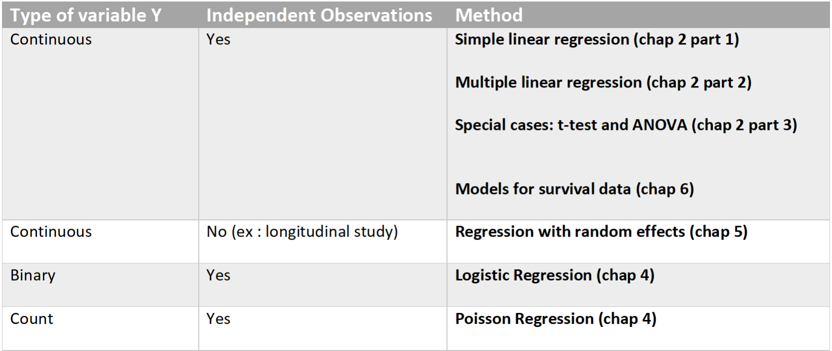
Chapter 2: Linear Regression
Part 1: Simple Linear Regression

MATH 60604: Statistical Modelling

HEC Montréal
Department of decision sciences

Overview of course material



Type of variable Y	Independent Observations	Method
Continuous	Yes	**Simple linear regression (chap 2 part 1)** **Multiple linear regression (chap 2 part 2)** **Special cases: t-test and ANOVA (chap 2 part 3)** **Models for survival data (chap 6)**
Continuous	No (ex : longitudinal study)	**Regression with random effects (chap 5)**
Binary	Yes	**Logistic Regression (chap 4)**
Count	Yes	**Poisson Regression (chap 4)**

Table of contents

1 Introduction

2 Correlation

3 Regression

4 Estimation

5 Prediction

6 Residual analysis

7 R^2

8 Binary predictor

9 Exam errors

Linear regression

- The goal of linear regression is, among others, to investigate the effect of one or more explanatory variables on the mean of the variable of interest.
- For example, suppose we're interested in assessing the effects of the following variables on a person's income:
  - age (continuous)
  - years of experience (categorized into different intervals - ordinal variable)
  - type of job (categorical)
  - education level (ordinal)
- ... and all of these variables could interact together in affecting the income level!

Linear regression terminology

- Dependent/response/outcome variable (Y): main variable of interest
- Independent/explanatory/predictor variables (X): the variables that are potentially associated with Y.
- Simple linear regression: regression model with only one explanatory variable
- Multiple linear regression: regression model with several explanatory variables (*part 2 of this chapter*)

Linear regression models

Objectives of linear regression

1. **Investigate** how the independent variables X explain / affect the dependent variable Y
   - quantify the effect of the explanatory variables X on the outcome variable Y
   - test hypotheses regarding the variable effects
2. **Prediction**: develop a model to predict future values of Y using the variables X
3. Both of the above simultaneously.

Regression models

Linear regression models and their counterparts are probably the most used of all statistical models.

Here are a few examples of different types of regression models:

- Linear regression: Y is continuous
  - Particular cases: t-test, ANOVA
- Logistic regression : Y is binary.
- Multinomial logistic regression: Y is nominal.
- Cumulative logistic model: Y is ordinal.
- Poisson regression: Y is a count variable

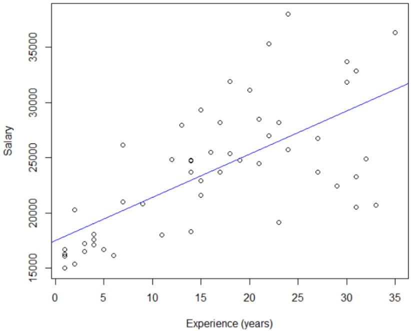
All of these models are particular examples of generalized linear models which we will see in chapter 3.

Other types of regression models

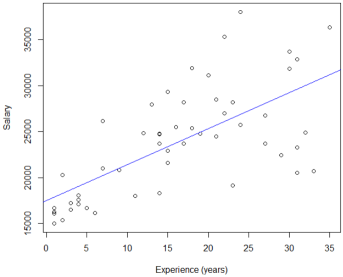
- If the observations are not independent, for example in the case of longitudinal data, linear mixed models could be considered. We will also see this later on.
- Finally, generalized linear mixed models encompass all of the above mentioned models. They allow us to model a response variable Y of any kind, with or without dependence between the observations.
- A solid understanding of linear regression is crucial as it serves as a basis for more complex models.

Basic principles of simple linear regression

- Simple linear regression allows to evaluate the effect of a single variable X on the mean of a variable of interest Y
- However in linear regression, we're not interested in just any kind of effect of X on the variable Y (e.g. the effect of the number of years of experience on salary)
- We're only interested in a linear relationship (that may or may not exist) between the two variables.



Basic principles of simple linear regression



- Using the data (X, Y) from a sample, simple linear regression allows us:
  - to find "the best possible fit" to model the linear relationship between X and Y
  - to interpret the estimated linear trend (eg: for each year of experience, salary increases, on average, by ...)
  - to see how well the model fits the data

Data analysis

Before starting any statistical analysis, it's important to understand the context:

- understand the research question and any relevant background information that may help guide the analysis
- understand the population (important for inference)
- understand the sample
  - how was the data collected?
  - what will the data allow you to conclude?
  - observational vs. experimental data
    - experimental: researcher has control over the experiment; any differences in responses from different groups can be attributed to the group assignment only; allows for causal interpretation
    - observational: researcher merely *observes* the data; never know if differences in responses from different groups are due to the group assignment or due to other "confounding" variables; causal interpretation difficult

Exploratory data analysis

An exploratory data analysis consists of an initial “analysis” of your data, with the goal of *understanding* your data

- understand the types of variables in the data:
  - quantitative vs. qualitative variables
  - quantitative: continuous vs. discrete
  - qualitative (categorical): nominal vs. ordinal
  - recall: the “type” of variables will help decide what kind of model(s) will be appropriate for the data
- explore the data:
  - get an overview of the variables
  - univariate summaries: each variable (individually)
  - bivariate summaries: relation between variables

Exploratory data analysis

- exploring the data: univariate summaries
  - qualitative data: frequency tables, barplots, etc.
  - quantitative data:
    - numerical summaries: central tendency (mean, median, mode), spread (range, variance/standard deviation)
    - graphical summaries: histogram (visualizing density), boxplot (to visualize center and spread of data)
- exploring the data: bivariate summaries:
  - two quantitative variables: scatterplots
  - a quantitative and qualitative variable: boxplots for each level of the qualitative variable
  - two qualitative: color-coded barplots
- be creative! There are endless ways to create graphical summaries of your data, remember, a picture is worth 1000 words!

Exploratory data analysis

It's always important to check for any "unusual" values in your data:

- missing data:
  - special care has to be taken when handling missing data
  - sometimes missingness is in fact meaningful
    - ex: in a questionnaire, sometimes the flow of questions can lead to missing values

      1) have you received your COVID vaccine?

      2) if yes, which vaccine did you receive (Pfizer / Moderna / etc)?
  - ignoring missing data can lead to biases in your analysis
  - there are techniques for handling missing data, e.g. imputation

Exploratory data analysis

- errors:
  - if you know that there is an error in your data and you know how to fix it, do so!
  - be careful - be sure that it is an error!
- outliers
  - are these values influential?
- be transparent: in practice, always report exploratory data analysis and any interesting findings
- reproducibility is really important!

Exploratory data analysis

Example: fixation-intention to buy

Table of contents

Linear correlation

- As linear regression only concerns the linear relationship between two variables, we will begin by exploring the concept of linear correlation between two variables.
- The concept of linear correlation is related to linear regression, as we will see.

Pearson's linear correlation coefficient

- The correlation coefficient allows to quantify the linear relationship between two variables X and Y .
- Suppose that we're studying n pairs of observations

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

where (X_i, Y_i) are observations of the pair (X, Y) for individual i.

- Pearson's correlation coefficient, denoted by r, measures the strength and direction of the linear relationship between two quantitative variables.
  - That is, the extent to which observations fit around a straight line.

Pearson's linear correlation coefficient

- The correlation coefficient is given by:

$$\boxed{r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of the two variables, respectively given by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

- Note: r is the *sample correlation*, which is an estimator of the theoretical correlation at the population level, denoted by ρ

Pearson's linear correlation coefficient

- The correlation at the population level ρ is given by

$$\rho = \frac{E\left\{(X - \mu_X)(Y - \mu_Y)\right\}}{\sqrt{\sigma_x^2 \sigma_Y^2}}$$

where μ_X and σ_X^2 are the mean and variance in the population for the variable X, and μ_Y and σ_Y^2 are that for Y.

Note: $E(\cdot)$ is the expectation - it is the operator that gives the mean of a random variable (population level). So, $E\left\{(X - \mu_X)(Y - \mu_Y)\right\}$ is the mean of $(X - \mu_X)(Y - \mu_Y)$ (population level).

- An estimator of ρ (based on a sample from the population) is r:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i = \bar{Y})^2}}$$

Pearson's linear correlation coefficient

Population vs. Sample:

Population	Sample
$\mathrm{cov}(X,Y) = \mathrm{E}\left\{(X-\mu_x)(Y-\mu_Y)\right\}$	$S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$
$\mathrm{var}(X) = \sigma_1^2 = \mathrm{E}\left\{(X-\mu_X)^2\right\}$	$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$
$\mathrm{var}(Y) = \sigma_2^2 = \mathrm{E}\left\{(Y-\mu_Y)^2\right\}$	$S_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$
$\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}$	$r = \frac{S_{12}}{S_1 S_2}$

Pearson's linear correlation coefficient

Properties of Pearson's linear correlation coefficient

- $-1 \leq r \leq 1$
- $r = 1$ if and only if the n observations fall exactly on a positively sloped line. In other words, if there exist two constants a and b ($b > 0$) such that
$$y_i = a + bx_i$$
for all i.
- $r = -1$ if and only if the n observations fall exactly on a negatively sloped line. In other words, if there exist a and b ($b < 0$) such that
$$y_i = a + bx_i$$
for all i.

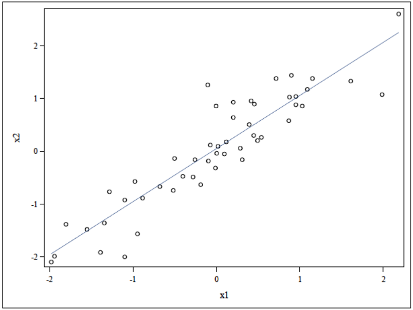
Pearson's linear correlation coefficient

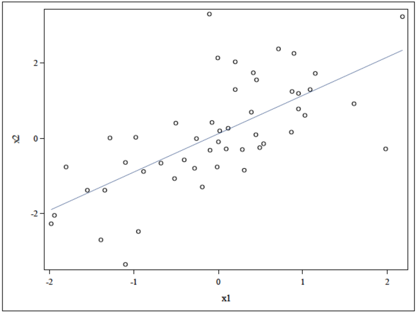
- The correlation coefficient measures the strength (strong/moderate/weak) and direction (positive/negative) of the linear relation between two variables.
- The closer the correlation is to 1, the closer the points are to a positively sloped line. Consequently, the more the value of X increases, the more the value of Y tends to increase (and vice-versa).
- Similarly, the closer the correlation is to -1, the closer the points are to a negatively sloped line. Consequently, the more the value of X increases, the more the value of Y tends to decrease, (and vice-versa).
- When the correlation is close to 0, the points will not tend to fall close to a line.
    - It's extremely important to note that this does not imply that there is no relationship between the two variables!!! It only means that there is no linear relationship between the two variables.

Pearson's linear correlation coefficient

- **Caution:** It can be misleading to interpret the correlation coefficient without examining the scatterplot, as the following examples will illustrate.
- The following slides show scatterplots as well as the correlation coefficient in several situations where you'll have to use your own judgement.
- The last two examples show the dangers of interpreting the correlation coefficient without having examined the scatterplot.

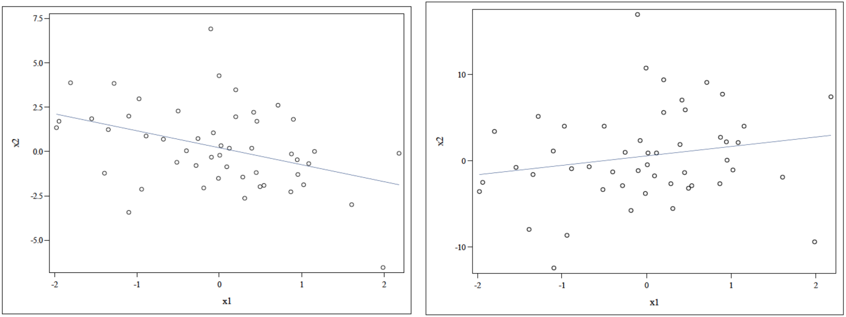
Pearson's linear correlation coefficient





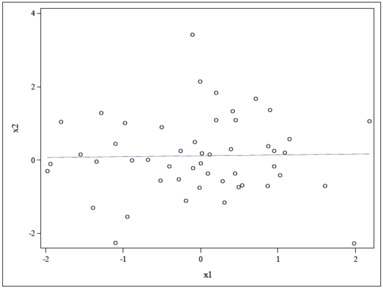
- Left: $r = 0.92$, very strong positive correlation.
- Right: $r = 0.68$, strong positive correlation.

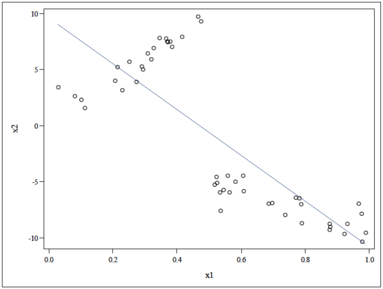
Pearson's linear correlation coefficient



- Left: $r = -0.4$, negative correlation.
- Right: $r = 0.2$, weak negative correlation.

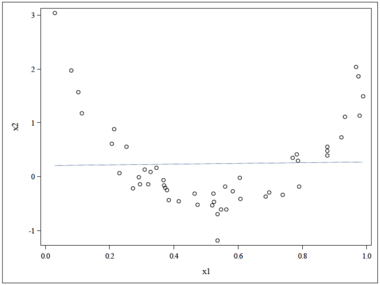
Pearson's linear correlation coefficient

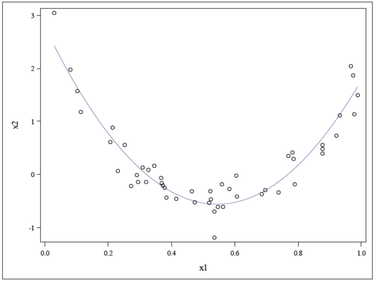




- left: $r = 0.02$, roughly no correlation, no obvious relation between the variables.
- right: $r = -0.81$, strong negative correlation, but this is misleading. Rather, it seems that there are two distinct patterns: X_2 increases with X_1 when $X_1 < 0.5$, but X_2 decreases with X_1 for $X_1 > 0.5$.

Pearson's linear correlation coefficient





- Correlation close to zero ($r = 0.02$), but the graph shows a clear strong relation between the variables. But because the relation is quadratic (and not linear), the correlation coefficient does not detect it.

Pearson's linear correlation coefficient

- We can carry out hypothesis tests to examine whether there is a significant correlation between two variables.

- The underlying hypothesis are

$$H_0 : \rho = 0 \quad \text{versus} \quad H_1 : \rho \neq 0$$

(where, recall that ρ is the true correlation at the population level, whereas r is the estimated value of ρ based on the sample).

  - Test statistic: $T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
  - Null distribution: $T \overset{H_0}{\sim} t_{n-2}$
  - P-value: $2 \times P(t_{n-2} > |T_{obs}|)$

Pearson's linear correlation coefficient

Example: fixation-intention to buy

Table of contents

1. Introduction
2. Correlation
3. **Regression**
4. Estimation
5. Prediction
6. Residual analysis
7. R^2
8. Binary predictor
9. Exam errors

Simple linear regression

- Simple linear regression is a special case of linear regression where there is only one predictor variable.

- Suppose that we're studying n pairs of observations

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

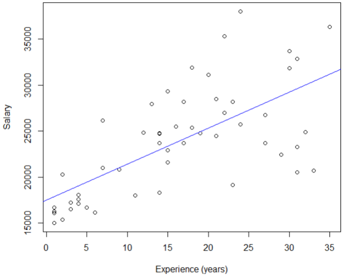
where (X_i, Y_i) denotes the observations of the variables X and Y for individual i

- Suppose further that the relationship between X and Y can be approximated by a line such that:

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \epsilon_i}$$

Simple linear regression

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \epsilon_i}$$



- β_0 is the intercept of the model.

- β_1 is the slope of the model (measures the effect of the variable X on Y)

- ϵ_i is the error term for individual i. This term accounts for the fact that there is not an exact relationship between X and Y.

Simple linear regression: fixed and random parts of the model

$$\boxed{\underbrace{Y_i}_{\text{random}} = \underbrace{\beta_0 + \beta_1 X_i}_{\text{fixed}} + \underbrace{\epsilon_i}_{\text{random}}}$$

- We're interested in modelling the random behaviour of Y, and thus we must assume that the variable Y is itself a random variable.
- The parameters β_0 and β_1 define the line representing the true relationship between X and Y in the population. These parameters are fixed, but we seek to estimate them using our sampled data.
- The error term ϵ is assumed to be random, as it represents the random deviation of the observed Y and the regression line $\beta_0 + \beta_1 X$
- **But what about X? Fixed or random?**

Simple linear regression: fixed and random parts of the model

$$\underbrace{Y_i}_{\text{random}} = \underbrace{\beta_0 + \beta_1 X_i}_{\text{fixed}} + \underbrace{\epsilon_i}_{\text{random}}$$

- The variable X is assumed to be fixed in the model:
  - It's possible that the X were truly fixed by the experimenter, and the Y values were observed afterwards.
- The variable X is assumed random in the model:
  - It's also possible that X was not actually fixed by the experimenter but was in fact observed (possibly at the same time as Y). We refer to this as observational data, and the variable X is also random in this case.
- However in regression models, we wish to describe the behaviour of Y for a given value of X and thus we always treat X as fixed.
  - we describe the behaviour Y conditional on given values for X

Simple linear regression

Recap

$$\boxed{Y_i = \beta_o + \beta_1 X_i + \epsilon_i}$$

- The regression model describes the behaviour of Y for a given value of X, i.e. conditional on the value of X.
- The part $\beta_0 + \beta_1 X_i$ is deterministic: it defines the line.
- The term ϵ_i is a random variable that allows the observations (X_i, Y_i) to not follow the line exactly.
- Consequently, the variable Y is a random variable.

Simple linear regression: model assumptions

The underlying assumptions in the model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ can be written in 2 equivalent ways:

Assumptions for the regression model

- **Assumptions on** ϵ: in the basic model, we assume that $\epsilon_1, \ldots, \epsilon_n$ are independent Normally distributed random variables with mean $E(\epsilon_i) = 0$ and variance $Var(\epsilon_i) = \sigma^2$:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- **Assumptions on** Y: in the basic model, we assume that $Y_1, \ldots, Y_n$ are independent Normally distributed random variables such that:

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i \; ; \; Var(Y_i|X_i) = \sigma^2$$

- We will revisit these assumptions later on.

Simple linear regression: model assumptions

- The linear regression model assumes that $\epsilon_1, \ldots, \epsilon_n$ are independent random variables with
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$
$\rightarrow$ for a given value of X_i, the mean of $Y_i = \beta_0 + \beta_1 X_i + \epsilon$ is simply $\beta_0 + \beta_1 X_i$ since
$$\begin{aligned} E(Y_i|X_i) &= E(\beta_0 + \beta_1 X_i + \epsilon|X_i) \\ &= \beta_0 + \beta_1 X_i + E(\epsilon) \\ &= \beta_0 + \beta_1 X_i + 0 \end{aligned}$$
and the variance of Y_i is σ^2 since
$$Var(Y_i|X_i) = Var(\epsilon_i) = \sigma^2$$

- Thus, the simple linear regression model assumes that the conditional mean of Y is a linear function of X and that the conditional variance of Y is constant (does not depend on X)

- The model thus has three parameters: $\beta_0, \beta_1, \sigma^2$ that we want to estimate

Parameter interpretations: slope

- β_1 is the slope of the model and represents the effect of the explanatory variable X on the dependent variable Y. It is interpreted as follows:

  For each 1 unit increase in X, Y increases on average by β_1 .

- Why?

$$\begin{aligned}\beta_1 &= E(Y|X=x+1) - E(Y|X=x)\\ &= \{\beta_0 + \beta_1(x+1)\} - \{\beta_0 + \beta_1 x\}\end{aligned}$$

- Ex: for the fitted regression model:

$$\widehat{\texttt{Salary}} = 10 + 1.5\ \texttt{Experience}$$

  - Y = salary (in millions of dollars), X = years of experience
  - $\hat{\beta}_1 = 1.5$ (estimated value of β_1 from the sample data)
  - For each additional year of experience, salary increases on average by \$1.5K.

Parameter interpretation: intercept

- It's rare that we're specifically interested in the intercept β_0. We're usually more interested in measuring the effect of the explanatory variables on the response variable.

- In general, the intercept is interpreted as follows:

  When $X = 0$, Y has an average of β_0.

- Why? $\beta_0 = E(Y|X = 0) = \beta_1 + \beta_1 \times 0$

- In some cases, this interpretation does not make sense, since:

  i) the value $X = 0$ may not be possible

  ii) there are no observed values in the range of $X = 0$, even if this value is possible. In this case, interpreting this value would be considered extrapolation.

- In our example $\widehat{\texttt{Salary}} = 10 + 1.5\ \texttt{Experience}$
  - $\hat{\beta}_0 = 10$ (estimated value β_0 from the sample data)
  - People having no experience have a mean salary of $10K.

Table of contents

1. Introduction
2. Correlation
3. Regression
4. Estimation
5. Prediction
6. Residual analysis
7. R^2
8. Binary predictor
9. Exam errors

Parameter estimation

- The simple linear regression model has three parameters: β_0, β_1, and σ^2
  - We will see how they are estimated (and how the line is estimated)
  - We will also see how to formally test hypotheses on these parameters

Parameter estimation

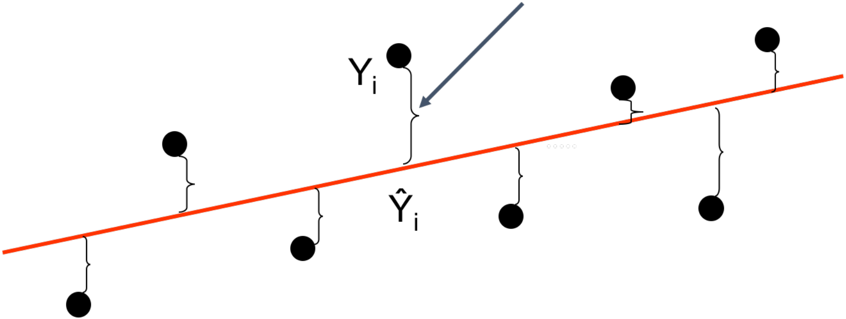
- One way of estimating the regression line is by choosing the one that "fits" the data the best.
- There are several ways to define the notion of "best fit". The most common one is called the least squares criterion.
- Note that $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators of β_0 and β_1
- Also note that $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ is the fitted or predicted value of the i^{th} observation according to the model.
- The least squares criterion finds values for the estimators that minimizes the sum of squares:

$$\boxed{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.}$$

- The resulting regression line is the line that corresponds to the smallest possible error among all possible lines.

Least squares criterion

Vertical distance between a data point and a line (**residual**) = $(Y_i - \hat{Y}_i)$



Why do you think we consider the squared errors and not simply the errors $Y_i - \hat{Y}_i$?

Simple linear regression: parameter estimation

- In the case of simple linear regression, there exist explicit formulas for these estimators:

$$\boxed{\hat\beta_1 = \frac{\sum_{i=1}^n (X_i - \bar X)(Y_i - \bar Y)}{\sum_{i=1}^n (X_i - \bar X)^2} = \frac{S_{XY}}{S_{XX}}}$$

$$\boxed{\hat\beta_0 = \bar Y - \hat\beta_1 \bar X}$$

- **If the model has been correctly specified**, these estimators are **unbiased** for their respective parameters, that is

$$\mathrm{E}(\hat\beta_0) = \beta_0, \quad \mathrm{E}(\hat\beta_1) = \beta_1$$

- It can be shown that the estimators have variance

$$\mathrm{var}(\hat\beta_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar X^2}{S_{XX}}\right), \quad \mathrm{var}(\hat\beta_1) = \frac{\sigma^2}{S_{XX}}$$

where $S_{XX} = \sum_{i=1}^n (X_i - \bar X)^2$

Simple linear regression: parameter estimation

- An unbiased estimator for σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- See supplementary information for chapter 2 part 1 for more details.

Simple linear regression: parameter estimation

Example: fixation-intention to buy

Hypothesis testing

- In an experimental research setting, it's often of interest to test whether the effect of the explanatory variable is significant.
- This can be done by testing whether the parameter corresponding to the variable effect is significantly different from 0, i.e.

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0$$

  If $\beta_1 = 0$ then for every 1 unit increase in X, Y does not change, on average - i.e. X does not have a (linear) effect on Y.
- The test is based on the following statistic:

$$\boxed{t = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)}}$$

  where $\hat{se}(\hat{\beta}_1)$ is an estimate of the standard error of $\hat{\beta}_1$, i.e. $\hat{se}(\hat{\beta}_1) = \hat{\sigma}/\sqrt{S_{XX}}$
- It can be shown that under H_0, t follows a Student t distribution with $n-2$ degrees of freedom.

Confidence intervals

- We can also obtain confidence intervals (CI) for the coefficients.
- For β_j, $j = 0, 1$, the CI is given by:

$$\boxed{\hat{\beta}_j \pm t_{n-2,\alpha/2}\hat{se}(\hat{\beta}_j)}$$

- Note that, as we saw in the previous chapter, we can use CI to test whether the regression parameters are significantly different from 0:
  - If 0 is NOT contained in the confidence interval, then we can conclude that β_j is significantly different from 0 (at the α significance level).

Hypothesis testing & Confidence intervals

Chapter 2 part 1 supplementary information

Hypothesis testing & Confidence intervals

Example: fixation-intention to buy

Simple linear regression: interpretation of the test

Note: there is a link between the correlation r and the estimator $\hat{\beta}_1$ in <u>simple</u> linear regression:

- Recall:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

and

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

so we can rewrite

$$\hat{\beta}_1 = r \times \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

→ if $r = 0$ then $\hat{\beta}_1 = 0$!

Simple linear regression: interpretation of the test

- It turns out that when there's one explanatory variable (simple linear regression), testing the effect of the explanatory variable on the dependent variable is equivalent to testing whether the correlation between the two variables differs from 0. Formally, in terms of hypotheses:

  $$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

  is equivalent to

  $$H_0 : \rho = 0 \quad \text{vs.} \quad H_1 : \rho \neq 0$$

  This is only true in the case of simple linear regression.

- We've already seen that the correlation between fixation and intention is significant (p-value 1.21×10^{-6}).

- Careful: this is no longer true when there is more than one explanatory variable in the model!

Simple linear regression: interpretation of the test

Careful:

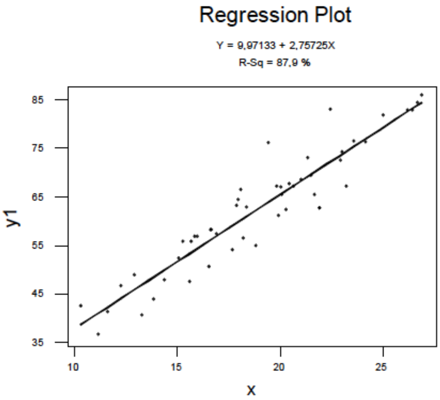
- In this case, fixation and intention appear to be linearly related. So the model is reasonable, and testing the effect of fixation on intention is justified. As we'll see with some examples, just like in the case of linear correlation, things are not always so simple with simple linear regression!

Simple linear regression: interpretation of the test

- **Recall**: $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$
- **If we reject** H_0, we can conclude that X is useful in explaining the behaviour of Y. There are two possibilities :
  - The relationship between X and Y is truly linear. See example (A). This is the case in our example.
  - We could do better with another model. See example (B).
- **If we do not reject** H_0, we can only conclude that X does not appear to be useful in explaining Y in a linear way. There are two possibilities:
  - There is no relationship between X and Y at all. See example (C).
  - The relationship between X and Y is not linear, but X could be useful in explaining the behaviour in Y through another model. See example (D).

Simple linear regression: interpretation of the test

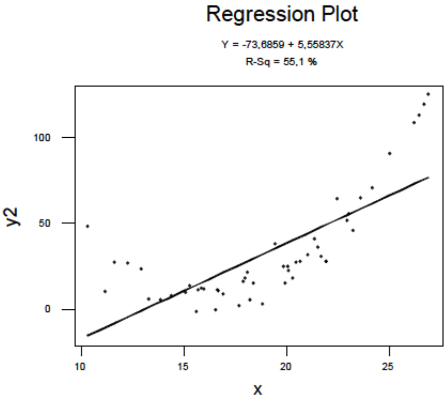
Example A:



- The slope is significantly different from 0 and the relationship between X and Y appears to be linear. The simple linear regression model is appropriate.

Simple linear regression: interpretation of the test

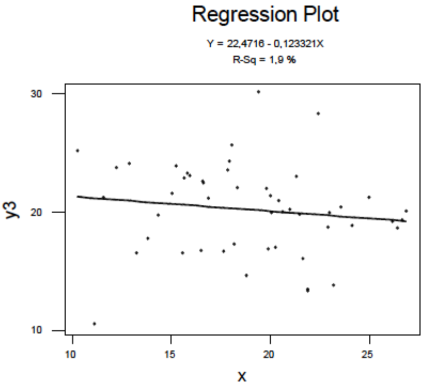
Example B:



- The slope is significantly different from 0. However, the relationship between X and Y is not linear. Another model would be more appropriate.

Simple linear regression: interpretation of the test

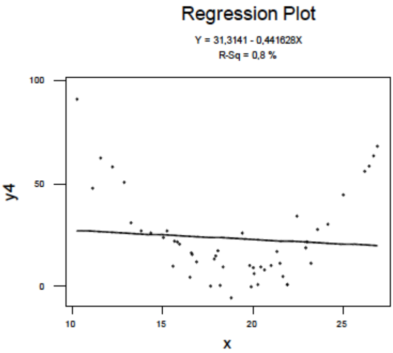
Example C:



- The slope is not significantly different from 0 and there does not appear to be any relationship between X and Y.

Simple linear regression: interpretation of the test

Example D:



- The slope is not significantly different from 0 but there is clearly a (quadratic) relationship between X and Y .

Table of contents

1. Introduction
2. Correlation
3. Regression
4. Estimation
5. **Prediction**
6. Residual analysis
7. R^2
8. Binary predictor
9. Exam errors

Prediction

- In many applications, the primary goal is to develop a model to make predictions of the dependent variable and ultimately use these predictions to make business decisions.
- For example, we might want to predict the amount of money spent if we were to send a promotional offer to a client.
- The usual way of proceeding would be to send offers to a sample of clients, build a model with the sample data, and then apply the model (i.e. obtain predictions) to the other clients in the dataset.

Prediction

- In fact, we may want to make two types of predictions:
  - We might want to estimate the mean of Y when $X = x$
  - We also might want to predict the value of the random variable Y when $X = x$
- If we wish to estimate the (conditioanl) mean of Y when $X = x$, we wish to estimate

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

- If we wish to predict the value of Y when $X = x$, we wish to predict

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

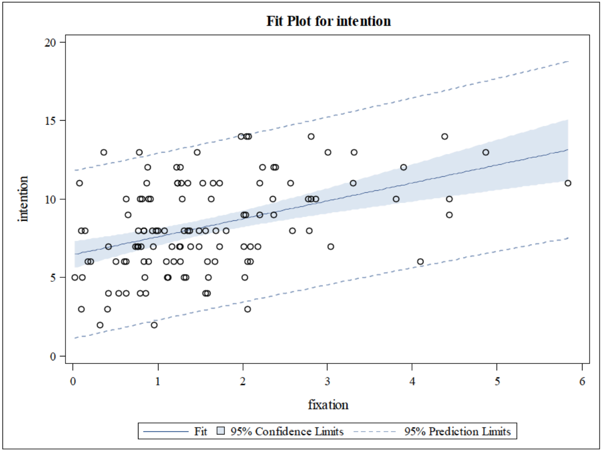
Prediction

- Whether we want to estimate the mean or predict the value of Y when $X = x$, the predicted values will actually be identical, and will be the point on the line corresponding to $X = x$.
- More formally, the prediction will be:

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

- However, the difference between these two situations will be in the precision of estimation of the predicted value:
  - Estimation will be more precise when we predict the mean of Y, compared to predicting an invidual value of Y

Prediction



- In this plot we see the confidence bands for estimating the mean (blue band) and those for the individual predictions of Y (dotted lines).
- (Note: this plot is created in SAS from the `proc glm` command, a similar graph can be produced in R using `ggplot` - see code for more details).

Prediction

- In our example, we estimated the line:

$$\widehat{\texttt{Intention}} = 6.45 + 1.14\,\texttt{Fixation}$$

- If we wanted to estimate the mean intention to buy score for people who fixated on the product for 3 seconds, we would get:

$$6.45 + 1.14 \times 3 = 9.87$$

- If we wanted to predict the value of the intention to buy score for someone who fixated on the product for 3 seconds, we would get exactly the same value.
- However, if we wanted to get a CI for this value, the interval would be wider for the individual value.

Prediction

Example: fixation-intention to buy

Table of contents

1. Introduction
2. Correlation
3. Regression
4. Estimation
5. Prediction
6. Residual analysis
7. R^2
8. Binary predictor
9. Exam errors

Analysis of residuals

- Until now, we have fit models and carried out statistical tests for the parameters (or calculated CIs for them).
- CAREFUL: the validity of the tests and CIs we've seen rely on certain assumptions on the model.

Necessary assumptions for valid analysis

- We'll now revisit the necessary assumptions in linear regression in detail.
- Recall: two equivalent ways to express the model assumptions are:

Assumptions for the regression model

- **Assumptions on** ϵ: in the basic model, we assume that $\epsilon_1, \ldots, \epsilon_n$ are independent Normally distributed random variables with mean $E(\epsilon_i) = 0$ and variance $Var(\epsilon_i) = \sigma^2$:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- **Assumptions on** Y: in the basic model, we assume that $Y_1, \ldots, Y_n$ are independent Normally distributed random variables such that:

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i \; ; \; Var(Y_i|X_i) = \sigma^2$$

Necessary assumptions for valid analysis

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \epsilon_i}$$

1. The error terms $\epsilon_1, \ldots, \epsilon_n$ are independent random variables
   - This also implies the Y_i are also independent
2. The expectation of the errors is $E[\epsilon_i] = 0$ for all $i = 1, \ldots, n$
   - This implies that the "mean" of the model is correctly specified:

     $$E(Y|X) = \beta_0 + \beta_1 X$$

   - This also implies that all important explanatory variables are included in the model and that their effect (assumed linear) is correctly modelled.
3. The variance of the error terms is constant for all i: $Var[\epsilon_i] = \sigma^2$
   - This also implies that the variance of the of the observations is constant; that is, there is homoscedasticity. If this is not true, then we say there is heteroscedasticity.
4. The error terms, ϵ, follow a normal distribution.

Necessary assumptions for valid analysis

See Chapter 2 Part 1 Supplementary Information

Verification of model assumptions

- The model assumptions involve the random error term ϵ. Thus, the validity of the assumptions can be tested using the residuals of the fitted model.
- These residuals are estimates of the errors ϵ and represent the difference between the observed value Y_i and the estimated value on the line $\hat{Y}_i$.
- Therefore, we define the residuals in the following way:

$$\boxed{e_i = Y_i - \hat{Y}_i,}$$

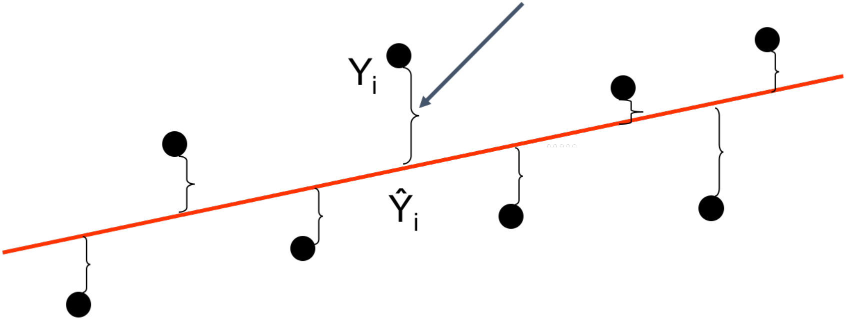
or equivalently:

$$\boxed{e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)}$$

for $i = 1, \ldots, n$.

Residuals

Vertical distance between a data point and a line (residual) = ($Y_i - \hat{Y}_i$)



Residuals

- For any sample data and fitted regression model, the sum of the residuals is always equal to 0. This comes from the way they're constructed (i.e. the way the line is fitted to the data).

$$\boxed{\sum_{i=1}^{n} e_i = 0}$$

- The problem with the residuals (called the ordinary residuals) is that they do not all have the same standard deviation. This can make certain comparisons difficult when we want to verify the model assumptions.

- It is often preferable to use a standardized version of the residuals instead. There are several different versions...

Residuals

- Standardized residuals (`rstandard` function in R)
  - Residuals are standardized to have unit variance by dividing each ordinary residual e_i by an estimate of its standard deviation (which differs from one residual to another).

$$e_i/\hat{\sigma}(1-h_{ii})$$

  - Note that these residuals are sometimes also referred to as internally studentized residuals.
- Studentized residuals (`rstudent` function in R)
  - Similar to the standardized residuals, only a leave-one-out measure of the variance is used. To calculate the residual for the i^{th} observation
    - remove the i^{th} observation and refit the model with the $n-1$ remaining observations, and obtain an estimate of the variance (σ^2) based on these $n-1$ observations ($\hat{\sigma}^2_{(i)}$)
  - The studentized residuals are obtained by standardizing with $\hat{\sigma}^2_{(i)}$
  - Note that these residuals are sometimes also referred to as externally studentized residuals or jackknife studentized residuals.

Analysis of residuals

- The residuals allow us to assess the validy of the model assumptions, and thus allow to verify the following:
  - Are the explanatory variables properly modelled (i.e. is the mean model correctly specified)?
  - Is there heteroscedasticity?
  - Is the distribution of the error terms close to normal?
  - Are there any extreme (or outlier) values?
- In general, we will rely on residual plots and descriptive statistics of the residuals.
- We will see how to test each of the model assumptions using plots and descriptive statistics.

Assumption (1): Independent observations

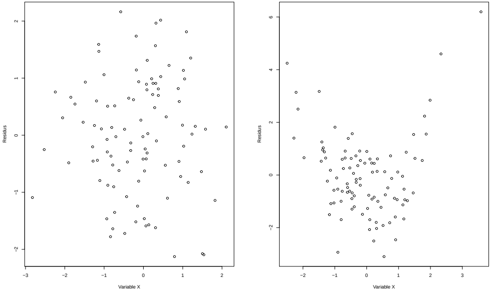
- Usually, the independence of observations follows directly from the type of sampling used.
- Generally this assumption is implicitly true if the observations were taken from a random sample from the population.
- One example where the assumption is generally not met is in longitudinal data. In this case, we have repeated measures from the same individuals across time.
- If we want to include all the time points in the analysis, we must take into account the possible dependence (correlation) between observations from the same individual. We will see how to do this in the chapter 5 with linear mixed models.

Assumption (2): zero mean of residuals

- As mentioned earlier, this assumption implies that the mean model is correctly specified
  - i.e. the effect of X (assumed linear) is correctly modelled and all important explanatory variables have been included in the model (we'll discuss linear regression with several explanatory variables in part 2 of this chapter)
- Thus, we want to check if the fitted line is a good model for the data (in comparison to some other model, such as a quadratic model or some other non-linear model)
- To check this assumption, we can plot the (standardized or studentized) residuals as a function of the explanatory variable X (scatterplot)
- If the assumption is met, we should not see any trend or pattern in the plot.

Verifying assumption (2): zero mean of residuals

- Here is an example where the model is good (left) and another where the model is bad (the true relationship between X and Y is quadratic).



- If this assumption is not met and we fit a linear regression model to the data where X in fact has a non linear relation with Y, then the estimated coefficients β are not meaningful since the model itself is incorrectly specified.

Assumption (2): zero mean of residuals (continued)

- **What happens if this assumption is not met and we add too many variables to the model?**
  - Usually the estimated parameters β and the model predictions will be unbiased. However, the variances of the estimators will be a bit higher, since we would be estimating extra unneccesary parameters.
- **What happens when this assumption is not met and there are variables missing from the model?**
  - Usually, the parameters β and the predictions $\hat{Y}$ are biased.
  - In most cases, it's worse to leave out important variables. That is, it's worse to have an under-specified model (missing variables) than one with too many (useless) variables included.
- But even if we've included all the important variables, the assumption requires that their effects have been correctly modelled. For example, if an explanatory variable X has a quadratic effect on Y, we need to include the terms X and X^2 in the model (we will see this in multiple linear regression in part 2 of this chapter).

Assumption (3): constant variance

- The constant variance assumptions implies that the variance of the error terms ϵ are constant for each subject, and thus constant across values of X

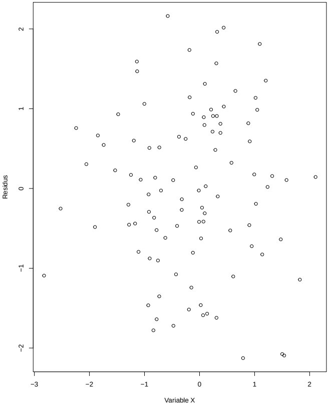
- To say that the variance of the residuals is constant for all subjects is equivalent to saying the variance of the observations Y is constant for all subjects.

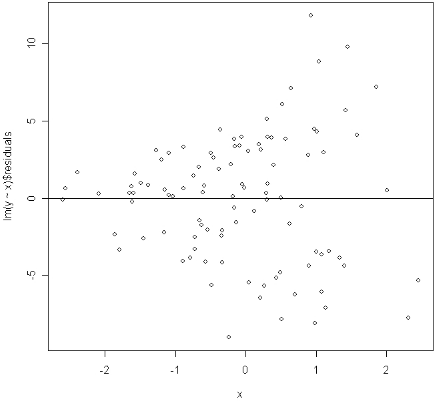
- To verify this assumption, we can plot the residuals versus the predicted values $\hat{Y}_i$. We can also look at same plot as before: the residuals versus X.

- We look at the variability of the residuals, and make sure there is no discernible pattern.
    - a cone-shaped pattern, or some other pattern, is an indication that the variance is not constant

Verification of assumption (3): constant variance

- Example: constant variance (left) and non-constant variance (right)





Verification of assumption (3): constant variance

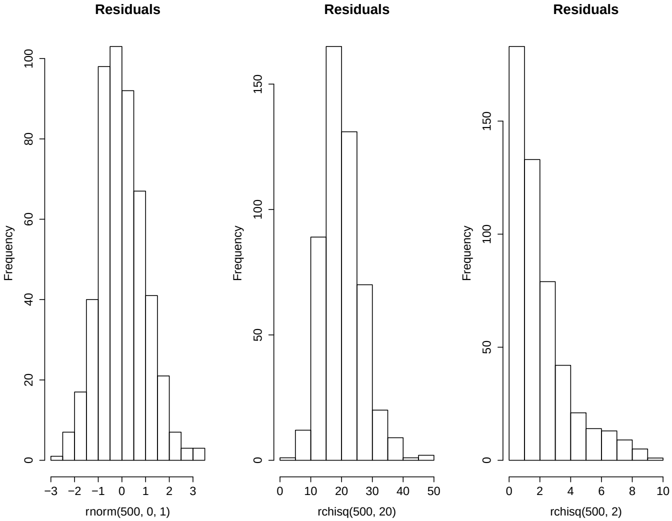
- When this assumption is not met, that is, when the variance is not constant over observations (as per the value of X or $\hat{Y}$), then the estimated effects of the variables (the β parameters) are still valid in the sense that they are unbiased.
- Conversely, the estimated standard deviations of the $\hat{\beta}$ are no longer valid.
- Consequently, the CIs and the hypothesis tests concerning the parameters will NOT be correct, since their calculations involve the standard deviations of $\hat{\beta}$.

Assumption (4): normality of the residuals

- This assumption is only important if we have a small sample size.
    - The analysis will still be valid for large samples even if the errors are not normally distributed.
- For small sample sizes, the assumption of Normally distributed error terms ϵ_i is necessary in order for the hypothesis tests and confidence intervals to be valid. It's precisely the fact that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ that allows us to use the Student-t distribution in the t-test and for computing CIs.
- In practice, for medium sample sizes, we only worry about whether the errors are symmetric about 0.

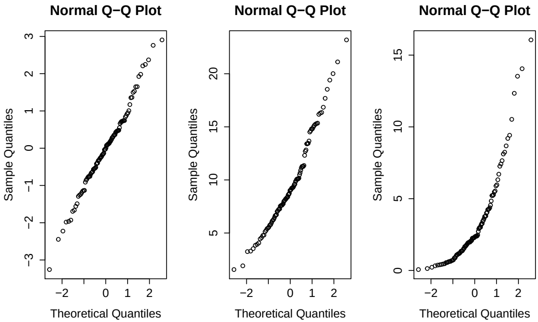
Verification of assumption (4): normality of the residuals

- To verify the normality of the residuals, we can plot a histogram of the residuals



Verification of assumption (4): normality of the residuals

- We can also make a “QQ-plot” of the residuals



- *A QQ-plot is a graph of the quantiles of some reference distribution (here the Normal distribution) against the quantiles of a sample of observations. The more the points fall on the diagonal line, the more the sample follows this reference distribution (Normal distribution here).*

Residual Analysis

Example: fixation-intention to buy

Remarks

- It's important to not try too hard to look for patterns.
- The human eye often has the tendency to look for patterns that aren't really there.
- Usually, we are concerned with a "clear" and obvious pattern.

Table of contents

Simple linear regression: coefficient of determination

- Once the model has been fitted, it is useful to have a measure that will tell us whether the model fits the data well.
- The coefficient of determination, R^2, measures the strength of the linear relationship between X and Y .
- It is interpreted as the proportion of the variation in Y explained by X.

Simple linear regression: coefficient of determination

- Suppose that we do not use the variable X. In this case, the best prediction of Y is simply the overall mean $\bar{Y}$. In this case,

$$\boxed{SS_T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

  is the sum of the squared errors when $\bar{Y}$ is used as the predicted value of Y_i.

- When we use X, the prediction of Y_i is given by $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. So,

$$\boxed{SS_E = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

  is the sum of the squared errors in the linear regression model including X

Simple linear regression: coefficient of determination

- It can be shown that

$$\begin{aligned} SS_T &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \\ &= SS_E + SS_R \end{aligned}$$

- SS_T is referred to as the total sum of squares - the total variability in the response variable.
- SS_R is referred to as the sum of squares due to the regression - measures the variability in the response variable that is explained by the model.
- SS_E is referred to as the sum of squares of the errors - the variability in the response variable that cannot be explained by the model.

Simple linear regression: coefficient of determination

- Recall:
  - SS_T is the model error <u>without</u> X (i.e. $Y_i = \beta_0 + \epsilon_i$)
  - SS_E is the model error <u>with</u> X (i.e. $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$)
- Consequently, $SS_T - SS_E$ is the reduction of the prediction error associated with including X in the model , or rather, the portion of the variation in Y explained by X.
- By dividing by SS_T, we end up with a proportion:

$$\boxed{R^2 = \frac{SS_T - SS_E}{SS_T}}$$

This gives the proportion of the variability in Y explained by X.

Simple linear regression: coefficient of determination

Properties of the coefficient of determination

- $0 \leq R^2 \leq 1$
- R^2 is the squared correlation coefficient between X and Y ($R^2 = r^2$). This is only true when there is one explanatory variable in the model (i.e. simple linear regression).
- R^2 is the squared correlation coefficient between the predicted values and the observed values of the dependent variable. That is, the correlation coefficient of $(Y_1, \hat{Y}_1), \dots, (Y_n, \hat{Y}_n)$
- The higher the value of R^2, the stronger the linear relationship between X and Y .

Simple linear regression: coefficient of determination

Example: fixation-intention to buy

Table of contents

1. Introduction
2. Correlation
3. Regression
4. Estimation
5. Prediction
6. Residual analysis
7. R^2
8. **Binary predictor**
9. Exam errors

Binary explanatory variable

- Though the response variable Y must be continuous (normality assumption), the X variables can be of any type.
- The case of a binary variable is the most simple, but up to now we've only seen examples with continuous variables.
- When it comes to including a binary predictor, we can either:
  - treat it as continuous, coding it as 0/1
  - treat it as categorical (we'll see how to do this exactly)

  Both approaches will lead to the same results, as we will see.

Coefficient interpretations

- Suppose we have a simple linear regression model

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

  where $X \in \{0, 1\}$.

  - e.g., $Y =$ intention, $X =$ sex

- In order to interpret the coefficients, we must first understand the model itself.

- Before now, we interpreted the slope β_1 as the increase in the mean of Y when X increases by one unit.

- This interpretation can be simplified in the case of a binary predictor - essentially what does it mean to say that “X increases by one unit”, when X is binary?

Coefficient interpretations

- We can write the model as:

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

- Since X takes only two values (0 or 1), we can say that:

$$\begin{aligned} E(Y|X=0) &= \beta_0 \\ E(Y|X=1) &= \beta_0 + \beta_1 \\ \Rightarrow \beta_1 &= E(Y|X=1) - E(Y|X=0) \end{aligned}$$

- In our example with Y = intention and X = sex (dichotomization), this means that
  - the mean intention to buy is equal to β_0 for male individuals (sex=0)
  - the mean intention to buy is equal to $\beta_0 + \beta_1$ for non-male individuals (e.g., females) (sex=1)
  - $\Rightarrow \beta_1$ (the effect of sex) represents the difference in the mean intention to buy between these two groups.

Binary explanatory variable

Example: fixation-intention to buy

Table of contents

Types of errors in exams and assignments

You will lose marks if you do the following...

Writing the regression model

When you're asked to write the regression model, you can either:

- write the model in terms of the random variable Y

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

  (Don't forget the error term! ϵ)

- or in terms of the mean

$$E(Y|X) = \beta_0 + \beta_1 X$$

- If you are asked for the fitted model, (i.e., you have the estimated values, e.g. $\hat{\beta}_0 = 1.2$ and $\hat{\beta}_0 = 45.9$) :

$$\hat{y} = 1.2 + 45.9x$$

  (without the ϵ)

- You should NOT write:
  - $Y = \beta_0 + \beta_1 X$
  - $\hat{y} = 1.2 + 45.9x + \epsilon$

Parameter interpretations

- Suppose that you're asked to interpret the regression model parameters
- Caution: the R^2 and the p-values in the regression model are not considered "parameters" in the model. Only the β terms are parameters.
- If one of the parameters is not significant, sometimes students do not interpret them....
- The significance of the parameters has nothing to do with the interpretation. Even if a coefficient is not significant, it can still be interpreted in the same way.

Describe the methods used

- A `lm` is not a statistical method. To say that you used the `lm` function doesn't sufficiently describe the method.
- Instead, you could say that you fit a linear regression model, for example.