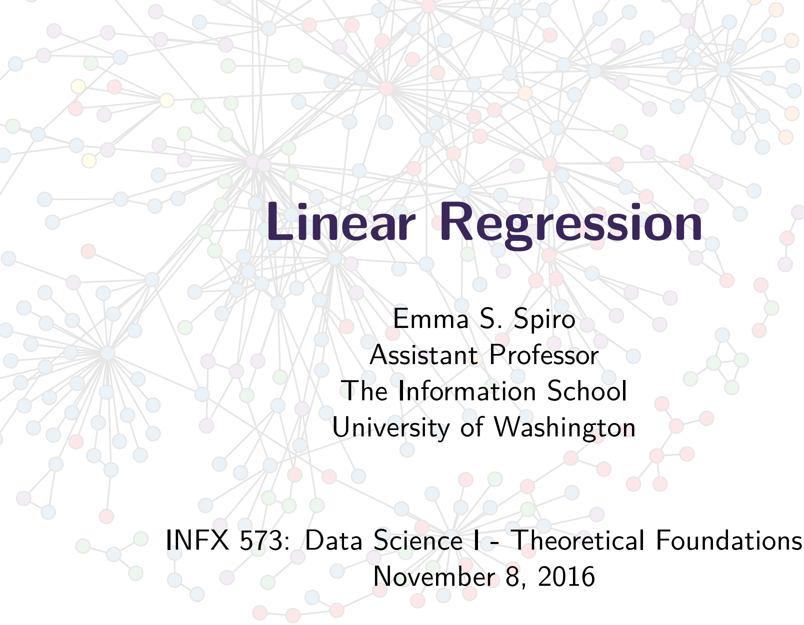
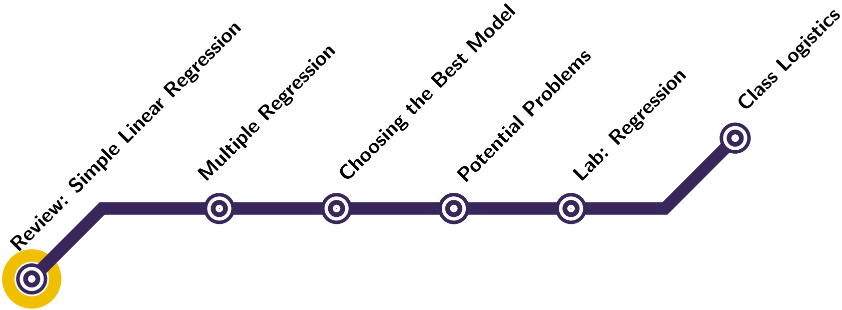


# Linear Regression

Emma S. Spiro
Assistant Professor
The Information School
University of Washington

INFX 573: Data Science I - Theoretical Foundations
November 8, 2016

# Today's Roadmap



Review: Simple Linear Regression

Multiple Regression

Choosing the Best Model

Potential Problems

Lab: Regression

Class Logistics

# Linear Regression

- Linear regression is a very powerful statistical technique
- Linear models can be used to predict a quantitative response

# Linear Regression

- Linear regression is a very powerful statistical technique
- Linear models can be used to predict a quantitative response
- Linear regression assumes that the relationship between two variables, x and y, can be modeled by a straight line:

$$Y = \beta X$$

where $\beta$ is a vector of model parameters, estimated from data.

# Linear Regression Motivating Example

- Advertising budgets are input variables, $X$ (predictors, independent variables, features)
- Sales is the output variable, $Y$ (response, dependent variable)

# Linear Regression Motivating Example

- Advertising budgets are input variables, $X$ (predictors, independent variables, features)
- Sales is the output variable, $Y$ (response, dependent variable)
- Suppose we observe a quantitative response $Y$, and predictor $X$, we assume there is some relationship between $Y$ and $X$,

$$Y = f(X) + \epsilon$$

# Linear Regression Motivating Example

- Advertising budgets are input variables, $X$ (predictors, independent variables, features)
- Sales is the output variable, $Y$ (response, dependent variable)
- Suppose we observe a quantitative response $Y$, and predictor $X$, we assume there is some relationship between $Y$ and $X$,

$$Y = f(X) + \epsilon$$

- Two main reasons that we may wish to estimate $f$: inference and prediction

# Linear Regression Motivating Example

On the basis of this data, suggest a marketing plan for next year that will result in high product sales.

# Linear Regression Motivating Example

On the basis of this data, suggest a marketing plan for next year that will result in high product sales.

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

# Simple Linear Regression: Prediction

- Very straightforward approach for predicting a quantitative response $Y$ on the basis of a single predictor variable $X$.
- It assumes that there is approximately a linear relationship between $X$ and $Y$.

$$\text{sales} \approx \beta_0 + \beta_1 \times \text{TV}$$

- $\beta_0$ and $\beta_1$ are two unknown constants that represent the intercept and slope terms in the linear model.
- Use training data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients.

# Simple Linear Regression: Estimating the Coefficients

- Want to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the resulting line is as close as possible to each of the data points

# Simple Linear Regression: Estimating the Coefficients

- Want to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the resulting line is as close as possible to each of the data points
- Many ways of measuring closeness, most common approach in SLR is to minimize the least squares criteron:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$$

$$\hat{\beta}_0 = \overline{Y}_n - \beta_1 \overline{X}_n$$

# Simple Linear Regression: Assessing Coefficient Estimates

- $Y = \beta_0 + \beta_1 X + \epsilon$ defines the population regression line, which is the best linear approximation of the true relationship between $X$ and $Y$
- The least squares coefficient estimates characterize the least squares line!

# Simple Linear Regression: Assessing Coefficient Estimates

- Even if we knew the true regression line (i.e. even if $\beta_0$ and $\beta_1$ were known), we would not be able to perfectly predict $Y$ from $X$!

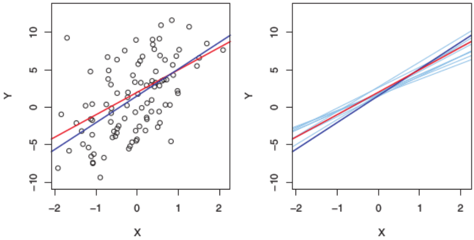


**FIGURE 3.3.** *A simulated data set.* Left: *The red line represents the true relationship, $f(X) = 2 + 3X$, which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for $f(X)$ based on the observed data, shown in black.* Right: *The population regression line is again shown in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.*

# Simple Linear Regression: Assessing Coefficient Estimates

- The distinction between the population regression line and the least squares line is a natural extention of using information about a sample to estimate characeristics of a population (e.g $\mu$ of a random variable $Y$)!

# Simple Linear Regression: Assessing Coefficient Estimates

- We can wonder how close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true values $\beta_0$ and $\beta_1$.
- We compute the standard errors for the estimates.

# Simple Linear Regression: Assessing Coefficient Estimates

- We can wonder how close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true values $\beta_0$ and $\beta_1$.
- We compute the standard errors for the estimates.
  - Standard errors can be used to compute confidence intervals.
  - Standard errors can also be used to perform hypothesis tests on the coefficients, e.g.

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

# Simple Linear Regression: Assessing Coefficient Estimates

- In practice we do this by computing a $t$- statistic,

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

- The $t$-distribution has a bell shape and for values of $n$ greater than approximately 30 it is quite similar to the normal distribution.
- Compute the probability of observing any value equal to $|t|$ or larger, assuming $\beta_1 = 0$.
- We call this the p-value.

# Simple Linear Regression: Prediction

- We can predict future sales on the basis of a particular value of TV advertising by computing

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

# Simple Linear Regression: Assessing Model Accuracy

- To what extent does the model fit the data?

# Simple Linear Regression: Assessing Model Accuracy

- To what extent does the model fit the data?
- Residual Standard Error (RSE): is an estimate of the standard deviation of $\epsilon$, in other words, it is the average amount that the response will deviate from the true regression line

# Simple Linear Regression: Assessing Model Accuracy

- To what extent does the model fit the data?
- Residual Standard Error (RSE): is an estimate of the standard deviation of $\epsilon$, in other words, it is the average amount that the response will deviate from the true regression line
- Large RSE indicates of lack of fit

# Simple Linear Regression: Assessing Model Accuracy

- To what extent does the model fit the data?
- Residual Standard Error (RSE): is an estimate of the standard deviation of $\epsilon$, in other words, it is the average amount that the response will deviate from the true regression line
- Large RSE indicates of lack of fit
- Not always clear what constitutes a “good” RSE value

# Simple Linear Regression: Assessing Model Accuracy

- $R^2$ statistic provides an alternative measure of fit
- Proportion of variance explained - independent of the scale of $Y$
- $R^2$ measures the proportion of variability in $Y$ that can be explained using $X$
- What is a good $R^2$ value will depend on the application

# Linear Regression Motivating Example

- In the advertising situation, we have examined the relationship between sales and TV advertising

# Linear Regression Motivating Example

- In the advertising situation, we have examined the relationship between sales and TV advertising
- But we also had data for the amount of money spent advertising on the radio and in newspapers

# Linear Regression Motivating Example

- In the advertising situation, we have examined the relationship between sales and TV advertising
- But we also had data for the amount of money spent advertising on the radio and in newspapers
- We may want to know whether either of any of these media is associated with sales

# Simple Linear Regression

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 7.0325 | 0.4578 | 15.36 | $< 0.0001$ |
| TV | 0.0475 | 0.0027 | 17.67 | $< 0.0001$ |

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 9.312 | 0.563 | 16.54 | $< 0.0001$ |
| radio | 0.203 | 0.020 | 9.92 | $< 0.0001$ |

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 12.351 | 0.621 | 19.88 | $< 0.0001$ |
| newspaper | 0.055 | 0.017 | 3.30 | $< 0.0001$ |

# Simple Linear Regression

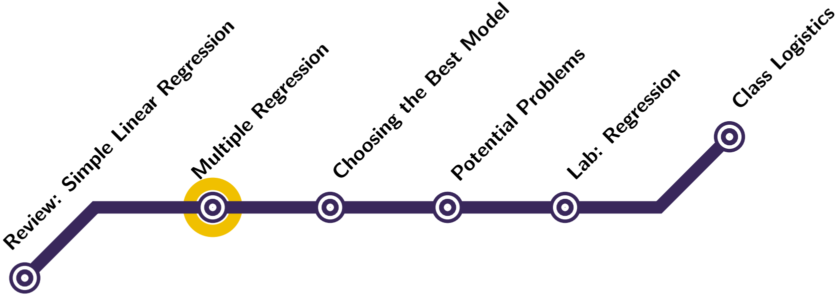
| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 7.0325 | 0.4578 | 15.36 | $< 0.0001$ |
| TV | 0.0475 | 0.0027 | 17.67 | $< 0.0001$ |

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 9.312 | 0.563 | 16.54 | $< 0.0001$ |
| radio | 0.203 | 0.020 | 9.92 | $< 0.0001$ |

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 12.351 | 0.621 | 19.88 | $< 0.0001$ |
| newspaper | 0.055 | 0.017 | 3.30 | $< 0.0001$ |

How would we make a single prediction for sales?

# Today's Roadmap



Review: Simple Linear Regression

Multiple Regression

Choosing the Best Model

Potential Problems

Lab: Regression

Class Logistics

# Multiple Regression

- Extend the simple linear regression model so that it can directly accommodate multiple predictors.

# Multiple Regression

- Extend the simple linear regression model so that it can directly accommodate multiple predictors.

- Multiple regression: possibility of more than one predictor

# Multiple Regression

- A multiple regression model is a linear model with many predictors.
- In general, we write the model as

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

when there are $p$ predictors.

# Multiple Regression

- A multiple regression model is a linear model with many predictors.
- In general, we write the model as

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

  when there are $p$ predictors.
- Unlike the simple linear regression estimates, the multiple regression coefficient estimates have somewhat complicated forms that are most easily represented using matrix algebra.
- The $\beta_i$ parameters are estimated using the same least squares approach, using a computer.

# Multiple Regression

- We interpret $\beta_i$ as the average effect on $Y$ of a one unit increase in $X_i$, holding all other predictors fixed.

# Simple vs. Multiple Regression

- Simple and multiple regression coefficients can be quite different.

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 12.351 | 0.621 | 19.88 | $< 0.0001$ |
| newspaper | 0.055 | 0.017 | 3.30 | $< 0.0001$ |

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 2.939 | 0.3119 | 9.42 | $< 0.0001$ |
| TV | 0.046 | 0.0014 | 32.81 | $< 0.0001$ |
| radio | 0.189 | 0.0086 | 21.89 | $< 0.0001$ |
| newspaper | -0.001 | 0.0059 | -0.18 | 0.8599 |

# Simple vs. Multiple Regression

- Simple and multiple regression coefficients can be quite different.

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 12.351 | 0.621 | 19.88 | $< 0.0001$ |
| newspaper | 0.055 | 0.017 | 3.30 | $< 0.0001$ |

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 2.939 | 0.3119 | 9.42 | $< 0.0001$ |
| TV | 0.046 | 0.0014 | 32.81 | $< 0.0001$ |
| radio | 0.189 | 0.0086 | 21.89 | $< 0.0001$ |
| newspaper | -0.001 | 0.0059 | -0.18 | 0.8599 |

# Multiple Regression

- In the simple regression case, the slope term represents the average effect of a $1,000 increase in newspaper advertising, ignoring other predictors such as TV and radio.

# Multiple Regression

- In the simple regression case, the slope term represents the average effect of a $1,000 increase in newspaper advertising, ignoring other predictors such as TV and radio.
- In the multiple regression setting, the coefficient for newspaper represents the average effect of increasing newspaper spending by $1,000 while holding TV and radio fixed.

# Relationships Between Variables

| | TV | radio | newspaper | sales |
|---|---|---|---|---|
| TV | 1.0000 | 0.0548 | 0.0567 | 0.7822 |
| radio | | 1.0000 | 0.3541 | 0.5762 |
| newspaper | | | 1.0000 | 0.2283 |
| sales | | | | 1.0000 |

# Relationships Between Variables

| | TV | radio | newspaper | sales |
|---|---|---|---|---|
| TV | 1.0000 | 0.0548 | 0.0567 | 0.7822 |
| radio | | 1.0000 | 0.3541 | 0.5762 |
| newspaper | | | 1.0000 | 0.2283 |
| sales | | | | 1.0000 |

- Tendency to spend more on newspaper advertising in markets where more is spent on radio advertising.

# Relationships Between Variables

| | TV | radio | newspaper | sales |
|---|---|---|---|---|
| TV | 1.0000 | 0.0548 | 0.0567 | 0.7822 |
| radio | | 1.0000 | 0.3541 | 0.5762 |
| newspaper | | | 1.0000 | 0.2283 |
| sales | | | | 1.0000 |

- Tendency to spend more on newspaper advertising in markets where more is spent on radio advertising.
- A SLR which only examines sales versus newspaper will observe that higher values of newspaper tend to be associated with higher values of sales, even though newspaper advertising does not actually affect sales.

# Relationships Between Variables

| | TV | radio | newspaper | sales |
|---|---|---|---|---|
| TV | 1.0000 | 0.0548 | 0.0567 | 0.7822 |
| radio | | 1.0000 | 0.3541 | 0.5762 |
| newspaper | | | 1.0000 | 0.2283 |
| sales | | | | 1.0000 |

- Tendency to spend more on newspaper advertising in markets where more is spent on radio advertising.
- A SLR which only examines sales versus newspaper will observe that higher values of newspaper tend to be associated with higher values of sales, even though newspaper advertising does not actually affect sales.
- Newspaper sales are a surrogate for radio advertising; newspaper gets “credit” for the effect of radio on sales.

# Multiple Regression: Important Questions

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
2. Do all the predictors help to explain $Y$, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

# Multiple Regression: Important Questions

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
2. Do all the predictors help to explain $Y$, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

# Multiple Regression: Is There a Relationship?

- We test the null hypothesis,

  $$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

  versus the alternative

  $$H_a : \text{ at least one } \beta_i \text{ is non-zero.}$$

- This hypothesis test is performed by computing the F-statistic,

  $$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}$$

- When there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1.

# Reminder: Sums of Squares

- RSS: residual sum of squares

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

- TSS: total sum of squares

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

# Multiple Regression: Is There a Relationship?

- Why not look at each predictor individually?

# Multiple Regression: Is There a Relationship?

- Why not look at each predictor individually?
- 5% of the p-values associated with each variable will be below 0.05 by chance!

# Multiple Regression: Is There a Relationship?

- Why not look at each predictor individually?
- 5% of the p-values associated with each variable will be below 0.05 by chance!
- If we use the individual t-statistics and associated p-values in order to decide whether or not there is any association between the variables and the response, there is a very high chance that we will incorrectly conclude that there is a relationship.

# Multiple Regression: Is There a Relationship?

- 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 $n$.
- If $p > n$ then there are more coefficients $\beta_i$ to estimate than observations from which to estimate them.
- In this case we cannot even fit the multiple linear regression model using least squares.

# Multiple Regression: Important Questions

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
2. Do all the predictors help to explain $Y$, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

# Multiple Regression: Which Predictors are Important?

- How might we select only the important variables and build a model with only those?

# Multiple Regression: Which Predictors are Important?

- How might we select only the important variables and build a model with only those?
- This is called variable selection.

# Multiple Regression: Which Predictors are Important?

- How might we select only the important variables and build a model with only those?
- This is called variable selection.
- We can try out a bunch of models with various combinations of predictors.

# Multiple Regression: Which Predictors are Important?

- How might we select only the important variables and build a model with only those?
- This is called variable selection.
- We can try out a bunch of models with various combinations of predictors.
- Which one is best?

# Multiple Regression: Model Selection

- Statistics to judge the quality of a model
  - Mallow's $C_p$
  - Akaike information criterion (AIC)
  - Bayesian information criterion (BIC)
  - Adjusted $R^2$.
- Also look at model diagnostics!

# Multiple Regression: Model Selection

- Statistics to judge the quality of a model
  - Mallow's $C_p$
  - Akaike information criterion (AIC)
  - Bayesian information criterion (BIC)
  - Adjusted $R^2$.
- Also look at model diagnostics!
- But there are $2^p$ possible subsets to consider...

# Multiple Regression: Model Selection

- Statistics to judge the quality of a model
  - Mallow's $C_p$
  - Akaike information criterion (AIC)
  - Bayesian information criterion (BIC)
  - Adjusted $R^2$.
- Also look at model diagnostics!
- But there are $2^p$ possible subsets to consider...
- Forward and backward model selection

# Multiple Regression: Important Questions

1. Is at least one of the predictors $X_{,}X_2, \ldots, X_p$ useful in predicting the response?
2. Do all the predictors help to explain $Y$, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

# Multiple Regression: Model Fit

- $R^2$ will always increase when more variables are added to the model, even if those variables are only weakly associated with the response.

# Multiple Regression: Model Fit

- $R^2$ will always increase when more variables are added to the model, even if those variables are only weakly associated with the response.
- Adjusted $R^2$ is modified to take model complexity into account (i.e. how many predictors are included)

# Multiple Regression: Important Questions

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
2. Do all the predictors help to explain $Y$, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

# Multiple Regression: Predictions

- Straightforward to predict the response $Y$ on the basis of a set of values for the predictors $X_1, X_2, \ldots, X_p$.

# Multiple Regression: Qualitative Predictors

- If a qualitative predictor only has two levels, we simply create an indicator or dummy variable that takes on two possible numerical values.
  For example, dichotomous gender variable:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if ith person is female} \\ \beta_0 + \epsilon_i & \text{if ith person is male} \end{cases}$$

# Multiple Regression: Qualitative Predictors

- $\beta_0$ can be interpreted as the average response among males, $\beta_0 + \beta_1$ as the average response among females, and $\beta_1$ as the average difference in between females and males.

# Multiple Regression: Qualitative Predictors

- $\beta_0$ can be interpreted as the average response among males, $\beta_0 + \beta_1$ as the average response among females, and $\beta_1$ as the average difference in between females and males.
- When a qualitative predictor has more than two levels, we can create additional dummy variables, one for each possible level.

# Multiple Regression: Qualitative Predictors

- $\beta_0$ can be interpreted as the average response among males, $\beta_0 + \beta_1$ as the average response among females, and $\beta_1$ as the average difference in between females and males.
- When a qualitative predictor has more than two levels, we can create additional dummy variables, one for each possible level.
- There will always be one fewer dummy variable than the number of levels.
- The level with no dummy variable is known as the baseline.

# Multiple Regression: Interaction Terms

- If we increase $X_1$ by one unit, then $Y$ will increase by an average of $\beta_1$ units. $X_2$ does not alter this statement.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

# Multiple Regression: Interaction Terms

- If we increase $X_1$ by one unit, then $Y$ will increase by an average of $\beta_1$ units. $X_2$ does not alter this statement.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

# Multiple Regression: Interaction Terms

- If we increase $X_1$ by one unit, then $Y$ will increase by an average of $\beta_1$ units. $X_2$ does not alter this statement.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

# Multiple Regression: Interaction Terms

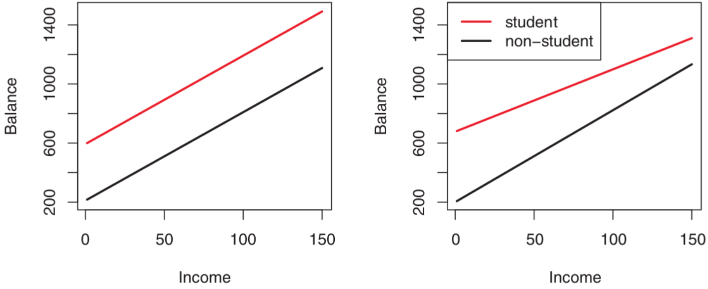


**FIGURE 3.7.** *For the* Credit *data, the least squares lines are shown for prediction of* balance *from* income *for students and non-students.* Left: *The model (3.34) was fit. There is no interaction between* income *and* student. Right: *The model (3.35) was fit. There is an interaction term between* income *and* student.

James et al. (2014)

# Multiple Regression: Confidence and Prediction Intervals

- Even if we knew $f(X)$, the true values for $\beta_0, \beta_1, \ldots, \beta_p$, the response value cannot be predicted perfectly because of the random error in the model

# Multiple Regression: Confidence and Prediction Intervals

- Even if we knew $f(X)$, the true values for $\beta_0, \beta_1, \ldots, \beta_p$, the response value cannot be predicted perfectly because of the random error in the model
- How much will $Y$ vary from $\hat{Y}$?

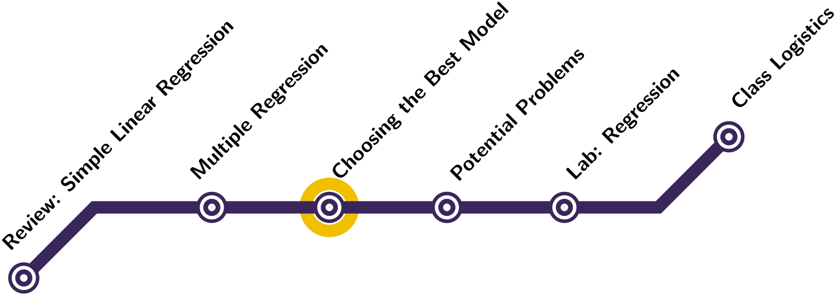
# Multiple Regression: Confidence and Prediction Intervals

- Even if we knew $f(X)$, the true values for $\beta_0, \beta_1, \dots, \beta_p$, the response value cannot be predicted perfectly because of the random error in the model
- How much will $Y$ vary from $\hat{Y}$?
- Prediction intervals: a range prediction, see the `predict()` function in R

# Multiple Regression: Confidence and Prediction Intervals

- Even if we knew $f(X)$, the true values for $\beta_0, \beta_1, \dots, \beta_p$, the response value cannot be predicted perfectly because of the random error in the model
- How much will $Y$ vary from $\hat{Y}$?
- Prediction intervals: a range prediction, see the `predict()` function in R
- Prediction intervals are always wider than confidence intervals, because they incorporate both the error in the estimate for $f(X)$ and the uncertainty as to how much an individual point will differ from the population regression plane.

# Today's Roadmap



# Regression: Variable Selection

- Possible that all of the predictors are associated with the response, but more often the response is only related to a subset of the predictors.
- Task of determining which predictors are associated with the response is referred to as variable selection.
- Ideally: perform variable selection by trying out a lot of different models, each containing a different subset of the predictors.
- We can then select the best model out of all of the models that we have considered.

# Regression: Variable Selection

- Possible that all of the predictors are associated with the response, but more often the response is only related to a subset of the predictors.
- Task of determining which predictors are associated with the response is referred to as variable selection.
- Ideally: perform variable selection by trying out a lot of different models, each containing a different subset of the predictors.
- We can then select the best model out of all of the models that we have considered.
- How do we determine which model is best?

# Regression: Model Selection

- Unfortunately, there are a total of $2^p$ models

# Regression: Model Selection

- Unfortunately, there are a total of $2^p$ models
- If $p = 2$, then there are $2^2 = 4$ models to consider

# Regression: Model Selection

- Unfortunately, there are a total of $2^p$ models
- If $p = 2$, then there are $2^2 = 4$ models to consider
- But if $p = 30$, then we must consider $2^{30} = 1,073,741,824$ models! This is not practical.

# Regression: Model 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 for the new two-variable model.
- This approach is continued until some stopping rule is satisfied.

# Regression: Model Selection

Backward selection:

- Start with all variables in the model, and 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.
- This procedure continues until a stopping rule is reached. For instance, we may stop when all remaining variables have a p-value below some threshold.

# Regression: Model Selection

Mixed selection:

- Combination of forward and backward selection.
- Start with no variables in the model, and as with forward selection, we add the variable that provides the best fit.
- Continue to add variables one-by-one.
- If at any point the p-value for one of the variables in the model rises above a certain threshold, then we 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.

# Regression: What is the best model?

- RSS of these models decreases monotonically, and the $R^2$ increases monotonically, as the number of features included in the models increases.
- Therefore, if we use these statistics to select the best model, then we will always end up with a model involving all of the variables.
- The problem is that a low RSS or a high $R^2$ indicates a model with a low *training error*, whereas we wish to choose a model that has a low *test error*

# Regression: Model Selection

- To select the best model with respect to test error, we need to estimate this test error.
- There are two common approaches:
  1. We can indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting.
  2. We can directly estimate the test error, using either a validation set approach or a cross-validation approach.

# Regression: Model Selection

How do we determine which model is best?

- Recall, MSE is generally an underestimate of the test MSE - we specifically estimate the least squares regression coefficients such that the training RSS is as small as possible

# Regression: Model Selection

How do we determine which model is best?

- Recall, MSE is generally an underestimate of the test MSE - we specifically estimate the least squares regression coefficients such that the training RSS is as small as possible
- Various statistics can be used to select from models of different sizes - adjusting the training error for the model size:
  - Mallow's $C_p$
  - Akaike information criterion (AIC)
  - Bayesian information criterion (BIC)
  - Adjusted $R^2$

# Regression: What is the best model?

- Mallow's $C_p$: For a fitted least squares model containing $d$ predictors, the $C_p$ estimate of test MSE is:

  $$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

  where $\hat{\sigma}^2$ is an estimate of the variance of the error
- $C_p$ statistic adds a penalty to the training RSS in order to adjust for the fact that the training error tends to underestimate the test error.

# Regression: What is the best model?

- AIC criterion is defined for a large class of models fit by maximum likelihood:

$$AIC = \frac{1}{2n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

- BIC is derived from a Bayesian point of view, but ends up looking similar to $C_p$ (and AIC) as well:

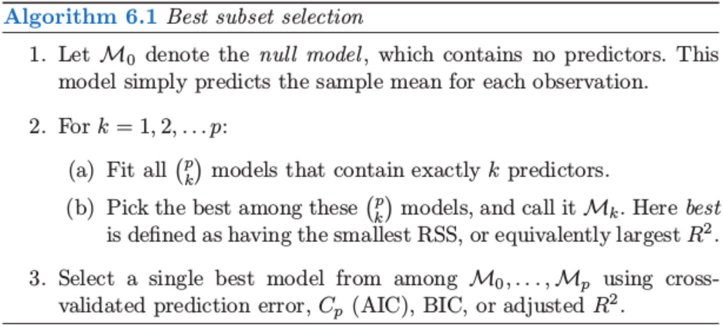
$$BIC = \frac{1}{n}(RSS + log(n)d\hat{\sigma}^2)$$

# Regression: What is the best model?

- Adjusted $R^2$ statistic is another popular approach for selecting among a set of models that contain different numbers of variables.

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

# Regression: Model Selection



**Algorithm 6.1** *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 1, 2, \ldots p$:

   (a) Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.

   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest $R^2$.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$.

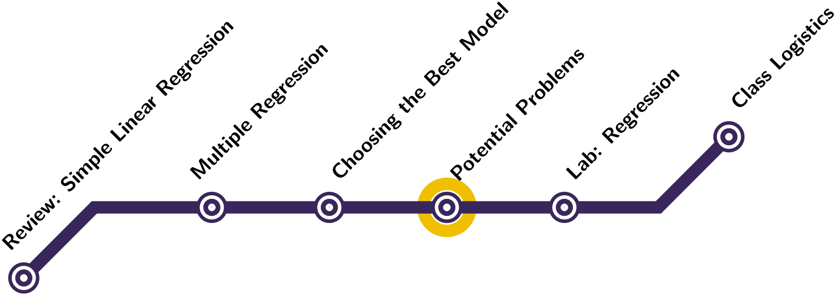
# Regression: Model Selection

- You can perform stepwise selection (forward, backward, both) using the `stepAIC()` function from the `MASS` package.
- Backward selection cannot be used if $p > n$, while forward selection can always be used.
- You can perform all-subsets regression using the `leaps( )` function from the `leaps` package

# Linear Model Selection

- Shrinkage Methods: Fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero.
  - ridge regression
  - lasso
- Dimension Reduction Methods: class of approaches that transform the predictors and then fit a least squares model using the transformed variables.
  - principal components regression
  - partial least squares

# Today's Roadmap



# Regression: Potential Problems

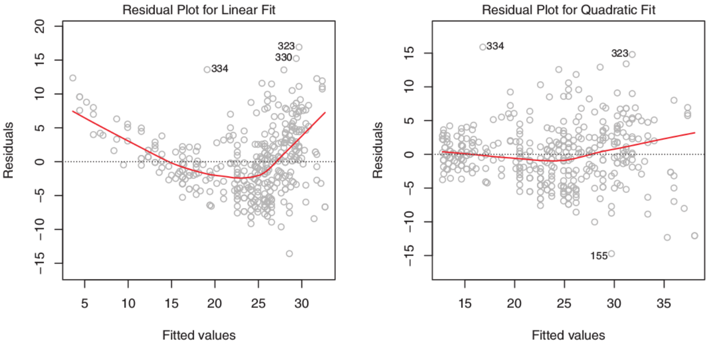
1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

# Non-Linearity of the Data

- Linear regression model assumes that there is a straightline relationship between the predictors and the response.

# Non-Linearity of the Data

- Linear regression model assumes that there is a straightline relationship between the predictors and the response.
- Residual plots are a useful graphical tool for identifying non-linearity, residuals versus fitted values.



James et al. (2014)

# Correlation of the Error Terms

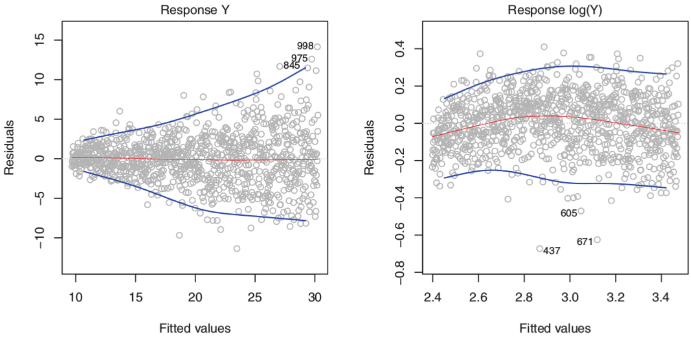
- Important assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$, are uncorrelated.

# Correlation of the Error Terms

- Important assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$, are uncorrelated.
- If in fact there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors.

# Non-Constant Variance in Error Terms

- Non-constant variances in the errors is called heteroscedasticity
- We can identify it from the presence of a funnel shape in the residual plot.



James et al. (2014)

# Outliers

- An outlier is a point for which $y_i$ is far from the value predicted by the model.

# Outliers

- An outlier is a point for which $y_i$ is far from the value predicted by the model.
- Residual plots can be used to identify outliers.

# Outliers

- An outlier is a point for which $y_i$ is far from the value predicted by the model.
- Residual plots can be used to identify outliers.
- If we believe that an outlier has occurred due to an error in data collection or recording, then one solution is to simply remove the observation.

# Outliers

- An outlier is a point for which $y_i$ is far from the value predicted by the model.
- Residual plots can be used to identify outliers.
- If we believe that an outlier has occurred due to an error in data collection or recording, then one solution is to simply remove the observation.
- But be careful with dropping data!

# Outliers

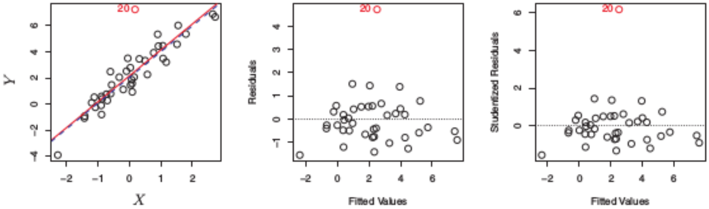


**FIGURE 3.12.** Left: *The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue.* Center: *The residual plot clearly identifies the outlier.* Right: *The outlier has a studentized residual of 6; typically we expect values between −3 and 3.*

James et al. (2014)

# High Leverage Points

- Observations with high leverage have an unusual value for $x_i$

# High Leverage Points

- Observations with high leverage have an unusual value for $x_i$
- High leverage observations tend to have a big impact on the estimated regression line

# High Leverage Points

- Observations with high leverage have an unusual value for $x_i$
- High leverage observations tend to have a big impact on the estimated regression line
- We can use a leverage statistics to find observations with high leverage.

# Outliers

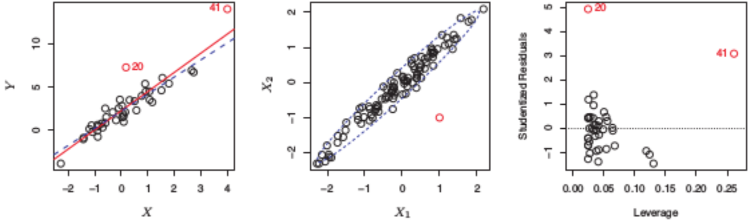


**FIGURE 3.13.** Left: *Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed.* Center: *The red observation is not unusual in terms of its $X_1$ value or its $X_2$ value, but still falls outside the bulk of the data, and hence has high leverage.* Right: *Observation 41 has a high leverage and a high residual.*

James et al. (2014)

# Collinearity

- Collinearity refers to the situation in which two or more predictor variables are closely related to one another - highly correlated.

# Collinearity

- Collinearity refers to the situation in which two or more predictor variables are closely related to one another - highly correlated.
- Can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response.

# Collinearity

- Collinearity refers to the situation in which two or more predictor variables are closely related to one another - highly correlated.
- Can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response.
- Collinearity reduces the accuracy of the estimates of the regression coefficients, and this the power of the hypothesis test is reduced.

# Collinearity

- Collinearity refers to the situation in which two or more predictor variables are closely related to one another - highly correlated.
- Can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response.
- Collinearity reduces the accuracy of the estimates of the regression coefficients, and this the power of the hypothesis test is reduced.
- Look at the correlation matrix of the predictors.
- Look at the Variance Inflation Factor (VIF) for each predictor.

# Regression: Errors

- The accuracy of $\hat{Y}$ as a prediction for $Y$ depends on two quantities, the *reducible error* and the *irreducible error*

# Regression: Errors

- The accuracy of $\hat{Y}$ as a prediction for $Y$ depends on two quantities, the *reducible error* and the *irreducible error*
- $\hat{f}$ will not be a perfect estimate for $f$, and this inaccuracy will introduce some error - reducible because we can potentially improve the accuracy of $\hat{f}$ by using the most appropriate statistical learning technique to estimate $f$

# Regression: Errors

- The accuracy of $\hat{Y}$ as a prediction for $Y$ depends on two quantities, the *reducible error* and the *irreducible error*
- $\hat{f}$ will not be a perfect estimate for $f$, and this inaccuracy will introduce some error - reducible because we can potentially improve the accuracy of $\hat{f}$ by using the most appropriate statistical learning technique to estimate $f$
- Variability associated with $\epsilon$ also affects the accuracy of our predictions - irreducible error

# Regression: Bias Variance Trade Off

- Consider a given estimate $\hat{f}$ and a set of predictors $X$, which yields the prediction $\hat{Y} = \hat{f}(X)$.
- Assume $\hat{f}$ and $X$ are fixed:

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2$$

$$= [f(X) - \hat{f}(X)]^2 + Var(\epsilon)$$

# Regression: Assessing Model Accuracy

- There is no free lunch in statistics: no one method dominates all others over all possible data sets.

# Regression: Assessing Model Accuracy

- There is no free lunch in statistics: no one method dominates all others over all possible data sets.
- Important task: for any given set of data which method produces the best results?

# Regression: Assessing Model Accuracy

- In the regression setting, the most commonly-used measure is the mean squared error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

- The MSE is computed using the training data that was used to fit the model, and so should more accurately be referred to as the training MSE

# Regression: Assessing Model Accuracy

- But in general, we do not really care how well the method works on the training data.
- We are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data!

# Regression: Assessing Model Accuracy

- But in general, we do not really care how well the method works on the training data.
- We are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data! We want to choose the method that gives the lowest test MSE, as opposed to the lowest training MSE.

# Regression: Assessing Model Accuracy

- How can we go about trying to select a method that minimizes the test MSE?

# Regression: Assessing Model Accuracy

- How can we go about trying to select a method that minimizes the test MSE?
- We may have a test data set available - observations that were not used to train the statistical learning method

# Regression: Assessing Model Accuracy

- How can we go about trying to select a method that minimizes the test MSE?
- We may have a test data set available - observations that were not used to train the statistical learning method
- No guarantee that the method with the lowest training MSE will also have the lowest test MSE.

# Regression: Assessing Model Accuracy

- How can we go about trying to select a method that minimizes the test MSE?
- We may have a test data set available - observations that were not used to train the statistical learning method
- No guarantee that the method with the lowest training MSE will also have the lowest test MSE.
- As model flexibility increases, training MSE will decrease, but the test MSE may not.

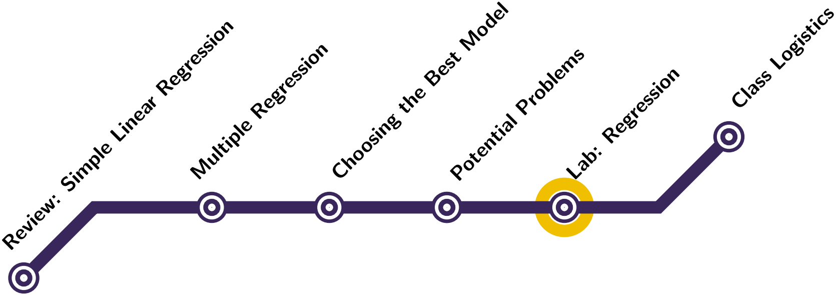
# Regression: Assessing Model Accuracy

- How can we go about trying to select a method that minimizes the test MSE?
- We may have a test data set available - observations that were not used to train the statistical learning method
- No guarantee that the method with the lowest training MSE will also have the lowest test MSE.
- As model flexibility increases, training MSE will decrease, but the test MSE may not.
- When a given method yields a small training MSE but a large test MSE, we are said to be overfitting the data.

# Regression: Assessing Model Accuracy

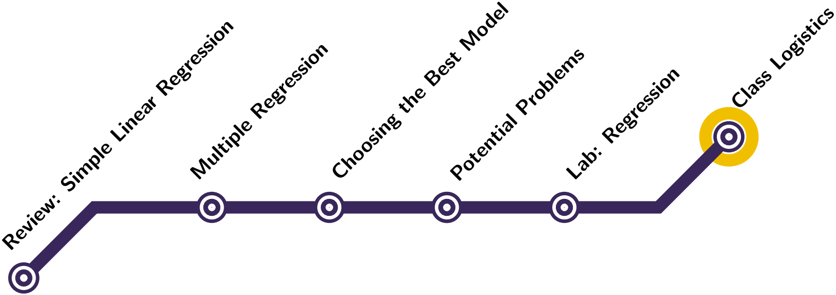
- How can we go about trying to select a method that minimizes the test MSE?
- We may have a test data set available - observations that were not used to train the statistical learning method
- No guarantee that the method with the lowest training MSE will also have the lowest test MSE.
- As model flexibility increases, training MSE will decrease, but the test MSE may not.
- When a given method yields a small training MSE but a large test MSE, we are said to be overfitting the data.
- One important method is cross-validation which is a method for estimating test MSE using the training data.

# Today's Roadmap



# Lab: Linear Regression

# Today’s Roadmap



# Questions?