BIG DATA ANALYTICS
Linear Regression

Olga Klopp
klopp@essec.edu

Linear Regression

- Useful and widely used statistical learning method
- A good jumping-off point for newer approaches: many fancy statistical learning approaches can be seen as generalizations or extensions of linear regression

Outline

Case study

Simple Linear Regression

Multiple Linear Regression

Extensions of the Linear Model

Potential Problems

Summary: The Marketing Plan

Parametric vs non-parametric approaches

Classification

Outline

Case study

- Provide advice to a client on how to improve sales of a particular product
- The Advertising data set consists of the sales of the product in 200 different markets, along with advertising budgets for the product for three different media: TV, radio, and newspaper
- A marketing plan for next year that will result in higher product sales

Case study

- The advertising budgets are **input variables** (predictors, independent variables, features, or just variables):
  - X_1 is the TV budget
  - X_2 is the radio budget
  - X_3 the newspaper budget
- Sales is an **output variable** (response or dependent variable): Y

Case study: important questions

1. *Is there a relationship between advertising budget and sales?*
   - First goal: determine whether the data provide evidence of an association between advertising expenditure and sales
   - If the evidence is weak, then one might argue that no money should be spent on advertising!
2. *How strong is the relationship between advertising budget and sales?*
   - Assuming that there is a relationship between advertising and sales: the strength of this relationship?
     - Given a certain advertising budget, can we predict sales with a high level of accuracy? $\rightarrow$ strong relationship
     - Is a prediction of sales based on advertising expenditure only slightly better than a random guess? $\rightarrow$ a weak relationship.

Case study: important questions

3. *Which media contribute to sales?*
   - Do all three media - TV, radio, and newspaper - contribute to sales, or do just one or two of the media contribute?
   - Find a way to separate out the individual effects of each medium when the money has been spent on all three media
4. *How accurately can we estimate the effect of each medium on sales?*
   - For every dollar spent on advertising in a particular medium, by what amount will sales increase?
   - How accurately can we predict this amount of increase?

Case study: important questions

5. *How accurately can we predict future sales?*
   - For any given level of television, radio, or newspaper advertising, what is our prediction for sales?
   - What is the accuracy of this prediction?
6. *Is the relationship linear?*
   - If there is approximately a linear relationship between advertising expenditure in the various media and sales, then linear regression is an appropriate tool
   - If not: it may be possible to transform the predictor or the response so that linear regression can be used.

Case study: important questions

7. *Is there synergy among the advertising media?*

   - Perhaps spending \$50,000 on television advertising and \$50,000 on radio advertising results in more sales than allocating \$100,000 to either television or radio individually
   - In marketing: **synergy effect**; in statistics: **interaction effect**

Outline

Simple Linear Regression

- A 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:

$$Y \approx \beta_0 + \beta_1 X$$

- Here "$\approx$" as is approximately modeled as
- We are **regressing** Y on X
- E.g., X may represent TV advertising and Y may represent sales
- β_0 and β_1 two unknown constants that represent the **intercept** and **slope** terms: the model coefficients or parameters.

Simple Linear Regression

- Training data: estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients
- We can predict future sales on the basis of a particular value of TV advertising:

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

- $\hat{y}$ indicates a prediction of Y on the basis of $X = x$
- Symbol $\widehat{\ }$ to denote the estimated value for an unknown parameter or the predicted value of the response.

Estimating the Coefficients

- In practice β_0 and β_1 are unknown
- We use data to estimate the coefficients:

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

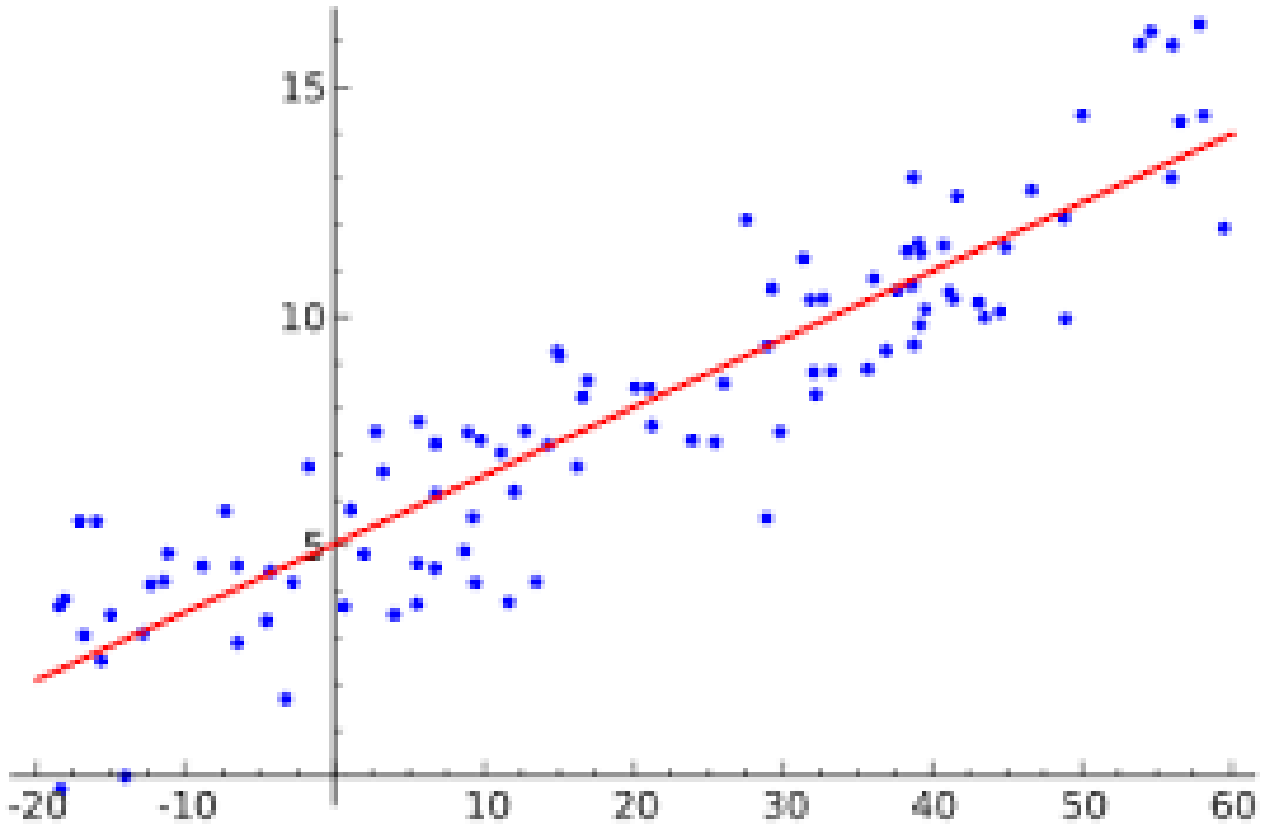
n observation pairs (measurement of X and a measurement of Y)

  - advertising example: the TV advertising budget and product sales in $n = 200$ different markets

- Goal: obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ s.t. the linear model fits the available data well

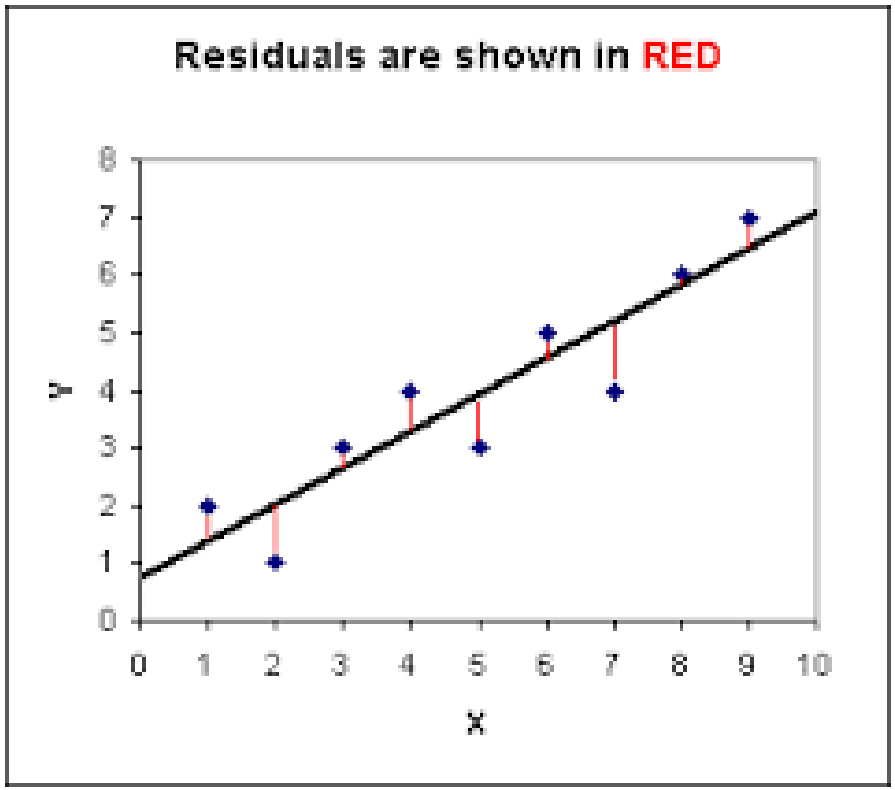
Estimating the Coefficients

- We want to find an intercept $\hat{\beta}_0$ and a slope $\hat{\beta}_1$ s. t. the resulting line is as close as possible to the data points:



Estimating the Coefficients

- There are a number of ways of measuring closeness
- The most common approach: minimizing the least squares criterion
- The fit minimize the sum of the squares of the errors:



Estimating the Coefficients

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X
- Then $e_i = y_i - \hat{y}_i$ is the ith *residual*
- **Residual Sum of Squares**:

$$\text{RSS} = e_1^2 + e_2^2 + ... + e_n^2$$

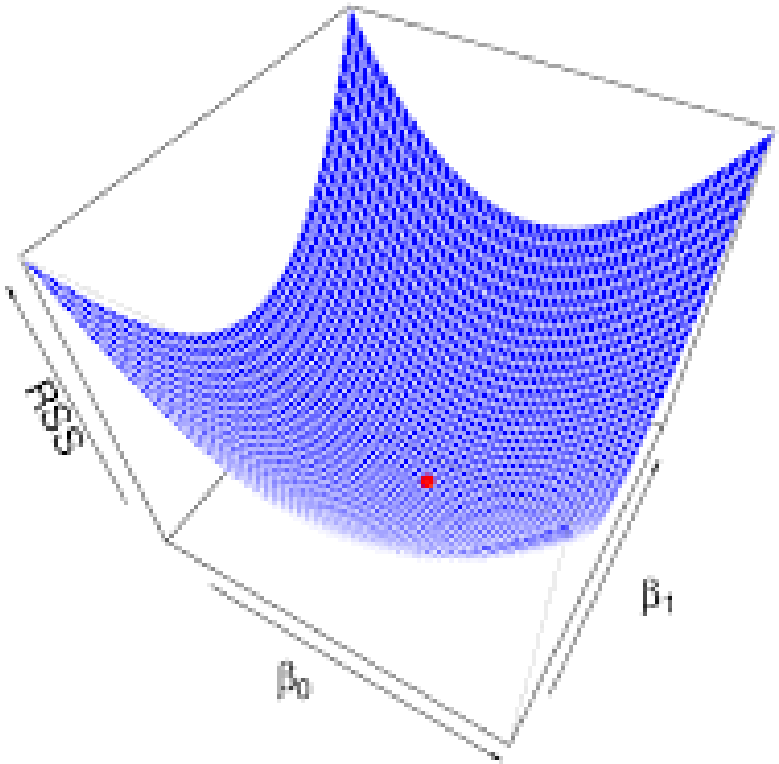
- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS

Estimating the Coefficients

The minimizers are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the sample means

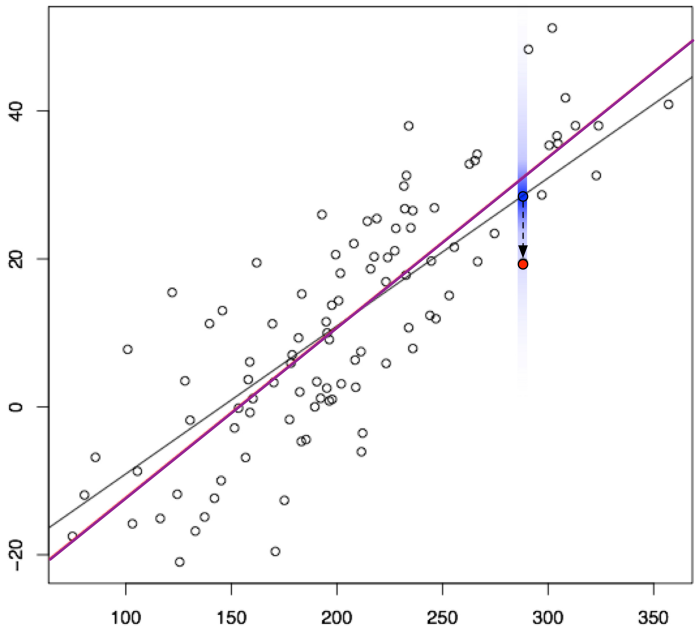


Assessing the Accuracy of the Coefficient Estimates

$$Y = \beta_0 + \beta_1 X + \epsilon$$

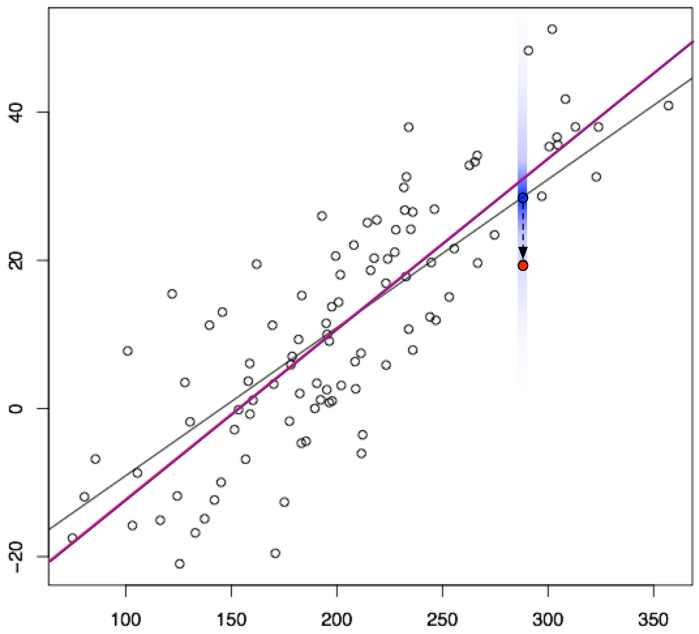
- ϵ is a mean-zero random error term
- A catch-all for what we miss with this simple model:
  - the true relationship is probably not linear
  - there may be other variables that cause variations in Y
  - measurement errors
  - we typically assume that the error term is independent of X

Population Regression Line



- The red line represents the true relationship $f(X) = -1 + 0.5X$ - the *population regression line*
- The grey line is the least squares estimate for $f(X)$ based on the observed data

Population Regression Line



- In real applications, we have access to a set of observations from which we can compute the least squares line
- The population regression line is unobserved.

Mean estimation

- A natural extension of the standard statistical approach of using information from a sample to estimate characteristics of a large population
- E.g., we are interested in knowing the mean μ of some random variable Y
- μ is unknown, but we do have access to n observations from Y: $y_1, ..., y_n$
- Estimate μ using the sample mean $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$
- In general $\mu \neq \bar{y}$ but close

Unbiased estimator

The sample mean $\bar{y}$ is an *unbiased* estimator of μ: on average, $\bar{y}$ is equal μ

- On the basis of one particular set of observations, $\bar{y}$ might overestimate μ, and on the basis of another set of observations, $\bar{y}$ might underestimate μ
- If we average a huge number of estimates of μ obtained from a huge number of sets of observations, then this average would exactly equal μ

  An unbiased estimator does not systematically over- or under-estimate the true parameter.
- The property of unbiasedness holds for the least squares coefficient estimates

Assessing the Accuracy of a Estimate

- The average of $\bar{y}$'s over many data sets will be very close to μ
- A single estimate $\bar{y}$ may be a substantial underestimate or overestimate of μ
- How far off will that single estimate be?

Computing the standard error of $\bar{y}$

Residual standard error

How close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true values β_0 and β_1?

- The standard errors associated with $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\mathrm{SE}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$$

$$\mathrm{SE}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $\sigma^2 = Var(\epsilon)$

Residual standard error

- We assume that the errors ϵ_i for each observation are uncorrelated with common variance σ^2
- When it is not true the formula still turns out to be a good approximation
- $\mathrm{SE}(\hat{\beta}_1)$ is smaller when the x_i are more spread out: we have more leverage to estimate a slope when this is the case
- In general, σ^2 is not known, but can be estimated from the data: *residual standard error*

$$\mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}$$

Confidence Interval

- Standard errors can be used to compute **confidence intervals**:

  A 95 % confidence interval is defined as a range of values such that with 95 % probability, the range will contain the true unknown value of the parameter

- For linear regression, the 95 % confidence interval for β_1 approximately takes the form

$$[\hat{\beta}_1 - 2\text{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2\text{SE}(\hat{\beta}_1)]$$

  there is approximately a 95% chance that this interval will contain the true value of β_1

Hypothesis tests

Standard errors can also be used to perform *hypothesis tests* on the coefficients:

- The most common hypothesis test:

  H_0 : There is no relationship between X and Y

  versus the alternative hypothesis

  H_a : There is some relationship between X and Y

- Mathematically: $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$
- To test the null hypothesis, we need to determine whether $\hat{\beta}_1$, our estimate for β_1, is sufficiently far from zero that we can be confident that β_1 is non-zero.

Hypothesis tests

How far is far enough?

- It depends on the accuracy of $\hat{\beta}_1$ - that is, it depends on $\mathrm{SE}(\hat{\beta}_1)$:
  - If $\mathrm{SE}(\hat{\beta}_1)$ is small, then even relatively small values of $\hat{\beta}_1$ may provide strong evidence that $\beta_1 \neq 0$
  - If $\mathrm{SE}(\hat{\beta}_1)$ is is large, then $\hat{\beta}_1$ must be large in absolute value in order to reject the null hypothesis.

Hypothesis tests

- In practice, we compute a t-statistic:

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

  It measures the number of standard deviations that $\hat{\beta}_1$ is away from 0

- If there is no relationship between X and Y, then we expect to have a t-distribution with $n-2$ degrees of freedom

  The degrees of freedom is the number of data rows minus the number of coefficients fit

- The t-distribution has a bell shape and for values of $n \geq 30$ it is quite similar to the normal distribution

Hypothesis tests

- The probability of observing any value equal to $|t|$ or larger, assuming $\beta_1 = 0$: **the p-value**
- *A small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance, in the absence of any real association between the predictor and the response*
- If we see a small p-value $\rightarrow$ there is an association between the predictor and the response (we reject the null hypothesis)
- Typical p-value cutoffs for rejecting the null hypothesis are 5 or 1%.

Assessing the Accuracy of the Model

- Assume that we have rejected the null hypothesis in favor of the alternative hypothesis, that is we accepted the hypothesis that there is some relationship between X and Y
- Quantify the extent to which the model fits the data?
- The quality of a linear regression fit is typically assessed using two related quantities: the residual standard error (RSE) and the R^2 statistic.

Residual Standard Error

- We have errors in our model, which implies that even if we knew the true regression line (i.e. even if β_0 and β_1 were known), we would not be able to perfectly predict Y from X
- The RSE is an estimate of the standard deviation of the noise
- It is the average amount that the response will deviate from the true regression line:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Residual Standard Error

The RSE is a measure of the lack of fit of the model to the data:

- If the predictions obtained using the model are very close to the true outcome values - that is, if $\hat{y}_i \approx y_i$ for $i = 1, \ldots, n$ - then RSE will be small: we can conclude that the model fits the data well
- If $\hat{y}_i$ is very far from y_i for one or more observations, then the RSE may be quite large: the model doesn't fit the data well.

R^2 Statistic

- The RSE provides an absolute measure of lack of fit of the model to the data
- It is measured in the units of $Y \implies$ it is not always clear what constitutes a good RSE
- R^2 statistic provides an alternative measure of fit
- It takes the form of a proportion: value between 0 and 1
  - It is independent of the scale of Y .

R^2 Statistic

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

- $\text{TSS} = \sum(y_i - \bar{y})^2$ is the total sum of squares
  - TSS measures the total variance in the response Y
  - The amount of variability inherent in the response before the regression is performed
- $\text{RSS} = \sum(y_i - \hat{y}_i)^2$
  - RSS measures the amount of variability that is left unexplained after performing the regression
- TSS - RSS measures the amount of variability in the response that is explained by performing the regression
- *R^2 measures the proportion of variability in Y that can be explained using X*

R^2 Statistic

- An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression
- A number near 0 indicates that the regression did not explain much of the variability in the response
- The R^2 statistic has an interpretational advantage over the RSE
- It can still be challenging to determine what is a good R^2 value: this will depend on the application.

R^2 Statistic

- In certain problems in physics, we may know that the data truly comes from a linear model
  - we would expect to see an R^2 value that is extremely close to 1
  - a substantially smaller R^2 value might indicate a serious problem with the experiment in which the data were generated
- In typical applications in biology, psychology, marketing, the linear model is an extremely rough approximation to the data
  - residual errors due to other unmeasured factors are often very large
  - we would expect only a very small proportion of the variance in the response to be explained by the predictor
  - A R^2 value below 0.1 might be quite realistic!

R^2 Statistic

- R^2 statistic is a measure of the linear relationship between X and Y
- Correlation r^2 is also a measure of the linear relationship between X and Y
- In the simple linear regression setting: $R^2 = r^2$
- For the multiple linear regression problem (several predictors simultaneously) the concept of correlation between the predictors and the response does not extend automatically
- We can still use R^2 statistics.

Outline

Case study

Simple Linear Regression

Multiple Linear Regression

Extensions of the Linear Model

Potential Problems

Summary: The Marketing Plan

Parametric vs non-parametric approaches

Classification

Multiple Linear Regression

- We often have more than one predictor
- E.g., in the advertising data, we have also the data for the amount of money spent advertising on the radio and in newspapers
- Either of these two media is associated with sales?
- Extend our analysis in order to accommodate these two additional predictors

Multiple Linear Regression

- Three separate simple linear regressions, each of which uses a different advertising medium as a predictor:
  - How to make a single prediction of sales given levels of the three advertising media budgets, since each of the budgets is associated with a separate regression equation?
  - Each of the three regression equations ignores the other two media in forming estimates for the regression coefficients
  - If the media budgets are correlated with each other, this can lead to very misleading estimates of the individual media effects on sales.

Multiple Linear Regression

Better approach:

- Extend the simple linear regression model to accommodate multiple predictors
- Each predictor is associated with a separate slope coefficient in a single model
- p distinct predictors:

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

- X_j represents the jth predictor and β_j quantifies the association between X_j and the response.
- β_j is the average effect on Y of a one unit increase in X_j (holding all other predictors fixed).

Estimating Regression Coefficients

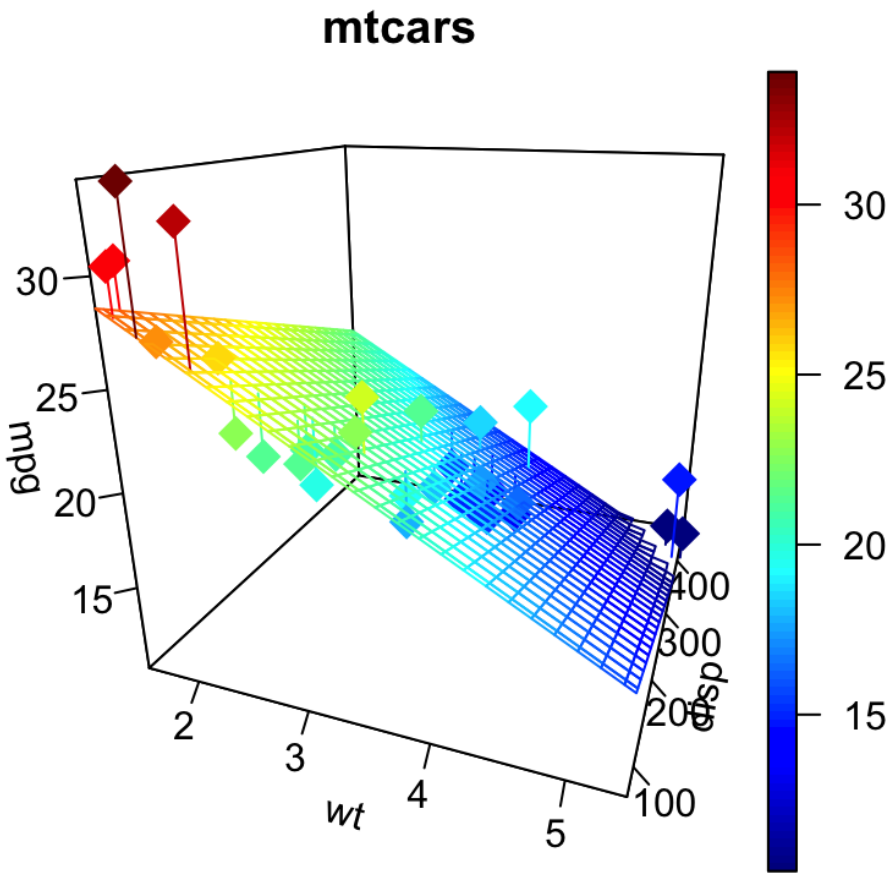
- The regression coefficients are unknown, and must be estimated
- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make predictions:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

- The parameters are estimated using the same least squares approach:
  - we choose $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ to minimize the sum of squared residuals

$$\text{RSS} = \sum (y_i - \hat{y}_i)^2$$

Estimating Regression Coefficients



Two predictors and one response:

- the least squares regression line becomes a plane
- it is chosen to minimize the sum of the squared vertical distances between each observation and the plane

Important Questions

1. Is at least one of the predictors $X_1, X_2, ..., X_p$ 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?

Is There a Relationship Between the Response and Predictors?

- In the simple linear regression setting: we check whether $\beta_1 = 0$
- In the multiple regression setting: p predictors
  - all of the regression coefficients are zero
- As in the simple linear regression setting, we use a **hypothesis test**:

  $\mathrm{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

  versus the alternative

  H_a : at least one β_j is non-zero.

F-statistic

- This hypothesis test is performed by computing the *F-statistic*:

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

where $\text{TSS} = \sum(y_i - \bar{y})^2$ and $\text{RSS} = \sum(y_i - \hat{y}_i)^2$

- F-statistic measures whether the linear regression model predicts outcome better than the constant mode (the mean value of y)

F-statistic

- The F-statistic gets its name from the F-test:

  F-test is the technique used to check if two variances - in this case, the variance of the residuals from the constant model and the variance of the residuals from the linear model - are significantly different

- The corresponding p-value is the estimate of the probability that we would've observed an F-statistic this large or larger if the two variances in question are in reality the same

F-statistic

F-statistic formula:

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

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

$$\mathbb{E}\left(\text{RSS}/(n-p-1)\right) = \sigma^2 \quad \text{and} \quad \mathbb{E}\left((\text{TSS} - \text{RSS})/p\right) = \sigma^2$$

- On the other hand, if H_a is true, then

$$\mathbb{E}(\text{TSS} - \text{RSS})/p > \sigma^2 \implies F > 1$$

F-statistic

- Large F-statistic suggests that at least one of the predictors must be related to the response variable
- How large does the F-statistic need to be in order to reject H_0?
- It depends on the values of n and p:
  - when n is large, an F-statistic that is just a little larger than 1 might still provide evidence against H_0
  - a larger F-statistic is needed to reject H_0 if n is small

F-statistic

- When H_0 is true and the errors ϵ_i have a normal distribution, the F-statistic follows an F-distribution
- If the errors are not normally-distributed, the F-statistic approximately follows an F-distribution provided that the sample size n is large
- For any given value of n and p, any statistical software package can be used to compute the p-value associated with the F-statistic using this distribution
- Based on this p-value, we can determine whether or not to reject H_0

Hypothesis test

- We can test that a particular subset of q of the coefficients are zero:

$$\mathrm{H}_0 : \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0$$

- F-statistic:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p-1)}$$

where RSS_0 is the residual sum of squares for the model that uses all the variables except last q

Hypothesis test

- Assume that we preformed simple linear regression for each variable
- Given individual p-values for each variable, do we still need to look at the overall F-statistic?
  - it seems likely that if any one of the p-values for the individual variables is very small, then at least one of the predictors is related to the response

This logic is erroneous (especially when the number of predictors p is large)!

Hypothesis test

- E.g., $p = 100$ and $\mathrm{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ is true, so no variable is truly associated with the response
- For $p = 100$ about 5 % of the p-values associated with each variable will be below 0.05 by chance
- We expect to see approximately five small p-values even in the absence of any true association between the predictors and the response
- 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
- F-statistic does not suffer from this problem because it adjusts for the number of predictors.

High-dimensional setting

- F-statistic works when p is relatively small, and small compared to n
- Sometimes we have a very large number of variables
- If $p > n$ then there are more coefficients β_j to estimate than observations from which to estimate them
- We cannot fit the multiple linear regression model using least squares and the F-statistic cannot be used!
- When p is large we are in **high-dimensional setting**.

Deciding on Important Variables

- Assume that we concluded that at least one of the predictors is related to the response $\rightarrow$ which are the guilty ones?
- We could look at the individual p-values but, if p is large we are likely to make some false discoveries
- It is possible that all of the predictors are associated with the response
- More often: the response is only related to a subset of the predictors

Variable selection: the task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors

Variable selection: classical approaches

- First attempt: to perform variable selection by trying out a lot of different models, each containing a different subset of the predictors:
  - E.g., if $p = 2$, then we can consider four models:
    1. a model containing no variables
    2. a model containing X_1 only
    3. a model containing X_2 only
    4. a model containing both X_1 and X_2
- Select the best model out of all of the models:

  Mallows C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R^2...

Variable selection: classical approaches

- With p variables, there are a total of 2^p models that contain subsets of p variables: if $p = 30$, then we must consider $2^{30} = 1,073,741,824$ models!
- Even for moderate p, trying out every possible subset of the predictors is infeasible
- We need an automated and efficient approach to choose a smaller set of models to consider
- There are three classical approaches for this task:
  - *Forward selection*
  - *Backward selection*
  - *Mixed selection*

Forward selection

- We begin with the null model - starting with no variables in the model
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS
- Repeating this process until none improves the model to a statistically significant extent
- Forward selection is a greedy approach and might include variables that later become redundant

Backward selection

- Start with all variables in the model
- Remove the variable with the largest p-value (the variable that is the least statistically significant)
- 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:
  - e.g., we may stop when all remaining variables have a p-value below some threshold
- Backward selection cannot be used if $p > n$

Mixed selection

- A 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
- We continue to add variables one-by-one
- The p-values for variables can become larger as new predictors are added to the model
- 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
  - all variables outside the model would have a large p-value if added to the model

Model Fit

- Two of the most common numerical measures of model fit: RSE and R^2
- They are computed and interpreted in the same fashion as for simple linear regression:
  - e.g., an R^2 value close to 1 indicates that the model explains a large portion of the variance in the response variable
- In addition it can be useful to plot the data: graphical summaries can reveal problems with a model that are not visible from numerical statistics.

Model Fit

- Example Advertising data:
  - The model that uses all three advertising media to predict sales has an $R^2 = 0.8972$
  - The model that uses only TV and radio to predict sales has an $R^2 = 0.89719$
  - There is only a small increase in R^2 if we include newspaper advertising in the model that already contains TV and radio advertising

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):
  - 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 also computed on the training data, must increase
- The fact that adding newspaper advertising to the model containing only TV and radio advertising leads to just a tiny increase in R^2 provides evidence that newspaper can be dropped from the model

Predictions

- Once we have fit the multiple regression model we can predict the response Y on the basis of a set of values for the predictors $X_1, X_2, ..., X_p$
- In practice assuming a linear model is always an approximation of reality
- Using a linear model, we are estimating the best linear approximation to the true surface.

Predictions

- For advertising data we use a *confidence interval* to quantify the uncertainty surrounding the average sales over a large number of cities
  - E.g., 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 $[10.985, 11.528]$
  - Interpretation: with probability 95 % this interval contains the true value of $f(X)$
- To quantify the uncertainty for sales for a particular city: a *prediction interval*
  - E.g. the 95% prediction interval is $[7.930, 14.580]$
  - Interpretation: with probability 95 % this interval contains the true value of Y for this city

Predictions

- The 95 % confidence interval is

$$[10.985, 11.528]$$

- The 95% prediction interval is

$$[7.930, 14.580]$$

- The prediction interval is substantially wider than the confidence interval: the increased uncertainty about sales for a given city in comparison to the average sales over many locations

Qualitative Predictors

- We have assumed that all variables in our linear regression model are quantitative
- Often some predictors are qualitative: gender, marital status...
- If a qualitative predictor (a factor) only has two possible values: create an *indicator or dummy variable*:
  - it takes on two possible numerical values:

$$x_i = \begin{cases} 0 & \text{if } i\text{th person is female} \\ 1 & \text{if } i\text{th person is male} \end{cases}$$

- Use this variable as a predictor in the regression equation

Qualitative Predictors

- When a qualitative predictor has more than two levels, a single dummy variable cannot represent all possible values
- We can create additional dummy variables:
  - e.g., for the ethnicity variable we create two dummy variables X_1 for Asian and X_2 for Caucasian
  - Both are used in the regression equation

Outline

Extensions of the Linear Model

- The standard linear regression model provides interpretable results and works quite well on many real-world problems
- It makes several highly restrictive assumptions that are often violated in practice
- Two most important assumptions:

The relationship between the predictors and response is additive and linear

Removing the Additive Assumption

- Standard linear regression model with two variables,

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

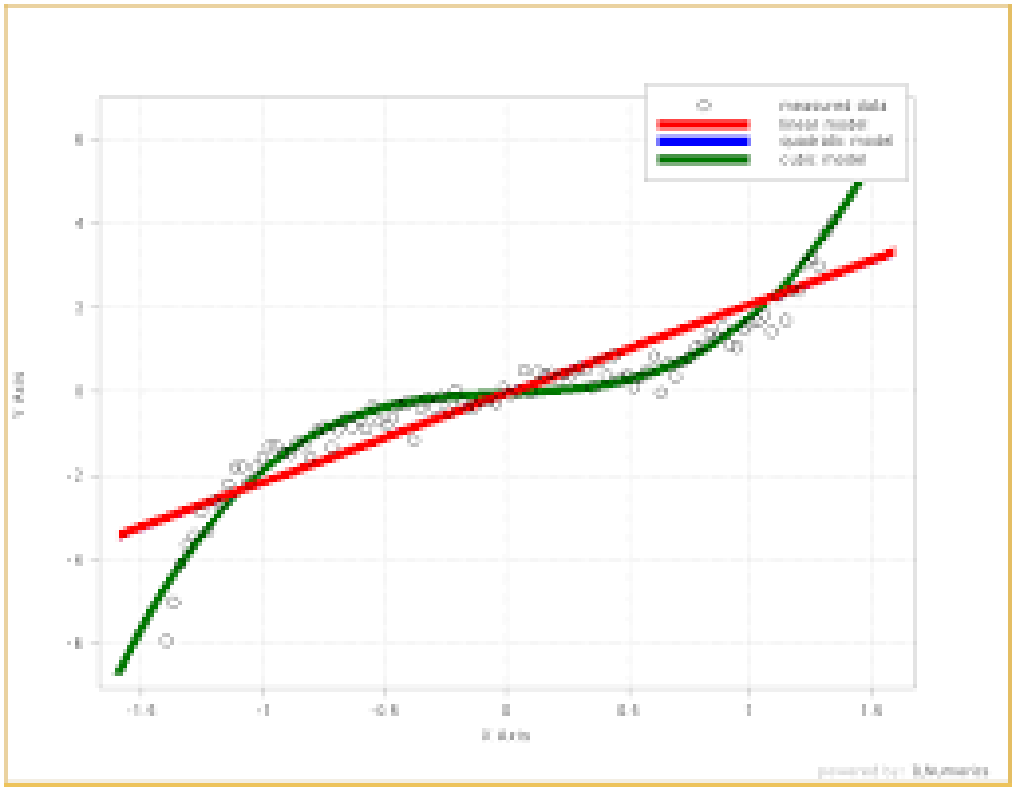
- If we increase X_1 by one unit, then Y will increase by an average of β_1 units
- The presence of X_2 does not alter this statement
- One way of extending this model is to allow for *interaction effects*:
  - a third predictor, called an interaction term:

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

Non-linear Relationships

- The linear regression model assumes a linear relationship between the response and predictors
- In some cases, the true relationship between the response and the predictors may be nonlinear
- One simple way to extend the linear model to non-linear relationships is using *polynomial regression*

Non-linear Relationships



- The red line represents the linear regression fit
- The relationship is in fact non-linear: the data suggest a curved relationship.

Non-linear Relationships

- A simple approach: to include transformed versions of the predictors:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

- This equation involves predicting Y using a non-linear function of X
- It is still a linear model: a multiple linear regression model with $X_1 = X$ and $X_2 = X^2$
- We can use standard linear regression software to estimate β_0, β_1, and β_2 in order to produce a non-linear fit.

Outline

Case study

Simple Linear Regression

Multiple Linear Regression

Extensions of the Linear Model

Potential Problems

Summary: The Marketing Plan

Parametric vs non-parametric approaches

Classification

Potential Problems

When we fit a linear regression model to a particular data set, many problems may occur:

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.

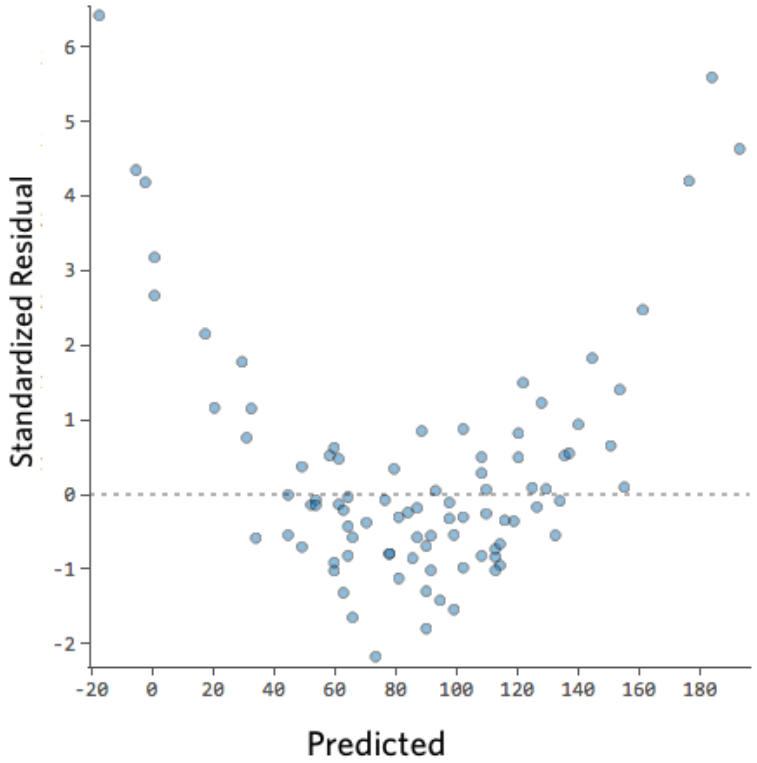
Non-linearity of the Data

- The linear regression model assumes that there is a straight-line relationship between the predictors and the response
- If the true relationship is far from linear, then eventually all of the conclusions that we draw from the fit are suspect
- The prediction accuracy of the model can be significantly reduced.

Non-linearity of the Data

- A useful graphical tool for identifying non-linearity: *residual plots*
- For a simple linear regression model, we can plot the residuals $e_i = y_i - \hat{y}_i$ versus the predictor x_i
- In the case of a multiple regression model we plot the residuals versus the predicted values $\hat{y}_i$

Non-linearity of the Data



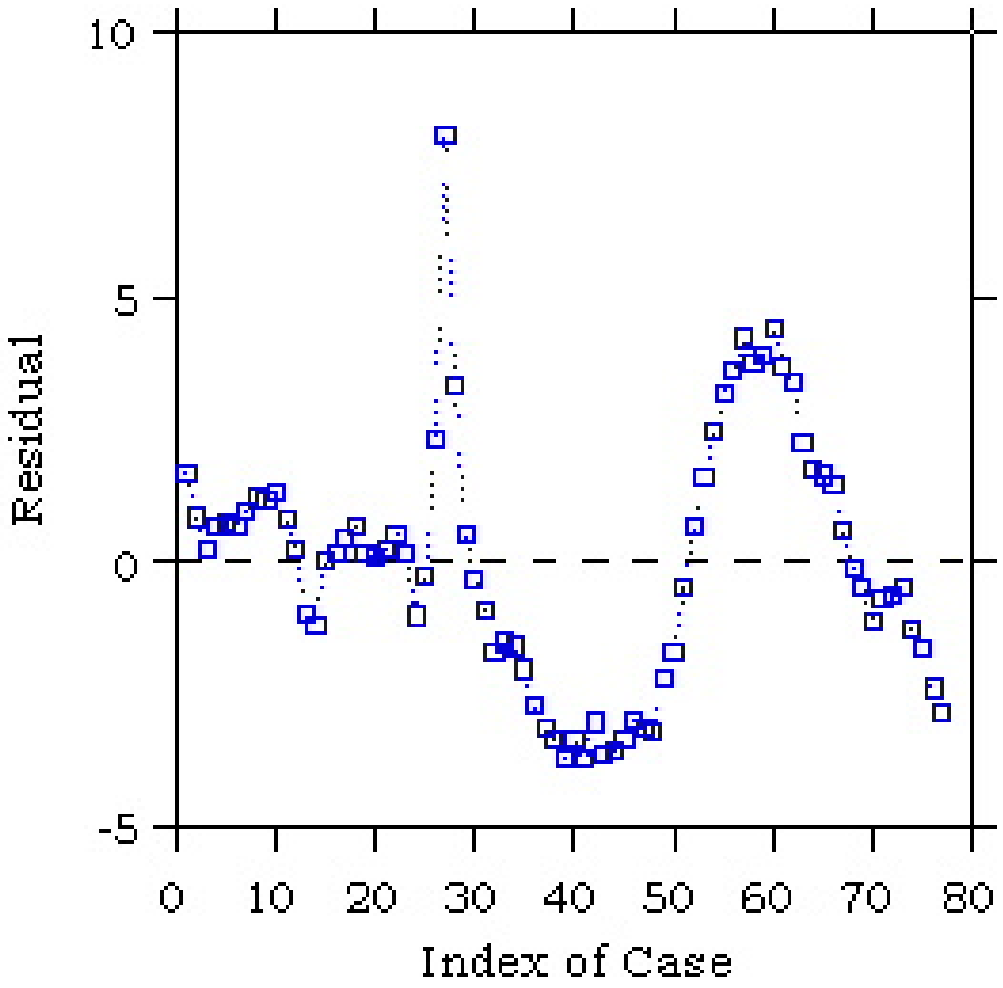
- The presence of a pattern may indicate a problem with some aspect of the linear model
- If the residual plot indicates that there are non-linear associations in the data: non-linear transformations of the predictors $\log X, \sqrt{X}$ and X^2

Correlation of Error Terms

- An important assumption of the linear regression model is that the error terms $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are uncorrelated
- The standard errors are based on the assumption of uncorrelated error terms
- Correlation among the error terms $\Longrightarrow$ the estimated standard errors will tend to underestimate the true standard errors $\Longrightarrow$ **confidence and prediction intervals will be narrower than they should be**
- p-values associated with the model will be lower than they should be $\Longrightarrow$ **could cause to erroneously conclude that a parameter is statistically significant**

Correlation of Error Terms

- Frequently occur in the context of time series data:
  - observations that are obtained at adjacent time points will have positively correlated errors
- To determine: plot the residuals as a function of time
- If the errors are uncorrelated, then there should be no discernible pattern
- If the error terms are correlated, then we may see tracking in the residuals - that is, adjacent residuals may have similar values.



Non-constant Variance of Error Terms

- Assumption: the error terms have a constant variance:

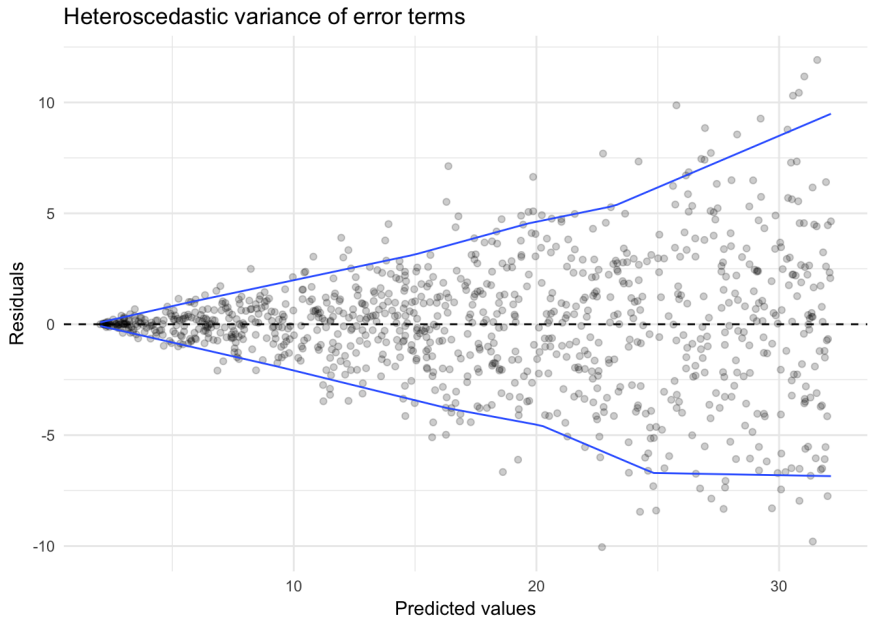
$$\text{Var}(\epsilon_i) = \sigma^2$$

- The standard errors, confidence intervals, and hypothesis tests associated with the linear model rely upon this assumption
- In practice: often the variances of the error terms are non-constant
  - e.g. the variances of the error terms may increase with the value of the response
- Non-constant variances in the errors:

heteroscedasticity

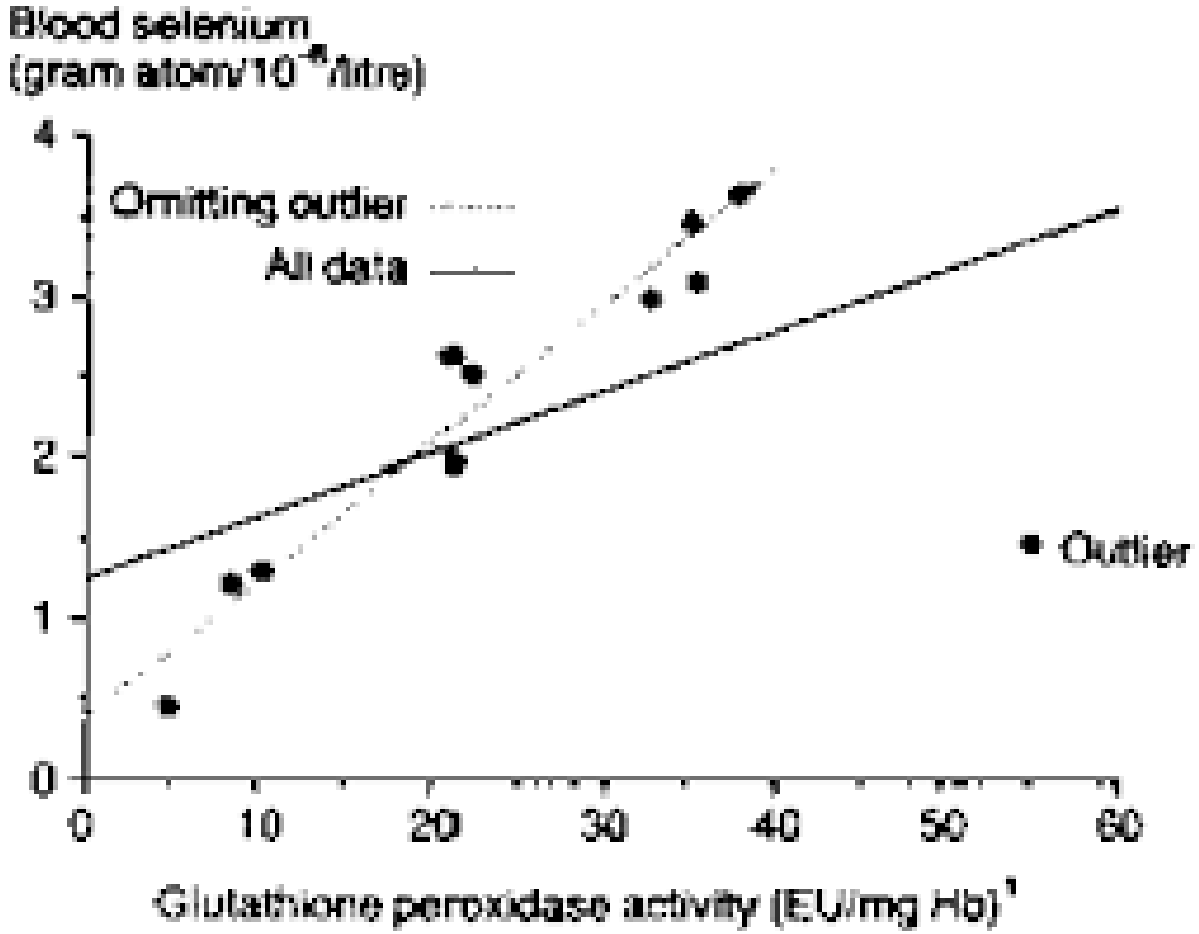
Non-constant Variance of Error Terms

- Identifying heteroscedasticity from the presence of a funnel shape in the residual plot:



- Possible solution: transform the response Y using a concave function such as $\log(Y)$ or $\sqrt{Y}$
  - we shrink the larger responses which leads to a reduction in heteroscedasticity.

Outliers

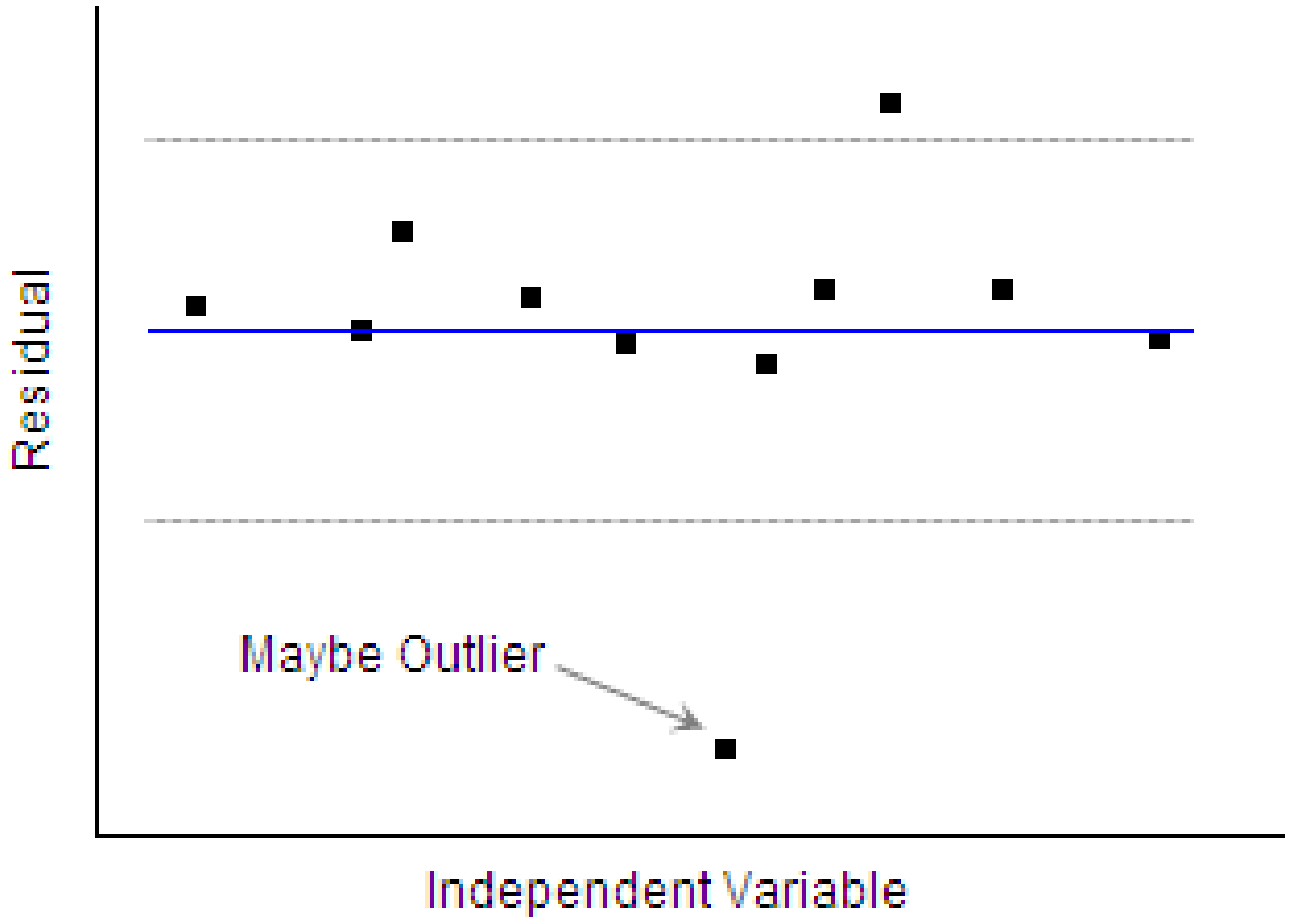


- The solid line is the least squares regression fit
- The dashed line is the least squares fit after removal of the outlier

Outliers

- Even if an outlier does not have much effect on the least squares fit, it can cause other problems (values of RSS and p-values)
- To identify outliers: residual plots
- It can be difficult to decide how large a residual needs to be before we consider the point to be an outlier
- Instead of plotting the residuals, we can plot the *studentized residuals*
- studentized residuals: divide each residual e_i by its estimated standard error

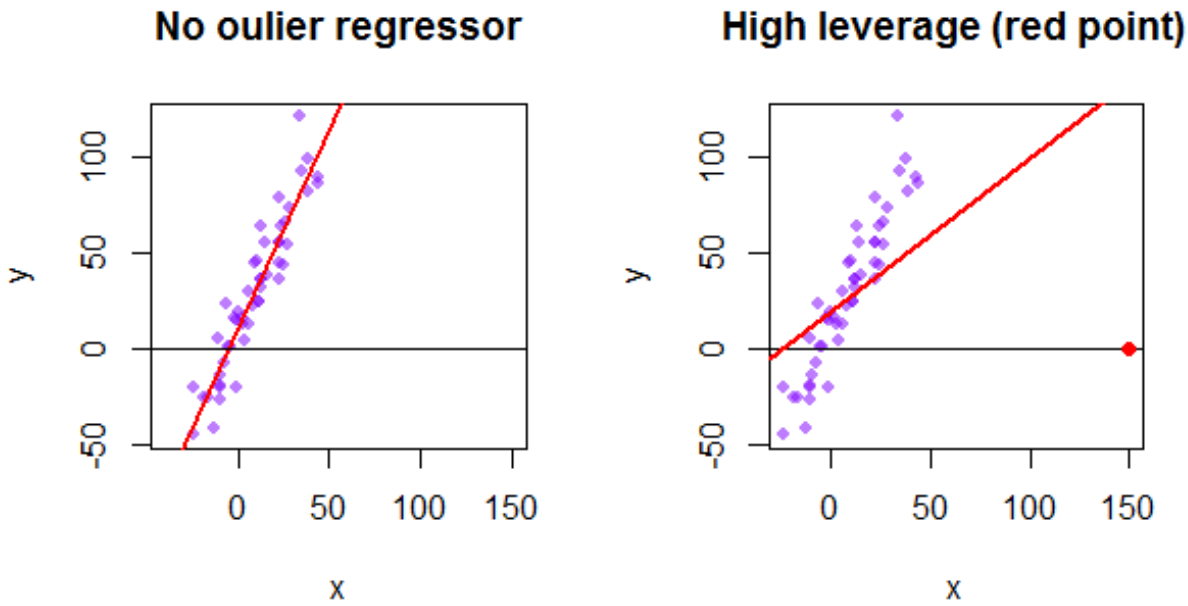
Outliers



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

High Leverage Points

- Outliers are observations for which the response y_i is unusual given the predictor x_i
- Observations with *high leverage* have an unusual value for x_i



Removing the high leverage observation has a substantial impact on the least squares line.

High Leverage Points

- In a simple linear regression, high leverage observations are fairly easy to identify: observations for which the predictor value is outside of the normal range of the observations
- In a multiple linear regression with many predictors, it is possible to have an observation that is well within the range of each individual predictor's values, but that is unusual in terms of the full set of predictors!

High Leverage Points

- To quantify an observations leverage, we compute the leverage statistic:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

- h_i increases with the distance of x_i from $\bar{x}$
- The leverage statistic h_i is between $1/n$ and 1
- The average leverage for all the observations is equal to $(p+1)/n$
- If a given observation has a leverage statistic that greatly exceeds $(p+1)/n$, then we may suspect that the corresponding point has high leverage

Collinearity

- *Collinearity* : two or more predictor variables are closely related to one another
- It can be difficult to separate out the individual effects of collinear variables on the response:
  - there is a broad range of values for the coefficient estimates that result in equal values for RSS
- Estimation is not robust: small changes in the data $\Longrightarrow$ big changes in the resulting coefficients

Collinearity

- Collinearity also lowers significance
- Sometimes, a predictive variable won't appear significant because it's collinear with another predictive variable:
  - E.g., we use both age and number of years in the workforce to predict income
  - Age tends to be correlated with number of years in the workforce
  - Neither variable may appear significant

Collinearity

- The collinearity reduces the accuracy of the estimates of the regression coefficients: it causes the standard error for $\hat{\beta}_j$ to grow
- The t-statistic for each predictor is calculated by dividing $\hat{\beta}_j$ by its standard error
- Collinearity results in a decline in the t-statistic: in the presence of collinearity, we may fail to reject $\mathrm{H}_0 : \beta_j = 0$
- This means that the *power* of the hypothesis test - *the probability of correctly detecting a non-zero coefficient* - is reduced by collinearity.

Collinearity

- The overall model can still predict income quite well, even when the inputs are correlated
- But it can't determine which variable deserves the credit for the prediction
- If you want to use the coefficient values as advice as well as to make good predictions, try to avoid collinearity

Detecting Collinearity

- Identify and address potential collinearity problems while fitting the model
- If you remove one of the variables and the other one gains significance $\implies$ a good indicator of correlation
- Another possible indication of collinearity in the inputs is seeing coefficients with an unexpected sign:
  - e.g. seeing that income is negatively correlated with years in the workforce.
- A simple way to detect collinearity is to look at the *correlation matrix* of the predictors:
  - an element of this matrix large in absolute value indicates a pair of highly correlated variables

VIF

- Not all collinearity problems can be detected by inspection of the correlation matrix
- It is possible for collinearity to exist between three or more variables even if no pair of variables has a particularly high correlation

multicollinearity

- A better way to assess multi-collinearity is to compute the *variance inflation factor (VIF)*:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors

VIF

- If $R^2_{X_j|X_{-j}}$ is close to one, then collinearity is present, and the VIF will be large
- The smallest possible value for VIF is 1, which indicates the complete absence of collinearity
- In practice there is a small amount of collinearity among the predictors
- A VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.

Collinearity

The problem of collinearity solved by:

- Dropping one of the problematic variables from the regression:
  - The presence of collinearity implies that the information that this variable provides about the response is redundant in the presence of the other variables
- Combining the collinear variables together into a single predictor
- Using regularization is helpful in collinear situations

Outline

Case study

Simple Linear Regression

Multiple Linear Regression

Extensions of the Linear Model

Potential Problems

Summary: The Marketing Plan

Parametric vs non-parametric approaches

Classification

Is there a relationship between advertising sales and budget?

- Fitting a multiple regression model of sales onto TV, radio and newspaper advertising budget
- Testing the hypothesis

$$\mathrm{H}_0 : \beta_{\mathsf{TV}} = \beta_{\mathsf{radio}} = \beta_{\mathsf{newspaper}} = 0$$

- The F-statistic can be used to determine whether or not we should reject this null hypothesis
- The p-value corresponding to the F-statistic will indicate if the is an evidence of a relationship between advertising and sales.

How strong is the relationship?

- Two measures of model accuracy:
  - the RSE estimates the standard deviation of the response from the population regression line
  - The R^2 statistic records the percentage of variability in the response that is explained by the predictors
- For example, for the Advertising data, the predictors explain almost 90 % of the variance in sales $\Longrightarrow$ strong relationship

Which media contribute to sales?

- We can examine the p-values associated with each predictor's t-statistic
- For the Advertising data, the p-values for TV and radio are low, but the p-value for newspaper is not $\Longrightarrow$ only TV and radio are related to sales

How large is the effect of each medium on sales?

We can answer this question constructing confidence intervals for β_j

How accurately can we predict future sales?

- The answer depends on whether we wish to predict an individual response, $Y = f(X) + \epsilon$, or the average response, $f(X)$
  - Individual response: use a prediction interval
  - Average response: a confidence interval
- Prediction intervals will always be wider than confidence intervals because they account for the uncertainty associated with errors.

Is the relationship linear?

- Residual plots can be used in order to identify non-linearity
- If the relationships are linear, then the residual plots should display no pattern.

Is there synergy among the advertising media?

- The standard linear regression model assumes an additive relationship between the predictors and the response
- It is easy to interpret because the effect of each predictor on the response is unrelated to the values of the other predictors
- This additive assumption may be unrealistic for certain data sets
- We can include an interaction term in the regression model in order to accommodate non-additive relationships
- A small p-value associated with the interaction term indicates the presence of such relationships

Outline

Parametric approach

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

- Linear regression is an example of a *parametric* approach because it assumes a linear functional form for $f(X)$
- Parametric methods have several advantages:
  - easy to fit: one need estimate only a small number of coefficients
- Disadvantage:
  - by construction, they make strong assumptions about the form of $f(X)$
  - If the specified functional form is far from the truth, then the parametric method will perform poorly.

Non-parametric approaches

- *Non-parametric* methods do not explicitly assume a parametric form for $f(X)$
- More flexible approach for performing regression
- There exists various non-parametric methods: one of the simplest and best-known non-parametric methods, K-nearest neighbors regression (KNN regression).

K-nearest neighbors regression

- Given a value for K and a prediction point x_0
  - KNN regression first identifies the K training observations that are closest to x_0: N_0
  - It estimates $f(x_0)$ using the average of all the training responses in N_0:

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

Parametric vs non-parametric approaches

- In what setting will a parametric approach such as least squares linear regression outperform a non-parametric approach such as KNN regression?
  - the parametric approach will outperform the nonparametric approach if the parametric form that has been selected is close to the true form of f
  - In practice, the true relationship between X and Y is rarely exactly linear.

Should KNN be favored over linear regression?

Parametric vs non-parametric approaches

In reality, even when the true relationship is highly non-linear, KNN may still provide inferior results to linear regression:

- This decrease in performance is associated with the increase of dimension:
  - In higher dimensions there is a reduction in sample size
  - Assume that we have 100 training observations
  - when $p = 1$, this provides enough information to accurately estimate $f(X)$
  - Spreading 100 observations over $p = 20$ dimensions results in a phenomenon in which a given observation has no nearby neighbors: **curse of dimensionality**

Curse of dimensionality

- The K observations that are nearest to a given test observation x_0 may be very far away from x_0 in p-dimensional space when p is large
- $\Longrightarrow$ a poor KNN fit
- Parametric methods will tend to outperform non-parametric approaches when there is a small number of observations per predictor

Parametric vs non-parametric approaches

Even in problems in which the dimension is small, we might prefer linear regression to KNN from an interpretability standpoint:

- described in terms of just a few coefficients
- available p-values

Outline

Classification

- The linear regression model assumes that the response variable Y is quantitative
- In many situations, the response variable is qualitative:
  - e.g. eye color is qualitative, taking on values blue, brown, or green
- Approaches for predicting qualitative responses are known as *classification*:
  - we are classifying the observation assigning the observation to a category, or class
- Often the methods used for classification first predict the probability of each of the categories $\Longrightarrow$ they behave like regression methods

Classification

- There are many possible classification techniques, or classifiers
- Just as in the regression setting, in the classification setting we have a set of training observations $(x_1, y_1), ..., (x_n, y_n)$ that we can use to build a classifier
- We want our classifier to perform well not only on the training data, but also on test observations that were not used to train the classifier.

Why Not Linear Regression?

- Suppose that we are trying to predict the medical condition of a patient in the emergency room on the basis of her symptoms
- There are three possible diagnoses:
  - stroke
  - drug overdose
  - epileptic seizure

Why Not Linear Regression?

- We could consider encoding these values as a quantitative response variable:

$$Y = \begin{cases} 1 & \text{if epileptic seizure} \\ 2 & \text{if stroke} \\ 3 & \text{if drug overdose} \end{cases}$$

- Using this coding, least squares could be used to fit a linear regression model to predict Y on the basis of a set of predictors $X_1, ..., X_p$
- This coding implies an ordering on the outcomes and that the difference between stroke and drug overdose is the same as the difference between drug overdose and epileptic seizure
- There is no particular reason that this needs to be the case!

Why Not Linear Regression?

- Another coding:

$$Y = \begin{cases} 1 & \text{if epileptic seizure} \\ 2 & \text{if drug overdose} \\ 3 & \text{if stroke} \end{cases}$$

  a different relationship among the three conditions

- Each of these codings would produce fundamentally different linear models that would lead to different sets of predictions on test observations

- It is preferable to use a classification method: Logistic Regression or Linear Discriminant Analysis (Machine Learning course)

Linear regression takeaways

- Linear regression is the bread and butter prediction method for statisticians and data scientists
- If you're trying to predict a numerical quantity like profit, cost, or sales volume, you should always try linear regression first
- If it works well, you're done
- If it fails, the detailed diagnostics produced give you a good clue as to what methods you should try next

Linear regression takeaways

- Linear regression will have trouble with problems that have a very large number of variables, or categorical variables with a very large number of levels
- You can enhance linear regression by adding new variables or transforming variables
- With linear regression, think in terms of residuals!
- Linear regression can predict well even in the presence of correlated variables, but correlated variables lower the quality of the variable selection

References

- Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The elements of statistical learning - Data Mining, inference, and prediction.* Springer.

- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning with applications in R.* Springer.