution if desired.

Remarks & intuitions

- **Noise** σ^2 cannot be reduced by any model — it is inherent in y.
- **Bias** measures how far the average model $\bar{f}$ is from the true f. High-bias models are underfitting.
- **Variance** measures how much the learned model $\hat{f}$ fluctuates around its mean when training data changes. High-variance models overfit.
- The bias–variance tradeoff: reducing one often increases the other. E.g. increasing model complexity typically lowers bias but raises variance.
- This decomposition holds **exactly** for squared error; for other loss functions (e.g., 0–1 classification loss) there is no simple analogous decomposition.

Quick examples (intuition)

- **KNN with small K**: low bias (can fit complex shapes), high variance → may overfit.
- **Linear regression** (simple linear model): low variance, possibly high bias if true relation is nonlinear.