University of Glasgow

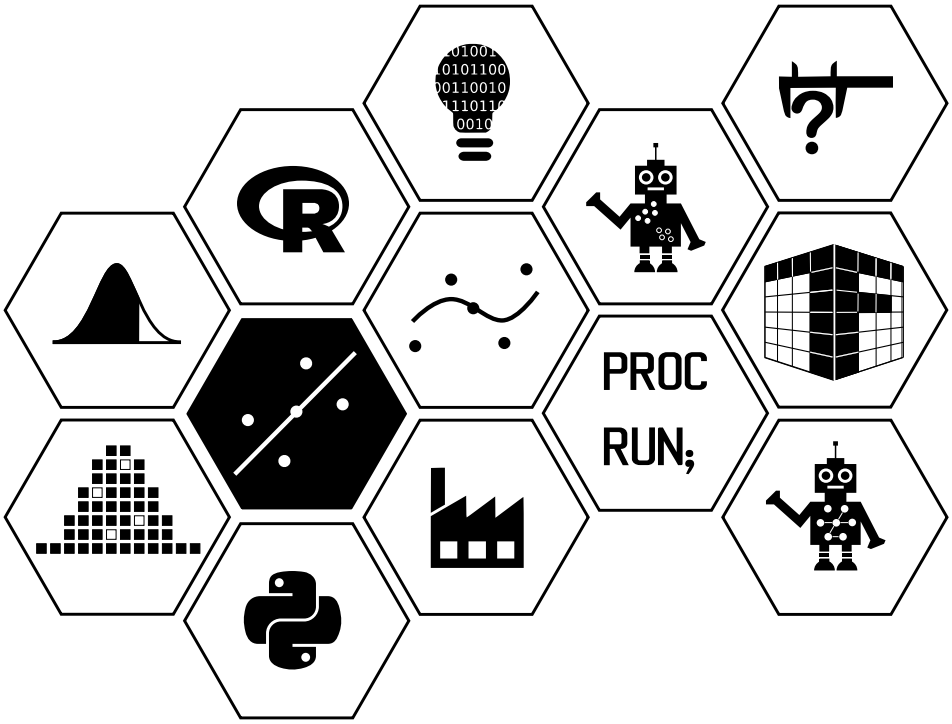
Predictive Modelling

Surajit Ray and Dimitra Eleftheriou

Academic Year 2021-22

Week 1:

Introduction and scope of this course



DATA ANALYTICS GLASGOW

What are Predictive Models?

Introduction and scope of this course

https://youtu.be/fjYnw6rUI6o

Duration: 7m56s

Predictive modeling is the collection of statistical techniques that have the common goal of finding a relationship between a target, "response" variable and various "predictors" or "covariates", with the goal in mind of predicting future values of the target variable given the values of the predictor. Equally important is to understand the effect that each of the predictors have on the response in a model. Predictive modeling is one of the most exciting areas in data analytics. It is the method by which data are used to guide important decisions for smart business operations.

This course is designed to provide a gentle introduction of some of the most useful statistical tools for predictive modeling. We will start with linear regression and introduce you to some extensions of linear regression. While discussing the concepts of predictive modeling, we will also mention other statistical and non-statistical approaches to predictive modeling; most of which will be taught in other courses during your master's studies (e.g. Advanced Predictive Models, and the two courses on Data Mining and Machine Learning). Along the way we will use concepts that you have already learnt in Probability and Stochastic Models, R Programming and even some that you will be learning in the course Learning from Data

Before we start with the theory of predictive models, we will first discuss a few examples where predictive modeling techniques are widely used. For each question of interest we will identify: the response variable, the predictors, and the type of the response and predictors whose relationship we are interested in. In each case, the dataset used can be loaded through `R` libraries. The relavant libaries are given in each example.

Example 1 (Alcohol consumed and blood alcohol content).

In a study of alcohol consumption and related blood alcohol content, 16 student volunteers at Ohio State University drank a randomly assigned number of cans of beer. Thirty minutes later, a police officer measured their percent blood alcohol content (BAC). The researcher is interested in finding out if the number of cans of beers influences the BAC measurement.

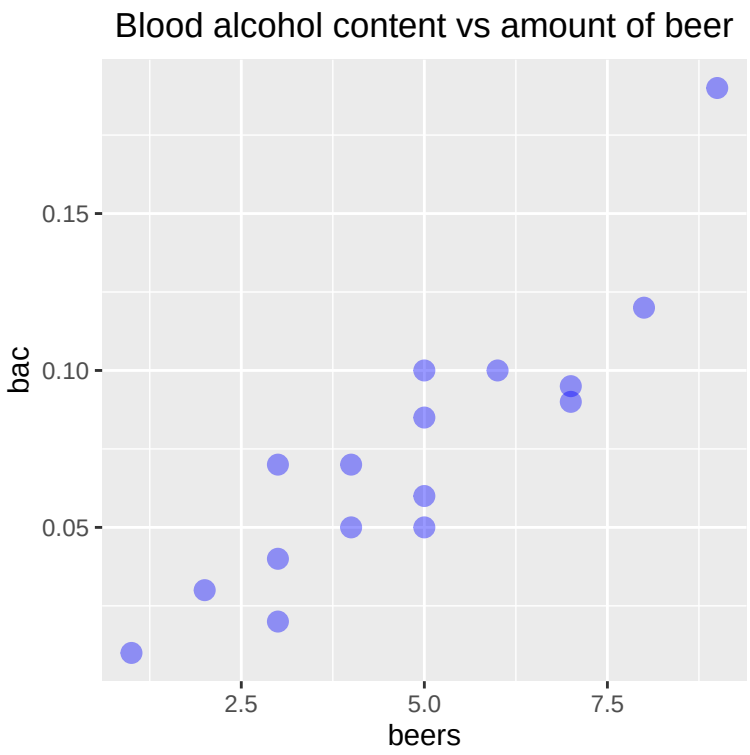
From the question it is clear that the **response** variable is `BAC` and the only **predictor** is the number of cans of `Beers`. The 16 measurements from the dataset and the corresponding scatterplot are given below.

student	beers	bac
1	5	0.100
2	2	0.030
3	9	0.190
4	8	0.120
5	3	0.040
6	7	0.095
7	3	0.070
8	5	0.060
9	3	0.020
10	5	0.050
11	4	0.070
12	6	0.100
13	5	0.085
14	7	0.090
15	1	0.010
16	4	0.050

In this example we see that both the variables are numeric and the relationship between the variable `BAC` and the number of cans of beers appear to be have a positive association. In fact, as the **response** variable `BAC` is a continuous

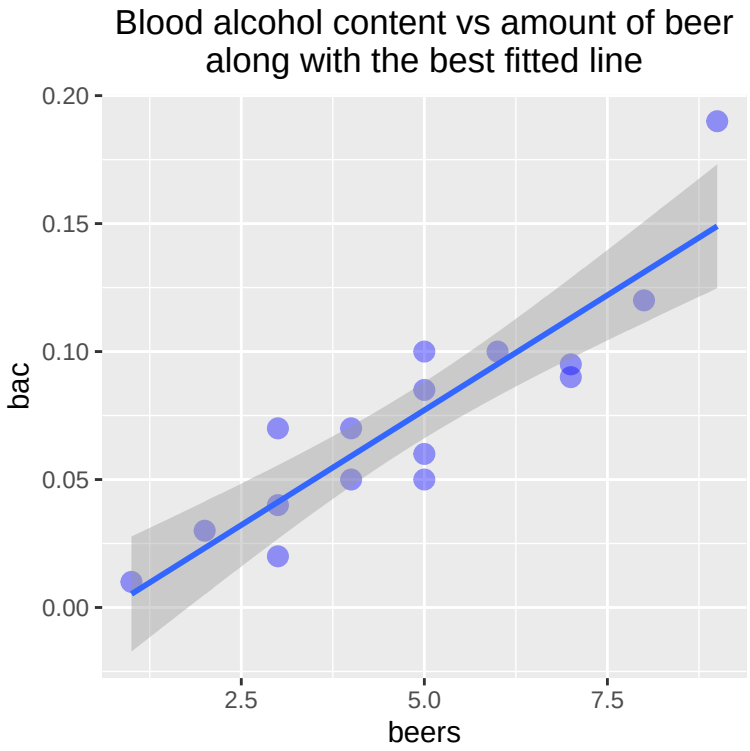
random variable and we have only one numeric **predictor**, `Beers`, and the relationship appears to be a straight line, we can use simple linear regression to model the relationship. In fact, the blue line in the second plot is the best fitted linear relationship and the shaded region gives the standard error.

```
library(openintro)
library(ggplot2)
data(bac)
bac.plot<-ggplot(bac, aes(x = beers, y = bac)) +
          geom_point(size=3.2, alpha = 0.4, col="blue") +
          ggtitle("Blood alcohol content vs amount of beer")
bac.plot
```




```
bac.plot.line<-bac.plot+
               geom_smooth(method = "lm",fullrange=TRUE)+
               ggtitle("Blood alcohol content vs amount of beer\nalong with the best fitted line")
bac.plot.line
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Example with multiple predictors

Let's discuss another example with two possible predictors.

Example 2 (Babies birthweight).

In the babies dataset we are interested in finding out if the birth-weight of newborns depends on both gestation period and the mother's age.

The variables we are interested in are:

- **bwt** - birth-weight, in ounces
- **gestation** - length of gestation, in days
- **age** - mother's age in years

It is clear from the research question that `bwt` is the response variable, and `gestation` and `age` are the predictors.

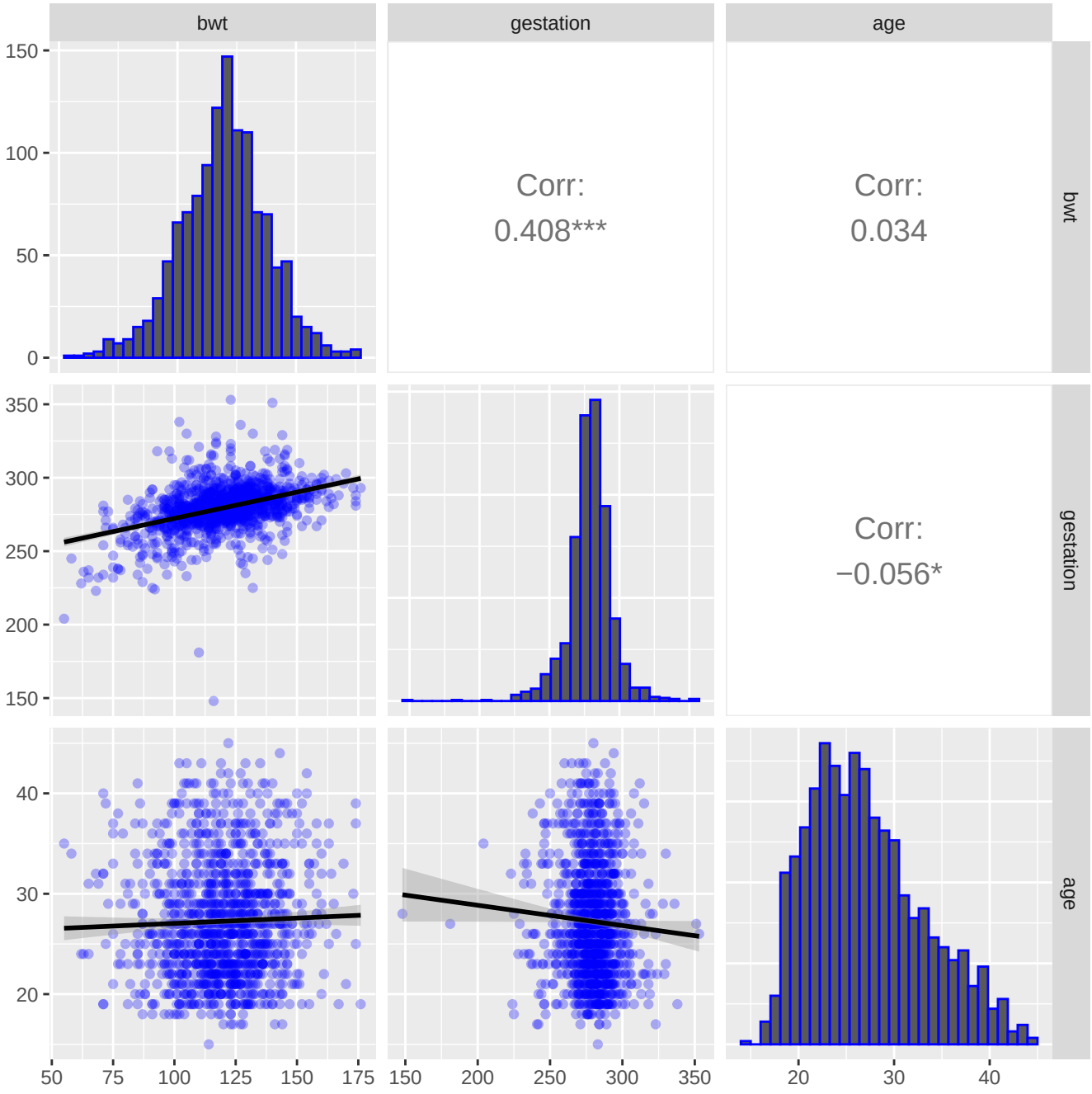
The first ten entries of the dataset are given below.

case	bwt	gestation	parity	age	height	weight	smoke
1	120	284	0	27	62	100	0
2	113	282	0	33	64	135	0
3	128	279	0	28	64	115	1
4	123	NA	0	36	69	190	0
5	108	282	0	23	67	125	1
6	136	286	0	25	62	93	0
7	138	244	0	33	62	178	0
8	132	245	0	23	65	140	0
9	120	289	0	25	62	125	0
10	143	299	0	30	66	136	1

We will now plot two variables at a time in order to explore their relationship. In fact all pairs can be plotted using the `pairs` or `ggpairs` command.

```
library(GGally)
```

```
ggpairs(babies[,c(2,3,5)], lower = list(continuous = wrap("smooth", alpha = 0.3, color = "blue")),
  diag = list(continuous = wrap("barDiag", colour = "blue")),
  upper = list(continuous = wrap("cor", size = 5)))
```



We can see that the response and two predictors are continuous variables, and that the relationship exhibited in the plot of `bwt` vs `gestation` shows a linear trend, while `bwt` vs `age` shows no clear relationship. One option here is to start with a linear regression model, that includes both `gestation` and `age` as predictors; then we can check whether we should drop the variable `age` from the model. This technique is known as variable selection, we will study this in greater detail in the later part of the course.

Example with binary response

Now that you have seen two examples where the response variables are continuous random variables, we will now look at an example where the response is a binary variable.

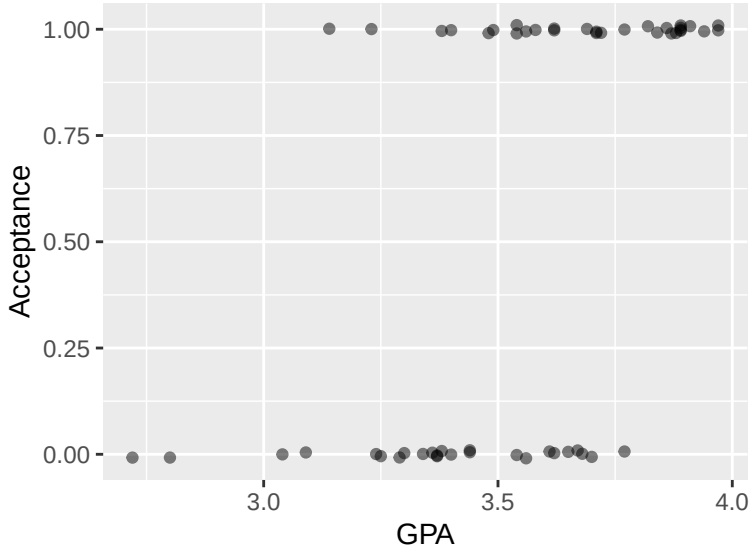
Example 3 (GPA and medical school).

To understand whether college GPA can predict whether a student is accepted for studying at a US medical school, we will explore a dataset on Medical school admission status which also has information on GPA and standardized test scores for 55 medical school applicants from a liberal arts college in the US Midwest.

First 6 rows of the dataset are given below.

Accept	Acceptance	Sex	BCPM	GPA	VR	PS	WS	BS	MCAT	Apps
D	0	F	3.59	3.62	11	9	9	9	38	5
A	1	M	3.75	3.84	12	13	8	12	45	3
A	1	F	3.24	3.23	9	10	5	9	33	19
A	1	F	3.74	3.69	12	11	7	10	40	5
A	1	F	3.53	3.38	9	11	4	11	35	11
A	1	M	3.59	3.72	10	9	7	10	36	5

```
library(Stat2Data)
data(MedGPA)
library(ggplot2)
medgpa.plot <- ggplot(data = MedGPA, aes(y = Acceptance, x = GPA)) +
               geom_jitter(width = 0, height = 0.01, alpha = 0.5)
medgpa.plot
```

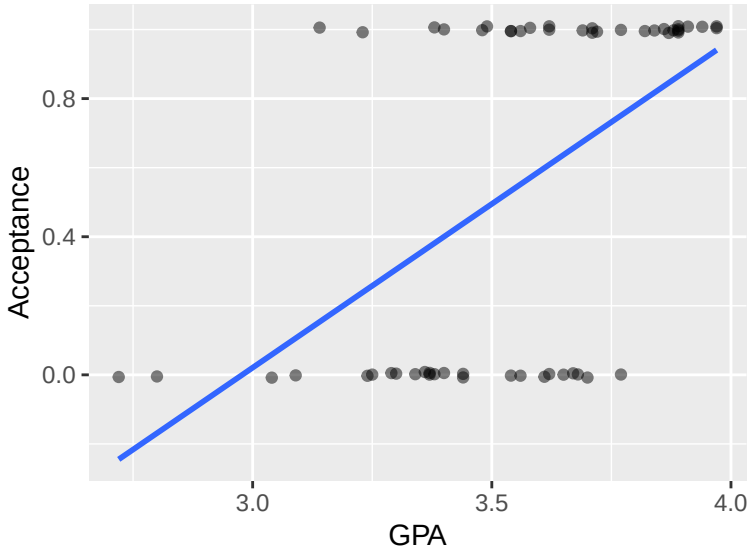


We first plot the two variables, acceptance on the y-axis and GPA on the x-axis. We jitter the points slightly so that multiple points are clearly visible. Note that the variable `acceptance` takes on only two values 0 and 1, which clearly indicates that it is a binary random variable. We also notice that high GPA is associated with acceptance, whereas students with relatively low GPA tends not to be accepted.

If we model the relationship between `acceptance` and `GPA` in a similar fashion as the previous examples with continuous response variables, the best fitted line will be given by

```
medgpa.plot +
  geom_smooth(method = "lm", se = FALSE)
```

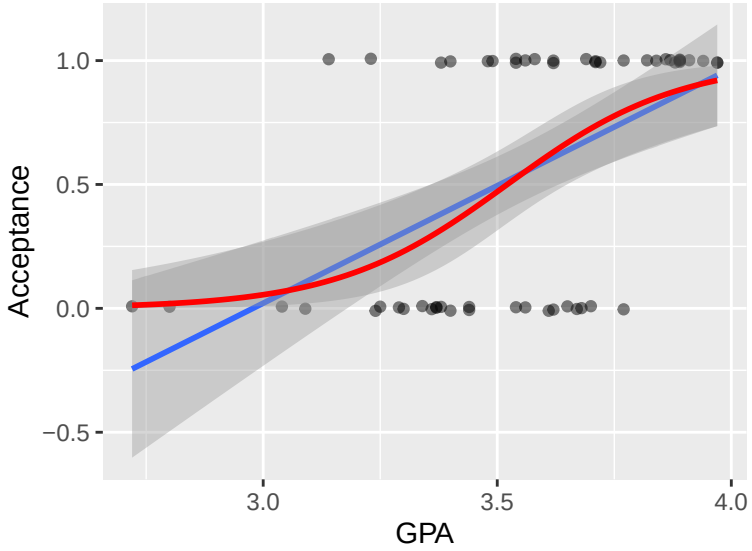
```
## `geom_smooth()` using formula 'y ~ x'
```



Of course, the fit is not very good. Moreover, the predicted values of the response can go from minus $-\infty$ to ∞, whereas the response variable can only take the value 0 or 1. Additionally, this model will fail many of the standard assumptions for a linear model, which will be discussed in detail later in this course. This is a clear classification problem with acceptance equal to 0 or 1. The best option is to use anthor type of model, a logistic regression, to model the probability of acceptance. In the plot below, the red curve gives the logistic regression, explaining how `GPA` predicts the probability of acceptance.

```
medgpa.plot +
  geom_smooth(method = "lm") +
  geom_smooth(method = "glm", color = "red",
              method.args = list(family = "binomial"))
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

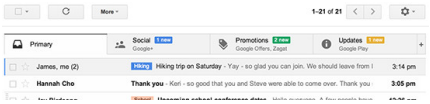


This is a relatively better fit of the data. Note that predicted values of probability are between 0 and 1, as desired. We might need to include other predictors to provide a better fit, but that can be done under the same modeling framework.

Another example where we might encounter categorical random variable as a response variable is an email application, predicting whether a particular email is a Spam or not. The response here is a binary random variable. You can look at this website discussing the spam filtering problem and some approaches to solve it

Example with categorical response

Expanding on the email filter example, **Gmail** categorizes the non-spam emails into one of several categories: primary, social, or marketing, depending on the content in the email. This is an example where the response is categorical with more than two categories.



Whenever you have more than two categories you should further investigate whether these categories are ordered (e.g. a Likert scale), or un-ordered (e.g. favourite colour) as we will usually need different techniques for the two subcases. For the gmail example it can be either, depending on the user. Some users may find all three categories to be equally important, but look at the categories of emails at different times of the day. On the other hand, some users will mainly look at the primary category, gloss over the social category after work, and possibly never look at the marketing category. For those users we can see that there is a clear ranking between the three categories which indicates an ordered set of categorical responses.

Example with several response variables

In some cases we might have several response variables of different data types, and we might need different predictive models for each of these response variables. You will be provided with a detailed discussion on variable types in the course **Learning from data**.

Now, I encourage you to complete the following task of extracting the possible response variables and to determining their data types and propose possible predictive models.

Task 1 (Car Insurance industry).

This dataset comes from one year of insurance policies from 2004 and 2005, there are over 60,000 observations. I have extracted 6 rows from the dataset to show you the different columns and the range of values they take

veh_value	exposure	clm	numclaims	claimcst0	veh_body	veh_age	gender	area	agecat	X_OBSTAT_
1.06	0.3039014	0	0	0	HBACK	3	F	C	2	01101 0 0 0
1.03	0.6488706	0	0	0	HBACK	2	F	A	4	01101 0 0 0
3.26	0.5694730	0	0	0	UTE	2	F	E	2	01101 0 0 0
4.14	0.3175907	0	0	0	STNWG	2	F	D	2	01101 0 0 0
0.72	0.6488706	0	0	0	HBACK	4	F	C	2	01101 0 0 0
2.01	0.8542094	0	0	0	HDTOP	3	M	C	4	01101 0 0 0

We can see that most of these individual policies are claim free.
To check the first six entries with a non-zero claim issue the command below. The command `?dataCar` will provide information about each of the variables in the dataset. You may also want to check the summary of each of the columns using the command `summary`.

```
library(insuranceData)
data(dataCar)
head(dataCar[dataCar$numclaims>0,]) # entries with non-zero claim
?dataCar
summary(dataCar)
```

The car insurance industry is interested in claim frequency, claim severity and no-claims bonus modeling. Your first task is to determine possible response variables that will help the insurance industry model claim frequency, claim severity and no claims bonuses.

Task 2 (Type of response).

Now based on your answer for Task 1, determine the possible data types for each of these response variables.

Note that these response variables might be modeled by predictors such as vehicle body, vehicle age, gender of the insurer, which area of the country they are from, their age-group, and several other variables.

In fact, a linear regression is not appropriate for any of the predictors in the above example and you will need to use generalized linear modeling techniques to model these responses. These techniques will be discussed in detail in the Advanced Predictive Models course.

You might also be familiar with several "model free" techniques, especially for categorical responses, e.g. neural networks, classification and regression trees and k-nearest neighbour. I will mention those alternative approaches while discussing specific examples, but you will see a detailed discussion of these techniques in the two courses on data mining and machine learning.

https://glasgow.rl.talis.com/courses/stats5076.html

If you are interested in seeing more examples of the applications of predictive modeling please read:

- Section 1.3 from Regression analysis by example - Samprit Chatterjee, Ali S. Hadi

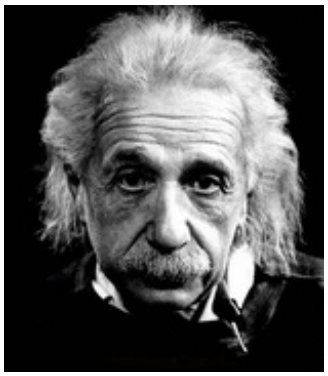
Scope of this course

In this course we will focus on linear regression characterized by the response variable being a continuous random variable, ideally following a normal distribution.

Steps in Predictive Modeling

While performing a statistical investigation you should start by asking the right research question, choosing the right variable(s) and proceeding to collect the right data. Often we already have the data, and thus our primary job would be to make sure that the data can be used to answer the question of interest. We then proceed to model building, model fitting and model checking. If the model assumptions are not true you might need to perform the last three steps again. Finally, we interpret the model and answer the research question that we started with.

Let us start this discussion with some words of wisdom.



"The formulation of the problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill." — Albert Einstein

The steps of model selection can be summarized as follows

1. Statement of problem
2. Selection of variables
3. Data collection (sometimes we already have data)
4. Model specification

5. Choice of model fitting
6. Model fitting
7. Model checking and criticism
8. Interepreting the model

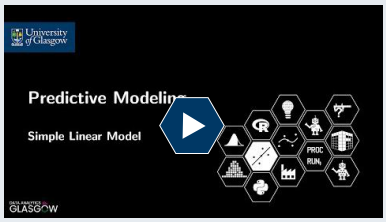
Reading material

https://glasgow.rl.talis.com/courses/stats5076.html

- Section 1.4 from Regression Analysis by Example - Samprit Chatterjee, Ali S. Hadi
- Section 1.1 from Linear Models with R - Julian James Faraway

In this course we will follow this framework and work through different examples. In different weeks of the course we will focus on different steps of this process; once we have covered the individual topics we will come back to look at the examples presented in this week, and analyze the data following the above framework.

Simple Linear Regression



Simple Linear Regression

https://youtu.be/5z72jz8y7Pw

Duration: 3m18s

We will start with the simplest predictive model — where we study the relationship between a single response, variable Y, and only one predictor, variable X. We will develop a "simple linear model" and show how to estimate and interpret the model. Don't worry, we will expand on why we call it a simple linear model.

To get us started let us discuss how we might use predictive modeling tools to answer the following questions:

- How does the price of a car depend on its age?
- Is there a linear relationship between a mother's height and her daughter's height?
- Is there a linear relationship between blood pressure measurements in the left and right arms of humans?
- How does air temperature depend on CO_2?
- Can the number of hospital admissions from respiratory illness be predicted by levels of air pollution?

Motivational Examples

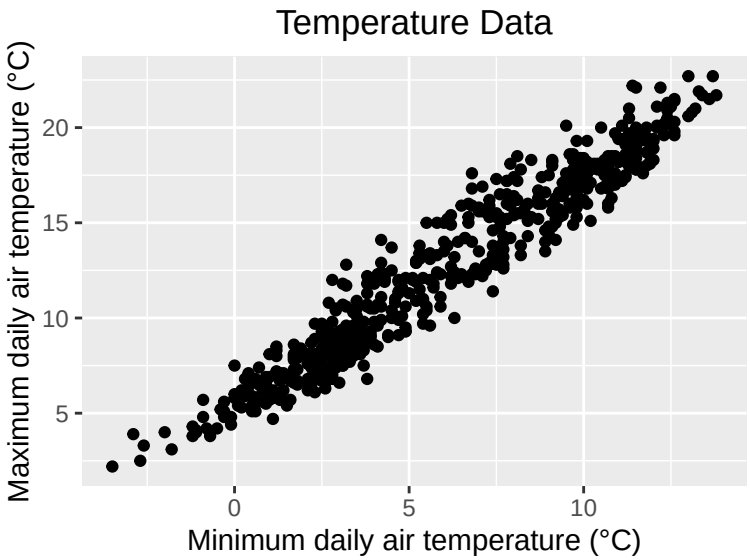
Firstly, we will discuss a series of examples to introduce and motivate the main ideas of linear models, and point out cases where a linear model might be appropriate and where they might not be applicable.

You can find all datasets used thoughout this course **here**.

Example 4 (Temperature).

The maximum and minimum daily air temperature was recorded at Paisley, Glasgow, over the last 50 years. These temperatures are displayed on the scatterplot below. What can we say about the relationship between maximum and minimum air temperature from the plot below?

```
#temp<-read.csv("temp.csv")
qplot(tmin, tmax, data=temp, xlab="Minimum daily air temperature (°C)",
      ylab="Maximum daily air temperature (°C)", main="Temperature Data")
```



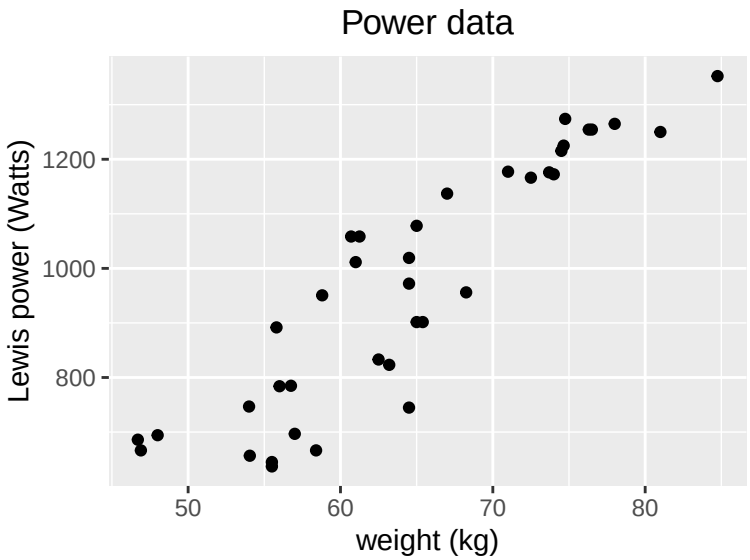
The scatter plot shows that:

1. Maximum temperature increases as minimum temperature increases i.e. there is a positive relationship.
2. The relationship appears linear.
3. The points lie in a fairly tight band indicating that the relationship is quite strong.
4. The variability appears reasonably constant. However, there is possibly less variability at the extremes.

Example 5 (Power).

A study of an individual's power (measured by a vertical jump and converted to power using the Lewis formula) and its relationship to their weight was undertaken by a sports scientist. A random sample of 38 users of the Stevenson Building facilities was selected and their power and weight measured.

```
#power <- read.table("power.dat")
names(power) <- c("Lewis", "weight", "sex")
qplot(weight,Lewis,data=power, xlab="weight (kg)",
      ylab="Lewis power (Watts)",main="Power data")
```



Is there a linear relationship between power and weight?

If so, what is the strength of the relationship?

A plot of Lewis power against the weight of the individual shows that:

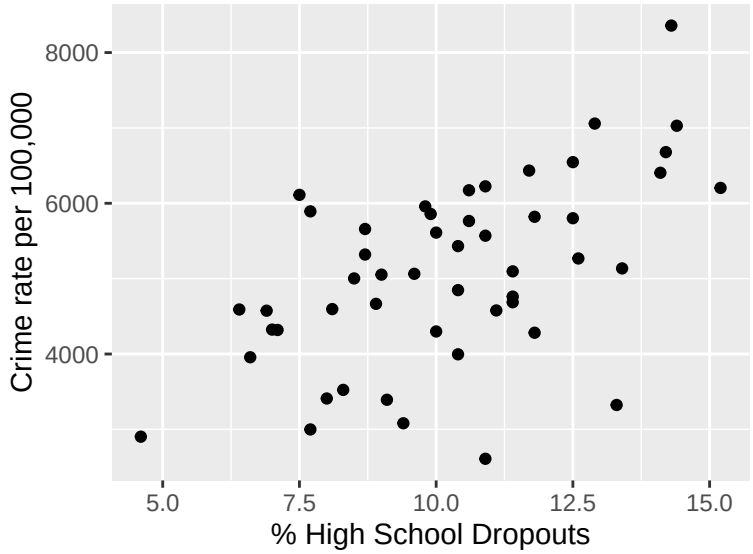
1. As the weight of an individual increases then their power also tends to increase. The plot highlights a positive, roughly linear relationship;

2. The relationship is not exact; the points do not lie in one straight line.
3. The variability changes; there appears to be more spread at lower values of weight.

Example 6 (Crime).

Fifty states in America were investigated in terms of their crime rates and percentage of high school dropouts. The crime rate per 100,000 people included: murder, rape, robbery, aggravated assault, burglary, larceny-theft and motor vehicle theft. The state high school dropout rate comprised the percentage of current 16-19 year olds who were not in school and had not finished the 12th grade. The data are recorded in the plot below.

```
#crime<-read.csv("crime.csv")
qplot(Dropout ,Crime,data=crime,xlab="% High School Dropouts",ylab="Crime rate per 100,000")
```



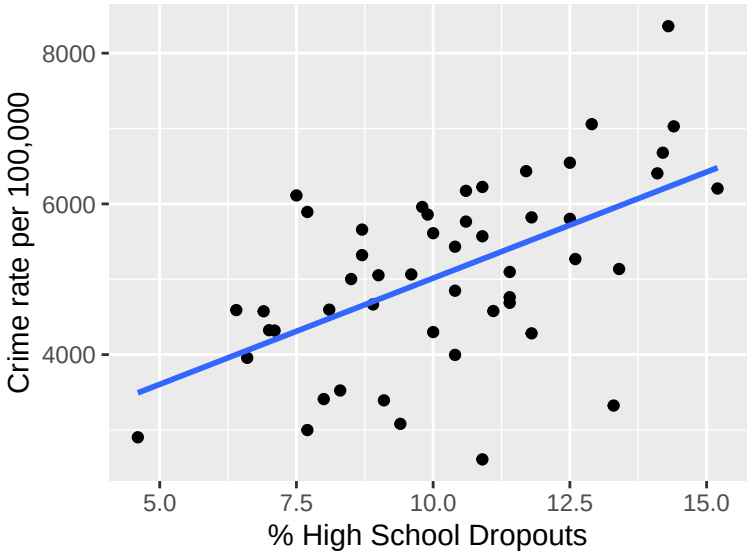
Now your task is to interpret the above graph:

Task 3 (Does crime rate depend on the percentage of dropouts?).

What can we say about the relationship between crime rate and the percentage of dropouts?

If we are interested in predicting the crime rate from the high school dropout rate one option is to fit a simple linear model to describe the relationship. The plot below gives such a fit.

```
crime.plot<-qplot(Dropout,Crime,data=crime,xlab="% High School Dropouts",ylab="Crime rate per 100,000")
crime.plot+stat_smooth(method="lm", se=FALSE)
```



Without discussing the details on how we fit the line, which we will cover later, the best fitted line on the plot (given in blue) is:

$$\text{Crime} = 2197 + 281.8 \times \text{Dropout}$$

Now let us see if you can comment on the relationship between crime and the percentage of dropouts, based on the above plot.

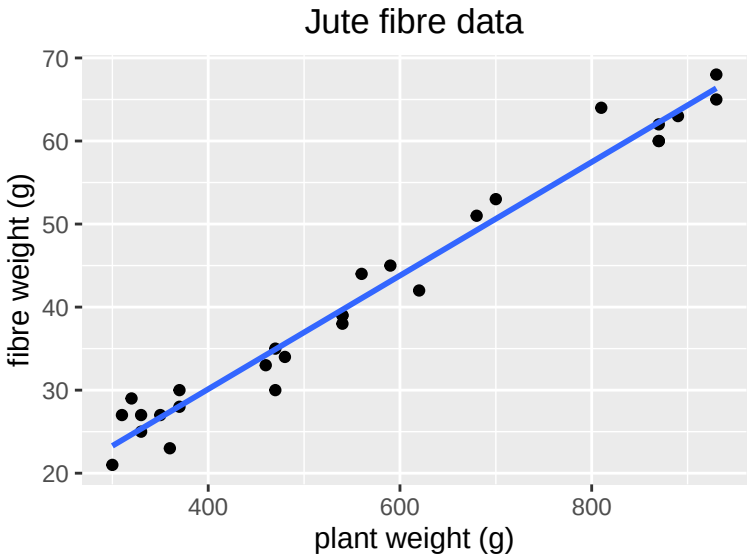
Task 4 (Interpret the above relationship).

Comment on the relationship between crime and percentage of dropouts based on the above plot.

Example 7 (Jute plants).

Twenty eight jute plants were weighed and the quantity of fiber extracted from each of them was also weighed. The goal was to estimate the total fiber production for a future jute plantation period.

```
#jute <- read.table("jute.dat")
names(jute) <- c("plant", "fibre")
jute.plot<-qplot(plant,fibre,data=jute, xlab="plant weight (g)",
                 ylab="fibre weight (g)", main="Jute fibre data")
jute.plot+stat_smooth(method="lm", se=FALSE)
```



A plot of fibre weight against plant weight reveals:

1. There is a strong positive linear relationship.
2. The pattern/scale of variability about the underlying relationship remains similar as plant weight increases.

It is of interest to predict the fiber weight from the plant weight.

The fitted line on the plot is :

$$\text{fibre weight} = 2.8465 + 0.06837 \times \text{plant weight}.$$

The fitted line describes the data reasonably well even though there is some variability about the line.

Task 5 (Interpret the above relationship).

What would be the expected increased yield in fiber for a 1 kg increase in weight?

Example 8 (X-ray on Bacteria).

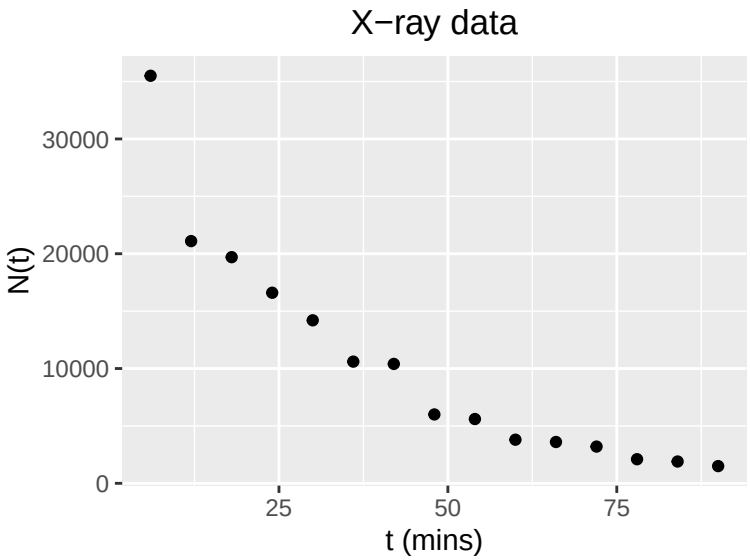
These data represent the number of surviving marine bacteria $N(t)$ following exposure to 200 kilovolt x-rays for periods of time t ranging from 6 to 90 minutes. In each case the initial number of bacteria was $N(0) = 40,000$. The goal is to understand of survival time of marine bacteria.

Biological theory suggests that

$$N(t) = N(0)\exp(\beta t)$$

for some underlying constant β.

```
#xray <- read.table("xray.dat")
names(xray) <- c("n.t", "time", "logNt")
qplot(time, n.t, data=xray, xlab="t (mins)", ylab="N(t)", main="X-ray data")
```



The first plot $N(t)$ vs t suggests that this is plausible, for some $\beta < 0$ (i.e. "exponential decay"). However it is not clear if the curve presented by the data is truly "exponential".

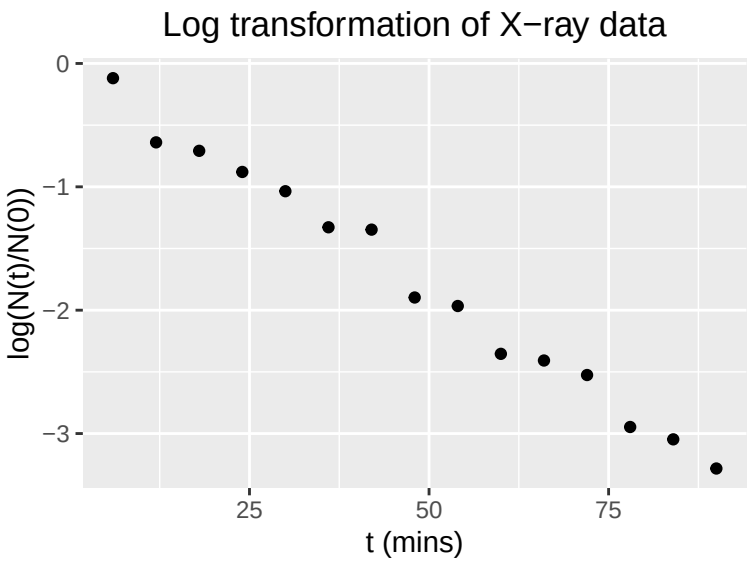
An alternative plot is useful.

$N(t)/N(0) = \exp(\beta t)$ so $\log(N(t)/N(0)) = \beta t$

Thus why not plot $y = \log(N(t)/N(0))$ vs t?

If the model is correct, then we should see a straight line (with a negative slope) that passes through the origin.

```
qplot(time, log(n.t/40000), data=xray, ylab="log(N(t)/N(0))",
      xlab="t (mins)", main="Log transformation of X-ray data")
```



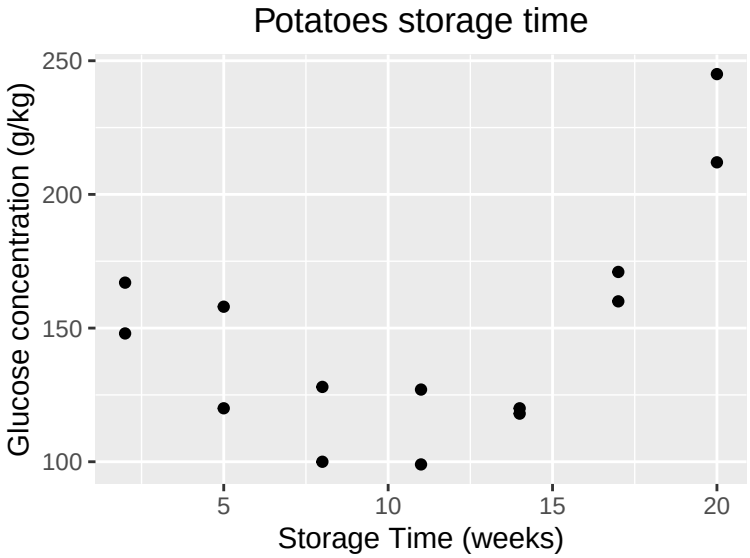
The second plot suggests that the relationship between y and t is linear, with the variability about the relationship not depending on t.

In this example, we have met the idea of a **transformation** which is something we will make further use of. As a result of the transformation, we are able to draw a plot such that the anticipated relationship is linear; which allows us to apply linear modelling techniques to this non-linear relationship.

Example 9 (Sugar in Potatoes).

The glucose level in potatoes is dependent on the length of time for which they have been stored. The scatterplot of glucose against storage time, below, shows a curvilinear relationship which is non-monotonic, and so cannot be transformed into a straight line. Possibly a quadratic curve would describe this relationship adequately.

```
#potatoes <- read.csv("potatoesstorageweeks.csv")
qplot(Weeks, Glucose, data=potatoes, xlab="Storage Time (weeks)",
      ylab="Glucose concentration (g/kg)", main="Potatoes storage time")
```



Summary comments on the 6 examples

Now we summarize the comments from the 6 examples:

- In examples 1 & 2, the air temperature and power examples, it is clear that the two variables are related in a linear fashion. The strength of the relationship can be measured even by using a **correlation coefficient**, as there is no clear indication of which variable is the response variable and which variable is a predictor.
- In each of the examples 3 to 6, one variable is a **response** (crime rate, fiber weight, number of bacteria surviving, glucose level) to the other variable which plays the role of a **predictor** (percent of dropouts, plant weight, time of irradiation, storage time).
- Alternative terms are frequently used and include:

  Response (Y): dependent variable, output

  Predictor (x): independent variable, explanatory variable, input, covariate
- In each of the examples 3 to 6, there is a fairly clear underlying relationship relating the response to the predictor. This may not be an exact mathematical formula, but clearly knowledge about the value of the predictor tells us a lot about the likely value of the response. **Regression analysis** is used to model the relationship between a response and one or more predictor variables.

The idea of regression is the focus of this course and will be developed further. We will now consider the definition of a statistical model.

Defining a statistical model

The statistical model is usually written as

$$\begin{aligned} Y &= \text{deterministic part + random error, or} \\ Y &= f(x) + \epsilon. \end{aligned}$$

The deterministic component $f(x)$ describes the relationship between the response and the predictor variables, this might take the form of a straight line or some other function. The random error term will also be described in some detail, and usually a number of assumptions will be made about its distribution.

We can consider some statistical models for the examples we discussed earlier in this material. When considering two variables we often use scatterplots to suggest the function or deterministic part of the model.

Example 10 (Crime Data (Continued)).

Let us now re-consider example 6.

Data: $(y_i, x_i), \quad i = 1, \ldots, n; \quad n = 50.$

y_i = crime rate for state i (vertical axis)

x_i = percentage of high school drop outs for state i (horizontal axis)

Possible model: $Y_i = \alpha + \beta x_i + \epsilon_i \qquad \text{for } i = 1, \ldots, n$

α and β are the intercept and slope of the line, respectively.

ϵ_i is an additive, unpredictable quantity.

Y_i is the response, regarded as a random variable.

$\alpha+\beta x_i$ is the deterministic part of the model, ϵ_i is the random part and βx_i is the part where the explanatory variable is incorporated.

α and β are called **model parameters**.

Assumptions

While describing the model for a linear regression we also make the following further assumptions:

We assume that

$$\mathrm{E}(\epsilon_i) = 0 \quad \text{and} \quad \mathrm{Var}(\epsilon_i) = \sigma^2,$$

for all i, where σ^2 does not depend on any other unknown or on x_i.

We might also assume that $\epsilon_i \sim N(0, \sigma^2)$ and usually that ϵ_i and ϵ_j are uncorrelated for $i \neq j$. We will discuss these assumptions in detail in Week 3.

The full probability model for the response Y_i given x_i can be written as

$$Y_i | x_i \sim N(\alpha + \beta x_i, \sigma^2)$$

The variable ϵ_i is often called the error, however it is not a mistake, just random variation.

Example 11 (X-rays (Continued)).

If we consider example 8

Data: $(y_i, x_i) \quad i = 1, \ldots, n; \quad n = 15.$

$y_i = \log(N(t)/N(0))$

x_i = time of irradiation

Possible model: $Y_i = \beta x_i + \epsilon_i$ for some β.

This is the equation of a straight line through the origin with slope β and intercept = $(0, 0)$. We can also plausibly assume $\mathrm{Var}(\epsilon_i) = \sigma^2, \epsilon_i \sim N(0, \sigma^2)$ and that the $\{\epsilon_i\}$ are uncorrelated.

Example 12 (Sugar in Potatoes).

Now, let us revisit example 9

Data: $(y_i, x_i), \quad i = 1, \ldots, n; \quad n = 14$

y_i = glucose concentration

x_i = storage time

Possible model: The scatterplot showed a curved relationship, which a quadratic function might describe. In mathematical terms this would be written as

$$Y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i, \quad i = 1, \ldots, n$$

What is a linear model?

We discussed the structure of a model and identified two components, the deterministic and the stochastic (or random) parts. We saw that the deterministic part involves the predictor or explanatory variable.

In each of the examples we looked at, the model can be described as a regression model, with Y regressed on x. We identified the parameters, using Greek symbols (α, β, γ etc). These are unknown quantities which must be estimated, consequently allowing us to estimate the deterministic part of the relationship, also called the **regression function.**

In every model we have discussed, the parameters (except for σ^2) must appear linearly in the deterministic component, so they are called **linear models**. (Note this does not necessarily imply a straight line).

Therefore, a **linear model** is one in which the parameters appear linearly in the deterministic part of the model. For example, a quadratic regression model is a linear model.

Definition 1.

The model is a linear model if $f(x)$ is a linear function.

Some examples of linear models:

$$Y = \beta_0 + \beta_1 x + \epsilon$$
$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$
$$Y = \beta_0 + \beta x_1 + \gamma x^2 + \delta \exp(5x) + \epsilon$$

The following are **not** linear models since one of the parameters, namely μ, enters the model non-linearly.

$$Y = \beta_0 + \beta_1 x + \beta_2 \exp(-\mu x) + \epsilon$$
$$Y = \beta_0 + \beta_1 x + \beta_2 \cos(\mu x) + \epsilon$$

Definition 2 (Simple linear regression).

A linear regression model with one explanatory variable is referred to as a simple linear regression model.

Definition 3 (Multiple linear regression).

A linear regression model with more than one explanatory variable is referred to as a multiple regression model.

Notation

Regression models can be expressed in several ways. For example, the quadratic regression model used in the sugar in potatoes example with respone variable Y and predictor variables x can be expressed as

$$Y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i, \qquad i = 1, \ldots, n.$$

We assume that

$$\mathrm{E}(\epsilon_i) = 0 \quad \text{and} \quad \mathrm{Var}(\epsilon_i) = \sigma^2,$$

therefore

$$\begin{aligned} E(Y_i|x_i) &= E(\alpha + \beta x_i + \gamma x_i^2 + \epsilon_i) \\ &= E(\alpha + \beta x_i + \gamma x_i^2) + E(\epsilon_i) \\ &= E(\alpha + \beta x_i + \gamma x_i^2) \\ &= \alpha + \beta x_i + \gamma x_i^2. \end{aligned}$$

To ease notation, we may expression our model as

$$E(Y_i) = \alpha + \beta x_i + \gamma x_i^2.$$

Learning Outcomes for Week 1:

- Understand the scope of predictive modeling.
- Understand types of predictive models with respect to data types for response and predictors.
- Understand the steps of predictive models.
- Define a statistical model.
- Define a linear regression model.

Answers to tasks

Answer to Task 1 (Car Insurance industry). R output

```
##    veh_value  exposure clm numclaims claimcst0 veh_body veh_age gender area agecat           X_OBST
## 15      1.66 0.4845996   1         1  669.5100    SEDAN       3      M    B      6 01101    0    0
## 17      1.51 0.9938398   1         1  806.6100    SEDAN       3      F    F      4 01101    0    0
## 18      0.76 0.5393566   1         1  401.8055    HBACK       3      M    C      4 01101    0    0
## 41      1.89 0.6543463   1         2 1811.7100    STNWG       3      M    F      2 01101    0    0
## 65      4.06 0.8514716   1         1 5434.4400    STNWG       2      M    F      3 01101    0    0
## 66      1.39 0.3175907   1         1  865.7900    HBACK       3      F    A      4 01101    0    0
```

```
##    veh_value         exposure            clm            numclaims         claimcst0          veh_bo
##  Min.   : 0.000   Min.   :0.002738   Min.   :0.00000   Min.   :0.00000   Min.   :    0.0   SEDAN  :2
##  1st Qu.: 1.010   1st Qu.:0.219028   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:    0.0   HBACK  :1
##  Median : 1.500   Median :0.446270   Median :0.00000   Median :0.00000   Median :    0.0   STNWG  :1
##  Mean   : 1.777   Mean   :0.468651   Mean   :0.06814   Mean   :0.07276   Mean   :  137.3   UTE    :
##  3rd Qu.: 2.150   3rd Qu.:0.709103   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:    0.0   TRUCK  :
##  Max.   :34.560   Max.   :0.999316   Max.   :1.00000   Max.   :4.00000   Max.   :55922.1   HDTOP  :
##                                                                                           (Other):
##     veh_age      gender    area          agecat                    X_OBSTAT_
##  Min.   :1.000   F:38603   A:16312   Min.   :1.000   01101    0    0    0:67856
##  1st Qu.:2.000   M:29253   B:13341   1st Qu.:2.000
##  Median :3.000             C:20540   Median :3.000
##  Mean   :2.674             D: 8173   Mean   :3.485
##  3rd Qu.:4.000             E: 5912   3rd Qu.:5.000
##  Max.   :4.000             F: 3578   Max.   :6.000
##
```

The three possible response variables are:

clm - whether a person claimed or not, for modeling 'no-claims bonus.'

numclaims - How many claims has one made, for modeling 'claim frequency.'

claims0 - claim amount (0 if no claim), for modeling claim severity.

Answer to Task 2 (Type of response). Possible data types are:

clm is a binary random variable.

numclaims is a count variable only taking non-negative integer values.

claims0 is non-negative continuous random variable, which often has a skewed distribution.

Answer to Task 3 (Does crime rate depend on the percentage of dropouts?).

We can observe the following:

- There appears to be a positive linear relationship between the percentage of dropouts and crime rate.
- Crime rate increases as the percentage of dropouts increases.
- There is quite a lot of variability and hence this relationship appears moderate to weak. (The variability appears fairly constant over the range of percentage of dropouts).

Answer to Task 4 (Interpret the above relationship).

Based on the plot and the fitted line we can observe:

- The fitted line appears to fit the data reasonably well.
- There is quite a lot of spread/variability about the line.
- For every 1% increase in percent of dropouts the average (or expected) crime rate increases by 281.8.

Answer to Task 5 (Interpret the above relationship). 0.06837 Kg or 68.37 grams