Lecture 13

Administrative:

Assignment 2 Q4 Typo: $i=1,\dots,392$ (not $i=1,\dots,10$).

Term test 2 next Wednesday in class.

- Covers materials from Lecture 8 to Lecture 14 (this Wednesday).
- Three Questions (Heavy on application)
- You should be able to, e.g.,
  - → interpret R outputs
  - → conduct hypothesis tests (both t- and F-test) $\boxed{\text{both can be used when we have 1 constraint}}$
  - → interpret parameters estimates, etc...
- No cheat sheet
- Bring a calculator!!!

Last class: General Multicollinearity

- When some covariates are highly correlated with other covariates.
- Variance Inflation Factor (VIF):

$$VIF_j = \frac{1}{1 - R_j^2} \quad \text{for } j = 1, \dots, p$$

where R_j^2 is the R^2 value from regressing x_j on other explanatory variables.

★ Rule of thumb: If $VIF_j \geq 10$, this implies strong multicollinearity. ($R_j^2 \geq 0.9$)

★ Procedure:

1. Remove the covariate with the largest $VIF \geq 10$
2. Recalculate VIF for each covariate in reduced model
3. Repeat steps 1 and 2 until no more strong multicollinearity.

Model Selection

Given p covariates, how do we find a subset ($k \le p$) that gives us the "best" model? What's considered the best model?

1. Interpretability
2. Goodness-of-fit
3. Predictive performance.

Notes on interpretability and goodness-of-fit:

1. Interpretability: If the goal of a model is to make inference about the relationship between a response and explanatory variables, it is only useful to the extent that it is interpretable.

e.g. study the relationship between viral load (Y) and CD4 count (x) in HIV individuals. Goal: publish in a clinical journal.

$\rightarrow Y_i = \beta_0 + \beta_1 CD4_i + \varepsilon_i$ vs. $Y_i = \beta_0 + \beta_1 CD4_i + \beta_2 CD4_i^2 + \beta_3 CD4_i^3 + \dots + \varepsilon_i$

- There is a trade-off between model complexity and interpretability.

2. Goodness-of-fit.

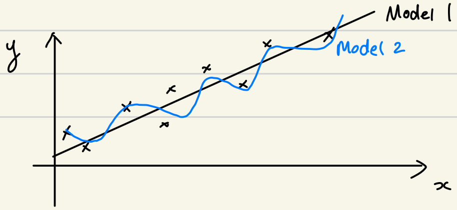
$$R^2 = SS(reg)/SS(tot) = 1 - SS(res)/SS(tot).$$

- calculates the proportion of variability explained by our model.
- Problem: R^2 is non-decreasing with addition of predictors.

Intuition: increasing the span(X) increases the space to find a better LS solution. Thus, in larger space, we could never do worse than in a smaller space.

→ If we fit a model w/ too many predictors, R^2 might be higher but we might be overfitting: explaining variations that are due to noise.

⇒ leading to imprecise predictions.



Model 1: $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x$

Model 2: $\hat{\mu} = \hat{\alpha}_0 + \hat{\alpha}_1 x + \hat{\alpha}_2 x^2 + \dots + \hat{\alpha}_j x^j$

(compare w/ underfitting: too few predictors $\Rightarrow$ leading to biased prediction).

Model Selection (Two main ingredients):

1. Model Selection Criteria: for comparing different models w/ potentially different # of predictors.
2. Model selection strategy: which model to fit?

Some common criteria

① Adjusted R^2: $R^2_{adj} = 1 - \frac{SS(res)/n-k-1}{SS(tot)/n-1}$: k = # of predictors in the model ($k \le p$)

- Denominator: sample variance of $y_1, \dots, y_n$
- Numerator: $\hat{\sigma}^2$ of model w/ k predictors.
- Compare w/ $R^2 = 1 - SS(res)/SS(tot)$.

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{SS(res)}{SS(tot)} = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2)$$
$$= 1 - \left(1 + \frac{k}{n-k-1}\right)(1-R^2) = R^2 - \underbrace{(1-R^2)\frac{k}{n-k-1}}_{\text{penalization for having too many predictors}}$$

Intuition: R^2_{adj} accounts for the number of predictors in our model, and penalizes when we include unimportant predictors (i.e. when $SS(res)$ decreases only slightly when these predictors are added).

- While R^2 always is non-decreasing w/ addition of predictors, R^2_{adj} may decrease. (e.g. when R^2 is not improving by much).
- Prefer model w/ a higher R^2_{adj}.
- No longer explaining the proportion of variability, but still used as a measure of goodness-of-fit. It is a model selection criteria (e.g. choose model w/ higher R^2_{adj}).

- $R^2_{adj} = 1 - \frac{SS(res)/n-k-1}{SS(tot)/n-1} = 1 - \hat{\sigma}^2 / (SS(tot)/n-1)$
  - → minimizing $\hat{\sigma}^2$ is equivalent to maximizing R^2_{adj}
  - → equivalently can choose a model w/ a smaller $\hat{\sigma}^2$.

e.g.

	Model 1 ($k=4$)	Model 2 ($k=6$)
SS(reg)	20	21
SS(res)	10	9
SS(tot)	30	30

$n=20$ ← sample size

Calculate

$R^2 = 20/30 = 0.67$	$R^2 = 21/30 = 0.7$.
$R^2_{adj} = 1 - \frac{10/15}{30/19} = 0.58$	$R^2_{adj} = 1 - \frac{9/13}{30/19} = 0.56$.
$\hat{\sigma}^2 = 10/15 = 0.67$	$\hat{\sigma}^2 = 9/13 = 0.69$.

∴ Model 1 is preferred.

Note:
- Generally $R^2_{adj} < R^2$ but $n \to \infty$ R^2_{adj} converges to R^2.
- Model w/ R^2_{adj} has a lower $\hat{\sigma}^2$

② AIC (Akaike Information Criterion).

Let n be the sample size, q be the number of parameters in a model (e.g. In MLR, $q = k+1+1$ where k is the number of predictors, $+1$ for intercept, $+1$ for σ^2).

$$AIC = -2[\ln L(\hat{\theta}) - q]$$
$$= 2q - 2\ln L(\hat{\theta})$$

where $L(\hat{\theta})$ is the likelihood evaluated at $\hat{\theta}$.

Note:
- $2q$ is the penalty for including more predictors.
- AIC penalizes the log likelihood function: w/ more parameter $L(\hat{\theta})$ will increase but will be offset by penalty $2q$.
- A model w/ a smaller AIC is preferred in general.

③ BIC (Bayesian Information Criterion).

$$BIC = \underbrace{q \ln(n)}_{\text{depends on sample size } n.} - 2\ln L(\hat{\theta})$$

- Similar to AIC, but w/ even more penalization on having more predictors.
- A model w/ a smaller BIC is preferred in general.

Remarks :
- R^2_{adj}, AIC and BIC all try to prevent overfitting, which can lead to poor predictions
- All three methods can be used to compare fitted models.

④ Mean Square Prediction Error (MSPE) : focuses on model's predictive performance.
- Consider predictive performance of a model on new data (not the one used to fit model) : out-of-sample performance.
- Asks if model is generalizeable to other data.
- Overfitted models tend to have higher MSPE (via cross-validation)
  $\to$ next week.