4. Criteria for what makes a fit good:
   a. $R^2$
      i. As we have seen, will generally favor the model with all possible variables.
      ii. Is a reasonable way to distinguish between models of the same size.
   b. Adjusted $R^2$
      i. Penalizes larger models.
      ii. Hence is a viable contender to choose between models of different sizes.
      iii. When comparing a model with $p$ to a model containing it as a subset, with $s$ additional regressors
         - Call adjusted values $R^2_{adj,p}$, $R^2_{adj,p+s}$ resp.
         - Let regression sum of squares be $SS_{R,p}$, $SS_{R,p+s}$.
         - Let total sum of squares be $SS_t$.
         - $R^2_{adj,p} = 1 - \frac{n-1}{n-p}(1 - \frac{SS_{R,p}}{SS_t})$.
         - $R^2_{adj,p+s} = 1 - \frac{n-1}{n-p-s}(1 - \frac{SS_{R,s+p}}{SS_t})$.
      iv. Adjusted $R^2$ drops increases if and only if associated $F$
statistic exceeds 1.

- Because
  \[
  R_{adj,p+s}^2 > R_{adj,p}^2
  \]
  \[
  \iff \frac{n-1}{n-p} \left(1 - \frac{SS_{Rp}}{SS_t}\right) > \frac{n-1}{n-p-s} \left(1 - \frac{SS_{Rs+p}}{SS_t}\right)
  \]
  \[
  \iff (n-p-s)(1 - SS_{Rp}/SS_t) > (n-p)(1 - SS_{Rs+p}/SS_t)
  \]
  \[
  \iff (n-p-s)(SS_t - SS_{Rp}) > (n-p)(SS_t - SS_{Rs+p})
  \]
  \[
  \iff SS_{Rs+p} > \frac{sSS_t + (n-p-s)SS_{Rp}}{n-p}.
  \]

- Recall \( F \) statistic for testing the last \( s \) variables:

  \[
  F > 1
  \]

- Substituting in for the marginal \( SS_{Rs+p} \), \( F > 1 \).

d. Mallows \( C_p \)

  i. Notation: Let

  - \( p \) be number of parameters in candidate model
    - Includes intercept if present, but not \( \sigma \)
  - \( \nu = X\beta \) are expectations for the true model.
  - \( \eta = (X_p^\top X_p)^{-1}X\beta \) are true expectations of too-small model
  - \( SSB = \sum_{i=1}^{n}(\nu_i - \eta_i)^2 \)
• $\hat{Y}_i$ be fitted value using the incorrect model.

ii. $E \left[ \sum_{i=1}^{n} (\hat{Y}_i - \nu_i)^2 \right] = SSB + \sum_{i=1}^{n} \text{Var} [\hat{Y}_i]$

• $\sum_{i=1}^{n} \text{Var} [\hat{Y}_i] = p\sigma^2$.

iii. Residual Sum of Squares:

• $SS_{Res} = \sum_{i=1}^{n} (Y_i - \hat{\eta}_i)^2$

• Then

$$E [SS_{Res}] = \sum_{i=1}^{n} \text{Var} [Y_i - \hat{\eta}_i] + \sum_{i=1}^{n} E [Y_i - \hat{\eta}_i]^2$$

$$= (n - p)\sigma^2 + SSB$$

▷ Method of moments estimator for SSB is

$$\hat{SSB} = SS_{Res} + (p - n)\hat{\sigma}^2$$

iv. Let $\Gamma_p = E \left[ \sum_{i=1}^{n} (\hat{Y}_i - \nu_i)^2 \right] / \sigma^2 = SSB / \sigma^2 - p$ be this expected squared error, scaled by variance.

v. $C_p = SS_{Res_p} / \hat{\sigma}^2 - n + 2p$ for SSE$_p$ the residual sum of squares using $p$ parameters (now not including $\hat{\sigma}$)

vi. Best when $C_p \approx p$.

• Note for models with all correct parameters, and extra parameters, $C_p$ goes back up.

d. Akaike Information Criterion (AIC):

i. Lower is better
Lecture 12

ii. AIC is defined as $2p - 2\ell(\hat{\beta}, \hat{\sigma})$

iii. $\ell$ is maximized Log likelihood

iv. For regression

$$\ell(\hat{\beta}, \hat{\sigma}) = -(n/2) \log(2\pi \hat{\sigma}^2) - \sum_{i=1}^{n} \frac{(Y_i - \hat{\beta}x_i)^2}{2\hat{\sigma}^2}$$

$$= -(n/2) \log(2\pi \hat{\sigma}^2) - n/2.$$  

v. AIC is $2p - 2\ell(\hat{\beta}, \hat{\sigma}) = 2p + n\log(2\pi \hat{\sigma}^2) + 1$

vi. $p$ is number of parameters, incl. $\sigma$

vii. Lower is better.

viii. Software reliably reports this only up to an additive constant.

e. Bayesian Information Criterion:

i. $p \ln(n) - 2\ell(\hat{\beta}, \hat{\sigma})$

ii. As above, $\ell(\hat{\beta}, \hat{\sigma}) = n\log(2\pi \hat{\sigma}^2) + 1)/2$

iii. Lower is better

iv. Large-sample approximation to Bayesian result with prior on number of models.

- As often happens, precise form of prior disappears in first-order approximation.

v. Penalty for larger model increases as $n$ increases.

5. Seach among models:
a. This is a big job, since there are \(2^p\) models to search through.
   i. There are algebraic tools do to this without fitting every model (ex. Furnival and Wilson)

b. Earlier principles involving model coherence doesn’t solve this.
   i. Still need to keep together factors, sine and cosine terms.
   ii. Still need not to remove lower order polynomial terms with higher-order terms still in model.

c. People often use stepwise:
   i. Start with an initial model.
   ii. Consider models with separate (groups of) parameters added or removed, one at a time.
      - Hypothesis test
        ▶ Typically set significance higher: 0.15?
        ▶ To ensure stability, put level for removal higher than that of inclusion.
      - Adjusted \(R^2\)
      - \(C_p\), AIC, BIC
   iii. Move to model with numerical criteria improved.
   iv. Gives a local, rather than guaranteed global, optimum.
v. At each step, one can add or remove variables.
   - Only considering additions: Forward stepwise.
   - Only considering deletions: Backwards stepwise.

6. Interpretation after selection:
   a. Model parameters measure effect of explanatory variable in light of all other variables in model.
   b. Hence interpretation of parameter changes as other variables move in and out of the model.
   c. Inference after selection has multiple-comparisons issue.
      i. Effect of variables in a best-fitting model will be exaggerated relative to a model selected a priori.
      ii. This exaggerated effect must be adjusted for when performing post-selection inference.
   iii. Solutions:
      - test and training set.
      - build model without explanatory variable for primary hypothesis.
   d. Model selection is impacted by coordinate system for variables.
      i. Ex., a model that contains a baseline value and a change from
baseline will be treated differently from a model that contains a baseline and later value.

7. Lasso
   a. Acronym for least absolute shrinkage and selection operator.
   b. Minimize \((Y - X\beta)^\top(Y - X\beta)\) subject to \(\sum_j |\beta_j| \leq t\) for some \(t\).
   c. Equivalently, minimize \((Y - X\beta)^\top(Y - X\beta) - \lambda \sum_j |\beta_j|\).
   d. Because of presence of absolute value signs, you can’t optimize using differentiation.
   e. Cf. minimize \((Y - X\beta)^\top(Y - X\beta) - \lambda \beta^\top\beta\)
      i. Derivative is \(2X^\top(Y - X\beta) - 2\lambda \beta\)
      ii. Estimate \(\tilde{\beta}\) sets this to zero.
      iii. \(X^\top Y = (X^\top X + \lambda I)\tilde{\beta}\)
      iv. \(\tilde{\beta} = (X^\top X + \lambda I)^{-1}X^\top Y\) : Ridge regression.
   f. Lasso is equivalent to Bayes approach with Laplace prior.

VI. Some Other Regression Models:

A. Quantile Regression
   1. Minimize Sum of Absolute Errors
      a. Minimize difference \(\sum_{j=1}^{n} |Y_j - \beta_1 - \beta_2 X_j|\).
b. Note the absence of $\cdot^2$, which would be present in regular least squares regression.

c. Equivalent to minimizing $\sum_j e_j^+ + e_j^-$, for $e_j^+ \geq 0$, $e_j^- \geq 0$,

$$e_j^+ - e_j^- = Y_j - \beta_1 - \beta_2 X_j \forall j.$$                  

i. This minimization is an example of linear programming.

ii. Solution is well-known, but computationally intensive.

- Harder than taking derivative, since function to be optimized is non-differentiable.

\[d. \] This solution is called $L^1$ regression, after the power on $|Y_j - \beta_1 - \beta_2 X_j|$.

e. Solution is also called quantile regression.

i. If $\beta_2 = 0$, best $\beta_1$ is median.

ii. Just as median is not always uniquely defined, these estimates are not necessarily uniquely defined.

2. Regression through other quantiles

a. Can replace $\sum_j e_j^+ + e_j^-$ by $\tau \sum_j e_j^+ + (1 - \tau) \sum_j e_j^-$. 

i. Runs regression line through $1 - \tau$ quantile of $Y|X$.

3. Geometry of minimum

a. At minimum $\hat{\beta}$, $\sum_{j=1}^n |Y_j - \beta_1 - \beta_2 X_j|$ should not decrease
as $\beta$ moves away from $\hat{\beta}$.

i. Since $\sum_{j=1}^{n} |Y_j - \beta_1 - \beta_2 X_j|$ is piecewise linear, function $S(\beta)$ giving “derivative” should be piecewise constant, with jumps.

ii. $\hat{\beta}$ should satisfy $S(\hat{\beta}) \approx 0$
   - Estimate either sets the function to 0, or is a point where it jumps across 0.

iii. $S(\beta)$ may be expressed as a rank statistic

iv. $H_0: \beta = \beta^o$ may be tested by comparing $S(\beta^o)$ to 0
   - Either exactly or asymptotically.

v. Test can be inverted to give confidence sets for $\beta$.

MPV: 15.1.1