Data Science II
(P8106)

Department of Biostatistics
Mailman School of Public Health
Columbia University

Spring 2024

Linear regression

- A simple approach for supervised learning
- Least Squares: Gauss (1809)
- Assumes that the dependence of Y on $X_1, \ldots, X_p$ is linear
- True regression functions are almost never linear
- But linear regression is still extremely useful both conceptually and practically
- Sometimes outperforms fancier nonlinear models

Simple linear regression

- Assume a model $Y = \beta_0 + \beta_1 X + \epsilon$
- Training data $\{(x_i, y_i), i = 1, \ldots, n\}$
- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we predict future response variables using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

- $\hat{y}$ indicates a prediction of Y on the basis of $X = x$

Estimation of the parameters by least squares

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the prediction for Y based on x_i
- $e_i = y_i - \hat{y}_i$ represents the i-th *residual*
- *Residual sum of squares* (RSS) as

$$RSS = e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2$$

- Least squares estimate: $\hat{\beta}_0, \hat{\beta}_1$ minimize the RSS
- The minimizing values are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \hat{\rho}_{XY} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

Assessing the overall accuracy of the model

- *R-squared* or fraction of variance explained is

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

  where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares
- In simple linear regression, $R^2 = \rho_{XY}^2$, where ρ_{XY} is the correlation between X and Y

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon,$$

- We estimate $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ as the values that minimize

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Least square estimates: $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ minimize RSS
- Given $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make predictions using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta} x_2 + \ldots + \hat{\beta}_p x_p$$

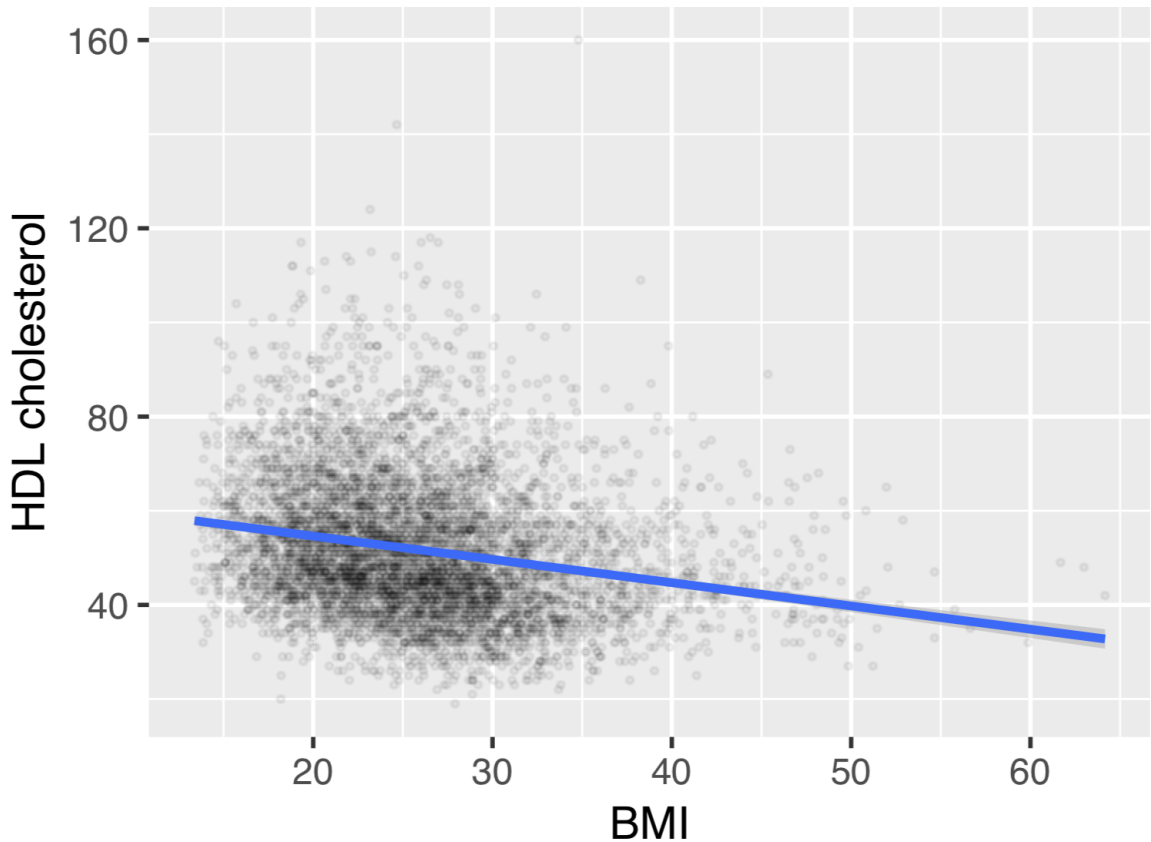
Multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon,$$

- We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed
- The ideal scenario is when the predictors are uncorrelated
  - Interpretations such as "a unit change in X_j is associated with a β_j change in Y, while all the other variables stay fixed", are possible

Example: the NHANES HDL cholesterol data

- Response: HDL cholesterol
- Predictors: BMI, age, gender, race, blood pressure



A small experiment

In R

- Fit a linear model `HDL` ~ `BMI + other predictors`

Then try the following steps

- Fit a linear model `BMI` ~ `other predictors` and get the residuals
- Fit a linear model `HDL` ~ `residual`

Any findings?

Interpreting regression coefficients

- Correlations amongst predictors can cause problems
  - The variance of coefficients tends to increase
  - Interpretations become hazardous – when X_j changes, everything else changes
- Claims of causality should be avoided for observational data

The Gauss-Markov Theorem

The least squares estimates of the parameters β have the smallest variance among all linear **unbiased** estimates

- Best linear unbiased estimator
- Is the restriction to unbiased estimates a wise one?
- A biased estimator with smaller mean squared error?
- Variable subset selection, ridge regression, ...

Deciding on the important variables

Why?

- Sacrifice a little bit of bias to reduce variance
- Sacrifice small details to get the big picture

How?

- All subsets or best subsets regression
  (Compute the least squares fit for all possible subsets and then choose between them based on some criterion)
- However we often can't examine all possible 2^p models
- We need an automated approach that searches through a subset of them
- Ad hoc ways: forward and backward selection

Forward selection

- Begin with the *null model* – only contains an intercept
- Fir p simple linear regressions and add to the null model the variable that results in the lowest RSS
- Add to that model the variable that results in the lowest RSS amongst all two-variable models
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold
- Can include variables early that later become redundant
- Can be used when $p > n$

Backward selection

- Start with all variables in the model
- Remove the variable with the largest p-value
- The new $(p-1)$ variable model is fit, and the variable with the largest p-value is removed
- Continue until a stopping rule is reached
- Can not be used when $p > n$

Model selection

- Keep things simple!
- Consider all subset regression
- Don't use forward/backward regression as the endgame
- More systematic criteria for model selection will be discussed later

Normally distributed errors?

- "*The regression assumption that is generally least important is that the errors are normally distributed. In fact, for the purpose of estimating the regression line (as compared to predicting individual data points), the assumption of normality is barely important at all.*"

  – Gelman, A., & Hill, J. (2006).

- When are normally distributed errors needed?

Prediction

Consider the prediction of the new response at input $X = x_0$

- $Y_0 = f(x_0) + \epsilon_0$
- The expected prediction error of $\widehat{f}(x_0)$ is

$$E(Y_0 - \widehat{f}(x_0))^2 = \sigma^2 + MSE(\widehat{f}(x_0))$$

- Prediction interval for a future observation Y_0
  - An interval in which one expects Y_0 to fall
  - Gaussian error!
  - Predicted value±(t-multiplier×standard error of the prediction)
- Compared to CI for the mean $f(x_0)$?

Prediction interval vs confidence interval

Take home messages

- Correlations amongst predictors can cause problems (e.g., large variance)
- The least square estimates are BLUE
- An optimal procedure is better than competitors in the same class, but that doesn't necessarily mean it is good
- BLUE can be better than other unbiased estimators, but all unbiased estimators may perform poorly
- One may consider estimators with some bias but much smaller variance
- How to determine which is better?