# Statistical Learning for Data Science

## Lecture 09

唐晓颖
电子与电气工程系
南方科技大学

March 20, 2023

# Multiple Linear Regression

1. Is there a relationship between the response and predictors?

- *The approach of using an F-statistic to test for any association between the predictors and the response works when p is relatively small, and certainly small compared to n.*

- *When p>n, we cannot fit the multiple linear regression model using least squares, so the F-statistic based approach does not work. This high-dimensional setting will be discussed in detail later.*

# Multiple Linear Regression

## 2. Deciding on important variables

If we conclude on the basis of the F-statistic based p-value that at least one of the predictors is related to the response, then it is natural to wonder which are the ones!

After selecting the important variables, we can fit a single model involving only those predictors.

***"Variable Selection"***

# Multiple Linear Regression

## 2. Deciding on important variables

*Variable selection*

Ideally, we would like to try out a lot of (and maybe all possibly) different models, each containing a different subset of the predictors, and then select the best one

- According to some criterion.
- Plotting various model outputs, such as the residuals, in order to search for patterns.

# Multiple Linear Regression

## 2. Deciding on important variables

*Variable selection*

Ideally, we would like to try out a lot of (and maybe all possibly) different models, each containing a different subset of the predictors, and then select the best one.

Not practical! We need an **automated** and **efficient** approach to choose a smaller set of models to consider!

# Multiple Linear Regression

## 2. Deciding on important variables

### *Variable selection*

#### Forward selection

- Begin with the *null model* – a model that contains an intercept but no predictors.
- Fit *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 among all new two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

# Multiple Linear Regression

## 2. Deciding on important variables

### *Variable selection*

#### Backward selection

- Start with all variables in the model.
- Remove the variable with the largest p-value – that is, the variable that is the least statistically significant.
- 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. For instance, we may stop when all remaining variables have a significant p-value defined by some significant threshold.

# Multiple Linear Regression

## 2. Deciding on important variables

*Variable selection*

Mixed selection (a combination of forward and backward selection)

- Start with no variables in the model, and then add the variable that provides the best fit (similar to forward selection).
- Add variables one-by-one until the p-value for one of the variables in the model rises above a certain threshold.
- Remove that variable from the model.
- Continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

# Multiple Linear Regression

## 2. Deciding on important variables

*Variable selection*

Comments on the three methods:

- Backward selection cannot be used if p>n.
- Forward selection can always be used.
- Forward selection might include variables early that later become redundant.

# Multiple Linear Regression

## 2. Deciding on important variables

*Variable selection*

- Later we discuss more systematic criteria for choosing an “optimal” member in the path of models produced by forward and backward stepwise selection.
- These include Mallow’s $C_p$ , Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted $R^2$ statistic, and Cross-validation (CV).

# Multiple Linear Regression

## 3. Assessing fit of the model

$R^2$ *Statistic (the fraction of variance explained)*

- In simple linear regression, we have $R^2 = [\mathrm{cor}(X, Y)]^2$

- In multiple linear regression, we have $R^2 = [\mathrm{cor}(Y, \hat{Y})]^2$ , the square of the correlation between the response and the fitted linear model; in fact, one property of the fitted linear model is that it maximizes this correlation among all possible linear models.

- An $R^2$ value close to 1 indicates that the model explains a large portion of the variance in the response variable.

# Multiple Linear Regression

## 3. Assessing fit of the model

$R^2$ *Statistic (the fraction of variance explained)*

Advertising data (sales versus TV, radio, and newspaper)

| Quantity | Value |
|---|---|
| Residual Standard Error | 1.69 |
| $R^2$ | 0.897 |
| F-statistic | 570 |

$R^2 = 0.8972$

# Multiple Linear Regression

## 3. Assessing fit of the model

$R^2$ *Statistic (the fraction of variance explained)*

Advertising data (sales versus TV and radio)

$R^2 = 0.89719$

- Even though the p-value for newspaper advertising is not significant (p=0.8599), there is still a small increase in $R^2$ if we include newspaper advertising in the model that already contains TV and radio advertising.

- The fact that there is only a tiny increase provides additional evidence that newspaper can be dropped from the model.

- Essentially, newspaper provides no real improvement in the model fit to the training samples, and its inclusion will likely lead to poor results on independent testing samples due to overfitting

# Multiple Linear Regression

## 3. Assessing fit of the model

$R^2$ *Statistic (the fraction of variance explained)*

- $R^2$ will always increase when more variables are added to the model, even if those variables are only weakly associated with the response.
- Adding another variable to the least squares equations allows us to fit the training data (though not necessarily the testing data) more accurately.
- The $R^2$ statistic, which is computed on the training data, must increase.

# Multiple Linear Regression

## 3. Assessing fit of the model

$R^2$ *Statistic (the fraction of variance explained)*

Advertising data (sales versus TV and radio)

$R^2 = 0.89719$

Advertising data (sales versus TV)

$R^2 = 0.61$

- Adding radio to the model leads to a substantial improvement in $R^2$ .
- A model that uses TV and radio to predict sales is substantially better than the one that uses only TV.

# Multiple Linear Regression

## 3. Assessing fit of the model

$RSE$

Advertising data (sales versus TV and radio)

$RSE = 1.681$

Advertising data (sales versus TV)

$RSE = 3.26$

This again shows that a model using TV and radio to predict sales is much more accurate (on the training data) than the one that only uses TV spending.

# Multiple Linear Regression

## 3. Assessing fit of the model

$RSE$

Advertising data (sales versus TV and radio)

$RSE = 1.681$

Advertising data (sales versus TV, radio, and newspaper)

$RSE = 1.686$

This again shows that there is no point in also using newspaper spending as a predictor in the model.

# Multiple Linear Regression

## 3. Assessing fit of the model

$RSE$

Advertising data (sales versus TV and radio)

$RSE = 1.681$

Advertising data (sales versus TV, radio, and newspaper)

$RSE = 1.686$

Why RSE increases when newspaper is added to the model given that RSS must decrease (more accurate)?

# Multiple Linear Regression

## 3. Assessing fit of the model

$$RSE$$

$$RSE = \sqrt{\frac{1}{n-p-1} RSS}$$

Models with more variables can have higher RSE if the decrease in RSS is small relative to the increase in p.

# Multiple Linear Regression

## 4. Assessing accuracy of the prediction

- Confidence intervals for the coefficient estimates
- Model bias (the assumption of a linear model may be biased)
- Prediction intervals (a combination of both the error in the estimation (the reducible error) and the uncertainty as to how much an individual point will differ from the population regression plane (the irreducible error))

# Multiple Linear Regression

## 4. Assessing accuracy of the prediction

The coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, .., \hat{\beta}_p$ are estimates for $\beta_0, \beta_1, .., \beta_p$.

That is, the *least squares plane*

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + .. + \hat{\beta}_p X_p$$

is only an estimate for the *true population regression plane*

$$f(X) = \beta_0 + \beta_1 X_1 + .. + \beta_p X_p$$

The inaccuracy in the coefficient estimates is related to the reducible error. We can compute a confidence interval in order to determine how close $\hat{Y}$ will be to $f(X)$

# Multiple Linear Regression

## 4. Assessing accuracy of the prediction

We use a confidence interval to quantify the uncertainty surrounding the average sales over a large number of cities.

For example, given that $100,000 is spent on TV advertising and $20,000 is spent on radio advertising in each city, the 95% confidence interval is [10,985, 11,528].

We interpret this to mean that 95% of intervals of this form will contain the true value of f(X).

In other words, if we collect a large number of data sets like the Advertising data set, and we construct a confidence interval for the average sales on the basis of each data set (given $100,000 in TV and $20,000 in radio advertising), then 95% of these confidence intervals will contain the true value of average sales.

# Multiple Linear Regression

## 4. Assessing accuracy of the prediction

In practice assuming a linear model for f(X) is almost always an approximation of reality, so there is an additional source of potentially reducible error which we call *model bias*.

# Multiple Linear Regression

## 4. Assessing accuracy of the prediction

Even if we knew f(X)—that is, even if we knew the true values for $\beta_0, \beta_1, .., \beta_p$ —the response value cannot be predicted perfectly because of the random error in the model (*irreducible error*).

How much will $Y$ vary from $\hat{Y}$? We use *prediction intervals* to answer this question.

Prediction intervals are always wider than confidence intervals, because they incorporate both the error in the estimate for f(X) (the reducible error) and the uncertainty as to how much an individual point will differ from the population regression plane (the irreducible error).

# Multiple Linear Regression

## 4. Assessing accuracy of the prediction

A prediction interval can be used to quantify the prediction uncertainty surrounding sales for a particular city.

Given that $100,000 is spent on TV advertising and $20,000 is spent on radio advertising in that city the 95% prediction interval is [7,930, 14,580].

We interpret this to mean that 95% of intervals of this form will contain the true value of Y for this city.

Note that both intervals are centered at 11,256, but that the prediction interval is substantially wider than the confidence interval, reflecting the increased uncertainty about sales for a given city in comparison to the average sales over many locations.

# Other considerations in the regression model

- Considerations in linear regression
  - Linear relationship
  - Variable selection
  - Data quality
  - Model evaluation
  - Model interpretation