Chapter 2: Linear Regression
Part 2: Multiple 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)**

Multiple linear regression

- Multiple linear regression is a generalization of the simple linear regression model, wherein several variables are included as predictors in the model.
- All the concepts we've seen so far in the context of simple linear regression are still valid. We will revisit these notions within the context of multiple linear regression, and see how there a few subtle differences, e.g.
  - linear relationship, linear correlation (pairwise*)
  - parameter interpretations (intercept, slope*)
  - method of least squares for parameter estimation
  - hypothesis testing, C.I.s
  - residual analysis
- We will also see some new concepts that are only relevant in the multiple linear regression model, e.g.
  - tests for global effects
  - interactions
  - multicollinearity

Table of contents

Multiple linear regression: model definition

- Multiple linear regression allows to examine the relationship between a dependent variable Y and p independent variables $X_1, X_2, \ldots, X_p$ simultaneously.
- Data:
  - n observations
  - Y_i denotes the value of Y for individual i
  - X_{ij} denotes the value of the j^{th} variable X_j for individual i
  - thus for each subject i we observe data

$$(Y_i, X_{i1}, X_{i2}, \ldots, X_{ip}), \quad i = 1, \ldots, n$$

- The multiple linear regression model is defined as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i$$

- As was the case for simple linear regression, we treat $X_1, \ldots, X_p$ as fixed, while Y and ϵ are considered random variables.

Linear regression: model assumptions

The underlying assumptions in the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i$$

can be written in 2 equivalent ways:

Assumptions for the linear regression model

- **Assumptions on** ϵ: $\epsilon_1, \ldots, \epsilon_n$ 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: $Y_1, \ldots, Y_n$ are independent Normally distributed random variables such that:

$$E(Y_i | X_{i1}, \ldots, X_{ip}) = \beta_0 + \beta_1 X_{i1} + \cdots \beta_p X_{ip}$$

$$Var(Y_i | X_{i1}, \ldots, X_{ip}) = \sigma^2$$

Linear regression: model assumptions

The assumptions of the multiple linear regression model are the same as those in simple linear regression:

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

     $$E(Y_i|X_{i1}, \ldots, X_{ip}) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

     This also means that all the important explanatory variables have been included in the model and their effects (presumed linear) have been properly captured by the model.
3. The variance of the errors is constant over i: $Var[\epsilon_i] = \sigma^2$
   - i.e. the variance of Y_i is constant (i.e. does not depend on X_i), meaning there is homoscedasticity. Otherwise, we have heteroscedasticity.
4. The error terms ϵ_i follow a normal distribution.
   - i.e. the Y_i follow a normal distribution

Parameter interpretations

- In multiple linear regression, the parameter β_j measures the effect of the variable X_j on the response variable Y while controlling for all other variables in the model.
- The interpretation of the β coefficients is the same as for simple linear regression, except for the following adjustment:
  - for every one unit increase in X_j , Y increases on average by β_j **when all other variables are held constant.**
- Another way of thinking about it is: β_j is the marginal contribution of X_j after all the other variables have been included in the model.
- Why?

$$\begin{aligned}\beta_1 &= E(Y|X_1 = x_1 + 1, X_2 = x_2, \ldots, X_p = x_p) \\ &\quad - E(Y|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) \\ &= \{\beta_1(x_1 + 1) + \beta_2 x_2 + \cdots \beta_p x_p\} \\ &\quad \{\beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p\}\end{aligned}$$

Interpretation of the intercept

- Recall the model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i$$

- The intercept β_0 represents the mean value of Y when all the variables in the model are set to zero.
- Why?

$$\begin{aligned} \beta_0 &= E(Y|X_1 = 0, X_2 = 0, \ldots, X_p = 0) \\ &= \beta_0 + \beta_1 \times 0 + \beta_2 \times 0 + \cdots + \beta_p \times 0 \end{aligned}$$

- Of course, it is possible that this interpretation does not make sense in the context of the study.

Parameter Estimation

- As was the case in simple linear regression, the coefficients are estimated according to the least squares criterion (i.e. we want to minimize the sum of the squared errors):

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

where

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_0 X_{ip}$$

We find estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ simultaneously such that the sum of the squared errors is minimized.

- In multiple linear regression, the parameters are estimated according to a system of equations. We can write out an explicit expression for these estimators, but it's more convenient to write this using matrix notation.

Parameter Estimation

- Matrix notation:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \ \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

- We can write the model in terms of matrix notation as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- We can show that the least squares estimators are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

These calculations are done numerically using statistical softwares such as R

- It can be shown that

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \quad \mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

Hypothesis tests on the parameters

- As we saw in the case of simple linear regression, we can evaluate whether the regression coefficients β_j are significant for $j = 1, \ldots, p$.
- To assess the significance for the parameter β_j, the underlying hypotheses are:

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

- It is important to note that this tests the marginal contribution of X_j when all other variables are included in the model. We're testing whether X_j can help explain the behaviour of Y on top of the other variables that are in the model.
- When $\beta_j = 0$ what does it mean?

$$\begin{aligned}\beta_j = E(Y|X_1 = x_1, \ldots, X_j = x_j + 1, \ldots, X_p = x_p) \\ - E(Y|X_1 = x_1, \ldots, X_j = x_j, \ldots, X_p = x_p)\end{aligned}$$

So if $\beta_j = 0$, it means that increasing X_j by one unit has no (linear) effect on Y, on average, *holding all other variables constant*

Hypothesis tests on the parameters

- Similar to simple linear regression, this test is based on the test statistic

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

- The interpretation of the test is similar to that of β_1 (the slope) in simple linear regression; although now we have to remember that we're testing the effect of a single variable X_j once all the other variables have been included in the model, that is, adjusting for the other variables in the model.
- If we have a small p-value ($p < \alpha$), we reject $H_0 : \beta_j = 0$ in favor of $H_1 : \beta_j \neq 0$ (for a given significance level α)

  $\rightarrow X_j$ has a significant (linear) effect on Y, even after adjusting for the other variables included in the model.

- Under H_0, the test statistic t follows a Student t distribution with $n - p - 1$ degrees of freedom.

Hypothesis tests on the parameters

- If we fail to reject $H_0 : \beta_j = 0$ (vs. $H_1 : \beta_j \neq 0$), it's important not to jump to the conclusion that the variable X_j is not related to the response variable Y. We can only say that there is not a significant linear association between X_j and Y once the other variables are included in the model, i.e. once the other variables have been adjusted for.

- Ex: Suppose we're interested in evaluating the effect of two explanatory variables X_1 and X_2 on a response variable Y. We have a sample of size $n = 50$.

- We fit 3 models:

  i) $Y = \beta_0 + \beta_1 X_1 + \epsilon$

  ii) $Y = \beta_0 + \beta_2 X_2 + \epsilon$

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

Hypothesis tests on the parameters

- Results:

  i) $Y = \beta_0 + \beta_1 X_1 + \epsilon$

  p-value < 0.0001 for $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$, thus X_1 has a significant (linear) effect on Y

  ii) $Y = \beta_0 + \beta_2 X_2 + \epsilon$

  p-value < 0.0001 for $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$, thus X_2 has a significant (linear) effect on Y

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

  p-value < 0.0001 for $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$, thus X_1 has a significant (linear) effect on Y, even after adjusting for X_2

  p-value 0.776 for $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$, thus X_2 does NOT has a significant (linear) effect on Y, after adjusting for X_1

Hypothesis tests on the parameters

- We see that once X_1 is included in the model, X_2 no longer contributes to explaining (linearly) the response variable Y!
- Yet, the model that only includes X_2 showed that X_2 had a significant (linear) effect on Y
- This could be explained by the fact that X_1 and X_2 are themselves correlated and in that sense X_1 and X_2 are explaining the same "part" of the behaviour of Y. We'll come back to this in detail and explain the implications of having explanatory variables that themselves are correlated (multicollinearity).

Example: fixation-intention to buy

Table of contents

Categorical explanatory variables

- In part 1 of chapter 2 (simple linear regression), we saw how to incorporate a specific type of categorical explanatory variable: binary variables
  - Recall: a binary variable can take on the value 0 or 1.
  - We saw how to include this variable as is, or by declaring it as categorical (in R).
- Including a categorical explanatory variable is not trivial, and it's imporant to understand the process of modelling this type of variable, otherwise we could misinterpret the effects of these variables
- In fact, including a categorical variable (with > 2 categories) in the model (eg: `educ`, education categorized into 3 levels) does not lead to a single predictor variable X (eg: `educ`) being included in the model...
  - Why? What would happen if we included the variable `educ` as is? How would we interpret the coefficient β for this variable?
  - What would happen if the variable is ordinal? nominal?

Categorical explanatory variables

- It turns out that a variable with k categories requires the inclusion of $k-1$ explanatory variables X in the model
- More precisely, if we have a categorical explanatory variable (nominal or ordinal) with multiple possible levels, we can include it in the model by creating several indicator or binary variables (dummy variables) to indicate the level
  - indicator or binary variable: $X \in \{0, 1\}$
- For a categorical variable with k possible values, we only need to create $k-1$ indicator variables.
- Note that the case of a binary variable that we've already seen is a special case of a categorical variable where $k = 2$ (ex. `sex`), so we only need a single indicator variable to represent the binary variable in the model (`sex`).

Categorical explanatory variables

- For example, take the variable `educ` categorized into three levels:
  - 1=Less than high school
  - 2=high school
  - 3=university
- Note that this is an ordinal variable.
- Since this variable has three levels, we will create two indicator variables, `educ1` and `educ2`:
  - `educ1` = 1 if `educ`=1 and 0 otherwise
  - `educ2` = 1 if `educ`=2 and 0 otherwise
- Note that there is no need to create a third indicator variable `educ3`
  - `educ3` = 1 if `educ`=3 and 0 otherwise

  as this would be redundant, since `educ1`+`educ2`+`educ3`=1

Categorical explanatory variables

- This can be seen in the following table:

educ	educ1	educ2
1	1	0
2	0	1
3	0	0

$\Rightarrow$ `educ=3` when `educ1`=`educ2`=0

- In this case, by including `educ1` and `educ2` in the model, the 3rd level of `educ` is called the reference category; it's when both dummy variables `educ1` and `educ2` are zero.
- We will see that this parameterization means we will directly compare the other categories to this reference category.

Interpretation

- The model can be written as:

$$E(Y|\texttt{educ}) = \beta_0 + \beta_1\ \texttt{educ1} + \beta_2\ \texttt{educ2}$$

- Since there are only 3 possibilities, we can easily write them all out:

$$\begin{aligned} E(Y|\texttt{educ=1}) &= E(Y|\texttt{educ1=1, educ2=0}) \\ &= \beta_0 + \beta_1 \\ E(Y|\texttt{educ=2}) &= E(Y|\texttt{educ1=0, educ2=1}) \\ &= \beta_0 + \beta_2 \\ E(Y|\texttt{educ=3}) &= E(Y|\texttt{educ1=0, educ2=0}) \\ &= \beta_0 \end{aligned}$$

- In this model, the mean of the variable `intention` takes a different value for each of the three possible values of `educ`.

Interpretation

$$\begin{aligned} E(Y|\texttt{educ=1}) &= \beta_0 + \beta_1 \\ E(Y|\texttt{educ=2}) &= \beta_0 + \beta_2 \\ E(Y|\texttt{educ=3}) &= \beta_0 \end{aligned}$$

- Thus the model allows the mean of the response variable `intention` to freely take on a different value for each of the three possible education categories.
- The parameter β_1 represents the difference in the mean `intention` when `educ=1` compared to `educ=3` (reference category):

$$\beta_1 = E(Y|\texttt{educ=1}) - E(Y|\texttt{educ=3})$$

- The parameter β_2 represents the difference in the mean `intention` when `educ=2` compared to `educ=3` (reference category):

$$\beta_2 = E(Y|\texttt{educ=2}) - E(Y|\texttt{educ=3})$$

- The parameter β_0 represents the mean `intention` when `educ=3`

Hypothesis tests

- As before, we can carry out tests of the form

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

- Here, however, we must be careful with the conclusions we make. For example, continuing with the model

$$E(Y|\texttt{educ}) \;=\; \beta_0 + \beta_1\ \texttt{educ1} + \beta_2\ \texttt{educ2}$$

  - $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ is testing whether there is a significant difference in the mean between education levels 1 and 3
  - $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$ is testing whether there is a significant difference in the mean between education levels 2 and 3
  - $\rightarrow$ the tests always involving comparisons with respect to the reference level, here level 3 (educ=3).

Categorical explanatory variables

- What about the difference between the groups educ=1 and educ=2?

$$E(Y|\texttt{educ=1}) - E(Y|\texttt{educ=2}) = \beta_1 - \beta_2$$

  - We can certainly *estimate* this difference simply as $\hat{\beta}_1 - \hat{\beta}_2$.
  - But we don't directly have a p-value to test whether this difference is significantly different from 0...

- One way to approach this is to consider a reparametrization of the model by changing the reference category
- Note that the model will be the same as that previously considered. We're just using a different parametrization for the binary (dummy) variables. For example, if we let 2 be the reference category:

educ	educ1	educ3
1	1	0
2	0	0
3	0	1

Categorical explanatory variables

- Note that this model is the same as the previous one, we just used a different parametrization wherein category 2 was used as the reference level:

$$\begin{aligned} E(Y|\texttt{educ=1}) &= \beta_0^* + \beta_1^* \\ E(Y|\texttt{educ=2}) &= \beta_0^* \\ E(Y|\texttt{educ=3}) &= \beta_0^* + \beta_2^* \end{aligned}$$

- The model is EQUIVALENT, it is simply a reparametrization.
- Note when fitting this model, we would obtain the same estimated means for each of the three groups as before, i.e. :

$$\begin{aligned} \widehat{E}(Y|\texttt{educ=1}) &= \hat{\beta}_0 + \hat{\beta}_1 \\ &= \hat{\beta}_0^* + \hat{\beta}_1^* \\ \widehat{E}(Y|\texttt{educ=2}) &= \hat{\beta}_1 + \hat{\beta}_2 \\ &= \hat{\beta}_0^* \\ \widehat{E}(Y|\texttt{educ=3}) &= \hat{\beta}_0 \\ &= \hat{\beta}_0^* + \hat{\beta}_2^* \end{aligned}$$

Categorical explanatory variables

- With this model, we can now test whether there is a significant difference in the mean intention score between the `educ=1` and `educ=2` groups:

$$E(Y|\texttt{educ=1}) - E(Y|\texttt{educ=2}) = \beta_1^*$$

by testing:

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

- When there are only 3 levels for the categorical variable, things are not too bad:
  - we only have to create two "dummy" variable;
  - we can obtain all possible pairwise comparisons by simply fitting two models (changing the reference level)
- When there are a large number of categories, however, this becomes cumbersome. Note that most statistical softwares have methods for doing this automatically:
  - creating the dummy variables
  - carrying out all possible pairwise comparisons

Categorical explanatory variables

Example: fixation-intention to buy

Remarks on nominal vs. categorical variables

- If the variable is nominal, we simply include its corresponding indicator variables since there is no order to the levels.
- If the variable is ordinal (as we saw for `educ`), we could actually treat it like a continuous variable with a single parameter to see if its effect is approximately linear.
- This allows us to use the intrinsic ordering of the values, as well as to reduce the number of parameters in the model.
- If the effect is not linear, we could then use indicator variables instead. This ignores the ordering of the levels, and means we're essentially treating the variable as nominal.

Table of contents

Hypothesis tests for several variables

- Up till now, we have seen hypothesis tests for individual parameters of the form

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

- There are also ways to test whether several parameters β's are equal to zero.
  - For example: $H_0 : \beta_1 = \beta_2 = 0$ vs. H_1 : at least one of β_1 or β_2 is different from 0.
  - This is useful if we want to test the joint contribution of 2 variables in a model. For example:

$$H_0 : \beta_{age} = \beta_{emotion} = 0$$

vs. H_1: at least one of these parameters is different from 0.

Hypothesis tests for several variables

- This can also be used to test the global effect of a categorical variable:
  - For example, we have seen that the variable `educ` with 3 levels is incorporated into a regression modelled by using two indicator variables `educ1` and `educ2`:

$$E(Y|\texttt{educ}) = \beta_0 + \beta_1 \texttt{educ1} + \beta_2 \texttt{educ2}$$

    Testing the global effect of `educ` means we're testing the hypothesis:

$$\begin{aligned} H_0 &: \beta_{educ1} = \beta_{educ2} = 0 \\ H_1 &: \text{at least one of the parameters is different from 0} \end{aligned}$$

  - If we fail to reject H_0, we can conclude that the variable `educ` does not have a significant effect on the response variable `intention`
  - This test is different from what we previously saw (ex. $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ allows us to conclude whether there is a significant different in the mean `intention` for those with `educ=1` in comparison to `educ=3`)

Hypothesis tests for several variables: test statistic

- Consider the "full" model which contains all p predictor variables:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \beta_{k+1} X_{k+1} + \ldots + \beta_p X_p + \epsilon$$

- Suppose that we want to test the following null hypothesis:

$$\begin{aligned} H_0 &: \beta_{k+1} = \beta_{k+2} = \ldots = \beta_p = 0 \\ H_1 &: \text{at least one of these parameters is different from 0} \end{aligned}$$

- This hypothesis specifies that $(p - k)$ of the β parameters are equal to zero.
- Consider the "reduced" model which contains all the variables except the ones we want to test:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon$$

(i.e. the true model if H_0 is true)

Hypothesis tests for several variables: test statistic

- Let $SS_E(full)$ be the sum of the squared errors for the full model:

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

where $\hat{Y}_i^{full}$ are the predicted values coming from the full model (with all p predictor variables)

- Similarly, define $SS_E(reduced)$ as the sum of the squared errors for the reduced model:

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

where $\hat{Y}_i^{reduced}$ are the predicted values coming from the reduced model (with only the k predictor variables)

Hypothesis tests for several variables: test statistic

- The hypothesis test for

$$H_0 \quad : \quad \beta_{k+1} = \beta_{k+2} = \ldots = \beta_p = 0$$
$$H_1 \quad : \quad \text{at least one of these parameters is different from 0}$$

is based on the test statistic

$$F = \frac{\{SS_E(reduced) - SS_E(full)\}/(p-k)}{SS_E(full)/(n-p-1)}$$

- When H_0 is true, the F statistic follows a Fisher distribution with $(p-k)$ and $(n-p-1)$ degrees of freedom.
  - This is a type of F-test

Hypothesis tests for several variables: test statistic

- Note that the tests we've seen thus far for the regression parameters in linear regression are actually particular applications of this test:
  - ex: if $k = p - 1 \rightarrow$, i.e. testing $H_0 : \beta_p = 0$ vs. $H_1 : \beta_p \neq 0$, this is equivalent to a t-test

$$F = \frac{\{SS_E(reduced) - SS_E(full)\}/1}{SS_E(full)/(n-p-1)} = t^2$$

  - ex: simple linear regression

$$F = \frac{\{SS_E(reduced) - SS_E(full)\}/1}{SS_E(full)/(n-2)} = t^2$$

- The formulation presented here for the F-test demonstrates that we can actually test any specific subset of the parameters using the same approach.

Hypothesis tests for several variables

- Note: in the special case where $k = 0$, i.e. when we test

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ H_1 &: \text{At least one } \beta_j \text{ is different from 0} \end{aligned}$$

  we're specifically testing whether at least one of the explanatory variables is useful in explaining the response variable (linearly)

- Under H_0, the reduced model is simply

$$Y = \beta_0 + \epsilon$$

  which leads to $\hat{\beta}_0 = \bar{Y}$ and hence $\hat{Y}_i^{reduced} = \bar{Y}$ for each i

Hypothesis tests for several variables

- In this case,

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

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

- This leads to the test statistic

$$F = \frac{(SS_T - SS_E)/p}{SS_E/(n-p-1)} = \frac{SS_R/p}{SS_E/(n-p-1)}$$

where F follows a Fisher distribution with p and $n-p-1$ degrees of freedom, i.e. $F \sim F(p, n-p-1)$

- This allows to globally test the model - i.e. are any of the explanatory variables useful in explaining the response variable Y (linearly).

Global F-test

Example: fixation-intention to buy

Table of contents

Making predictions from the fitted model

- Once the parameters have been estimated, we can get predictions for Y for a given set of predictors $X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p$:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p$$

- The idea is exactly the same as for simple linear regression.
- We will see how to make these kinds of predictions in the context of multiple regression through an example.

Predictions

Example: fixation-intention to buy

Table of contents

Recall: model assumptions in multiple linear regression

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

     $$E(Y_i|X_{i1}, \ldots, X_{ip}) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

     This also means that all the important explanatory variables have been included in the model and their effects (presumed linear) have been properly captured by the model.
3. The variance of the errors is constant over i: $Var[\epsilon_i] = \sigma^2$
   - i.e. the variance of Y_i is constant (i.e. does not depend on X_i), meaning there is homoscedasticity. Otherwise, we have heteroscedasticity.
4. The error terms ϵ_i follow a normal distribution.
   - i.e. the Y_i follow a normal distribution

Residual analysis

- As was the case in simple linear regression, we must verify the model assumptions
  - if the model assumptions are violated, it could lead to biased estimates and any conclusions about the variable effects could be invalid
- The idea is exactly the same as for simple linear regression
  - the model assumptions can be verified by using the model residuals:

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

  - as before, we'll consider a detailed residual analysis (using the jackknife studentized residuals)

Residual analysis

What to verify:

- Histogram and qq-plot of the residuals
  - allows us to check the normality of the residuals
- Plot of the residuals vs. the X values
  - (should do this for each X variable included in the model)
  - allows to verify whether the model is correctly specified, particularly if X is linearly related to Y
  - when X is categorical or binary, we can look at boxplots for the residuals for each level of X
  - also allows to verify whether the variance is constant
- Plot the residuals vs. the predicted values $\hat{Y}$
  - allows to verify whether the model is correctly specified and if the variance is constant.

* The model is correctly specified if

$$E(Y|X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Residual analysis

Example: fixation-intention to buy

Table of contents

R^2 coefficient of determination

- As we saw in the context of simple linear regression, the coefficient of determination R^2 is obtained by decomposing the variance of Y
- For multiple linear regression, R^2 is defined in the same way as in simple linear regression.

R^2 coefficient of determination

- Suppose that we do not include any of the explanatory variables $X_1, X_2, \ldots, X_p$ in the model
  - (i.e. we assume a model of the form $Y = \beta_0 + \epsilon$)
  - In this case, the best prediction for Y is simply the sample mean $\bar{Y}$.
  - As in simple linear regression, we define

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

    that is, the sum of the squared errors when we use $\bar{Y}$ as the predicted value for each Y_i.

  - SS_T is referred to as the total sum of the squares and represents the total variability in the response variable.

R^2 coefficient of determination

- When we do include the explanatory variables $X_1, X_2, \ldots, X_p$ in the model
  - (i.e. we assume a model of the form $Y = \beta_0 + \beta_1 X_1 + \cdot + \beta_p X_p + \epsilon$)
  - In this case, the predicted value Y_i is given by $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_p X_{ip}$.
  - As in simple linear regression, we define

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

    as the sum squared errors in the regression model including all the explanatory variables $X_1, \ldots, X_p$, i.e. when we use $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_p X_{ip}$ as the predicted value for Y_i.
  - SS_E is referred to as the sum of squares of the errors and measures the variability in Y due to error, that cannot be explained by the model

R^2 coefficient of determination

- The remaining portion

$$SS_R = SS_T - SS_E = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

  is the sum of squares due to the regression model, it measures the portion of the variability in Y that is explained by the model.

- When $SS_R >> SS_E$, this is an indication that the model is doing a good job at explaining the variation in the response variable Y.

R^2 coefficient of determination

- Recap:
  - SS_T is error for the model without the variables $X_1, \ldots, X_p$ (i.e. when we assume a model of the form $Y = \beta_0 + \epsilon$)
  - SS_E is error for the model with $X_1, \ldots, X_p$ (i.e. when we assume a model of the form $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$)
- Consequently, $SS_T - SS_E$ is the reduction in model error associated with the explanatory variables $X_1, \ldots, X_p$, or rather the amount of variation in Y that is explained by the explanatory variables X.
- If we divide by SS_T, we get a proportion:

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

- Like in simple linear regression, this gives the proportion of the variability in Y explained by the set of predictor variables $X_1, \ldots, X_p$.

R^2 coefficient of determination

- We saw that if we have a single explanatory variable (as in simple linear regression), R^2 is equal to the square of the correlation coefficient r between the explanatory variable X and the response variable Y .
- When there is more than one explanatory variable (as in multiple linear regression), the square root of R^2 is called the multiple correlation coefficient.
- R^2 is the square of the correlation coefficient between the predicted values and the original values $(\hat{Y}_1, Y_1), \ldots, (\hat{Y}_n, Y_n)$.
- R^2 always takes a value between 0 and 1, i.e. $0 \leq R^2 \leq 1$

Remark: R^2 and model selection

- R^2 is a measure of goodness of fit with $0 \leq R^2 \leq 1$.
  - the closer R^2 is to 1, the better the model fit
  - if $R^2 = 0$ then the model is of no use, it does not help explain Y
  - if $R^2 = 1$ then the model fit is perfect
- However, R^2 cannot be used for *model selection*, that is, to select the model with the best set of predictor variables:
  - the problem with R^2 is that as you add predictors to the model, its value can never go down!
  - thus, for model selection purposes, it is better to use an adjusted version of R^2:

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-(p+1)} \right) (1 - R^2)$$

  - There also exists other measures for model selection (ex. *Akaike information criterion* (AIC) and *Bayesian information criterion* (BC)). We will discuss these later in the course.
  - Model selection is usually done in the context of prediction (best model to predict or explain Y), and we will not discuss this here.

R^2 and partitioning the total variability

- The variability due to error SS_E is also used to estimate the variance parameter σ^2 in the regression model (recall $\epsilon \sim \mathcal{N}(0, \sigma^2)$).
- In particular, an estimator for the variance is given by

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

  where $p+1$ is the total number of the parameters in the regression model (since there are p explanatory variables $X_1, \ldots, X_p$ and $p+1$ parameters $\beta_0, \beta_1, \ldots, \beta_p$).

- Recall: $\hat{\sigma}^2$ is also used to calculate $\hat{se}(\hat{\beta}_j)$

$$\hat{se}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}^\top \mathbf{X})^{-1}$$

Example: fixation-intention to buy

Table of contents

Modelling non-linear relationships

- Sometimes, a line is not appropriate to model the relationship between the explanatory variable X and the response variable Y.
- It turns out that the linear model can be used to model non-linear relationships.
- This is done by considering transformations of the X variables.
- There are several ways to proceed, and one of the most popular models is a polynomial involving X.

Modelling non-linear relationships

- For example, consider the following quadratic model:

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

- This model fits a second-order polynomial (a parabola). If the fit is not good enough, we could also fit a cubic model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

- ...or even a polynomial of order k (usually, $k \leq 3$ in practice)

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

Modelling non-linear relationships

- Additionally, there's nothing preventing us from adding other predictor variables in the model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 Z + \epsilon$$

- The effect of X is modeled using two terms, a linear term and a quadratic term. The effect of Z is only modeled through a linear term.

Modelling non-linear relationships: interpreting the coefficients

- An additional difficulty that arises when modelling non-linear relations with a predictor variable X is interpretability...
  - Interpreting the effect of a variable X becomes more difficult when it's included in a non-linear term
- When the variable only has a linear term included, e.g.

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

the interpretation of its effect is straightforward, since the interpretation is the same regardless of the value of the variable:

> *when X increases by 1 unit (and when all other variables remain constant), the variable Y increases by β_1 on average.*
>
> $$\beta_1 = E(Y|X = x + 1) - E(Y|X = x)$$
>
> This interpretation holds regardless of the "starting value" of $X = x$.

- What if we have a quadratic model?

Modelling non-linear relationships: interpreting the coefficients

- Suppose we fit a model of the form

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

or, equivalently

$$E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$$

- In this case, when X increases by 1, we have that

$$\begin{aligned} E(Y|X = x+1) - E(Y|X = x) &= \left\{\beta_0 + \beta_1(x+1) + \beta_2(x+1)^2\right\} \\ &\quad - \left\{\beta_0 + \beta_1 x + \beta_2 x^2\right\} \\ &= \beta_1(x + 1 - x) + \beta_2((x+1)^2 - x^2) \\ &= \beta_1 + \beta_2(x^2 + 2x + 1 - x^2) \\ &= \beta_1 + \beta_2(2x + 1) \end{aligned}$$

- $\rightarrow$ thus, for every one unit increase in X, Y increases, on average, by $\beta_1 + \beta_2(2X + 1)$.

Modelling non-linear relationships: interpreting the coefficients

- When the effect of an explanatory variable is modelled through a non-linear term, it's often helpful and even necessary to visualize the effect graphically
- For example, take the fitted model:

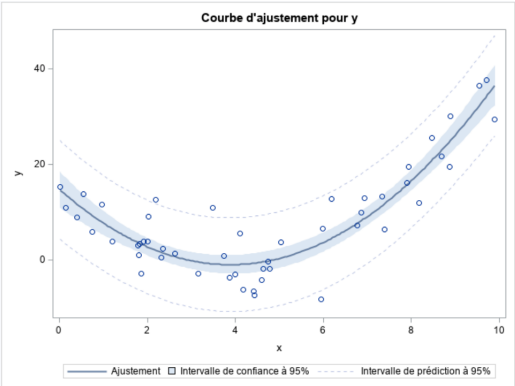
$$\hat{y} = 14.77 - 8.14x + 1.05x^2$$

- There is no single value describing the increase in the mean of Y associated with an increase of 1 in X.
  - The effect of increasing the value of X by 1 depends on the starting value of X:

$$\widehat{E}(Y|X = x + 1) - \widehat{E}(Y|X = x) = -8.14 + 1.05(2x + 1)$$

  - The effect of X on Y is more evident when examining the plot of Y vs X.

Modelling non-linear relationships: interpreting the coefficients



- We can still describe the estimated relationship:

  Y tends to decrease slightly when X increases from 0 to 5; however, Y tends to increase when X increases from 5 to 10.

Other kinds of non-linear relationships

- It is also possible to model other non-linear effects by considering different transformations of the variables.
- For example, we often use the logarithmic transformation, that is $\ln(X)$, when the distribution of a positive variable appears to be skewed to the right (salaries of individuals, for example).
- In this case, the model could be:

$$Y = \beta_0 + \beta_1 \ln(X) + \epsilon$$

- The interpretation of the effect of X can be made on the ln scale:
  - When $\ln(X)$ increases by 1, Y increases by β_1 on average.
- But on the original scale of X, the effect is non-linear:

  for each increase of 1% of X , Y increases by approximately $\beta_1/100$, on average

See the document "Linear Regression Models with Logarithmic Transformations" by Kenneth Benoit available at

http://www.kenbenoit.net/courses/ME104/logmodels2.pdf

Modelling non-linear relationships

- In general, we can use any type of transformation on the explanatory variables in the context of linear regression:

$$Y = \beta_0 + \beta_1 g_1(X_1) + \cdots + \beta_p g_p(X_p) + \epsilon$$

- This is still a linear regression model because it's linear in the parameters β!
- As we previously discussed, whenever we consider transformations of the explanatory variables X, the interpretations become less straightforward...

Transformations of Y

- We can also consider transformations of the response variable Y.
- In fact, this is often a technique used to correct for heteroscedasticity:
  - The transformation $\ln(Y)$ is often used when the variance is not constant. In this case, the model has the form

$$\ln(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

  - Other transformations can also be used, ex: $\sqrt{Y}, 1/Y$.
- As in the case for transformations on X, when we fit a linear regression model on a transformation of the response variable Y, parameter interpretations because more difficult...

Transformations of Y

- Ex: in the model

$$\ln(Y) = \beta_0 + \beta_1 X + \epsilon$$

the interpretation of β_1, as we've seen, is *for every one unit increase in* X, $\ln(Y)$ *increases by* β_1 *on average*

- However, with the logarithm transformation, we can actually say a little more...
- An equivalent way to express the model is*

$$Y = e^{\beta_0 + \beta_1 X + \epsilon} = e^{\beta_0} \times e^{\beta_1 X} \times e^{\epsilon}$$

and thus

$$E(Y|X) = e^{\beta_0} \times e^{\beta_1 X} \times E(e^{\epsilon}|X)$$

*since $e^{\ln(y)} = y$

- *Note: for* $\epsilon \sim \mathcal{N}(0, \sigma^2)$, *it can be shown that*

$$E(e^{\epsilon}|X) = E(e^{\epsilon}) = e^{\sigma^2/2}$$

Transformations of Y

- We can then compare the ratio of $E(Y|X = x + 1)$ to $E(Y|X = x)$:

$$\frac{E(Y|X = x + 1)}{E(Y|X = x)} = \frac{e^{\beta_0} \times e^{\beta_1(x+1)} \times E(e^{\epsilon})}{e^{\beta_0} \times e^{\beta_1(x)} \times E(e^{\epsilon})} = \frac{e^{\beta_1(x+1)}}{e^{\beta_1 x}} = e^{\beta_1}$$

- Thus, $\exp(\beta_1)$ represents the ratio of the mean of Y when $X = x + 1$ in comparison to $X = x$.
- We can interpret $\exp(\beta_1)$ as the multiplicative effect of X on the mean of Y: increasing X by one unit causes Y to increase by a factor of $\exp(\beta_1)$, on average.

Non-linear effects

Example: quadratic relation

Table of contents

Interactions between explanatory variables

- Interactions are very important concept in data science.
- An interaction implies 3 variables:
  - The two explanatory variables which interact with one another (ex: X_1 and X_2)
  - The variable Y which is affected by the interaction between X_1 and X_2.
- We say that X_1 and X_2 interact on Y when the effect of X_1 on Y depends on the value of X_2, and vice-versa.
- Ex: suppose that the variables age and sex interact on the salary:
  - The effect of age on salary depends on the person's sex. So the effect of age on salary is not the same between women and men.
  - Vice versa: the effect of sex on salary depends on the age of the person. So the difference in salary between men and women is different depending on age.

Caution: interaction does not imply correlation!

- Just because there is an interaction between two predictor variables X_1 and X_2, it does not mean that they're correlated!
  - an interaction between X_1 and X_2 means that the effect of each variable on the outcome Y depends on the value of the other
- Ex: age and sex can interact together on salary... However, age and sex are completely independent from one another!
- Recall that a correlation occurs between 2 variables, an interaction requires a 3rd variable.

Illustration of interaction vs. variable effect

- Y = Salary, X_1 = Age (3 categories), X_2 = Sex

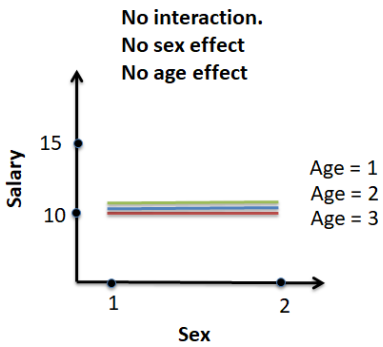
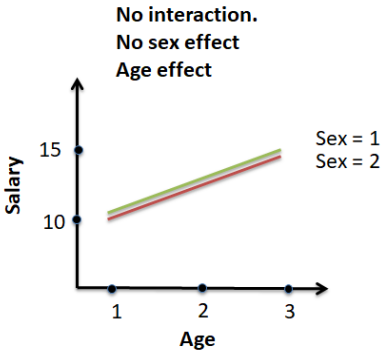


Illustration of interaction vs. variable effect

- Y = Salary, X_1 = Age (3 categories), X_2 = Sex



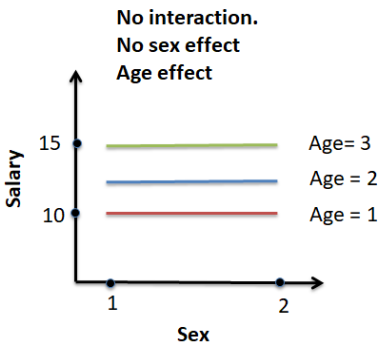
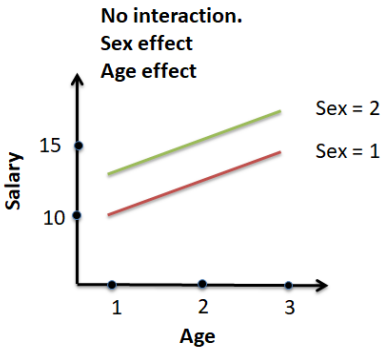


Illustration of interaction vs. variable effect

- Y = Salary, X_1 = Age (3 categories), X_2 = Sex



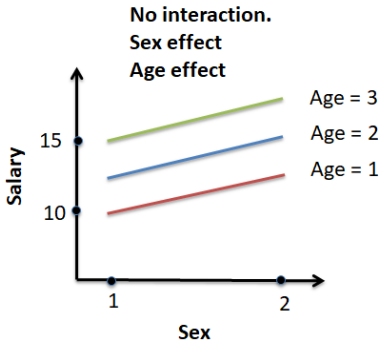
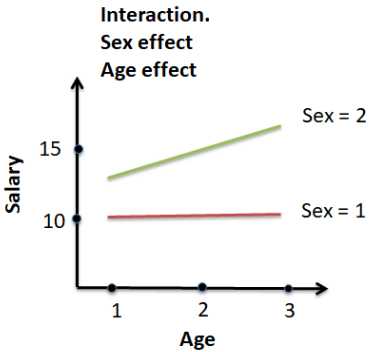


Illustration of interaction vs. variable effect

- Y = Salary, X_1 = Age (3 categories), X_2 = Sex



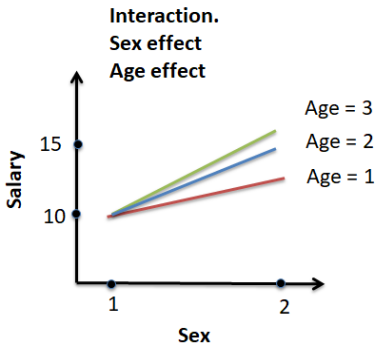


Illustration of interaction vs. variable effect

- The interaction between sex and age implies that:
  - the relationship between salary and age is different for men and women; i.e. the effect of age on salary depends on the sex of the person
  - the relationship between salary and sex is different for each age group; i.e. the effect of sex on salary depends on the age of the person

Interaction modelling

- There are several ways to model the interaction between two variables X_1 and X_2. The simplest and most common way involves creating a new variable equal to their product and adding it to a model already containing X_1 and X_2:

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

- To test whether the interaction is significant, you only need to test the significance of the parameter β_3:

$$H_0 : \beta_3 = 0 \quad \text{vs.} \quad H_1 : \beta_3 \neq 0$$

Interaction between a continuous and a binary variable

- The case of an interaction between a continuous and a binary variable is useful for visualizing and understanding the notion of interaction.
- For simplicity, we'll assume that there are only two predictor variables in the model, which interact with each other.
  - It's possible to have other additional variables in a model besides the two that interact. (We'll see an example of this later, where we'll include other variables in the model.)
- To illustrate this, recall the fixation-intention to buy example; we'll focus on the variables fixation and sex only.

Interaction between a continuous and a binary variable

- The main effects model, i.e., the model without the interaction, is given by:

$$Y = \beta_0 + \beta_1 sex + \beta_2 fixation + \epsilon$$

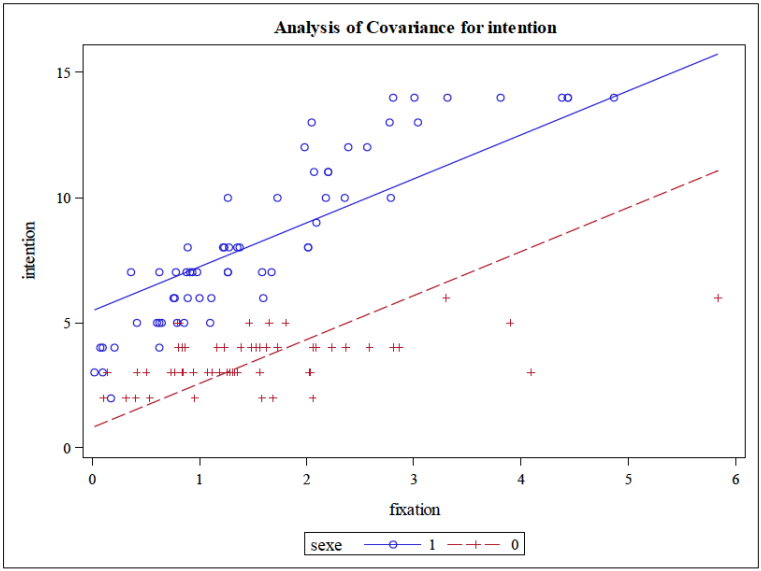
- We can split up the model based on the level of the variable sex:

$$\begin{aligned} E(Y|male, fixation) &= E(intention|sex = 0, fixation) \\ &= \beta_0 + \beta_2 fixation \\ E(Y|female, fixation) &= E(intention|sex = 1, fixation) \\ &= (\beta_0 + \beta_1) + \beta_2 fixation \end{aligned}$$

- $\Rightarrow$ this model assumes the effect of fixation is the same, and only the intercept changes according to the level sex.

Interaction between a continuous and a binary variable

Visually, we have two parallel lines



Interaction between a continuous and a binary variable

- In this main effects model, we assume that the effect of the continuous variable (fixation) is the same for the two values of the binary variable (sex).
  - Here, this means that the effect of `fixation` on `intention` is the same for both sex groups $\Rightarrow$ same slope β_2
- Likewise, the effect of the binary variable is assumed to be the same for all possible values of the continuous variable.
  - Here, this means that the effect of `sex` on `intention` is the same for all possible values of `fixation`. The model implies that for any value of `fixation`, the average difference in the intention to buy between females and males is β_1.
- We can see this on the graph, as the difference between the lines, which represents the effect of `sex`, is the same for all values of `fixation`; the lines are parallel. This model assumes that there is no interaction between `fixation` and `sex`.

Interaction between a continuous and a binary variable

- The model with the interaction is:

$$Y = \beta_0 + \beta_1 Sex + \beta_2 Fixation + \beta_3 Sex * Fixation + \epsilon$$

- We can decompose the model into two parts according to the value of sex:

$$\begin{aligned} E(Y|male, fixation) &= E(intention|sex = 0, fixation) \\ &= \beta_0 + \beta_2 fixation \\ E(Y|female, fixation) &= E(intention|sex = 1, fixation) \\ &= (\beta_0 + \beta_1) + (\beta_2 + \beta_3) fixation \end{aligned}$$

- Thus, the assumed model allows for both the intercept and the slope to differ for the two levels of sex.
- The parameter β_1 is the difference in the intercepts while β_3 (the parameter for the interaction term) is the difference in slopes.

Interaction between a continuous and a binary variable

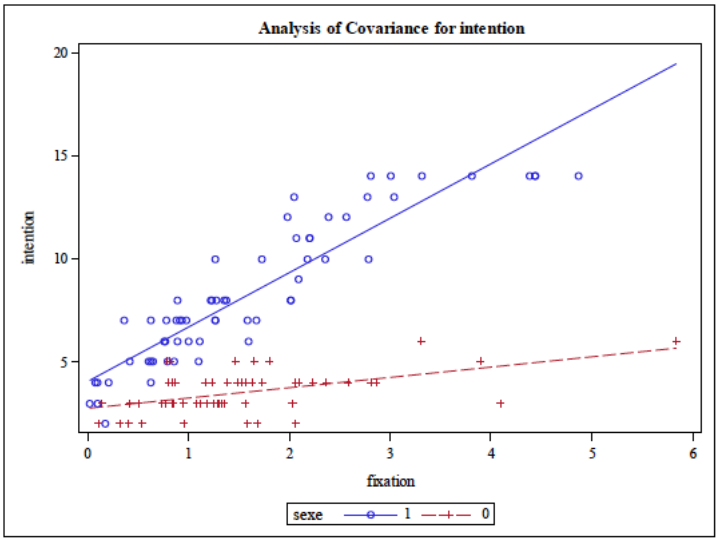
- If we want to test whether the slopes are different, that is, if the effect of `fixation` is different according to the value of `sex`, we only need to test whether the parameter β_3 is significantly different from 0.
- That is, testing whether there's a significant interaction between the two variables involves testing

$$H_0 : \beta_3 = 0 \quad \text{vs.} \quad H_1 : \beta_3 \neq 0$$

- If we reject H_0, then there is a significant interaction between the two variables.
  - If there's a significant interaction, it means that the effect of `fixation` on `intention` depends on `sex`, and vice versa...

Interaction between a continuous and a binary variable

Visually: two non-parallel lines



Interaction between a continuous and a binary variable

$$Y = \beta_0 = \beta_1 sex + \beta_2 fix + \beta_3 sex * fix + \epsilon$$

- In the interaction model, the effect of fixation on intention depends on sex:
  - for a male (sex=0), for each increase of 1 for `fixation`, `intention` increases on average by β_2
  - for females (sex=1), for each increase of 1 for `fixation`, `intention` increases on average by $\beta_2 + \beta_3$
  - the difference between these two effects is precisely the interaction coefficient β_3!
- But what about the effect of `sex`?

Interaction between a continuous and a binary variable

- In fact, an interaction goes in both directions...
- We saw that the effect of `fixation` is different for `sex=0` and `sex=1`. There were only two effects since `sex` is a binary variable.
- Just as the effect of `fixation` depends on the value of `sex`, likewise the effect of `sex` will depend on the value of `fixation`.
  - But for the effect of sex as a function of fixation, there are an infinite number of possible values for `fixation`.
- Visually, the effect of `sex` is the difference between the two lines on the plot.
  - Since the two lines are not parallel, the effect of `sex` changes as a function of `fixation`.

Interaction between a continuous and a binary variable

- For the model

$$Y = \beta_0 + \beta_1 sex + \beta_2 fixation + \beta_3 sex * fixation + \epsilon$$

the effect of the variable sex is

$$\begin{aligned} & E(Y|sex = 1, fixtaion) - E(Y|sex = 0, fixation) \\ & = \{(\beta_0 + \beta_1) + (\beta_2 + \beta_3) fixation)\} - \{\beta_0 + \beta_2 fixation\} \\ & = \beta_1 + \beta_3 fixation \end{aligned}$$

- Thus, the effect of `sex` depends on the value of `fixation`
- The parameter for `sex` (β_1) represents the effect of `sex` when `fixation=0`:

$$\beta_1 = E(Y|sex = 1, fixtaion = 0) - E(Y|sex = 0, fixation = 0)$$

But this is meaningless since `fixation=0` is not a possible value in this example.

Interaction between a continuous and a binary variable

Going a little further with the parameter interpretations...

- The model:

$$E(Y|sex, fixation) = \beta_0 + \beta_1\, sex + \beta_2\, fixation + \beta_3\, sex * fixation$$

- The parameters:
  - β_0:

$$E(Y|male, fixation = 0) = \beta_0$$

$=$ mean intention for males when fixation$=0$ (meaningless here!)

  - β_1:

$$E(Y|female, fixation = 0) - E(intention|male, fixation = 0) = \beta_1$$

$=$ effect of sex (i.e. difference in the mean intention for female vs. male) when fixation$=0$ (meaningless here!)

Interaction between a continuous and a binary variable

Going a little further with the parameter interpretations (continued)...

- β_2:

$$E(Y|male,\ fixation = x+1) - E(Y|male\ fixation = x) = \beta_2$$

  = effect of fixation for males

- β_3:

$$\Big[E(Y|female,\ fixation = x+1) - E(Y|female\ fixation = x)\Big] - \Big[E(Y|male,\ fixation = x+1) - E(Y|male\ fixation = x)\Big] = (\beta_2 + \beta_3) - \beta_2 = \beta_3$$

  = difference in the effect of fixation for females $(\beta_2 + \beta_3)$ vs. males (β_2)

Review of interactions: what you should know

- It's important to understand what we mean by "interaction"...
- You should understand how and why the interpretation of the model coefficients change when including an interaction term in the model

Interactions between two variables when other explanatory variables are in the model

- Note that in the previous examples, we only considered models with the two individual variables along with their interaction (sex, fixation, sex*fixation).
- In practice, there will generally be other variables included in the model as well.
- The effects of the variables that are directly included (without interactions) in the model can be interpreted as usual (e.g., for every one unit increase in X, Y will increase/decrease, on average, by... , holding all other variables constant).
- To correctly interpret the effects of the variables involved in the interaction, we need to consider the effect of each as a function of the other, but with the additional remark *holding all other variables constant*.

Higher-order interaction

- The preceding examples showed second-order interactions; that is, interaction between two variables.
- In theory, we could consider an interaction between any number of variables. However, in practice we rarely go higher than third-order because it quickly becomes difficult to interpret the effects. Additionally, estimating an interaction between several variables requires a very large sample size.
- The basic principle is still the same. To create an interaction of a given order between several variables, we also need to include all the lower-order terms between the variables included in the higher-order interaction term.
- Next, we interpret the variable effects while fixing the values of all the other variables in the interaction term.

Final remarks on interactions

- Always be sure to start by fitting a model without interaction (simple model) before proceeding to fit a model with interaction.
- Often, the variables on which we study interactions are chosen based on prior knowledge of the phenomenon we're studying. In other cases, the research question itself requires looking at specific interaction terms.
- As we saw, when there is an interaction, we cannot easily interpret the individual parameters in the model. We have to be careful and proceed as we did in the examples, i.e. by splitting up the model by fixing the values of certain variables.

Final remarks on interactions

- It's usually preferable to include all possible interactions of order less than the maximum order interaction included in the model.
- Don't remove a lower-order term, even if it's not significant. The lower-order terms are needed for proper inference!
- One exception is when we're developing a predictive model and we don't plan on testing individual variables. We would then include all the variables, as well as several (or even all) interaction terms and let an algorithm choose the best model.
- This algorithm will not necessarily respect the above rule, but this won't matter as we only care about predictive performance.

Interactions

Example: fixation-intention to buy

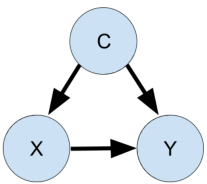
Table of contents

Multicollinearity

- In regression, the explanatory variables X are also referred to as the independent variables.
- However, in practice, these variables are rarely independent among themselves.
- We'll now see the effect that this kind of correlation amongst explanatory variables can have...

Confounding variable

- Once such example is the case of a confounder or confounding variable:
  - When two explanatory variables are correlated, it's possible that one variable is a confounder of the other variable.
  - A confounder (or confounding variable) is a variable C that is associated with the repsonse variable Y and is also correlated with the explanatory variable X (which is of interest), that is, C has an effect on both X and Y.



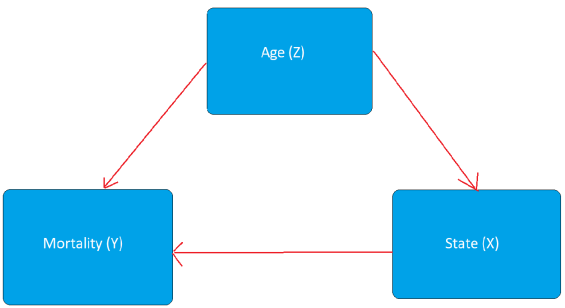
  - The confounding variable C can bias the observed relationship between X and Y, thus complicating the interpretations and conclusions of our analyses.
  - The concept of confounding is especially important in causal inference.

Example

- Suppose we're interested in comparing mortality rates in Florida and Michigan.
- After collecting data, we observe that the mortality rate in Florida is much higher than that in Michigan, so we conclude that Florida is a riskier place to live than Michigan. But is this really true...?
- Before drawing any conclusions, we need to investigate possible confounding variables, such as age:
  - Florida has a much higher proportion of older and retired people in comparison to Michigan
  - ... and older people obviously have a higher risk of dying in any given time period.
  - So, is it really the fact that someone lives in Florida that causes the mortality rate to be higher than in Michigan? Or is it rather the fact that people in Florida are generally older than people in Michigan?

Example

- In this example, age is a confounder: age is associated with the mortality rate (Y) AND it's also correlated with the state variable (X, Florida vs. Michigan).
- Thus, we must take into account the age variable in order to make proper conclusions regarding the relation between the state and mortality...



Stratification

How to handle confounding variables?

- One way of discovering and accounting for a possible confounder is through stratification. This kind of analysis is one of the first steps we can take to detect confounders.
  - In the Florida example, we could compare the mortality rate between Florida and Michigan separately for each age group (each age group consists of a stratum)
- Another way we can do this is by including age in a multiple regression model (along with other possible confounders):
  - In this case, the effect of X on Y, as measured through the multiple linear regression model, will account for all other variables. We'll be measuring the effect of X, adjusting for the other explanatory variables, which are possible confounders.
- Note that the concept of confounders is really only an issue in the context of observational studies. In a randomized or experimental design, the randomization process ensures balance across all other variables that could affect Y. In this case, we can make causal interpretations of the effect of X on Y without having to adjust for possible confounders.

Multicollinearity (or collinearity)

- Multicollinearity can be seen as an extreme case of a confounding variable.
- We say that two variables X_1 and X_2 are collinear if:
  - X_1 and X_2 are both correlated with Y
  - X_1 and X_2 are strongly correlated with each other - so much so that they contain essentially the same information.

Multicollinearity

- There could be multicollinearity between more than two variables... in the same way that there could be more than one confounding variable.
- In such a case multicollinearity (or simply collinearity) describes the case where one (or even several) explanatory variable is strongly correlated with a linear combination of other explanatory variables.
  - Ex: $X_p = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_{p-1} X_{p-1} + \delta$
- One potential harm of multicollinearity is a decrease in precision in parameter estimation, as it can increase the standard errors of the parameter estimates.
- Moreover, the fact that one or more variables are correlated, can render individual parameter interpretations difficult or even impossible.
  - Recall: in a multiple linear regression model

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

    we interpret the parameter β_j as the marginal (linear) effect of X_j on Y, holding all other variables constant.

Multicollinearity: illustration

- It's obvious that including the same variable twice in a model does not make sense...
  - Ex: including both temperature in both Celsius and Fahrenheit as predictors in the same model.
- Suppose that Y (say, height) is regressed on $X_1 =$ age, and the (fictional) linear equation between the two is

$$height = 20 + 3 \times age$$

- Now, suppose that $X_2 = age2$, an exact copy of age and that we try to fit the model:

$$height = \beta_0 + \beta_1 age + \beta_2 age2$$

Multicollinearity: illustration

- Note that there is no unique solution for this model. For example, all of these models are equivalent:

$$\begin{aligned} height &= 20 + 3 \times age + 0 \times age2, \\ height &= 20 + 0 \times age + 3 \times age2 \\ height &= 20 + 1 \times age + 2 \times age2 \\ height &= 20 + 1.5 \times age + 1.5 \times age2 \end{aligned}$$

- All of these models will give exactly the same predictions for height...
- In fact, any model such that $\beta_1 + \beta_2 = 3$ will be equivalent since $age = age2$!

Multicollinearity: illustration

- Here, we saw an extreme case where the variables *age* and *age*2 are perfectly collinear: they contain exactly the same information since they're in fact identical
- When this happens, most statistical softwares will recognize that the variables are equal (or perfectly linearly dependent). In this case, the software may either carry out estimation using only one of the two variables, or simply produce an error...
- The variables do not have to be perfectly collinear to cause estimation issues. However, in these cases, the statistical software may not recognize the collinearity - that's why it's important to look for certain signs...

Multicollinearity: another illustration

- Suppose we have 3 explanatory variables X_1, X_2, X_3 and that there is collinearity amongst the explanatory variables, for example

$$X_3 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \delta$$

(where α are the regression coefficients and δ is a random error term)

- The model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

is then equivalent to

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3(\alpha_0 + \alpha_1 X_1 + \alpha_2 + \delta) + \epsilon \\ &= (\beta_0 + \alpha_0) + (\beta_1 + \alpha_1) X_1 + (\beta_2 + \alpha_2) X_2 + (\epsilon + \beta_3 \delta) \\ &= \beta_0^* + \beta_1^* X_1 + \beta_2^* X_2 + \epsilon^* \end{aligned}$$

Multicollinearity: another illustration

- If we fit the model with all three variables X_1, X_2, X_3, we'll run into some difficulties...
  - If $\delta = 0$, then X_3 is an exact linear combination of X_1 and X_2 and we will not be able to fit the model that includes all three predictors. (Ex: In R, the estimate and corresponding statistics for β_3 will be set to NA, since R recognizes that X_3 is exactly a combination of X_1 and X_2).
  - If δ is not identically 0 but rather a random error term, the correlation between the explanatory variables X_1, X_2, X_3 will still have an adverse impact on the estimation...
  - One way to see this is that the variable X_3 is redundant in the model since X_3 can be explained by X_1 and X_2, and this causes problems in the model.

Effect of collinearity

- Generally, collinearity or strong correlation between the explanatory variables (e.g. X_1, X_2, X_3) will have the following effects:
  - The estimates of the regression coefficients change drastically between the simple linear regression models (one variable at a time) and the multiple linear regression model (including all explanatory variables, e.g. X_1, X_2, X_3)
  - The standard errors of the estimated coefficients in the multiple linear regression model will be very high, since β cannot be precisely estimated.
  - The confidence intervals for these coefficients will be very wide

Effect of confounders on the model coefficients

- Since collinearity can be seen as an extreme case of a confounding variable, the problems encountered for confounders are similar, but are moderated to some extent:
  - Moderate change in the estimated values of β between between the simple and multiple regression models
  - Moderate increase in the standard errors of the estimated coefficients

Multicollinearity

Example: fictional example height-age

How to detect collinearity (or confounders)?

- The first thing to do is to look at the correlations between the X variables as well as between each X variable and the Y variable.
  - The collinear variables (or confounding variables) will be correlated amongst themselves, as well as with Y
- Collinearity can also be found by looking at the difference between the simple regression results (one variable at a time) and a multiple regression model
  - If the variables are "independent", we should not see much of a difference between the coefficients from the simple and multiple regression models.
  - If there's collinearity, the changes will be more noticeable, and the standard errors will increase by quite a bit.
  - If there's a confounding variable present, these changes will be weaker.
- It's very important to start by fitting simple linear regression models for each predictor variables one at a time, before fitting a more complex model

How to detect collinearity (or confounders)?

- When more than 2 variables are collinear, it is more difficult to detect in the way we previously explored...
- In fact, one explanatory variable could be strongly correlated with a linear combination of the other variables, even though the individual correlations between the variables are not high.
  - Ex: If there are 10 explanatory variables $X_1, \ldots, X_{10}$ and X_{10} is correlated with a linear combination of the other variables, e.g.

    $$X_{10} = \alpha_0 + \alpha_1 X_1 + \cdots \alpha_9 X_9 + \delta$$

    the correlation between X_{10} and each of the variables $X_1, \ldots, X_9$ individually will not necessary be high...
    (α are the regression coefficients and δ is the random error term)
- Another tool we can use is the variance inflation factor (VIF).

Multicollinearity: the VIF

- For a given explanatory variable X_j , the VIF is defined as:

$$VIF(j) = \frac{1}{1 - R^2(j)}$$

where $R^2(j)$ is the R^2 of the model obtained by regressing X_j on all the other explanatory variables. More precisely, we get the R^2 for the model:

$$X_j = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_{j-1} X_{j-1} + \alpha_{j+1} X_{j+1} + \ldots + \alpha_p X_p + \delta$$

- $R^2(j)$ represents the proportion of the variance of X_j that is explained by all the other predictor variables.

Multicollinearity: the VIF

- The name itself, VIF, reflects the fact that multicollinearity causes an *inflation* of the variances of the estimators $\hat{\beta}_j$
- There is no generally accepted rule to declare that a model has a collinearity problem.
- Some have said that there's a multicollinearity problem when one of the $VIF(j)$ is > 10, while others use 4 or even 5 as the cutoff. These cutoffs are completely arbitrary.
- $VIF(j) > 4$ implies that $R^2(j) > 0.75$, meaning that 75% of the variability in X_j is explained by the other predictor variables.
- $VIF(j) > 5$ suggests that $R^2(j) > 0.8$ and $VIF(j) > 10$ suggests that $R^2(j) > 0.9$
- Note that the VIF is an individual measure. It does not tell us which particular variables are correlated with each other.

Multicollinearity

Example: fictional example

What to do if we have multicollinearity?

- This depends on the goal of the study.
- If the goal of the study is to develop a model to make predictions and we're not interested in estimating the parameters themselves, then we don't need to do anything.
- Multicollinearity is not a problem for the overall model.
- It's only a problem for the individual effects of the variables. Their joint effect is still present in the model, regardless of how it is separated amongst the individual effects in the model.

What to do if we have multicollinearity?

- If we're interested in obtaining precise estimates for the individual parameters, for example, to see how (and to what extent) the predictor variables explain the behaviour of Y, then it's a bit more complicated (inference).
- Collinearity only affects the variables that are strongly correlated with one another. In the previous example, if the goal of the study is to test the effects of X_4 and X_5, and the other three variables (which are correlated with one another) are only included for the sake of adjustment, then multicollinearity will not be a problem.
- In fact, we saw that the estimates for X_4 and X_5 were not affected by multicollinearity in the three remaining variables.
- However, if the goal of the study concerns one (or several) of the variables implicated in the multicollinearity, then there is really no entirely satisfactory solution to the problem.

Some options to handle multicollinearity

1. Obtain more data. This can help to reduce the effects of multicollinearity if it's specific to the sample.
2. Create a new variable that consists of a combination of the variables showing multicollinearity. Here, we get rid of the individual variables and rather use a composite score (ex. see R code).
3. Remove one (or more) of the multicollinear variables. You need to be careful when doing this, since you could end up with a misspecified model.

Whatever the method, it's important to understand that it can be very difficult (and sometimes impossible) to isolate the individual effect of a predictor variable when it's involved in multicollinearity with other predictors.

Confounders

- When we detect strong (but not extreme) correlations between the X variables and we also see moderate changes in the coefficients between the simple and multiple regression models, we usually have a confounding variable.
- In this case, it's very important to keep all the variables in the model and to take this into account when interpreting the results.
  - Ex: The state (Florida vs. Michigan) has a significant effect on the mortality rate (Y), but this effect disappears when we fit a model that includes state AND age.
  - Ex: Coffee consumption (X) has a significant effect on the appearance of a lung tumour (Y); however, this effect disappears when adjusting for smoking status. Here, the variable `smoker` is a confounder for the effect of `coffee` on the appearance on a lung tumour.

Summary: how to detect collinearity

In practice, it's a good idea to consider the following steps...

1. Calculate the correlation between each explanatory variables X and Y, as well as the correlation between the explanatory variables X themselves.
2. Fit simple linear regression models with each of the variables X
3. Fit a multiple linear regression model and evaluate the changes in the estimate regression coefficients as well as the corresponding standard errors in comparison to the simple linear regression models
4. If a problem is detected in the previous steps, this could be further confirmed by calculating the VIFs

Table of contents

Summary: carrying out a regression analysis

1. **Look at descriptive statistics for all the variables**
   - mean, median, quantiles, standard deviations for the continuous variables
   - frequency tables for categorical variables, histograms/boxplots for continuous variables
   - correlations between continuous variables
   - are there any missing values? outliers? errors?
   - etc... always report interesting findings
2. **Correct or re-code variables when necessary**
   - ex: erroneous values can either be deleted or corrected
   - ex: outliers – understand the impact of outlier values, and decide how to handle
   - ex: a variable with a large variability can be transformed onto the log-scale to reduce the variance
   - ex: for categorical variables, if there are levels with few observations this can perhaps be merged into another group/level
   - always report any modifications / transformations done to the data

Summary: carrying out a regression analysis

3. **Produce graphs to visually inspect the relation between Y and each explanatory variable X**
   - for a continuous X: scatterplots of Y vs. X
   - for a categorical X: side-by-side boxplots of Y for each category of X
4. **Fit initial regression models one variable at a time, taking into account the results from step 3**
   - Ex: if a quadratic pattern (or other) was clearly shown in the scatterplot in step 3, then consider a model which includes a quadratic (or other) term as an initial model. (It's useless to consider a linear model, for example, if there's clearly a quadratic relation between the variables!)
   - Ex: based on what is observed for ordinal categorical variables in step 3 (boxplots), you could decide to treat the variable as continuous in the model (if the boxplots showed a relatively linear pattern across levels, you could treat the ordinal variable as continuous).

Summary: carrying out a regression analysis

5. **Fit a multiple regression model including all the variables / interactions necessary**
   - Compare the results for the multiple regression model with the simple models from step 4, taking into account the correlation information obtained in step 1, in order to detect potential issues of multicollinearity
   - If you suspect that there is a problem with collinearity, calculate the VIFs. Take possible measures (depending on the objective of the analysis to begin with) to resolve the collinearity issues (e.g. remove variables, consider linear combinations of predictors, etc...)
   - Do not systematically test all possible interactions... rather, base yourself on the context or domain knowledge, or on results from the descriptive analysis.
   - Always remember – the model specification must allow you to answer the research question!
6. **Check the model assumptions**
   - Carry out a residual analysis
   - Based on the assessment of the model assumptions, potentially modify the model. (Ex: add a non-linear term, consider a log-transformation of Y for issues with non-normality or non-constant variance, ...)

Summary: carrying out a regression analysis

7. **Answer the research question that motivated the analysis**
   - Interpret the regression coefficients from the final model
   - Carry out relevant tests
   - Etc!

Some important remarks

All models are wrong, but some models are useful.
- G. E. P. Box

A fundamental question that always arises in regression analysis: how should the model be specified – that is, what covariates should be included, and *how* should they be included?

- It depends on the underlying objective of the analysis!
  - explanatory (inference): confirmatory / exploration – understanding, quantifying and assessing the relations between the covariates $X_1, \ldots, X_p$ and the response variable Y
  - prediction: predict new outcomes – predictions of Y given set of $X_1, \ldots, X_p$

Some important remarks

- When the underlying goal is inference:
  - First and foremost, the model should allow you to address the research question that motivated the analysis to begin with!
  - Generally, we want to find a model which provides a good description of the random phenomena we are studying. This requires careful consideration of the context and domain-specific knowledge, and the data itself.
  - Causal inference: with observational data, need to adjust for all possible confounders in order to establish causal relations. (In fact, causal inference relies on a set of identifiability assumptions, which go way beyond the scope of this course)

Some important remarks

- Typically, several model refinements are considered in the analysis, e.g., model diagnostics may lead to variable transformations, etc.
- CAUTION: inference after model selection
  - Ignoring the model selection process (generally) invalidates "traditional" statistical inference – the model selection process leads to a final model which itself is random, and this random aspect is not accounted for in most classical statistical inference. Typically, uncertainty measures are underestimated as they do not reflect the model selection process.
  - There are methods for adequately reflecting model selection in post-selection inference.
- Good practices:
  - Think very carefully about how the model should be specified before even beginning the analysis.
  - Sometimes, a more reasonable approach may be to report and discuss the results from several models.
  - Be **transparent** in the entire modeling process – your analysis should always be reproducible!

Some important remarks

- Some references and further readings:
  - Wakefield, J., *Bayesian and Frequentist Regression methods*. Springer, New York, 2013. : Sections 4.7 – 4.11
  - Leeb, H., and B. M. Pötscher. "Model Selection and Inference: Facts and Fiction." Econometric Theory, vol. 21, no. 1, 2005, pp. 21–59.
  - Berk, R., et al. "Valid post-selection inference." The Annals of Statistics, vol. 41, no. 2, 2013, pp. 802–37.