Variable and Model Selection

Setting:

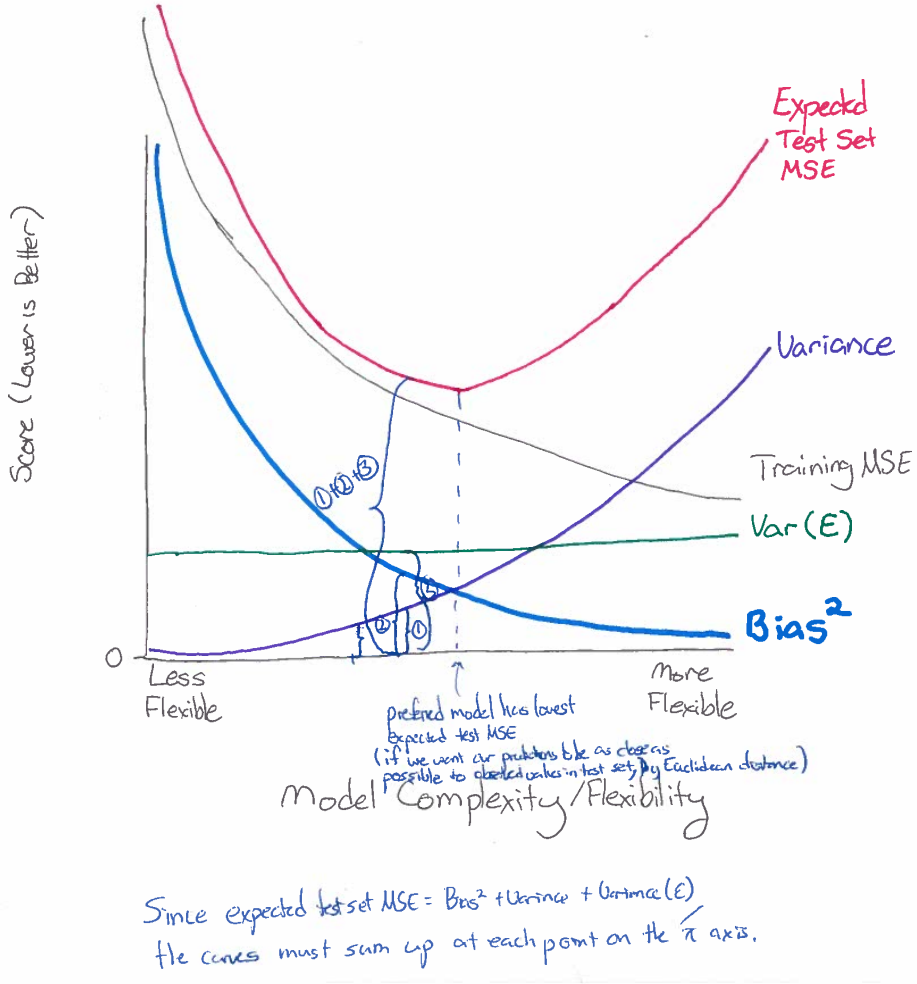
We are performing regression or classification, and we have many possible explanatory variables.

Which explanatory variables should be included in our model, and in what form (polynomial, interaction, log, . . .)?

Challenge:

Training set error (training MSE or training classification error) **always go down** if we add more explanatory variables higher degree polynomial terms, interactions, ...

Test set error goes down for a while, then back up.



We care about test set error.

Approaches to model comparison, ordered from most useful to least useful

Most Useful: Plots!

- training set residuals
- validation set residuals

Second Most Useful: (Cross-)Validated estimation of test set error

Third Most Useful: Theoretically-justified adjustments to training set error, to estimate test set error (presented here for the case of regression)

Recall that training set MSE is $MSE = \frac{1}{n}RSS$

Three most common examples:

1. Mallow's C_p: $\frac{1}{n}(RSS + 2p\hat{\sigma}^2)$
2. Akaike Information Criterion (AIC): $\frac{1}{n\hat{\sigma}^2}(RSS + 2p\hat{\sigma}^2)$
3. Bayesian Information Criterion (BIC): $\frac{1}{n}(RSS + 2\log(n)p\hat{\sigma}^2)$

In all three cases:

- Start with training set MSE and add a penalty to try to get to test set MSE.
- Size of penalty increases as p increases.
- Theoretically, for large n these approaches are equivalent and lead to selection of a model with best test set MSE.

Fourth Most Useful (in Evan's opinion; others disagree): Hypothesis Tests

Fifth Most Useful: Adjusted R^2

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

Adjusted $R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$

If d increases (and RSS doesn't change much), * $n-d-1$ is small * $RSS/(n-d-1)$ is large * Subtracting $RSS/(n-d-1)$ is small

Not well justified theoretically.

Strategies for Developing a Model:

If the number of candidate explanatory variables is relatively small

1. Pick a starting model based on intuition and background knowledge about context
   - Often main effects only unless you know something that suggests interactions will be needed.
2. Repeat the following until satisfied:
   a. Look at plots for current model(s) to try to identify limitations
   b. Identify one or more candidate changes to make
      - Add an explanatory variable
      - Remove an explanatory variable
      - Transformation of response or explanatory variables
      - Add polynomial term
      - Add interaction
   c. Make the identified changes.
      - Evaluate performance of each candidate model via cross-validation, C_p, AIC, or BIC
   d. Pick a small number of models to continue exploring and refining

If the number of candidate explanatory variables is large

The manual process described above can be infeasible if we have many possible explanatory variables. Some ideas:

Forward Stepwise Selection

1. Start with the *null* model, which has no explanatory variables. Denote this model by $\mathcal{M}_0$.
2. For each $k = 0, \ldots, p-1$:
   a. Consider all possible models obtained by adding one explanatory variable to $\mathcal{M}_k$
   b. Select the one of these models with best performance as measured by cross-validation, C_p, AIC, or BIC. Denote this model by M_{k+1}
3. Select one or more "best" models from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validation, C_p, AIC, or BIC.

Backward Stepwise Selection

1. Let $\mathcal{M}_p$ denote the full model, with all p explanatory variables
2. For each $k = p, p-1, \ldots, 1$:
   a. Consider all possible models obtained by removing one explanatory variable from $\mathcal{M}_k$
   b. Select the one of these models with best performance as measured by cross-validation, C_p, AIC, or BIC. Denote this model by M_{k-1}
3. Select one or more "best" models from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validation, C_p, AIC, or BIC.

Best Subsets Selection

1. Fit every possible model with 0 or more predictive variables (there will be 2^p such models)
2. Select one or more "best" models using cross-validation, C_p, AIC, or BIC.