ACTL3142 Week 3 - Multiple Linear Regression

Tadhg Xu-Glassop

2025T2

Multiple Linear Regression

Multiple Linear Regression

Relate p covariates linearly and additively to a quantitative response. That is,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

Estimation of coefficients, (most) assumptions, interpretation of coefficients all stays the same!

Estimating Coefficients (same as SLR)

Given a training set, we need assumptions for the ϵ term to estimate the β params.

Ordinary Least Squares (OLS) (weak assumptions)

- Assume errors have zero mean, constant variance and conditionally uncorrelated.
- Don't need distributional assumption for OLS.

MLE (strong assumptions)

- Assume $\epsilon_i \mid \mathbf{x} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$.
- Need a distributional assumption to construct our likelihood function and later construct confidence intervals.

Encoding Categorical Variables

Suppose a *categorical* predictor z can take on q different values. We need an alternative way to represent this in our model.

Encoding Categorical Variables

Suppose a *categorical* predictor z can take on q different values. We need an alternative way to represent this in our model.

Dummy-Encoding

1. Pick one possible value of z and make this the "base case." It will be reflected in β_0.
2. With the remaining $q - 1$ possible classes, make new predictors $x_1, x_2, \ldots, x_{q-1}$, which have value 1 if z is the respective category, otherwise 0.
3. Estimate $q - 1$ coefficients.

It is important to remember we only estimate $q - 1$ coefficients even though z can take on q unique values. **The resulting estimates are with respect to the base case.**

Comparing Different Models

We need ways to compare models with different predictors, and determine whether including features is actually adding valuable information or if we're just fitting to noise.

Comparing Different Models

We need ways to compare models with different predictors, and determine whether including features is actually adding valuable information or if we're just fitting to noise.

- F-test - test $\mathcal{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$, i.e. if the entire model is actually significant;
- Adjusted R^2 - similar to R^2, but penalty for adding more predictors (higher is better);
- C_p is an estimate for test error, adding penalty for more predictors (lower is better).
- AIC and BIC are alternates to log-likelihood which adds penalty for more predictors (lower is better).

Feature Selection

Forward Stepwise Selection

1. Start with the null model $\mathcal{M}_0$.
2. For $k \in [0, p-1]$: Consider all $p-k$ models which add one predictor to $\mathcal{M}_k$ and pick the best one via R^2.
3. Pick the best out of all the models found above via adjusted R^2, C_p and BIC.

Backward Stepwise selection is the reverse, where we begin with $\mathcal{M}_p$ and remove predictors that cause the least reduction in R^2.
Best subset selection considers all possible models.

Feature Selection

Forward Stepwise Selection

1. Start with the null model $\mathcal{M}_0$.
2. For $k \in [0, p-1]$: Consider all $p-k$ models which add one predictor to $\mathcal{M}_k$ and pick the best one via R^2.
3. Pick the best out of all the models found above via adjusted R^2, C_p and BIC.

Backward Stepwise selection is the reverse, where we begin with $\mathcal{M}_p$ and remove predictors that cause the least reduction in R^2.
Best subset selection considers all possible models.

Why do we expect the 'best' model to be somewhere in between $\mathcal{M}_0$ and $\mathcal{M}_p$?

Shortfalls in Linear Regression

We make a lot of assumptions in linear regression! Below are common violations/shortfalls and how to identify them.

1. Non-linearities between predictors and response.
   $\rightarrow$ Patterns in residual plots.
2. Correlations in error terms.
   $\rightarrow$ Correlations in error terms across fitted values.
3. Heteroskedastic errors (as opposed to homoskedastic).
   $\rightarrow$ Funnel shape in residual plots.
4. Outliers (violation of normal errors)
   Very different values in residual plots.
5. High-leverage points
   $\rightarrow$ Large Cook's distance.
6. Collinearity (violation of additive assumption)
   $\rightarrow$ High values of VIF.