Model building

Linear Regression

Linear regression models

- Suppose we have an input vector $X^T = (X_1, \ldots, X_p)$ and want to predict a output Y. The linear regression model has the form

$$Y = f(x) + \epsilon = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

- The model is very easy to interpret, e.g., β_j is the average increase in Y when X_j increases by one unit holding all others constant.
- It is relatively flexible in the sense that X_j can come from different sources:
  - transformations of quantitative inputs, such as log;
  - be basis expansions, such as $X_2 = X_1^2$, X_3 X_1^3 , leading to a polynomial fit
  - dummy variable coding of the levels of qualitative inputs
  - interactions between variables, e.g. $X_3 = X_1 \cdot X_2$.

Linear regression model in a matrix form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np} + \varepsilon_p \end{bmatrix}$$

$$y = X\beta + \varepsilon \quad \text{where}$$

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & & x_{np} \end{bmatrix} \quad \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\varepsilon = Y - X\beta$$

$$RSS = \varepsilon^T \varepsilon = (Y - X\beta)^T (Y - X\beta)$$

Ordinary least squares estimates (OLS)

- The residual sum of squares (RSS) is given by

$$\mathrm{RSS}(\beta) = (Y - X\beta)^T(Y - X\beta)$$

- Differentiating with respect to β we obtain

$$\frac{\partial}{\beta}\mathrm{RSS}(\beta) = -2XT(Y - X\beta) = 0 \Rightarrow \hat{\beta}$$

- Assuming a full column rank on *X* and setting the first derivative to zero, we obtain the unbiased estimate

$$\hat{\beta} = (X^T\mathrm{X})^{-1}X^TY$$

OLS

- The fitted values at the training inputs are

$$\hat{Y} = X^T\hat{\beta} = X^T(X^T\mathrm{X})^{-1}X^TY$$

$\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY$ predicted value

projection matrix

Measures of fit: R^2 Statistic

- Some of the variation in *Y* can be explained by variation in the *X*'s and some cannot.
- R^2 tells you the fraction of variance that can be explained by *X*.
- R^2 is always between 0 and 1. Zero means no variance has been explained. One means it has all been explained (perfect fit to the data).
- However, it can still be challenging to determine what is a good R^2 values, and in general, this will depend on the application.

Remarks on OLS

- Easy to compute and interpretable

$$\hat{\beta} = (X'X)^{-1}X'y$$

must be invertible

Potential problems

When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

- The number of predictors p larger than the sample size n. (high dimensional setting)
- Collinearity.
- Non-linearity of the response-predictor relationships.
- Correlation of error terms.
- Non-constant variance of error terms.
- Outliers.
- High-leverage points.
- ……

$p > n$ ⇓ $(X'X)^{-1}$ is NOT unique.

$y = X\beta + \varepsilon$

$\varepsilon \sim N(0, \sigma^2)$ ↑ not constant

p > n and/or collinearity

(High-dimensional setting)

ill-conditioned setting.

- When p > n or there exist multicollinearity among the predictors, $(X^TX)^{-1}$ does not exist and/or is unstable; ⇒ OLS may fail.
- The regression coefficients $\hat{\beta}$ to determine these predictions are not unique.
- We lose our ability to meaningfully interpret the coefficients.
- What should we do for this case (Data pre-processing)

dimension reduction for p>n

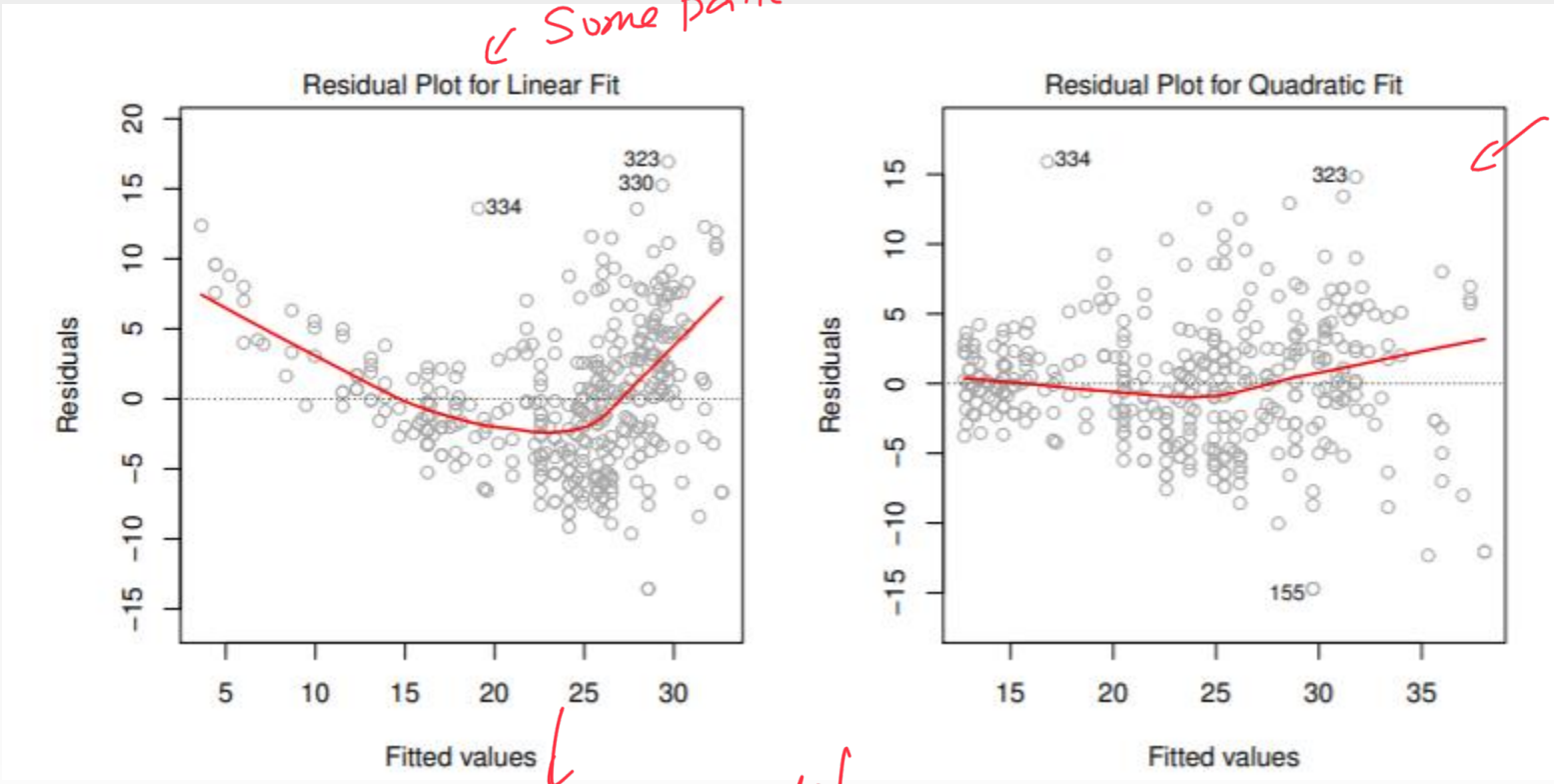
  - For p > n, we consider subsect selection or dimension reduction, such as PCA, etc.
  - For multicollinearity, we eliminate collinearity and/or diagnose multicollinearity using the variance inflation factor (VIF)
  - As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.

Non-linearity of the data

(Residual analysis)

- Based on the residual plots of the two models,

Some pattern for residual plot

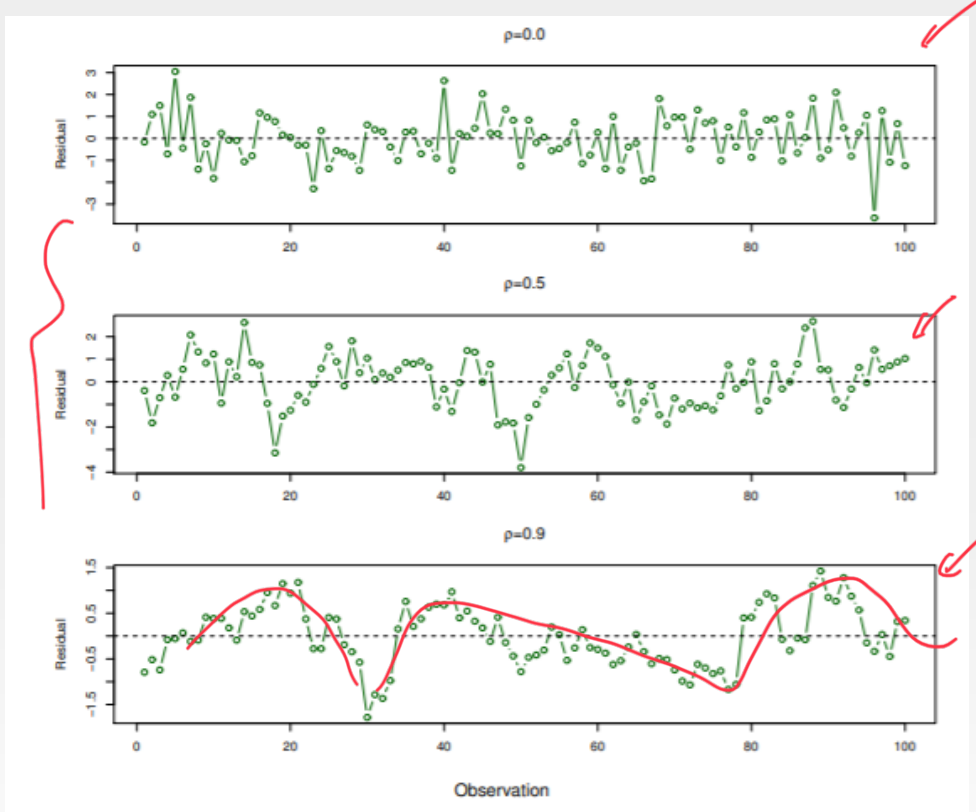


No significant pattern, indicating the linear model is appropriate.

- Which model is more linear?

Linear model may not appropriate!

Correlation of error terms



ρ=0

when ρ=0, no significant pattern

residual

ρ=0.5

some pattern

ρ=0.9

clear pattern

- If in fact there is correlation among the error terms, then the estimated standard errors will tend to underestimate the truth

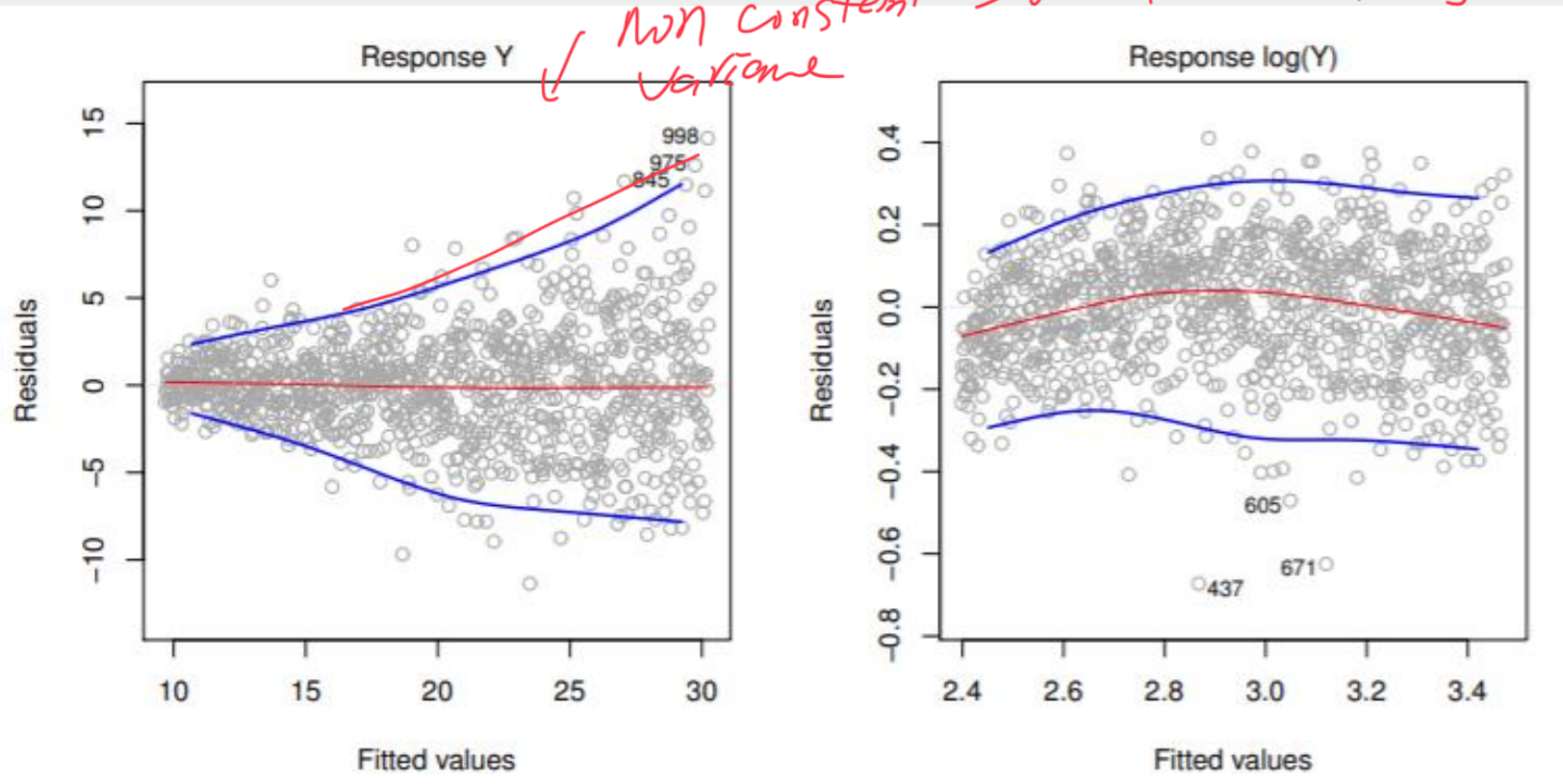
Non-constant variance of error terms

$y = x\beta + \varepsilon$

$Var(\varepsilon) = \sigma^2$

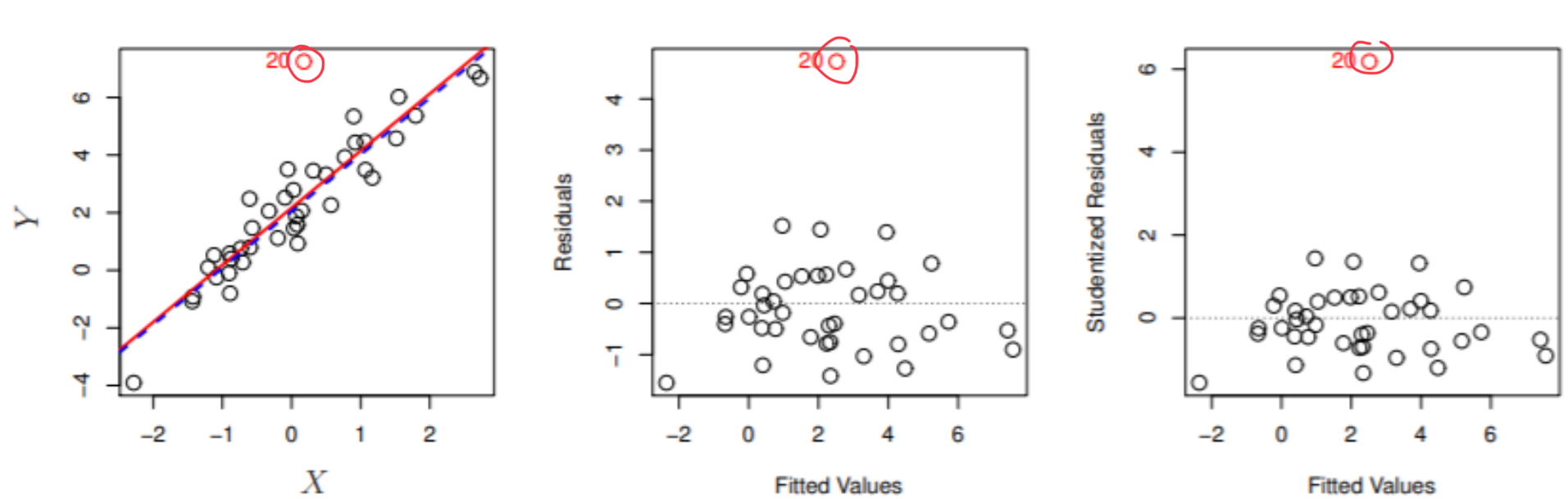
$= Var(y)$

Non constant variance → transformation for Y is required.



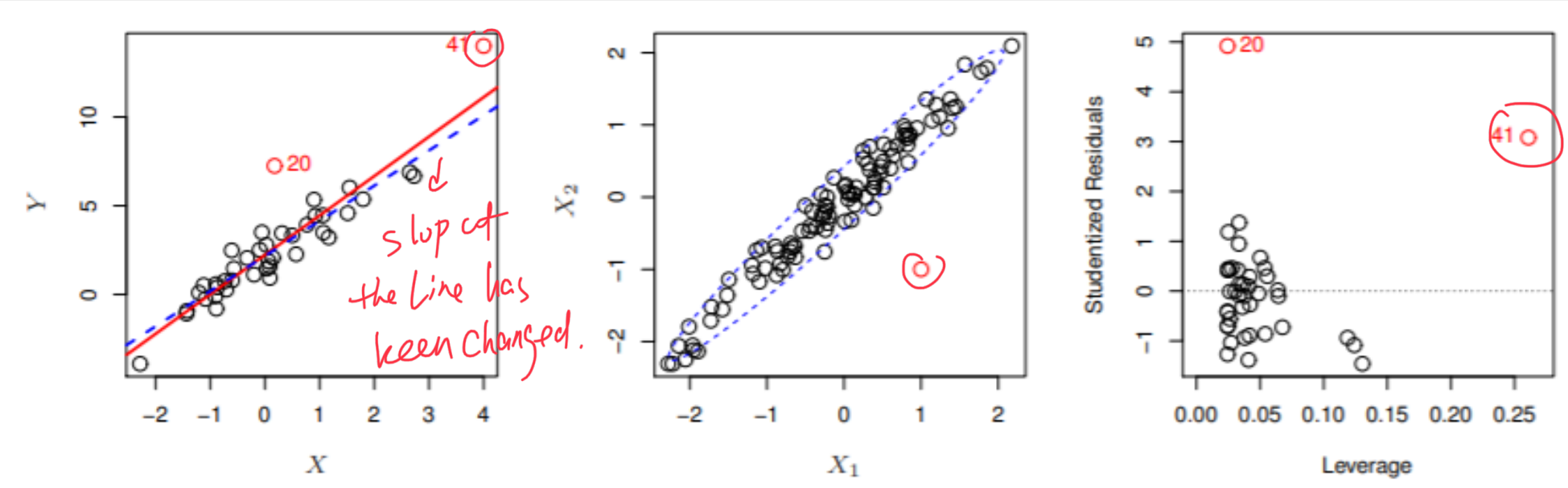
- A transformation of the response Y usually helps. Or consider more general methods that support heteroscedastic errors

Outliers



- Observations whose studentized residuals are greater than 3 in absolute value are possible outliers.

High-leverage points (influential points)



- LEFT: Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed.
- CENTER: High leverage
- RIGHT: Observation 41 has a high leverage and a high residual.

Conclusions

- These observations motivate to come alternative predictive models that can improve the linear model performance.
- There are 2 reasons we might not prefer to just use the ordinary least squares (OLS) estimates: *prediction accuracy* and *model interpretability*.

Prediction accuracy

- If n >>p, —that is, if *n*, the number of observations, is much larger than *p*, the number of variables—then the least squares estimates tend to also have low variance, and hence will perform well on test observations.
- But, when $n \approx p$, then the least squares fit can have high variance and may result in over fitting and poor estimates on unseen observations
- And, when $n < p$, then there is no longer a unique least squares coefficient estimate: the variance is infinite so the method cannot be used at all

Model interpretability

p is large (unimportant predictors in the model)

- When we have a large number of predictors *X* in the model there will generally be many that have little or no effect on *Y*
- Leaving these variables in the model makes it harder to see the "big picture", i.e., the effect of the "important variables".
- The model would be easier to interpret by removing (i.e. setting the coefficients to zero) the unimportant variables.

AIC, BIC, Mallow, Cp sequential F

Alternatives to least squares

- Subset Selection (*Chapter 19*)
  - Identifying a subset of all p predictors X that we believe to be related to the response Y, and then fitting the model using this subset
  - E.g. best subset selection and stepwise selection
- Dimension Reduction
  - Involves projecting all p predictors into an M-dimensional space where M < p, and then fitting linear regression model
  - E.g. Principal component regression, partial least squares regression
- Shrinkage (Penalization)
  - Involves shrinking the estimates coefficients towards zero
  - This shrinkage reduces the variance
  - Some of the coefficients may shrink to exactly zero, and hence shrinkage methods can also perform variable selection
  - E.g. Lasso or ENET