Machine Learning (Semester 1 2024)

Regression

Nina Hernitschek
Centro de Astronomía CITEVA
Universidad de Antofagasta

May 7, 2024

Motivation

In the previous session, we saw an overview about the objectives of statistical learning.

We also saw some applications of classification and regression algorithms, along with methods on how to quantify the quality of fit.

Today we will focus on **regression**.

recap: Regression Versus Classification Problems

Supervised learning as a function approximation from samples:

The basic goal of supervised learning is to use the training set S to learn a function f_S that looks at a new x value x_{new} and predicts the associated value of y:

$$y_{pred} = f_S(x_{new})$$

recap: Regression Versus Classification Problems

Supervised learning as a function approximation from samples:

The basic goal of supervised learning is to use the training set S to learn a function f_S that looks at a new x value x_{new} and predicts the associated value of y:

$$y_{pred} = f_S(x_{new})$$

If y is a real-valued random variable (thus being a quantitative value), we have **regression**.

If y takes values from an unordered finite set (thus being a qualitative categorial variable), we have **classification**.

In two-class (binary) classification problems, we can assign one class a y value of 1, and the oter class a y value of 0.

recap: Regression

In regression, instead of having training data with discrete labels, the "truth" is a continuous property and we are trying to predict the values of that property for the test data.

Mathematically:
We are given a training set $(x_1, y_1), ..., (x_n, y_n)$ where the y_i are real-valued. The goal is to learn a function f to predict the y values associated with new observed x values.

recap: Regression

In regression, instead of having training data with discrete labels, the "truth" is a continuous property and we are trying to predict the values of that property for the test data.

Mathematically:
We are given a training set $(x_1, y_1), ..., (x_n, y_n)$ where the y_i are real-valued. The goal is to learn a function f to predict the y values associated with new observed x values.

example:
The task of determining whether an object is a star, a galaxy, or a quasar is a classification problem: the label is from three distinct categories. In addition, we might want to estimate the age of an object based on such observations: this would be a regression problem, because the label (age) is a continuous quantity.

recap: Regression

In regression, instead of having training data with discrete labels, the "truth" is a continuous property and we are trying to predict the values of that property for the test data.

Mathematically:
We are given a training set $(x_1, y_1), ..., (x_n, y_n)$ where the y_i are real-valued. The goal is to learn a function f to predict the y values associated with new observed x values.

example:
The task of determining whether an object is a star, a galaxy, or a quasar is a classification problem: the label is from three distinct categories. In addition, we might want to estimate the age of an object based on such observations: this would be a regression problem, because the label (age) is a continuous quantity.

example (outside of astronomy):
In a financial application, one may attempt to predict the price of goods as a function of e.g. currency exchange rates, availability and demand.

Linear Regression

Linear regression is a very useful tool for **predicting a quantitative response**.

Despite the existence of more complex approaches in modern statistical learning, understanding this approach (and using it) is still of great value:

- it can serve as a first tool for analysis
- understanding linear regression aids the understanding of more sophisticated statistical learning approaches which generalize or extend linear regression.

Linear Regression

Linear regression is a very useful tool for **predicting a quantitative response**.

Despite the existence of more complex approaches in modern statistical learning, understanding this approach (and using it) is still of great value:

- it can serve as a first tool for analysis
- understanding linear regression aids the understanding of more sophisticated statistical learning approaches which generalize or extend linear regression.

Here we use linear regression as a starting point into a deeper understanding of regression methods.

Linear Regression

Example: We might want to know which kind of advertisement works best for a product, where we have advertising budgets for TV, radio, and internet. We measure the success in product sales.

Linear Regression

Example: We might want to know which kind of advertisement works best for a product, where we have advertising budgets for TV, radio, and internet. We measure the success in product sales.

In this context, we might want to address the following questions:

1. *Is there (at all) a relationship between advertising budget and sales?*

Our first goal should be to determine whether the data provide evidence of an association (*correlation*) between advertising and sales. If the evidence is weak, then one might argue that no money should be spent on advertising!

Linear Regression

Example: We might want to know which kind of advertisement works best for a product, where we have advertising budgets for TV, radio, and internet. We measure the success in product sales.

In this context, we might want to address the following questions:

1. *Is there (at all) a relationship between advertising budget and sales?*
2. *How strong is the relationship between advertising budget and sales?*

Assuming that we found a relationship between advertising and sales, we are interested in the *strength* of this relationship. I.e.: Is it worth to use a higher budget on advertisement?

Linear Regression

Example: We might want to know which kind of advertisement works best for a product, where we have advertising budgets for TV, radio, and internet. We measure the success in product sales.

In this context, we might want to address the following questions:

1. *Is there (at all) a relationship between advertising budget and sales?*
2. *How strong is the relationship between advertising budget and sales?*
3. *Which media are associated with sales?*

Are all three media (TV, radio, internet) associated with sales? To answer this question, we must find a way to separate out the individual contribution of each medium to sales when we have spent money on all three media.

Linear Regression

Example: We might want to know which kind of advertisement works best for a product, where we have advertising budgets for TV, radio, and internet. We measure the success in product sales.

In this context, we might want to address the following questions:

1. *Is there (at all) a relationship between advertising budget and sales?*
2. *How strong is the relationship between advertising budget and sales?*
3. *Which media are associated with sales?*
4. *How large is the association between each medium and sales?*

For every amount spent on advertising in a particular medium, by what amount will sales increase? How accurately can we predict this amount of increase?

Linear Regression

Example: We might want to know which kind of advertisement works best for a product, where we have advertising budgets for TV, radio, and internet. We measure the success in product sales.

In this context, we might want to address the following questions:

1. *Is there (at all) a relationship between advertising budget and sales?*
2. *How strong is the relationship between advertising budget and sales?*
3. *Which media are associated with sales?*
4. *How large is the association between each medium and sales?*
5. *How accurately can we predict future sales?*

For any given level of television, radio, or internet advertising, what is our prediction for sales, and what is the accuracy of this prediction?

Linear Regression

Example: We might want to know which kind of advertisement works best for a product, where we have advertising budgets for TV, radio, and internet. We measure the success in product sales.

In this context, we might want to address the following questions:

1. *Is there (at all) a relationship between advertising budget and sales?*
2. *How strong is the relationship between advertising budget and sales?*
3. *Which media are associated with sales?*
4. *How large is the association between each medium and sales?*
5. *How accurately can we predict future sales?*
6. *Is the relationship linear?*

If there is approximately a straight-line relationship between advertising expenditure in the various media and sales, then linear regression is an appropriate tool. If not, then it may still be possible to transform the predictor or the response so that linear regression can be used.

Linear Regression

Example: We might want to know which kind of advertisement works best for a product, where we have advertising budgets for TV, radio, and internet. We measure the success in product sales.

In this context, we might want to address the following questions:

1. *Is there (at all) a relationship between advertising budget and sales?*
2. *How strong is the relationship between advertising budget and sales?*
3. *Which media are associated with sales?*
4. *How large is the association between each medium and sales?*
5. *How accurately can we predict future sales?*
6. *Is the relationship linear?*
7. *Is there synergy among the advertising media?*

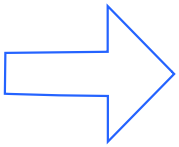
Perhaps spending \$50,000 on television advertising and \$50,000 on radio advertising is associated with higher sales than allocating \$100,000 to either television or radio individually. In marketing, this is known as a synergy effect, while in statistics it is called an interaction effect.

Linear Regression

Example: We might want to know which kind of advertisement works best for a product, where we have advertising budgets for TV, radio, and internet. We measure the success in product sales.

In this context, we might want to address the following questions:

1. *Is there (at all) a relationship between advertising budget and sales?*
2. *How strong is the relationship between advertising budget and sales?*
3. *Which media are associated with sales?*
4. *How large is the association between each medium and sales?*
5. *How accurately can we predict future sales?*
6. *Is the relationship linear?*
7. *Is there synergy among the advertising media?*



Linear regression can be used to answer each of these questions

Simple Linear Regression

Simple linear regression is a very straightforward approach for predicting a quantitative response Y based on **a single predictor** X. It assumes that there is an (approximately) linear relationship between X and Y.

Mathematically:

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

Simple Linear Regression

Simple linear regression is a very straightforward approach for predicting a quantitative response Y based on **a single predictor** X. It assumes that there is an (approximately) linear relationship between X and Y.

Mathematically:

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

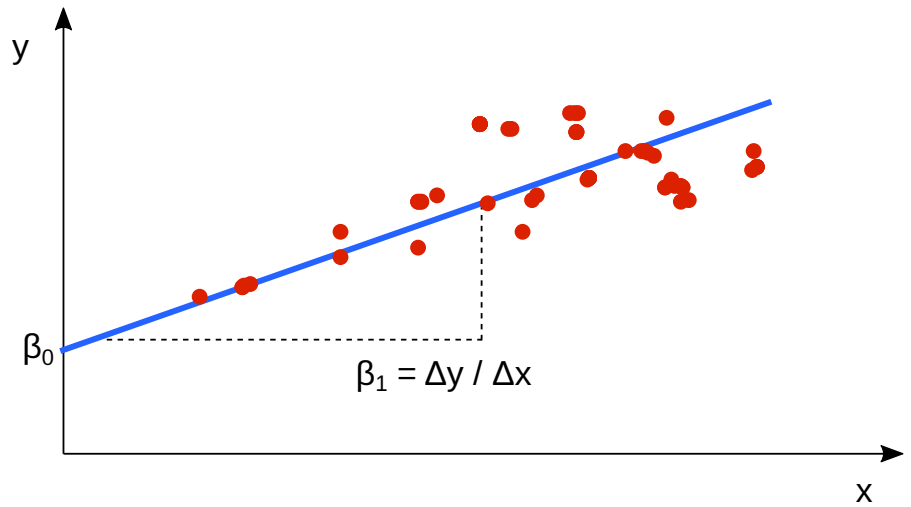
We often say we are **regressing** Y **on** X.

Estimating the Coefficients

Going back to the previous `example`, X may represent `TV` advertising and Y may represent `sales`.

Then we can regress `sales` on `TV` by fitting the following model:

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



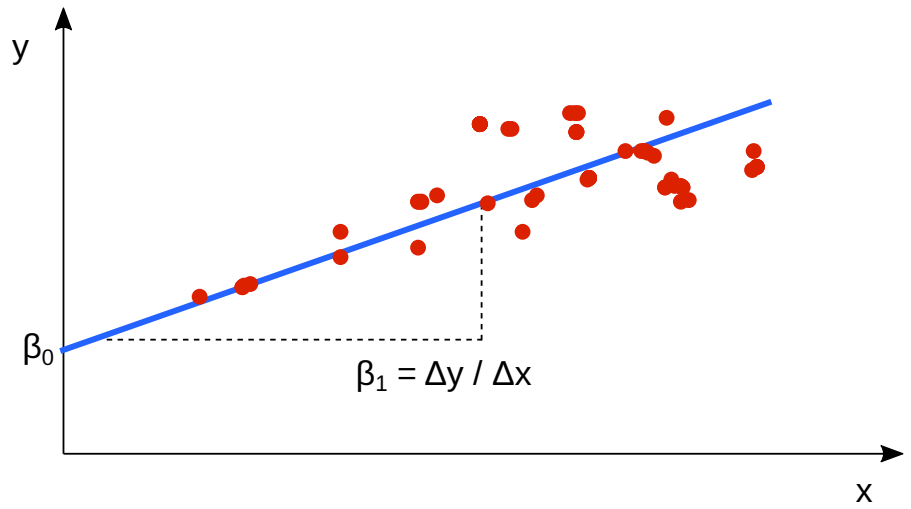
Estimating the Coefficients

Going back to the previous `example`, X may represent `TV` advertising and Y may represent `sales`.

Then we can regress `sales` on `TV` by fitting the following model:

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

In this equation, β_0 and β_1 are two unknown constants representing the intercept and slope terms in the linear model. Together, β_0 and β_1 are known as the **model coefficients** (or **parameters**).



Estimating the Coefficients

In practice, β_0 and β_1 are unknown.

As a first step, we use our **training data** to compute model coefficient estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$.

As a second step, we can **predict** future sales based on a particular value of internet advertising by computing

$$\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$$

where $\widehat{y}$ indicates a prediction of Y on the basis of $X = x$.

Estimating the Coefficients

In practice, β_0 and β_1 are unknown.

As a first step, we use our **training data** to compute model coefficient estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$.

As a second step, we can **predict** future sales based on a particular value of internet advertising by computing

$$\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$$

where $\widehat{y}$ indicates a prediction of Y on the basis of $X = x$.

We now look deeper into how to estimate β_0 and β_1 from training data.

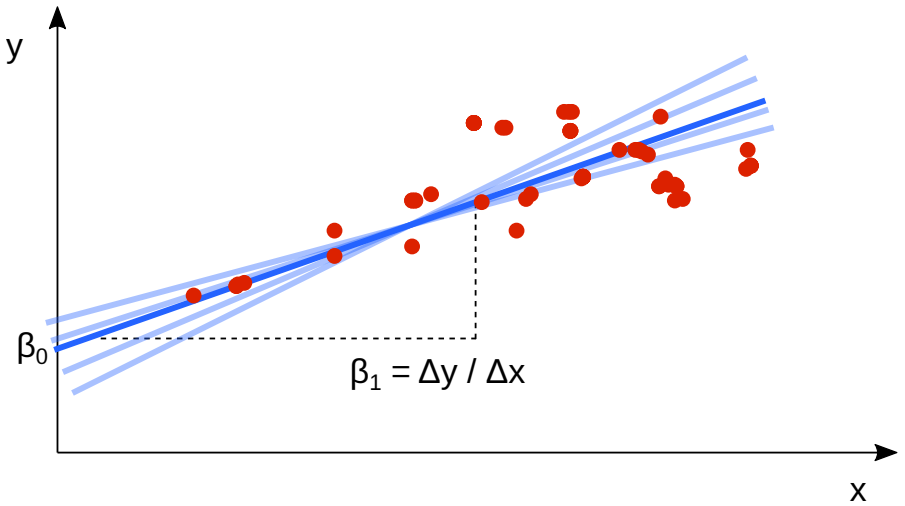
Estimating the Coefficients

Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ represent n observation pairs, each of which consists of a measurement of X and Y.

In the above example, this data set consists of the internet advertising budget and product sales in $n = 200$ different markets.

Our goal for the fit is obtaining coefficient estimates $\widehat{\beta_0}$, $\widehat{\beta_1}$ such that the linear model fits the available data well, i.e. $y_i \approx \widehat{\beta_0} + \widehat{\beta_1} x_i$ for $i = 1, ..., n$.

Graphically this means we want to find an intercept $\widehat{\beta_0}$ and slope $\widehat{\beta_0}$ such that the resulting line is as close as possible to the $n = 200$ data points.



Estimating the Coefficients

How to measure "close"?

Estimating the Coefficients

How to measure "close"?

Let $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \widehat{y_i}$ represents the ith **residual**, i.e. the difference between the ith observed response value and the predicte ith response value. We define the **residual sum of squares (RSS)** as

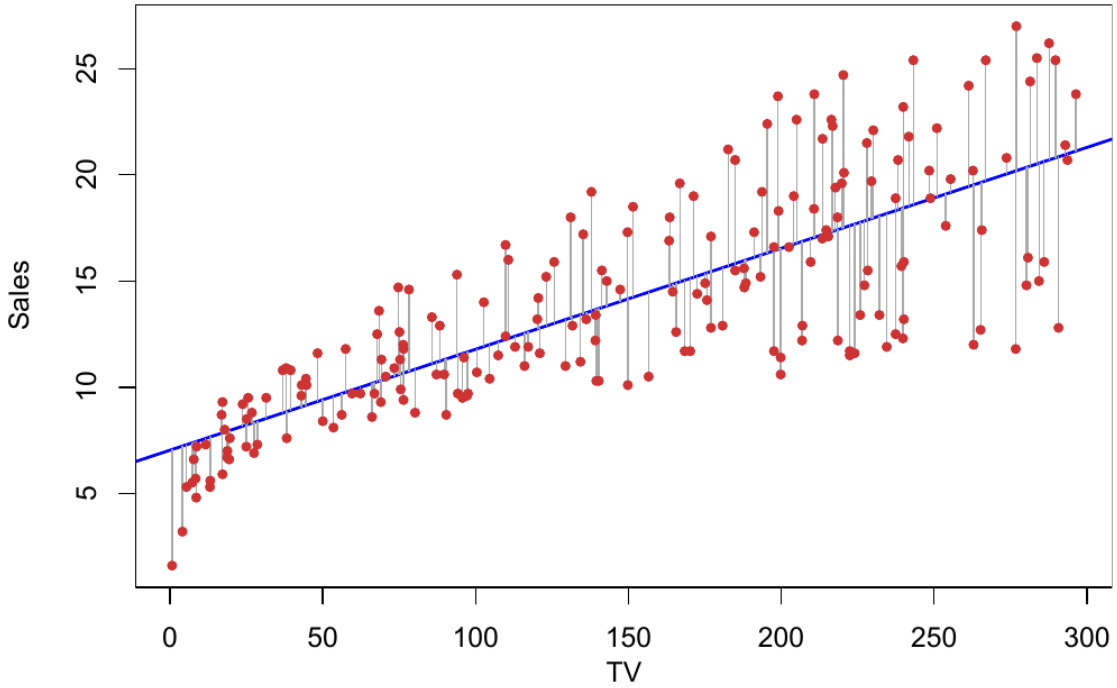
$$RSS = e_1^2 + ... + e_n^2$$

The least squares approach chooses $\widehat{\beta_0}$ and $\widehat{\beta_1}$ to minimize RSS. The minimizers are:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

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

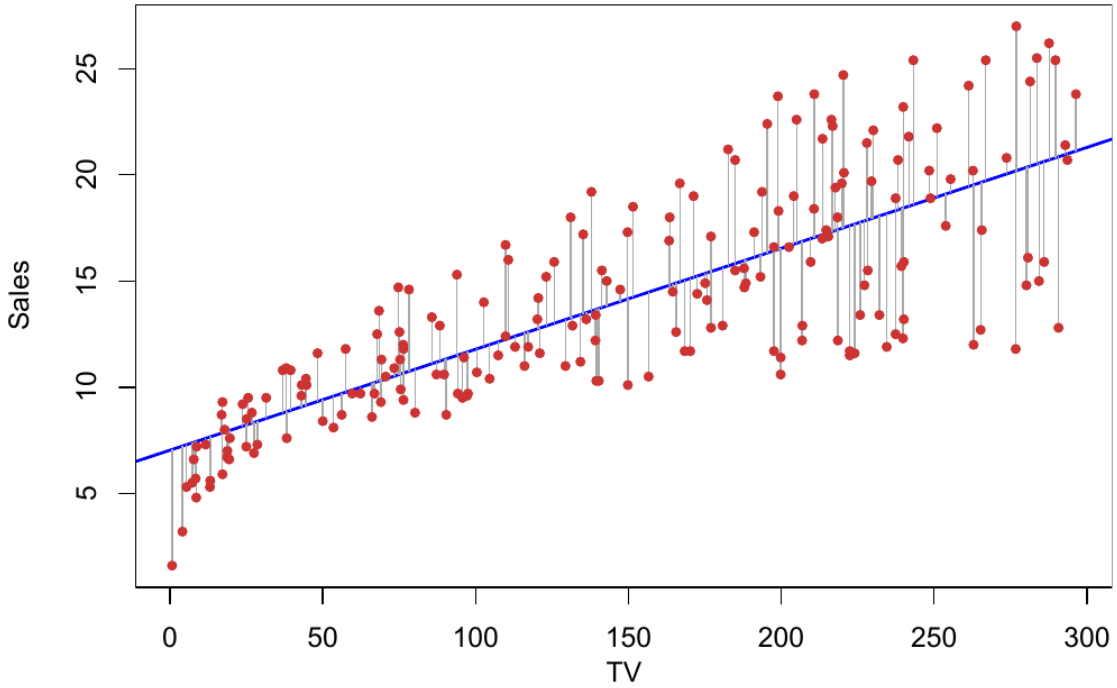
A Least Squares Fit



The least squares fit for the regression of `sales` on `TV`. Each grey line represents a residual. The observations are shown in red.
Source: Fig. 3.1 from `https://www.statlearning.com/`

What is your impression?

A Least Squares Fit



The least squares fit for the regression of `sales` on `TV`. Each grey line represents a residual. The observations are shown in red.
Source: Fig. 3.1 from `https://www.statlearning.com/`

What is your impression?

The linear fit captures the essence of the relationship, but overestimates the trend for small budgets.

Fit Accuracy

We can carry out the fit - but how accurate is it?

Remember: We assume that the true relationship between X and Y is o fthe form $Y = f(X) + \epsilon$ for some unknown function f, where ϵ is a mean-zero random error term.

If f is to be approximated by a linear function:

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

Fit Accuracy

We can carry out the fit - but how accurate is it?

Remember: We assume that the true relationship between X and Y is o fthe form $Y = f(X) + \epsilon$ for some unknown function f, where ϵ is a mean-zero random error term.

If f is to be approximated by a linear function:

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

ϵ will account for what we miss with this simple model:

- non-linear relationship
- influence on other variables on Y
- measurement errors.

Fit Accuracy

We can carry out the fit - but how accurate is it?

Remember: We assume that the true relationship between X and Y is o fthe form $Y = f(X) + \epsilon$ for some unknown function f, where ϵ is a mean-zero random error term.

If f is to be approximated by a linear function:

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

ϵ will account for what we miss with this simple model:

- non-linear relationship
- influence on other variables on Y
- measurement errors.

Typically it is assumed that the error term ϵ is **independent** of X.

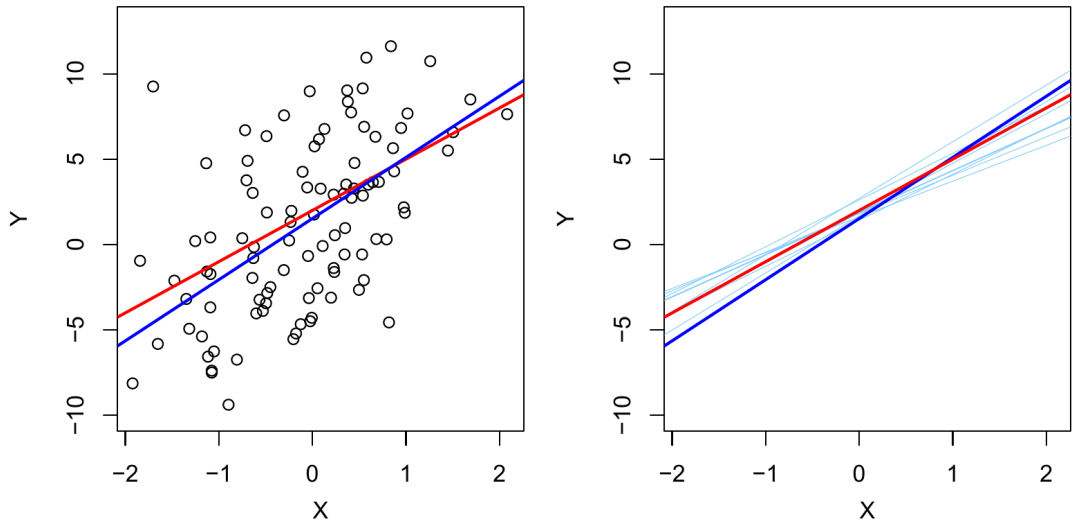
Fit Accuracy

It is important to understand that we have a set of **observations** for which we compute the fit; however, the **population regression line** is unobserved.

Fit Accuracy

It is important to understand that we have a set of **observations** for which we compute the fit; however, the **population regression line** is unobserved.

The following plots show fits for a simulated data set (known distribution).



Left: The red line gives the true relationship, $f(X) = 2 + 3X$ (population regression line). The blue line is the estimate from the observations.
Right: In addition to the population regression line (red) and least squares line (dark blue), ten least squares lines are shown (light blue), each based on a separate set of observations drawn from the distribution.
Source: Fig. 3.3 from `https://www.statlearning.com/`

Bias

From the plot, we have seen that our estimates are always **biased**:
They are influenced by the observations being drawn from an underlying population which is unknown (otherwise we would not have to estimate it).

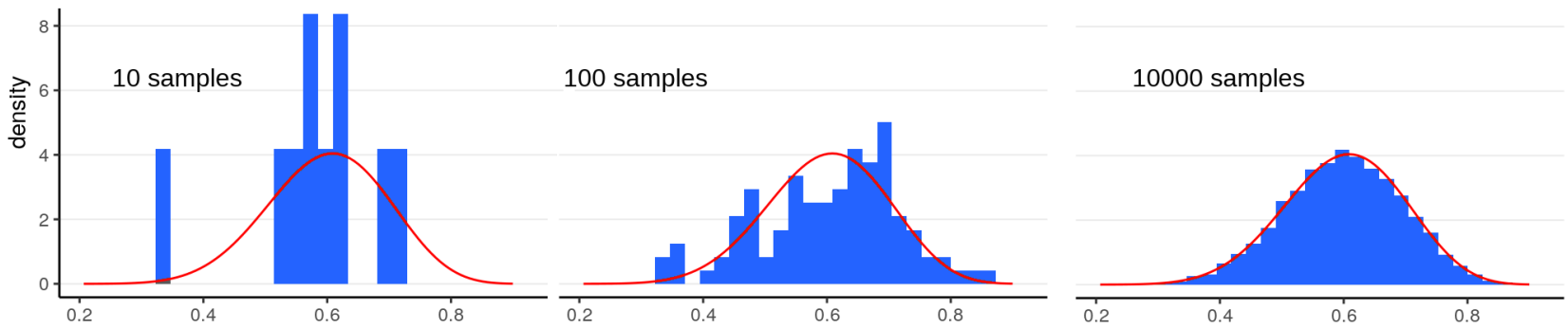
Bias

From the plot, we have seen that our estimates are always **biased**:
They are influenced by the observations being drawn from an underlying population which is unknown (otherwise we would not have to estimate it).

We know this from trying to estimate the population mean μ of some random variable X:

We have access to n observations from X, $x_1, ..., x_n$, which we can use to estimate μ. A reasonable estimate for μ is the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

If we could **average** a huge number of estimates obtained from different observations, then this average would be the same as for the population: the sample mean **converges** to the population mean.



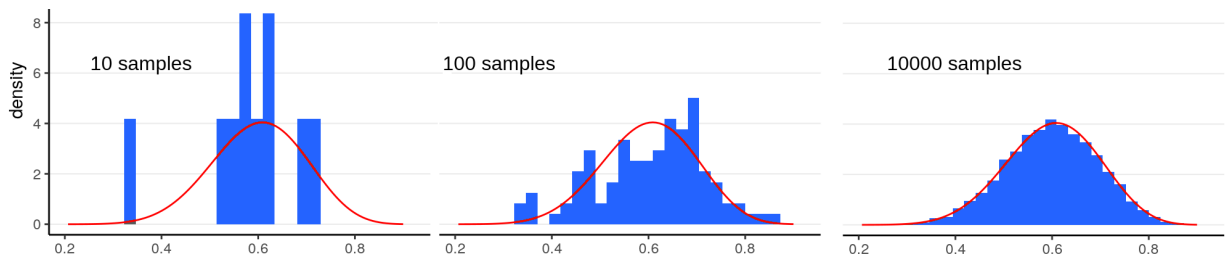
Bias

From the plot, we have seen that our estimates are always **biased**:
They are influenced by the observations being drawn from an underlying population which is unknown (otherwise we would not have to estimate it).

We know this from trying to estimate the population mean μ of some random variable X:

We have access to n observations from X, $x_1, ..., x_n$, which we can use to estimate μ. A reasonable estimate for μ is the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

If we could **average** a huge number of estimates obtained from different observations, then this average would be the same as for the population: the sample mean **converges** to the population mean.



Thus, an **unbiased** estimator like the mean does not systematically over- or under-estimate the true parameter.

Bias

The property of **unbiasedness** holds for the least squares coefficient estimates:

Just the same way, when the unknown coefficients β_0 and β_1 are estimated by $\widehat{\beta_0}$, $\widehat{\beta_1}$: Based on a particular set of observations, $\widehat{\beta_0}$ might be overestimated and $\widehat{\beta_1}$ underestimated, whereas based on another set of observations, the opposite might be true.

Bias

The property of **unbiasedness** holds for the least squares coefficient estimates:

Just the same way, when the unknown coefficients β_0 and β_1 are estimated by $\widehat{\beta_0}$, $\widehat{\beta_1}$: Based on a particular set of observations, $\widehat{\beta_0}$ might be overestimated and $\widehat{\beta_1}$ underestimated, whereas based on another set of observations, the opposite might be true.

If we estimate β_0 and β_1 on the basis of a particular data set, they won't be exactly β_0 and β_1, but if we could average the estimates from a huge number of data sets, they would.

Standard Errors

How **accurate** are our estimates?

Standard Errors

How **accurate** are our estimates?

We again look at analogy of estimating the population mean μ of a random variable X.
We have seen that the average of $\widehat{\mu}$ over many data sets will be very close to μ, but that a single estimate $\widehat{\mu}$ may substantially under- oder overestimate μ. By how much?
In general, we answer this question by computing the **standard error** of $\widehat{\mu}$, written as $SE(\widehat{\mu})$:

$$Var(\widehat{\mu}) = SE(\widehat{\mu})^2 = \frac{\sigma^2}{n}$$

where σ is the **standard deviation**.

Standard Errors

Similarly, we can compute the standard errors associated with $\widehat{\beta_0}$ and $\widehat{\beta_1}$:

$$SE(\widehat{\beta_0})^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad SE(\widehat{\beta_1})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Standard Errors

Similarly, we can compute the standard errors associated with $\widehat{\beta_0}$ and $\widehat{\beta_1}$:

$$SE(\widehat{\beta_0})^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad SE(\widehat{\beta_1})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

These equations are strictly valid under the assumption that the errors e_i have common variance σ^2 and are uncorrelated. In cases where this is not true, the equations can still serve as a good approximation.

Standard Errors

Similarly, we can compute the standard errors associated with $\widehat{\beta_0}$ and $\widehat{\beta_1}$:

$$SE(\widehat{\beta_0})^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad SE(\widehat{\beta_1})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

These equations are strictly valid under the assumption that the errors e_i have common variance σ^2 and are uncorrelated. In cases where this is not true, the equations can still serve as a good approximation.

What about the variance σ^2?
In general, σ^2 is not known but can be estimated from the data as the **residual standard error**

$$RSE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y_i}^2}$$

It can be interpreted as the average amount that the response will deviate from the true regression line.

Confidence Intervals

Standard errors can be used to calculate **confidence intervals**.

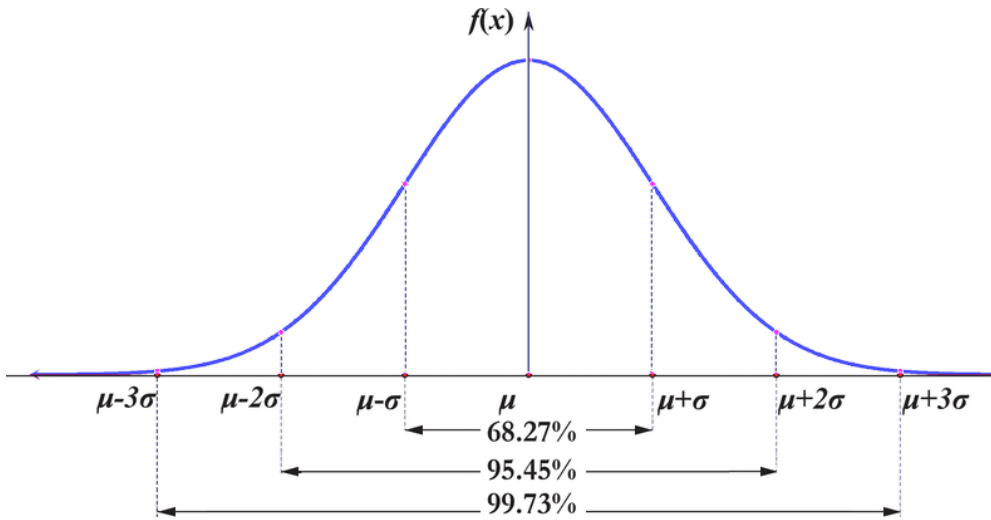
Recap: Gaussian Distribution

Gaussian confidence levels

The probability of a measurement drawn from a Gaussian distribution that is between $\mu - a$ and $\mu + b$ is

$$\int_{\mu-a}^{\mu+b} p(x|\mu,\sigma)dx$$

examples:
for a = b = 1σ, we get 68.3%
for a = b = 2σ, we get 95.4%
for a = b = 3σ, we get 99.7%



We refer to the ranges $\mu \pm 1\sigma$, $\mu \pm 2\sigma$, $\mu \pm 3\sigma$ as the 68%, 95% and 99% **confidence limits**, respectively. Note: These numbers are only valid for Gaussian distributions (but can be calculated for other distributions).

Confidence Intervals

Standard errors can be used to calculate **confidence intervals**.
For linear regression, the 95% confidence interval for β_1 approximately takes the form

$$\widehat{\beta_1} \pm 2 \times SE(\widehat{\beta_1})$$

That is, there is approximately a 95 % chance that the interval

$$\left[\widehat{\beta_1} - 2 \times SE(\widehat{\beta_1}), \widehat{\beta_1} + 2 \cdot SE(\widehat{\beta_1})\right]$$

contains the true value of β_1.

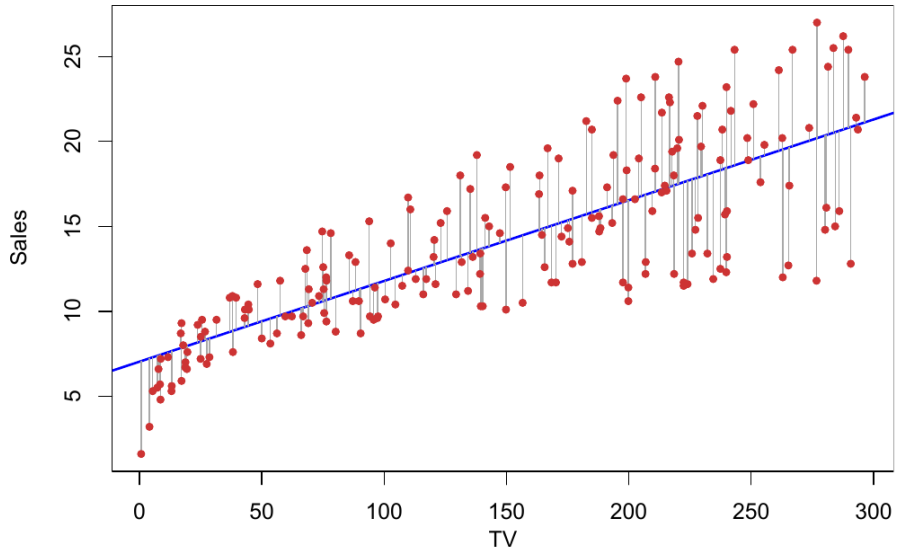
Similarly, the 95% confidence interval for β_0 is

$$\widehat{\beta_0} \pm 2 \times SE(\widehat{\beta_0}).$$

Confidence Intervals

Going back to our advertising **example**:

The 95 % confidence interval for β_0 is [6.130, 7.935] and the 95 % confidence interval for β_1 is [0.042, 0.053]. Therefore, we can conclude that in the absence of any advertising, sales will, on average, fall somewhere between 6,130 and 7,935 units.
Furthermore, for each \$1,000 increase in TV advertising, there will be an average increase in sales of between 42 and 53 units.



The least squares fit for the regression of `sales` on `TV`.
Source: Fig. 3.1 from `https://www.statlearning.com/`

Hypothesis Testing

Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the **null hypothesis** of

H_0: There is no relationship between X and Y

versus the **alternative hypothesis**

H_a: There is some relationship between X and Y.

Hypothesis Testing

Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the **null hypothesis** of

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, this corresponds to testing

$H_0 : \beta_1 = 0$

versus

$H_a : \beta_1 \neq 0$,

since if $\beta_1 = 0$, then our model reduces to $Y = \beta_0 + \epsilon$ and there is no association between X and Y.

Hypothesis Testing

To **test** the null hypothesis, we have to determine whether $\widehat{\beta_1}$ is sufficiently far from zero so we can be confident that $\beta_1 \neq 0$.

How far is enough? This depends on the accuracy of $\widehat{\beta_1}$, so it depends on $SE(\widehat{\beta_1})$. For small $SE(\widehat{\beta_1})$, even relatively small values of $\widehat{\beta_1}$ may prove strong evidence that $\beta_1 \neq 0$ and there is a relationship between X and Y.

Hypothesis Testing

To **test** the null hypothesis, we have to determine whether $\widehat{\beta_1}$ is sufficiently far from zero so we can be confident that $\beta_1 \neq 0$.

How far is enough? This depends on the accuracy of $\widehat{\beta_1}$, so it depends on $SE(\widehat{\beta_1})$. For small $SE(\widehat{\beta_1})$, even relatively small values of $\widehat{\beta_1}$ may prove strong evidence that $\beta_1 \neq 0$ and there is a relationship between X and Y.

In practice, for hypothesis testing we compute the **t-statistic**, which is given by

$$t = \frac{\widehat{\beta_1} - 0}{SE(\widehat{\beta_1})}.$$

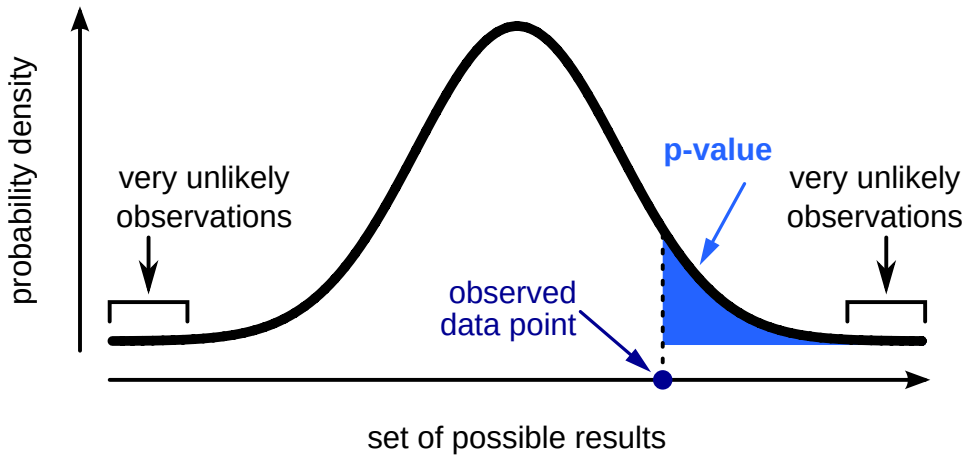
It measures the number of standard deviations that $\widehat{\beta_1}$ is away from 0.

Hypothesis Testing

Related to hypothesis testing is the *p*-**value**.
A *p*-**value** is the calculated **probability of obtaining an effect at least as extreme as from the sample data**, assuming the null hypothesis holds.

A small *p*-value means there is a small chance that the results could be completely random. A large *p*-value means that the results have a high probability of being random and not due to anything from the experiment. The smaller the *p*-value, the more statistically significant the result.

example: A *p*-value of 0.05 means that 5% of the time you would see a test statistic at least as extreme as found if the null hypothesis was true.



The *p*-value

The *p*-value is often **misinterpreted**.

The *p*-value is essentially the probability of a false positive based on the data in the experiment. It does not tell the probability of a specific event actually happening and it does not tell the probability that a variant is better than the control. *p*-values are **probability statements about the data sample** not about the hypothesis itself.

Multiple Linear Regression

We have seen **simple linear regression** as an useful approach for predicting a response on the basis of a **single predictor variable**.

However, in practice we often have more than one predictor.

Going back to our **example** with the advertising data: We have, so far, examined the relationship between sales and TV advertising.

But what is about the radio and internet advertising?

How can we extend our analysis of the advertising data in order to accommodate these two additional predictors?

Multiple Linear Regression

The simplest option would be: Run three separate simple linear regressions, each of for a different advertising medium (radio, TV, internet).

Multiple Linear Regression

The simplest option would be: Run three separate simple linear regressions, each of for a different advertising medium (radio, TV, internet).

However there are **problems**:

- it is unclear how to make a single prediction of sales given the three advertising media budgets, since each of the budgets is associated with a separate regression equation
- if the media budgets are correlated with each other, this can give very misleading estimates of the association between each media budget and sales

Multiple Linear Regression

The simplest option would be: Run three separate simple linear regressions, each of for a different advertising medium (radio, TV, internet).

However there are **problems**:

- it is unclear how to make a single prediction of sales given the three advertising media budgets, since each of the budgets is associated with a separate regression equation
- if the media budgets are correlated with each other, this can give very misleading estimates of the association between each media budget and sales

A **better approach**:

Extend the simple linear regression model to directly work with multiple predictors.

Multiple Linear Regression

We can do this by giving each predictor a separate slope coefficient in a single model. In general, suppose that we have p distinct predictors. Then the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$$

where X_j represents the jth predictor and β_j quantifies the association between that variable and the response.

We **interpret** β_j as the average effect on Y of a one-unit increase in X_j while **holding all other predictors fixed**.

Multiple Linear Regression

We can do this by giving each predictor a separate slope coefficient in a single model. In general, suppose that we have p distinct predictors. Then the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$$

where X_j represents the jth predictor and β_j quantifies the association between that variable and the response.

We **interpret** β_j as the average effect on Y of a one-unit increase in X_j while **holding all other predictors fixed**.

Going back to our **example**, this becomes:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{internet} + \epsilon.$$

Estimating the Regression Coefficients

How to estimate the coefficients?

As for simple linear regression, the regression coefficients β_j are unknown and must be estimated. We can make predictions using the equations

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

The parameters are estimated using the same least squares approach that we saw for the simple linear regression. We choose the $\widehat{\beta}_j$ to minimize the sum of squared residuals

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_p x_{ip})^2 \end{aligned}$$

Estimating the Regression Coefficients

How to estimate the coefficients?

As for simple linear regression, the regression coefficients β_j are unknown and must be estimated. We can make predictions using the equations

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

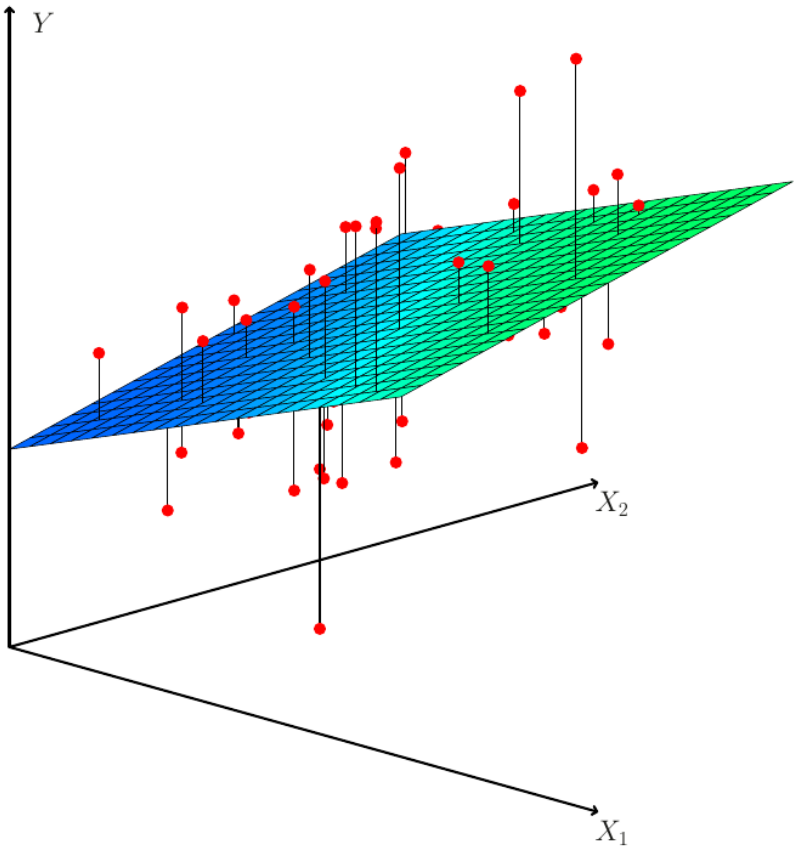
The parameters are estimated using the same least squares approach that we saw for the simple linear regression. We choose the $\widehat{\beta}_j$ to minimize the sum of squared residuals

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_p x_{ip})^2 \end{aligned}$$

The actual values can be computed by using **statistical software packages** like such provided for Python.

Multiple Linear Regression

In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.



Source: Fig. 3.4 from
https://www.statlearning.com/

Multiple Linear Regression Use Cases

When performing multiple linear regression, we usually are interested in answering a few important **questions**.

1. Is at least one of the predictors $X_1, ..., X_p$ 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 Linear Regression Use Cases

When performing multiple linear regression, we usually are interested in answering a few important **questions**.

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

We will now take a look at how to **answer** them.

1. Is There a Relationship Between the Response and Predictors?

Recall:
In the simple linear regression setting, in order to determine whether there is a relationship between the response and the predictor we can simply check whether $\beta_1 = 0$.

In the multiple regression setting with p predictors, we need to ask whether all of the regression coefficients are zero, i.e. whether $\beta_1 = ... = \beta_p = 0$. As in the simple linear regression setting, we use a hypothesis test to answer this question.

We test the null hypothesis,

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

versus the alternative

H_a: at least one β_j is non-zero.

Hypothesis Testing

For Multiple Linear Regression, the hypothesis test is performed by computing the **F-statistic**,

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

where we have, just like with simple linear regression:

$$TSS = \sum (y_i - \bar{y})^2, \qquad RSS = \sum (y_i - \widehat{y_i})^2.$$

Hypothesis Testing

For Multiple Linear Regression, the hypothesis test is performed by computing the **F-statistic**,

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

where we have, just like with simple linear regression:

$$TSS = \sum (y_i - \bar{y})^2, \qquad RSS = \sum (y_i - \widehat{y_i})^2.$$

If the assumptions from the linear model are correct, one can show that

$$E\{RSS/(n-p-1)\} = \sigma^2$$

and that, provided the hypothesis H_0 is true,

$$E\{(TSS - RSS)/p\} = \sigma^2$$

Hypothesis Testing

For Multiple Linear Regression, the hypothesis test is performed by computing the **F-statistic**,

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

where we have, just like with simple linear regression:

$$TSS = \sum (y_i - \bar{y})^2, \qquad RSS = \sum (y_i - \widehat{y_i})^2.$$

If the assumptions from the linear model are correct, one can show that

$$E\{RSS/(n-p-1)\} = \sigma^2$$

and that, provided the hypothesis H_0 is true,

$$E\{(TSS - RSS)/p\} = \sigma^2$$

From this: In the case of no relationship between the response and predictors, the F-statistic will be ≈ 1. On the other hand, if the alternative hypothesis H_a is true, then $E\{(TSS - RSS)/p\} > \sigma^2$, so $F > 1$.

Multiple Linear Regression

In some cases it's clear.
However, what if the F-statistic is been very close to 1? How large does the F -statistic need to be before we can reject H_0 and conclude that there is a relationship?

Multiple Linear Regression

In some cases it's clear.
However, what if the F-statistic is been very close to 1? How large does the F -statistic need to be before we can reject H_0 and conclude that there is a relationship?

The answer 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.

In contrast, a larger F-statistic is needed to reject H_0 if n is small. When H_0 is true and the errors ϵ_i are normally distributed, the F-statistic follows an F-distribution.

For any given value of n and p, statistical software packages can compute the p-value associated with the F -statistic using this distribution.

Multiple Linear Regression

In some cases it's clear.
However, what if the F-statistic is been very close to 1? How large does the F -statistic need to be before we can reject H_0 and conclude that there is a relationship?

The answer 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.

In contrast, a larger F-statistic is needed to reject H_0 if n is small. When H_0 is true and the errors ϵ_i are normally distributed, the F-statistic follows an F-distribution.

For any given value of n and p, statistical software packages can compute the p-value associated with the F -statistic using this distribution.

In the case of a very large number of variables ($p > n$), there are more coefficients β_j than observations from which to estimate them. In this case we can't fit the multiple linear regression model using least squares. Here, other approaches like forward selection (not discussed yet) are applied.

2. Deciding on Important Variables

We assume here that we conclude based on the p-value that at least one of the predictors is related to the response. But which one(s)?

2. Deciding on Important Variables

We assume here that we conclude based on the p-value that at least one of the predictors is related to the response. But which one(s)?

Ideally, we would try out different models, each containing a different predicator subset E.g., if $p = 2$, then we can consider four models:

2. Deciding on Important Variables

We assume here that we conclude based on the p-value that at least one of the predictors is related to the response. But which one(s)?

Ideally, we would try out different models, each containing a different predicator subset E.g., if $p = 2$, then we can consider four models:

- a model containing no variables
- a model containing X_1 only
- a model containing X_2 only
- a model containing both X_1 and X_2.

2. Deciding on Important Variables

We assume here that we conclude based on the p-value that at least one of the predictors is related to the response. But which one(s)?

Ideally, we would try out different models, each containing a different predicator subset E.g., if $p = 2$, then we can consider four models:

- a model containing no variables
- a model containing X_1 only
- a model containing X_2 only
- a model containing both X_1 and X_2.

We can then select the best model out of all of the models that we have considered using statistics appropriate for that, e.g. the Akaike information criterion (AIC), Bayesian information criterion (BIC).

Multiple Linear Regression

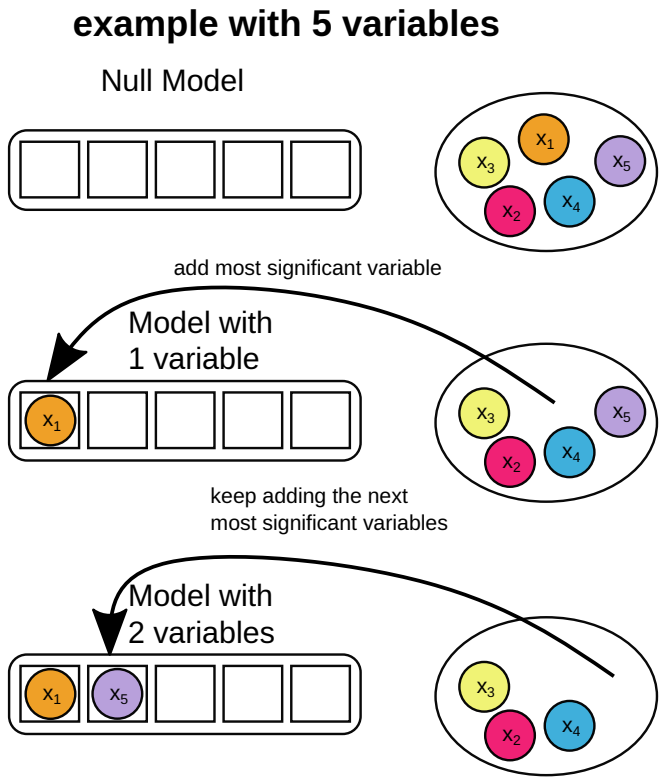
However: Unfortunately, there are a total of 2^p models that contain subsets of p variables. This means that even for moderate p, trying out every possible subset of the predictors is infeasible.

2. Deciding on Important Variables

The three classical approaches suitable for this task are:

1. Forward selection

1. Start with a model that contains no predictors (called the Null Model).
2. Fit p simple linear regressions.
3. Start adding the most significant variables one after the other (based on the lowest RSS). Fit.
4. This is done until a pre-specified stopping condition is reached.

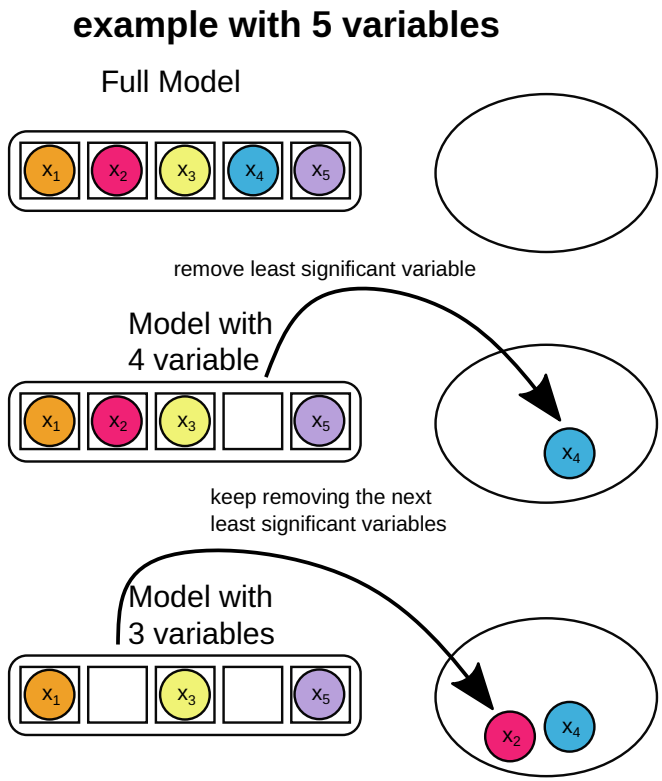


2. Deciding on Important Variables

The three classical approaches suitable for this task are:

2. Backward selection

1. Fit a model that contains all variables under consideration (called the Full Model).
2. Then starts removing the least significant variables (starting with the largest largest p-value) one after the other.
3. Fit the new $(p-1)$-variable model.
4. This is done until a pre-specified stopping rule is reached or until no variable is left in the model.



2. Deciding on Important Variables

The three classical approaches suitable for this task are:

3. Mixed selection

A combination of forward and backward selection: We start with the null model, and as with forward selection, we add the most significant variable. We continue to add variables one-by-one with increasing p-values.

If at any point the p-value for one of the variables in the model is above a certain threshold, we remove that variable from the model.

We 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.

2. Deciding on Important Variables

The three classical approaches suitable for this task are:

3. Mixed selection

A combination of forward and backward selection: We start with the null model, and as with forward selection, we add the most significant variable. We continue to add variables one-by-one with increasing p-values.

If at any point the p-value for one of the variables in the model is above a certain threshold, we remove that variable from the model.

We 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.

When to use which?
Backward selection can't be used if $p > n$, while forward selection can always be used. Forward selection might include variables early that later become redundant. Mixed selection can remedy this.

3. Accuracy of Model Fit

Two of the most common numerical measures of Multiple Linear Regression model fit are the **RSE and** R^2. These quantities are computed and interpreted like for simple linear regression.

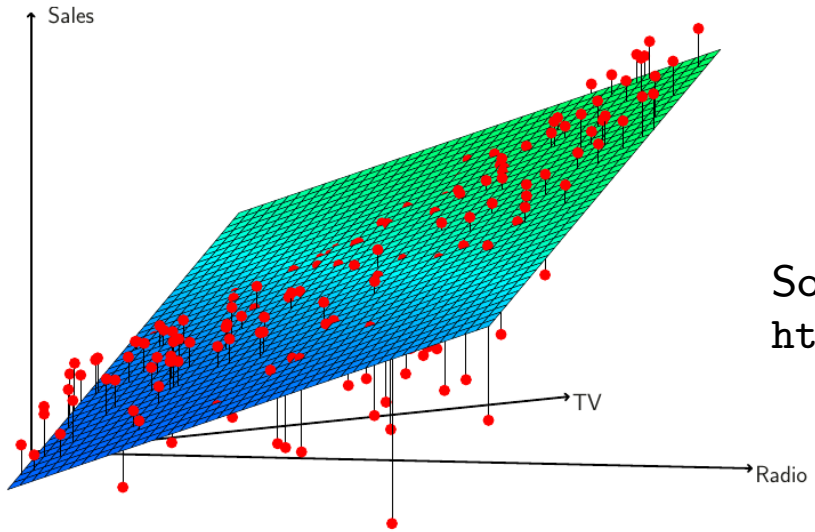
In simple regression, R^2 is the square of the correlation of the response and the variable. In multiple linear regression, it equals $Corr(Y, \widehat{Y})^2$, which is the square of the correlation between response and fitted linear model.

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

In addition to looking at the RSE and R^2 statistics just discussed, it can be useful to **plot** the data. Graphical summaries can reveal problems with a model that are not visible from numerical statistics.

Multiple Linear Regression

example: A linear regression fit to `sales` using `TV` and `radio` as predictors.



Source: Fig. 3.5 from `https://www.statlearning.com/`

From the pattern of the residuals we can see a **pronounced non-linear relationship** in the data.
The positive residuals (above the surface) tend to lie along the 45-degree line, where TV and radio budgets are split evenly. The negative residuals (most not visible) tend to lie away from this line.
The linear model seems to overestimate sales when most of the budget was spent only on either TV or radio. It underestimates sales when the budget was split between the two media. This pronounced non-linear pattern suggests an **interaction effect** between advertising media, whereby combining the media together boosts sales more than any single medium.

4. Predictions

Once we have fit the multiple regression model, it is straightforward to **apply it** in order to predict the response Y on the basis of a set of predictor values $X_1, ..., X_p$.

4. Predictions

Once we have fit the multiple regression model, it is straightforward to **apply it** in order to predict the response Y on the basis of a set of predictor values $X_1, ..., X_p$.

However, there are three kinds of uncertainties associated with this prediction:

- The coefficient estimates $\widehat{\beta}_0, .., \widehat{\beta}_p$ are estimates for $\beta_0, .., \beta_p$. That is, the least squares plane $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + .. + \widehat{\beta}_p X_p$ is only an estimate for the true population regression plane $f(x) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$. We can compute a confidence interval in order to determine how close $\widehat{Y}$ will be to $f(X)$.
- Of course, in practice assuming a linear model for $f(X)$ is almost always an approximation. So there is an additional source of potential reducible error which we call *model bias*.
- Even if we knew $f(x)$, the response vlaue cannot be predicted perfectly because of the random error ϵ in the model.

Qualitative Predictors

So far, we have assumed that all variables in our linear regression model are quantitative. But in practice, this is not necessarily the case; often some predictors are qualitative.

An **example**: A data set that records variables for a number of credit card holders. The response is credit card balance. There are several quantitative predictors: age, number of credit cards, years of education, income, credit limit, credit rating. In addition, we have four qualitative variables: house ownership, student statusus, marital status, region (East, West or South).

Qualitative Predictors with Two Levels

Suppose that we wish to investigate differences in credit card balance between those who own a house and those who don't, ignoring the other variables for the moment.

If a qualitative predictor (also known as a factor) only has two **levels**, or possible values, then incorporating it into a regression model is very simple. We create an **indicator or dummy variable** that takes on two possible numerical values.

Qualitative Predictors with Two Levels

For example, based on the `houseowner` variable:

$$x_i = \begin{cases} 1, & \text{if } i\text{th person owns a house} \\ 0, & \text{if } i\text{th person does not own a house} \end{cases}$$

Using x_i as a predictor in the regression equation results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i, & \text{if } i\text{th person does not own a house} \end{cases}$$

Qualitative Predictors with Two Levels

For example, based on the `houseowner` variable:

$$x_i = \begin{cases} 1, & \text{if } i\text{th person owns a house} \\ 0, & \text{if } i\text{th person does not own a house} \end{cases}$$

Using x_i as a predictor in the regression equation results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i, & \text{if } i\text{th person does not own a house} \end{cases}$$

Now β_0 can be interpreted as the average balance among those who do not own, $\beta_0 + \beta_1$ as the average balance among those who do own a house, and β_1 as the average difference in balance between owners and non-owners.

Qualitative Predictors with Two Levels

For example, based on the `houseowner` variable:

$$x_i = \begin{cases} 1, & \text{if } i\text{th person owns a house} \\ 0, & \text{if } i\text{th person does not own a house} \end{cases}$$

Using x_i as a predictor in the regression equation results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i, & \text{if } i\text{th person does not own a house} \end{cases}$$

Now β_0 can be interpreted as the average balance among those who do not own, $\beta_0 + \beta_1$ as the average balance among those who do own a house, and β_1 as the average difference in balance between owners and non-owners.

The creation of dummy variables that work that way to handle qualitative predictors is known as **one-hot encoding**.

Qualitative Predictors with Two Levels

For example, based on the `houseowner` variable:

$$x_i = \begin{cases} 1, & \text{if } i\text{th person owns a house} \\ 0, & \text{if } i\text{th person does not own a house} \end{cases}$$

Using x_i as a predictor in the regression equation results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i, & \text{if } i\text{th person does not own a house} \end{cases}$$

Now β_0 can be interpreted as the average balance among those who do not own, $\beta_0 + \beta_1$ as the average balance among those who do own a house, and β_1 as the average difference in balance between owners and non-owners.

The creation of dummy variables that work that way to handle qualitative predictors is known as **one-hot encoding**.

The decision to code owners as 1 and non-owners as 0 is arbitrary regarding the regression fit, but does alter the interpretation of the coefficients.

Qualitative Predictors with Two Levels

Alternatively, instead of a 0/1 coding scheme, we could create a dummy variable

$$x_i = \begin{cases} 1, & \text{if } i\text{th person owns a house} \\ -1, & \text{if } i\text{th person does not own a house} \end{cases}$$

It is important to note that again, the final predictions for the credit card balances of owners and non-owners will be identical regardless of the coding scheme used. The only difference is in the way that the coefficients are interpreted.

Qualitative Predictors with More than Two Levels

When a qualitative predictor has more than two levels, we can create additional dummy variables.

Example: for the region variable which can take on South, East, West we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1, & \text{if } i\text{th person is from the South} \\ 0, & \text{if } i\text{th person is not from the South} \end{cases}$$

and the second

$$x_{i2} = \begin{cases} 1, & \text{if } i\text{th person is from the West} \\ 0, & \text{if } i\text{th person is not from the West} \end{cases}$$

Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if } i\text{th person is from the South} \\ \beta_0 + \beta_2 + \epsilon_i, & \text{if } i\text{th person is from the West} \\ \beta_0 + \epsilon_i, & \text{if } i\text{th person is from the East} \end{cases}$$

Qualitative Predictors with More than Two Levels

Now β_0 can be interpreted as the average credit card balance for individuals from the East, β_1 can be interpreted as the difference in the average balance between people from the South versus the East, and β_2 can be interpreted as the difference in the average balance between those from the West versus the East.

You have recognized we encode South, East, West (we don't have North) with two variables:
There will always be one fewer dummy variable than the number of levels. The level with no dummy variable (East in this example) is known as the **baseline**.

Extensions of the Linear Model

The standard linear regression model provides interpretable results and works well on many real-world problems.

Downside: it makes several very restrictive assumptions which are usually violated in practice, two of them being that the relationship between the predictors and response are linear and additive.

The **linearity assumption** states that the change in the response Y associated with a one-unit change in X_j is constant, regardless of the value of X_j.

The **additivity assumption** means that the association between a predictor X_j and the response Y does not depend on the values of the other predictors.

Extensions of the Linear Model

The standard linear regression model provides interpretable results and works well on many real-world problems.

Downside: it makes several very restrictive assumptions which are usually violated in practice, two of them being that the relationship between the predictors and response are linear and additive.

The **linearity assumption** states that the change in the response Y associated with a one-unit change in X_j is constant, regardless of the value of X_j.

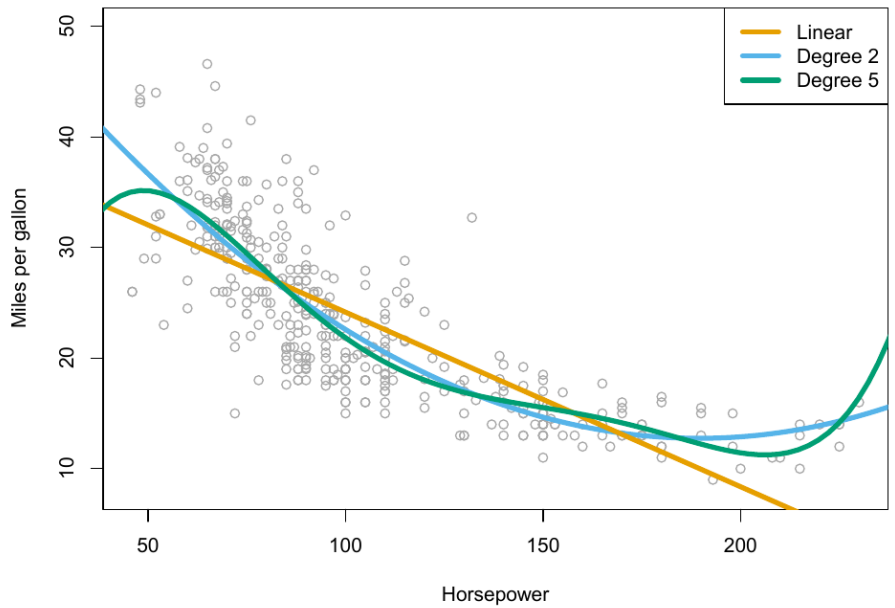
The **additivity assumption** means that the association between a predictor X_j and the response Y does not depend on the values of the other predictors.

We can overcome some of those limitations by extending the linear model.

Non-linear Relationships

The true relationship between predictors and response may be non-linear.

We can directly extend the linear model to accomodate non-linear relationshops by using **polynomial regression**.



This figure gives the `mpg` (gas mileage in miles per gallon) versus `horsepower` for a number of cars. The orange line represents the linear regression fit. A linear regression fit for a model that includes $\texttt{horsepower}^2$ is shown in blue. A linear regression fit for a model that includes all polynomials of horsepower up to a fifth-degree is shown in green.
Source: Fig. 3.8 from `https://www.statlearning.com/`

Extensions of the Linear Model

The approach that we have just described for extending the linear model to accommodate non-linear relationships is known as polynomial regression, since we have included polynomial functions of the predictors in the regression model.

A simple approach for incorporating non-linear associations in a linear model is to include transformed versions of the predictors. For example, the points in in the figure seem to have a quadratic shape, suggesting that a model

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

may provide a better fit.

Extensions of the Linear Model

The approach that we have just described for extending the linear model to accommodate non-linear relationships is known as polynomial regression, since we have included polynomial functions of the predictors in the regression model.

A simple approach for incorporating non-linear associations in a linear model is to include transformed versions of the predictors. For example, the points in in the figure seem to have a quadratic shape, suggesting that a model

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

may provide a better fit.

Despite a non-linear function of `horsepower` is used, this is still a linear model. It is simply a multiple linear regression model with $X_1 =$ `horsepower` and $X_2 =$ `horsepower`^{2}. We can estimate β_0, β_1, β_2 with the methods for linear regression as discussed before.

Potential Problems

When we fit a linear regression model to a particular data set, many **problems** may occur. The **most common** ones are the following:

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.

Potential Problems

When we fit a linear regression model to a particular data set, many **problems** may occur. The **most common** ones are the following:

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.

In the following, we will see a brief summary of these problems.

1. Non-linearity of the Data

The linear regression model assumes that there is a straight-line relationship between the predictors and the response.

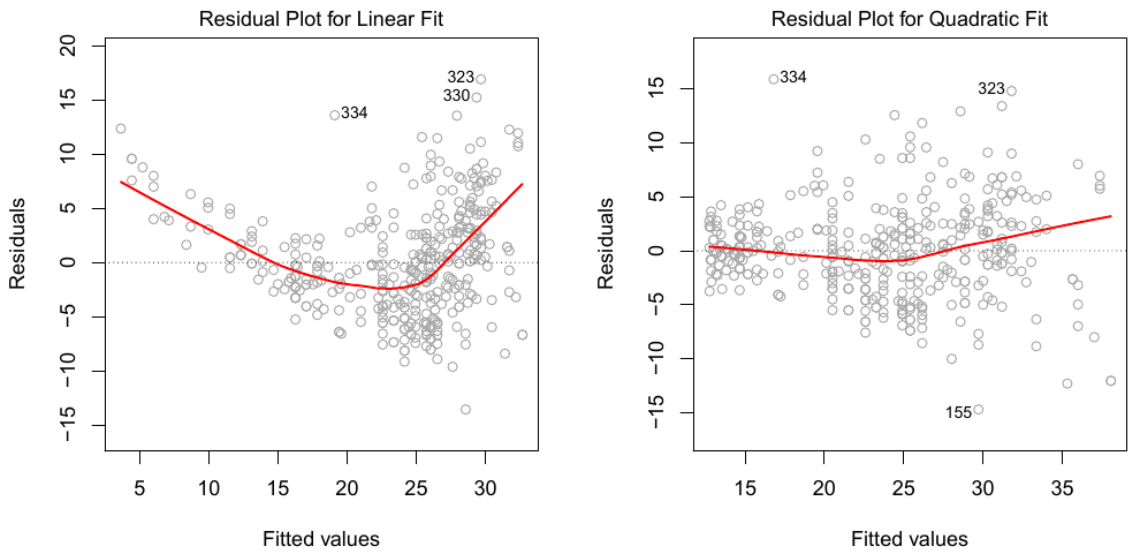
We have already seen how we can overcome non-linearity in our data while still carrying out linear regression.

Residual plots are a useful graphical tool for identifying non-linearity. Given a simple linear regression model, we can plot the residuals, $e_i =$ vs. the predictor x_i. In the case of a multiple regression model, we plot the residuals vs. the predicted values $\widehat{y}_i$.

If the residual plot indicates nonlinearity, then a simple approach is to include transformed versions of the predictors as described.

1. Non-linearity of the Data

In the following plots, pattern in the residuals provide a stong indication of non-linearity in the data.



Plots of residual vs. predicted value. In each plot, the red line is a smooth fit to the residuals, which aids in identfying a trend.
Left: A linear regression of `mpg` on `horsepower`. A strong pattern in the residuals inidcates non-linearity in the data.
Right: A linear regression of `mpg` on `horsepower` and `horsepower`^{2}. There is little pattern in the data.
Source: Fig. 3.9 from `https://www.statlearning.com/`

2. Correlation of Error Terms

An important assumption of the linear regression model is that the error terms are uncorrelated.

2. Correlation of Error Terms

An important assumption of the linear regression model is that the error terms are uncorrelated.

What does this mean?

2. Correlation of Error Terms

An important assumption of the linear regression model is that the error terms are uncorrelated.

What does this mean?

For instance, if the errors are uncorrelated, then *the fact that i is positive provides little or no information about the sign of $i + 1$.*

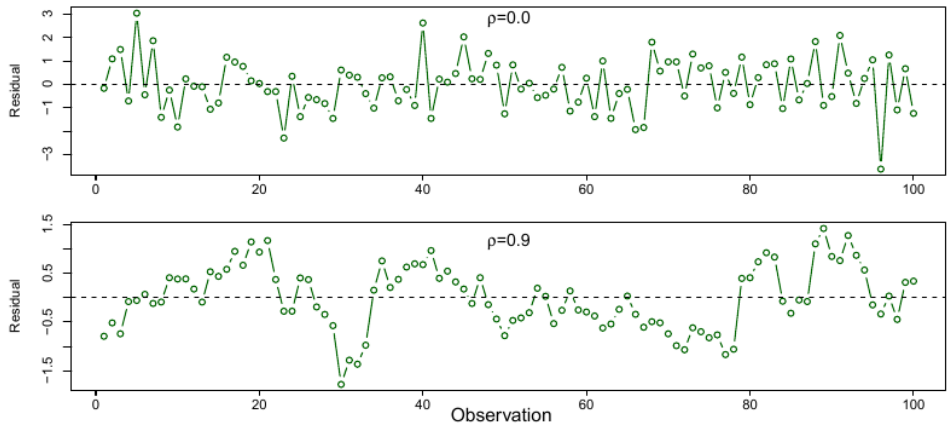
The standard errors that are computed for the estimated regression coefficients or the fitted values are based on the assumption of uncorrelated error terms. If in fact there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors. As a result, confidence and prediction intervals will be narrower than they should be.

2. Correlation of Error Terms

Correlated terms can lead to a wrong sense of confidence in our model.

For example, a 95 % CI may in reality have a much lower probability of containing the true parameter value. Underestimated p-values will give the erroneous impression that a parameter is statistically significant.

Error term correlations frequently occur in **time series data**. Often, observations from adjacent time points have positively correlated errors. Residual plots help to determine if this is the case for a given data set.

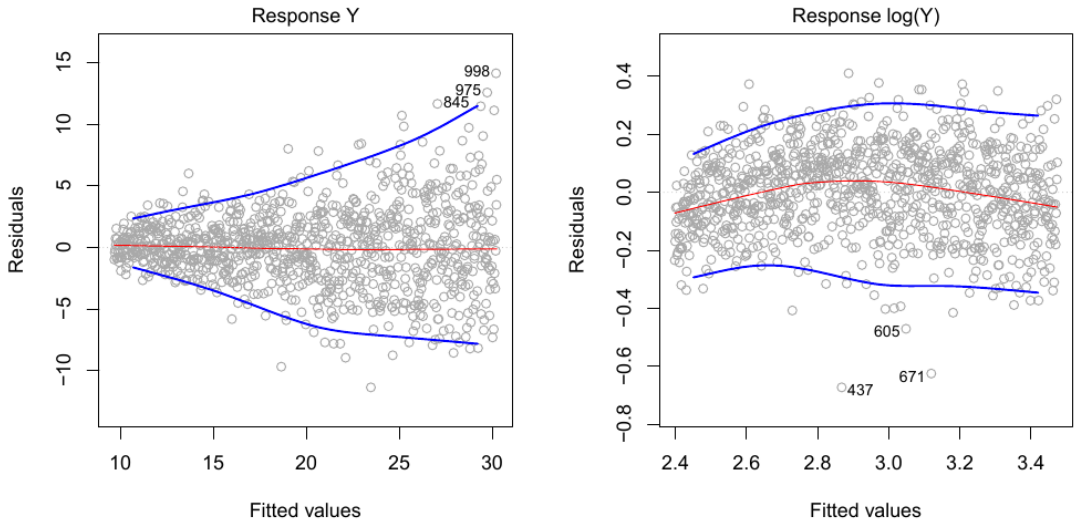


Top panel: There is no evidence of a time-related trend in the residuals.
Bottom panel: Residuals from data in which adjacent errors have a correlation of 0.9. There is a clear pattern: adjacent residuals tend to take on similar values.

Source: Fig. 3.10 from `https://www.statlearning.com/`

3. Non-constant Variance of Error Terms

Another important assumption of the linear regression model is that the error terms have a constant variance, $Var(\epsilon_i) = \sigma^2$. The standard errors, confidence intervals, and hypothesis tests associated with the linear model rely on this assumption.

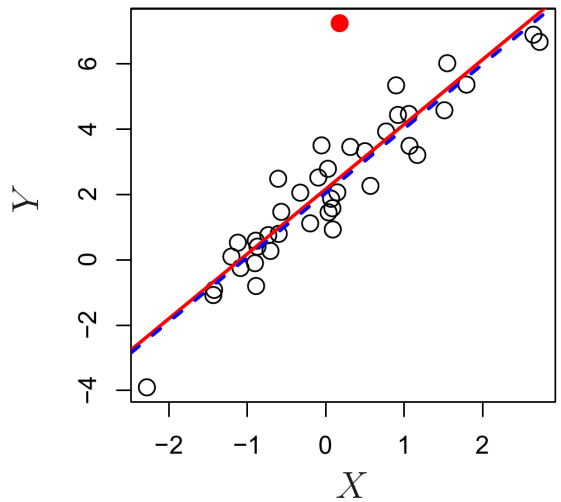


Plots of residual vs. predicted value. In each plot, the red line is a smooth fit to the residuals, which aids in identfying a trend. The blue lines track the outer quantiles of the residuals.
Left: the funnel shape indicates heteroscedasticity.
Right: The response has been log transformed; no there is no evidence for heteroscedasticity.
Source: Fig. 3.11 from `https://www.statlearning.com/`

4. Outliers

An outlier is a point for which y_i is far from the value predicted by the model. Outliers can arise for a variety of reasons, such as incorrect recording of an observation during data collection.

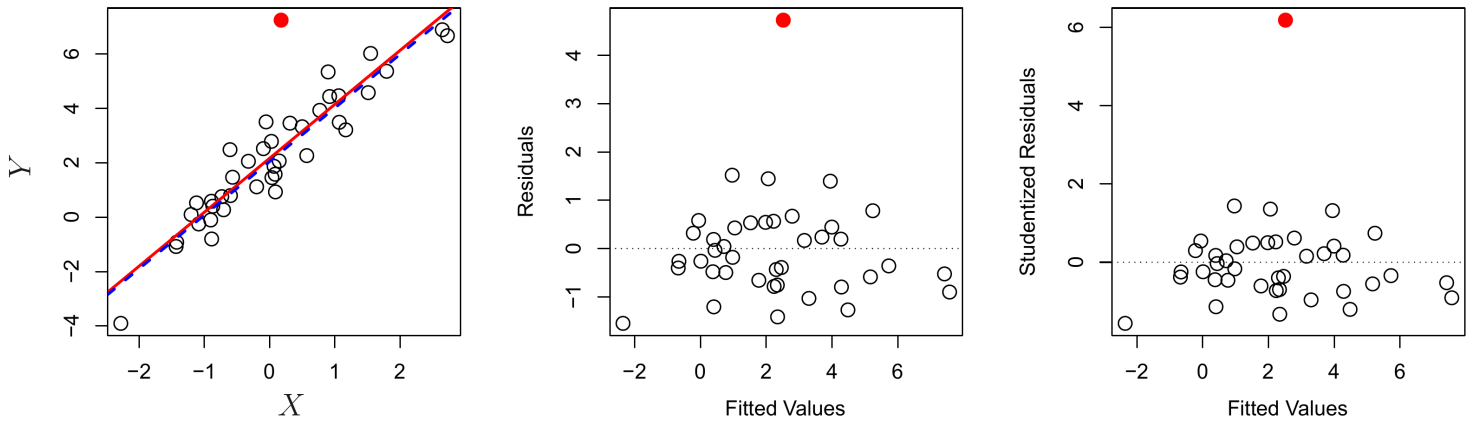
Even if an outlier does not have much effect on the least squares fit, it can cause other problems, such as wrong confidence intervals.



Left: The outlier is shown as a red data point. The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue. In this case, removing the outlier has little effect on the least squares line: it leads to almost no change in the slope, and a tiny reduction in the intercept.
Source: Fig. 3.12 from `https://www.statlearning.com/`

4. Outliers

Residual plots can be used to identify outliers. To address the problem of deciding on residual cut-offs, this problem, instead of plotting the residuals, we can plot the **studentized residuals**, computed by dividing each residual e_i by its estimated standard error. Observations whose studentized residuals are greater than 3 in absolute value are possible outliers.



Left: The outlier is shown as a red data point. The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue. In this case, removing the outlier has little effect on the least squares line: it leads to almost no change in the slope, and a miniscule reduction in the intercept.
Center: The residual plot clearly identifies the outlier.
Right: Also the studentized residuals clearly identify the outlier.
Source: Fig. 3.12 from `https://www.statlearning.com/`

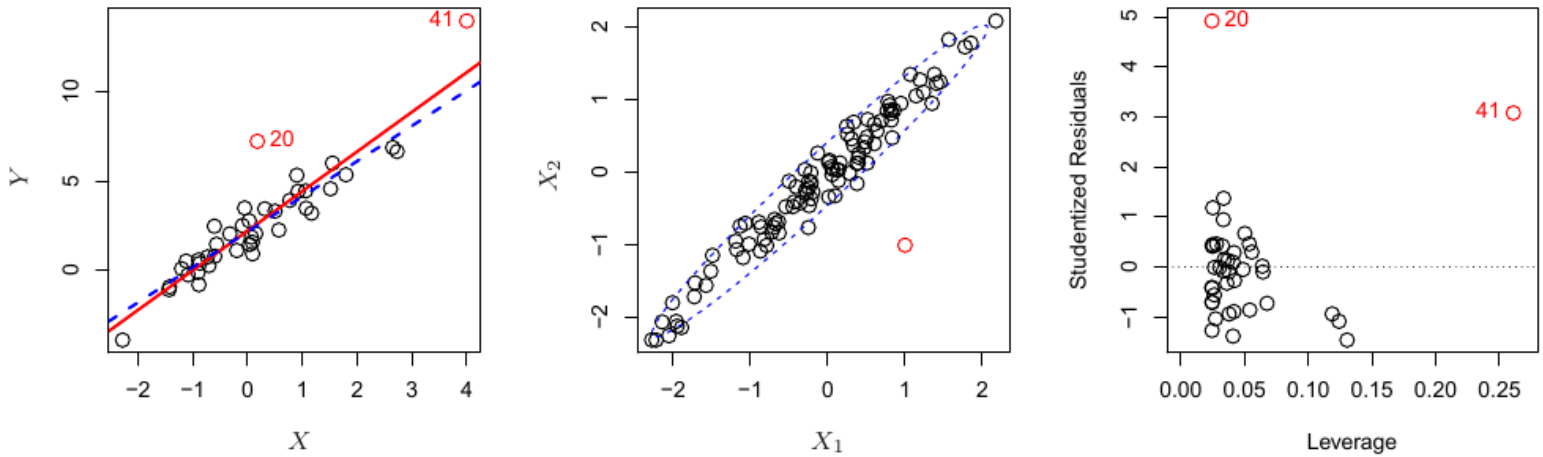
4. Outliers

If we believe that a data point is an outlier, one solution is to simply remove it. However, care should be taken, since an outlier may instead indicate a deficiency with the model, such as a missing predictor.

5. High Leverage Points

In contrast to outliers which have an unusual response y_i given the predictor x_i, observations with high leverage have an unusual value for x_i.

High leverage observations tend to have a sizable impact on the estimated regression line. It is thus important to identify high leverage observations.



Left: Observation 41 is a high leverage point; its predictor value is large relative to the other observations. Observation 20 is an outlier. The red line is the fit to all data, and the blue one is the fit with observation 41 removed.
Center: The red observation is not unusual in terms of its X_1 or 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.
Source: Fig. 3.13 from `https://www.statlearning.com/`

5. High Leverage Points

In order to quantify an observation's leverage, we compute the **leverage statistic**. A large value of this statistic indicates an observation with high leverage. For a simple linear regression, the leverage statistic is

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

5. High Leverage Points

In order to quantify an observation's leverage, we compute the **leverage statistic**. A large value of this statistic indicates an observation with high leverage. For a simple linear regression, the leverage statistic is

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

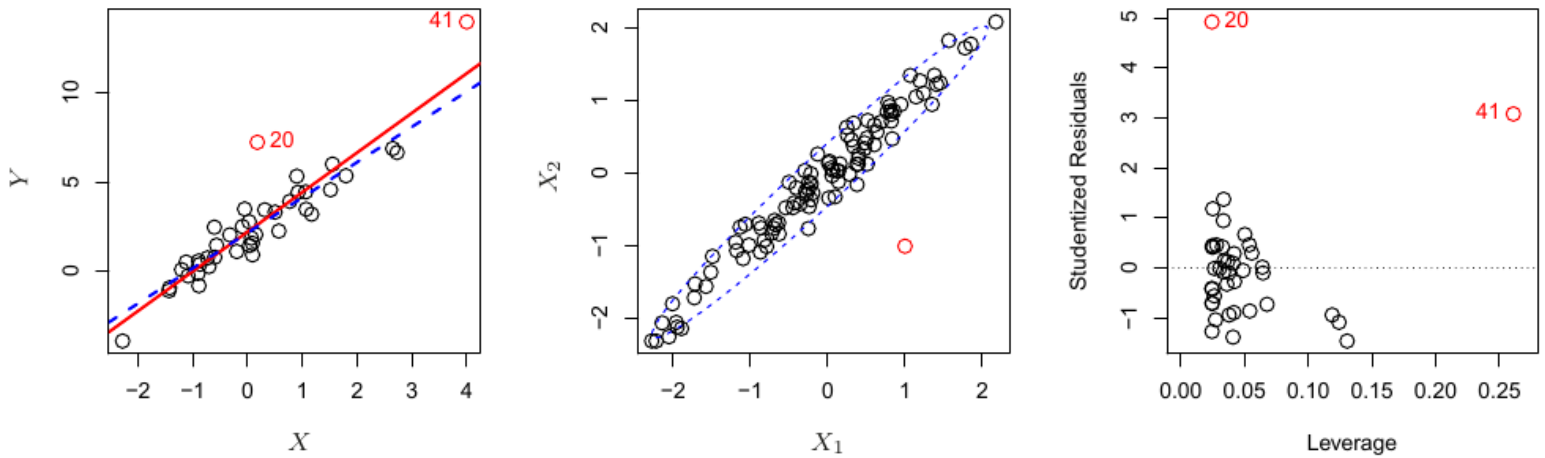
It is easy to see from this equation that h_i increases with the distance of x_i from $\bar{x}$.

The leverage statistic h_i is always between $1/n$ and 1, and the average leverage for all the observations is always equal to $(p+1)/n$. So 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.

5. High Leverage Points

Going back to the figure:

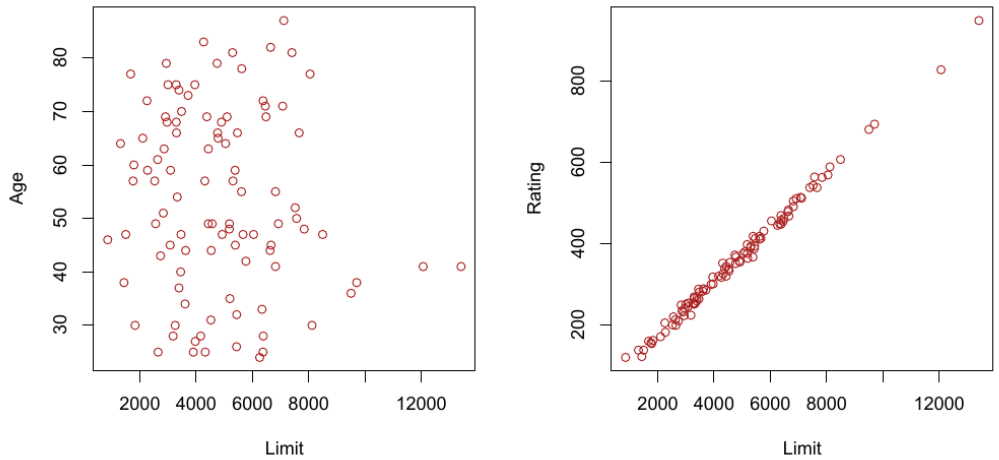
The right panel shows the studentized residuals versus h_i for the data in the left-hand panel. Observation 41 has a very high leverage statistic as well as a high studentized residual: It is an outlier as well as a high leverage observation. The plot also reveals why observation 20 has relatively little effect on the least squares fit the left panel: it has low leverage.



Source: Fig. 3.13 from `https://www.statlearning.com/`

6. Collinearity

Collinearity refers to the situation in which two or more predictor variables are closely related to one another. The concept of collinearity is illustrated in the figure below using the credit card data set.

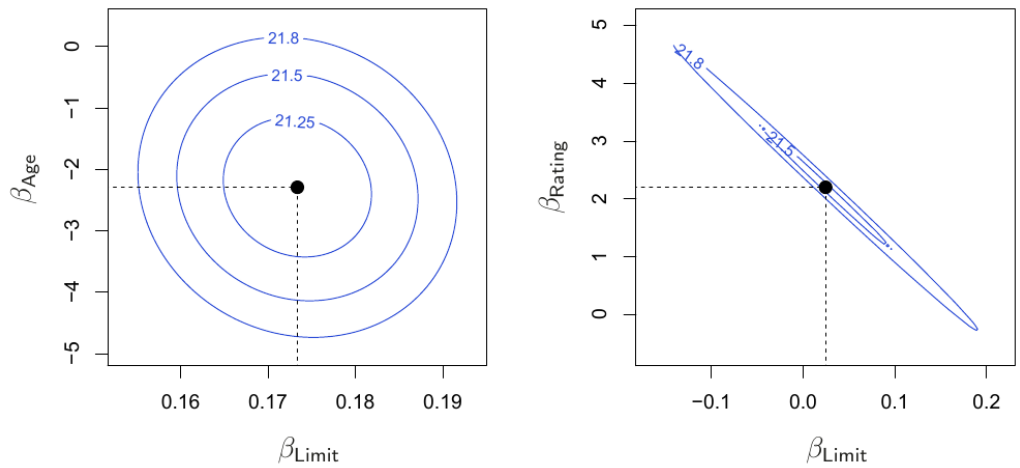


Scatter plots of the observations from the credit card data set.
Left: A plot of age vs. `limit`. These two variables are not collinear.
Right: A plot of `rating` vs. `limit`. These two variables are highly collinear.
Source: Fig. 3.14 from `https://www.statlearning.com/`

In the left-hand panel, the two predictors limit and age show no obvious relationship. In contrast, in the right-hand panel, the predictors limit and rating are very highly correlated with each other - that they are collinear.

6. Collinearity

The presence of collinearity can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response. In other words, since limit and rating tend to increase or decrease together, it can be difficult to determine how each one separately is associated with the response, balance.



Contour plots for the RSS values as a function of the parameters β. Each ellipse represents a set of coefficients that correspond to the same RSS. The black dots represent the coefficient values corresponding to the minimum RSS.
Left: RSS for the regression of `balance` on age and `limit`, well-defined minimum.
Right: RSS for the regression of `balance` on `rating` and `limit`. Because of the collinearity, there are many pairs $(\beta_{Limit}, \beta_{Rating})$ with a similar value for RSS.
Source: Fig. 3.15 from `https://www.statlearning.com/`

6. Collinearity

Since collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for $\widehat{\beta}_j$ to grow. Recall that the t-statistic for each predictor is calculated by dividing $\widehat{\beta}_j$ by its standard error. Consequently, collinearity results in a decline in the t-statistic. As a result, in the presence of collinearity, we may fail to reject $H_0 : \beta_j = 0$. This means that the power of the hypothesis test, the probability of correctly detecting a non-zero coefficient, is reduced by collinearity

A simple way to **detect collinearity** is to look at the correlation matrix of the predictors. An element of this matrix that is large in absolute value indicates a pair of highly correlated variables, and therefore a collinearity problem in the data.

6. Collinearity

Unfortunately, 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. We call this situation **multicollinearity**.

Instead of inspecting the correlation matrix, a better way to assess multi-collinearity is to compute the variance inflation factor (VIF). The VIF is the ratio of the variance of $\widehat{\beta}_i$ when fitting the full model divided by the variance of $\widehat{\beta}_i$ if fit on its own. (More on this can be found e.g. in the book at `https://www.statlearning.com/`)

Possible **solutions** in the case of collinearity are:

- drop one of the problematic variables from the regression

This can be done as the presence of collinearity implies redundancy.

- combine the collinear variables into a single predictor:

example: Combine the average of standardized versions of limit and rating in order to create a new variable that measures credit worthiness.

Summary

Today we have learned about the basics of Linear Regression.

We also looked into the topics of assessing our model accuracy, and which possible problems with our data can make linear regression results unreliable.

We have, in addition, seen possible ways to extend our model without deviating from the principle of linear regression.

Outlook

Next time we will see more on **Linear Model Selection and Regularization**.