Lecture 4 | 17.03.2026

Multiple regression model

with categorical and mixed predictor variables

Overview: Linear regression model

- Theoretical (population model) for a continuous dependent (random) variable $Y \in \mathbb{R}$ and independent covariates $\boldsymbol{X} \in \mathbb{R}^p$ where the intercept is included in the model (i.e., $X_1 = 1$ with probability one) is of the form

$$Y = \boldsymbol{X}^\top \beta + \varepsilon$$

Overview: Linear regression model

- Theoretical (population model) for a continuous dependent (random) variable $Y \in \mathbb{R}$ and independent covariates $\boldsymbol{X} \in \mathbb{R}^p$ where the intercept is included in the model (i.e., $X_1 = 1$ with probability one) is of the form

$$Y = \boldsymbol{X}^\top \boldsymbol{\beta} + \varepsilon$$

- More generally, for $Y \in \mathbb{R}$ and some $\boldsymbol{X} \in \mathbb{R}^q$ the **linear** regression model with unknown parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ can be also specified as

$$Y = \beta_1 t_1(\boldsymbol{X}) + \beta_2 t_2(\boldsymbol{X}) + \cdots + \beta_p t_p(\boldsymbol{X}) + \varepsilon$$

for the set of unknown parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ and some known transformation functions $t_j : \mathbb{R}^q \to \mathbb{R}$ for $j = 1, \ldots, p$, such that the transformations $t_1, \ldots, t_p$ do not depend on the unknown parameters

(for simplicity of the notation and the presentation of the methods, it is often assumed that $\boldsymbol{X} \in \mathbb{R}^p$)

Overview: Linear regression model

- Theoretical (population model) for a continuous dependent (random) variable $Y \in \mathbb{R}$ and independent covariates $\boldsymbol{X} \in \mathbb{R}^p$ where the intercept is included in the model (i.e., $X_1 = 1$ with probability one) is of the form

$$Y = \boldsymbol{X}^\top \boldsymbol{\beta} + \varepsilon$$

- More generally, for $Y \in \mathbb{R}$ and some $\boldsymbol{X} \in \mathbb{R}^q$ the **linear** regression model with unknown parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ can be also specified as

$$Y = \beta_1 t_1(\boldsymbol{X}) + \beta_2 t_2(\boldsymbol{X}) + \cdots + \beta_p t_p(\boldsymbol{X}) + \varepsilon$$

  for the set of unknown parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ and some known transformation functions $t_j : \mathbb{R}^q \to \mathbb{R}$ for $j = 1, \ldots, p$, such that the transformations $t_1, \ldots, t_p$ do not depend on the unknown parameters
  (for simplicity of the notation and the presentation of the methods, it is often assumed that $\boldsymbol{X} \in \mathbb{R}^p$)

- **Linearity** of the regression model refers to the linearity wrt. the unknown parameters $\beta_1, \ldots, \beta_p \in \mathbb{R}$; it does not specify anything about $\boldsymbol{X}$ (or $t_1, \ldots, t_p$)
- **Categorical covariates** (i.e., factors) represent (qualitative) information and the corresponding values, typically integers, do not have any numerical interpretation

Transformations of continuous covariates

- Let $Y \in \mathbb{R}$ and $\boldsymbol{X} \in \mathbb{R}^q$ and let a general (regression) model formulation be of the form

$$Y = \beta_1 t_1(\boldsymbol{X}) + \beta_2 t_2(\boldsymbol{X}) + \cdots + \beta_p t_p(\boldsymbol{X}) + \varepsilon$$

for some (known) transformation functions $t_1, \ldots, t_p$ (where $_j : \mathbb{R}^q \to \mathbb{R}$)

Transformations of continuous covariates

- Let $Y \in \mathbb{R}$ and $\boldsymbol{X} \in \mathbb{R}^q$ and let a general (regression) model formulation be of the form

$$Y = \beta_1 t_1(\boldsymbol{X}) + \beta_2 t_2(\boldsymbol{X}) + \cdots + \beta_p t_p(\boldsymbol{X}) + \varepsilon$$

for some (known) transformation functions $t_1, \ldots, t_p$ (where $_j : \mathbb{R}^q \to \mathbb{R}$)

What can be reasonable (practically useful) forms of transformations?

- What would be the consequnces for the model with $t_j(x) = 0$ for all $j = 1, \ldots, p$?
- What would it mean for the model to assume that $t_1(x) = 0$?
- For the model with the intercept parameter it is typically assumed, that $t_1(\cdot) \equiv 1$
- For practical ilustrations it often holds, that $t_{j+1}(\boldsymbol{x}) = x_j$, for $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$
- Very common transformations are of a linear type: $t_j(\boldsymbol{x}) = \mathbb{A}_j \boldsymbol{x} + \boldsymbol{c}_j$
- A specific form of such linear transformation is $t_{j+1}(\boldsymbol{x}) = a_j x_j + c_j$
- Many other options can be considered... *(flexibility, interpretation, efficiency, ...)*

Transformations of continuous covariates

- Let $Y \in \mathbb{R}$ and $\boldsymbol{X} \in \mathbb{R}^q$ and let a general (regression) model formulation be of the form

$$Y = \beta_1 t_1(\boldsymbol{X}) + \beta_2 t_2(\boldsymbol{X}) + \cdots + \beta_p t_p(\boldsymbol{X}) + \varepsilon$$

for some (known) transformation functions $t_1, \ldots, t_p$ (where $_j : \mathbb{R}^q \to \mathbb{R}$)

What can be reasonable (practically useful) forms of transformations?

- What would be the consequnces for the model with $t_j(x) = 0$ for all $j = 1, \ldots, p$?
- What would it mean for the model to assume that $t_1(x) = 0$?
- For the model with the intercept parameter it is typically assumed, that $t_1(\cdot) \equiv 1$
- For practical ilustrations it often holds, that $t_{j+1}(\boldsymbol{x}) = x_j$, for $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$
- Very common transformations are of a linear type: $t_j(\boldsymbol{x}) = \mathbb{A}_j \boldsymbol{x} + \boldsymbol{c}_j$
- A specific form of such linear transformation is $t_{j+1}(\boldsymbol{x}) = a_j x_j + c_j$
- Many other options can be considered... *(flexibility, interpretation, efficiency, ...)*

Example 1 (`mtcars`)

- Dependent variable $Y \in \mathbb{R}$ respresents a car's consumption efficiency [*mpg*]
- Two covariates: X_1 – the number of cylinders; X_2 – car weight [$\times 1000 lbs$]

Example 1 (mtcars)

- Dependent variable $Y \in \mathbb{R}$ respresents a car's consumption efficiency [*mpg*]
- Two covariates: X_1 – the number of cylinders; X_2 – car weight [$\times 1000 lbs$]

- **Model 1:** $Y = a + bX_1 + cX_2 + \varepsilon$ (models fitted in R)

```
> summary(lm(mpg ~ cyl + wt, data = mtcars))

             Estimate  Std.Error   t value    Pr(>|t|)
(Intercept)  39.6863     1.7150    23.141     < 2e-16 ***
cyl          -1.5078     0.4147    -3.636    0.001064 **
wt           -3.1910     0.7569    -4.216    0.000222 ***
```

Example 1 (mtcars)

- Dependent variable $Y \in \mathbb{R}$ respresents a car's consumption efficiency [*mpg*]
- Two covariates: X_1 – the number of cylinders; X_2 – car weight [$\times 1000 lbs$]

- **Model 1:** $Y = a + bX_1 + cX_2 + \varepsilon$ (models fitted in R)

```
> summary(lm(mpg ~ cyl + wt, data = mtcars))
```

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	39.6863	1.7150	23.141	< 2e-16 ***
cyl	-1.5078	0.4147	-3.636	0.001064 **
wt	-3.1910	0.7569	-4.216	0.000222 ***

- **Model 2:** $Y = a + b_1 \mathbb{I}_{\{X_1=6\}} + b_2 \mathbb{I}_{\{X_1=8\}} + cX_2 + \varepsilon$

```
> summary(lm(mpg ~ as.factor(cyl) + wt, data = mtcars))
```

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	33.9908	1.8878	18.006	< 2e-16 ***
as.factor(cyl)6	-4.2556	1.3861	-3.070	0.004718 **
as.factor(cyl)8	-6.0709	1.6523	-3.674	0.000999 ***
wt	-3.2056	0.7539	-4.252	0.000213 ***

Example 1 (mtcars)

- The average weight of a car – the sample mean of X_2's is (cca) 3.22
- The sample median of the weight is 3.325 (i.e., 3325 *pounds* $\equiv$ 1460 *kg*)

Example 1 (mtcars)

- The average weight of a car – the sample mean of X_2's is (cca) 3.22
- The sample median of the weight is 3.325 (i.e., 3325 *pounds* $\equiv$ 1460 *kg*)

- **Model 3:** $Y = a + b_1\mathbb{I}_{\{X_1=6\}} + b_2\mathbb{I}_{\{X_1=8\}} + c(X_2 - 3.2) + \varepsilon$

```
> summary(lm(mpg ~ as.factor(cyl) + I(wt - 3.2), data = mtcars))

                Estimate  Std.Error t value   Pr(>|t|)
(Intercept)      23.7328     1.0341  22.949    < 2e-16 ***
as.factor(cyl)6  -4.2556     1.3861  -3.070   0.004718 **
as.factor(cyl)8  -6.0709     1.6523  -3.674   0.000999 ***
I(wt - 3.2)      -3.2056     0.7539  -4.252   0.000213 ***
```

Example 1 (mtcars)

- The average weight of a car – the sample mean of X_2's is (cca) 3.22
- The sample median of the weight is 3.325 (i.e., 3325 *pounds* $\equiv$ 1460 *kg*)
- One kilogram is (roughly) 2.2 US pounds

- **Model 3:** $Y = a + b_1\mathbb{I}_{\{X_1=6\}} + b_2\mathbb{I}_{\{X_1=8\}} + c(X_2 - 3.2) + \varepsilon$

```
> summary(lm(mpg ~ as.factor(cyl) + I(wt - 3.2), data = mtcars))

                Estimate  Std.Error t value   Pr(>|t|)
(Intercept)      23.7328     1.0341  22.949    < 2e-16 ***
as.factor(cyl)6  -4.2556     1.3861  -3.070   0.004718 **
as.factor(cyl)8  -6.0709     1.6523  -3.674   0.000999 ***
I(wt - 3.2)      -3.2056     0.7539  -4.252   0.000213 ***
```

- **Model 4:** $Y = a + b_1\mathbb{I}_{\{X_1=6\}} + b_2\mathbb{I}_{\{X_1=8\}} + c(\frac{1000}{2.2}X_2 - 1460) + \varepsilon$

```
> summary(lm(mpg ~ as.factor(cyl) + I(1000 * wt / 2.2 - 1460), data = mtcars))

                          Estimate  Std.Error t value   Pr(>|t|)
(Intercept)              23.694364   1.040186  22.779    < 2e-16 ***
as.factor(cyl)6          -4.255582   1.386073  -3.070   0.004718 **
as.factor(cyl)8          -6.070860   1.652288  -3.674   0.000999 ***
I(1000 * wt/2.2 - 1460) -0.007052   0.001659  -4.252   0.000213 ***
```

Regression model with categorical covariates

- For a model $Y \sim X$ with just one binary variable (i.e., a two sample problem), there are infinitely many parametrizations that can be used to encode the information in X to model the conditional expectation $E[Y|X = x] = f(x)$ *(e.g., covariate X can take values ± 1, or $X \in \{0, 1\}$, or $X \in \{1, 2\}$, ...)*
- A particular choice of the model parametrization (i.e., which two values are taken by $X \in \mathbb{R}$) only effects the model interpretation... not the model itself!

Regression model with categorical covariates

- For a model $Y \sim X$ with just one binary variable (i.e., a two sample problem), there are infinitely many parametrizations that can be used to encode the information in X to model the conditional expectation $E[Y|X = x] = f(x)$ *(e.g., covariate X can take values ± 1, or $X \in \{0, 1\}$, or $X \in \{1, 2\}$, ...)*
- A particular choice of the model parametrization (i.e., which two values are taken by $X \in \mathbb{R}$) only effects the model interpretation... not the model itself!
- For a model $Y \sim X$ with one categorical variable (a general K sample problem), the most common approach for implementing categorical labels into a linear regression model is to use **dummy variables** (i.e., factorization into binary cases)
- At some point, the dummy variables can be seen as a partial adjustments of the model intercept parameter depending on a particular factor level (category)

Regression model with categorical covariates

- For a model $Y \sim X$ with just one binary variable (i.e., a two sample problem), there are infinitely many parametrizations that can be used to encode the information in X to model the conditional expectation $E[Y|X = x] = f(x)$ *(e.g., covariate X can take values ± 1, or $X \in \{0, 1\}$, or $X \in \{1, 2\}$, ...)*
- A particular choice of the model parametrization (i.e., which two values are taken by $X \in \mathbb{R}$) only effects the model interpretation... not the model itself!
- For a model $Y \sim X$ with one categorical variable (a general K sample problem), the most common approach for implementing categorical labels into a linear regression model is to use **dummy variables** (i.e., factorization into binary cases)
- At some point, the dummy variables can be seen as a partial adjustments of the model intercept parameter depending on a particular factor level (category)

Illustration

- The dependent (random) variable $Y \in \mathbb{R}$ is assumed to be continuous...
- Let $X_1 \in \mathbb{R}$ be discrete, taking only values in $\{v_1, \ldots, v_K\}$, for $v_k \in \mathbb{R}$ and $K \in \mathbb{N}$

Regression model with categorical covariates

- For a model $Y \sim X$ with just one binary variable (i.e., a two sample problem), there are infinitely many parametrizations that can be used to encode the information in X to model the conditional expectation $E[Y|X = x] = f(x)$ *(e.g., covariate X can take values ± 1, or $X \in \{0, 1\}$, or $X \in \{1, 2\}$, ...)*
- A particular choice of the model parametrization (i.e., which two values are taken by $X \in \mathbb{R}$) only effects the model interpretation... not the model itself!
- For a model $Y \sim X$ with one categorical variable (a general K sample problem), the most common approach for implementing categorical labels into a linear regression model is to use **dummy variables** (i.e., factorization into binary cases)
- At some point, the dummy variables can be seen as a partial adjustments of the model intercept parameter depending on a particular factor level (category)

Illustration

- The dependent (random) variable $Y \in \mathbb{R}$ is assumed to be continuous...
- Let $X_1 \in \mathbb{R}$ be discrete, taking only values in $\{v_1, \ldots, v_K\}$, for $v_k \in \mathbb{R}$ and $K \in \mathbb{N}$
- The goal is to find some reasonable linear function f (linear wrt. some unknown parameters) that will properly describe the relationship $Y \approx f(X_1)$ or, alternatively, the conditional expectation $E[Y|X_1] = f(X_1)$

$$f : \{v_1, \ldots, v_K\} \to \mathbb{R}$$

Regression model with categorical covariates

- For a model $Y \sim X$ with just one binary variable (i.e., a two sample problem), there are infinitely many parametrizations that can be used to encode the information in X to model the conditional expectation $E[Y|X = x] = f(x)$ *(e.g., covariate X can take values ± 1, or $X \in \{0, 1\}$, or $X \in \{1, 2\}$, ...)*
- A particular choice of the model parametrization (i.e., which two values are taken by $X \in \mathbb{R}$) only effects the model interpretation... not the model itself!
- For a model $Y \sim X$ with one categorical variable (a general K sample problem), the most common approach for implementing categorical labels into a linear regression model is to use **dummy variables** (i.e., factorization into binary cases)
- At some point, the dummy variables can be seen as a partial adjustments of the model intercept parameter depending on a particular factor level (category)

Illustration

- The dependent (random) variable $Y \in \mathbb{R}$ is assumed to be continuous...
- Let $X_1 \in \mathbb{R}$ be discrete, taking only values in $\{v_1, \ldots, v_K\}$, for $v_k \in \mathbb{R}$ and $K \in \mathbb{N}$
- The goal is to find some reasonable linear function f (linear wrt. some unknown parameters) that will properly describe the relationship $Y \approx f(X_1)$ or, alternatively, the conditional expectation $E[Y|X_1] = f(X_1)$

$$f : \{v_1, \ldots, v_K\} \to \mathbb{R}$$

- Adding another covariate $X_2 \in \mathbb{R}$ that is countinuous, the problem can be formulated as finding a parametric function $f : \{v_1, \ldots, v_K\} \times \mathbb{R} \to \mathbb{R}$

Dummy variables in a regression model

- **Dummy variables** for the categorical covariate X_1 can be defined as
  - $X_1^{\mathrm{D1}} = \mathbb{I}_{\{X_1 = v_1\}}, X_1^{\mathrm{D2}} = \mathbb{I}_{\{X_1 = v_2\}}, X_1^{\mathrm{D3}} = \mathbb{I}_{\{X_1 = v_3\}}, \ldots, X_1^{\mathrm{D}K} = \mathbb{I}_{\{X_1 = v_K\}}$
  - its clear, that each $X_1^{\mathrm{D1}}, X_1^{\mathrm{D2}}, \ldots, X_1^{\mathrm{D}K}$ can only take value zero or one *(therefore, the "**factorization into binary cases**" as mentioned before)*
  - the principle is analogous to a situation with the binary variable (which takes only two different values and just one dummy is needed)
  - but, also, analogous problems occur—the **over-parametrization** issues

Dummy variables in a regression model

- **Dummy variables** for the categorical covariate X_1 can be defined as
  - $X_1^{\mathrm{D}1} = \mathbb{I}_{\{X_1=v_1\}}, X_1^{\mathrm{D}2} = \mathbb{I}_{\{X_1=v_2\}}, X_1^{\mathrm{D}3} = \mathbb{I}_{\{X_1=v_3\}}, \ldots, X_1^{\mathrm{D}K} = \mathbb{I}_{\{X_1=v_K\}}$
  - its clear, that each $X_1^{\mathrm{D}1}, X_1^{\mathrm{D}2}, \ldots, X_1^{\mathrm{D}K}$ can only take value zero or one *(therefore, the "**factorization into binary cases**" as mentioned before)*
  - the principle is analogous to a situation with the binary variable (which takes only two different values and just one dummy is needed)
  - but, also, analogous problems occur—the **over-parametrization** issues
- The linear regression model with $X_1 \in \{v_1, \ldots, v_K\}$ can be expressed, using the dummy variables $X_1^{\mathrm{D}1}, \ldots, X_1^{\mathrm{D}K}$ as

$$Y = a + \beta_1 X_1^{\mathrm{D}1} + \cdots + \beta_K X_1^{\mathrm{D}K} + \varepsilon = a + \sum_{k=1}^{K} \beta_k X_1^{\mathrm{D}k} + \varepsilon \qquad (\text{❖})$$

but the meaning of the intercept $a \in \mathbb{R}$ parameter may not be clear now...

(note, that $E[Y|X_1^{\mathrm{D}1} = 0, \ldots, X_1^{\mathrm{D}K} = 0] = a$, but this implies that $X_1 \notin \{v_1, \ldots v_K\}$, which can not happened)

Dummy variables in a regression model

- **Dummy variables** for the categorical covariate X_1 can be defined as
  - $X_1^{\mathrm{D}1} = \mathbb{I}_{\{X_1=v_1\}}, X_1^{\mathrm{D}2} = \mathbb{I}_{\{X_1=v_2\}}, X_1^{\mathrm{D}3} = \mathbb{I}_{\{X_1=v_3\}}, \ldots, X_1^{\mathrm{D}K} = \mathbb{I}_{\{X_1=v_K\}}$
  - its clear, that each $X_1^{\mathrm{D}1}, X_1^{\mathrm{D}2}, \ldots, X_1^{\mathrm{D}K}$ can only take value zero or one *(therefore, the "**factorization into binary cases**" as mentioned before)*
  - the principle is analogous to a situation with the binary variable (which takes only two different values and just one dummy is needed)
  - but, also, analogous problems occur—the **over-parametrization** issues
- The linear regression model with $X_1 \in \{v_1, \ldots, v_K\}$ can be expressed, using the dummy variables $X_1^{\mathrm{D}1}, \ldots, X_1^{\mathrm{D}K}$ as

$$Y = a + \beta_1 X_1^{\mathrm{D}1} + \cdots + \beta_K X_1^{\mathrm{D}K} + \varepsilon = a + \sum_{k=1}^{K} \beta_k X_1^{\mathrm{D}k} + \varepsilon \qquad (\text{❖})$$

but the meaning of the intercept $a \in \mathbb{R}$ parameter may not be clear now...

(note, that $E[Y|X_1^{\mathrm{D}1} = 0, \ldots, X_1^{\mathrm{D}K} = 0] = a$, but this implies that $X_1 \notin \{v_1, \ldots v_K\}$, which can not happened)

- Moreover, there are $1 + K$ "intercept" parameters in the model but only K different sub-populations that can be used for the estimation process

Over-parametrization problem

Using Model (❖), it is clear that the whole (unknown) population is split into $K \in \mathbb{N}$ (disjoint) subpopulations according to the value of $X_1 \in \{v_1, \ldots, v_K\}$ – having $K \in \mathbb{N}$ different groups for which we can estimate the unknown means – **over-parametrization** (but there are $K + 1$ unknown parameters all together explicitly included in (❖))

Different parametrizations for dummy variables

- the intercept parameter $a \in \mathbb{R}$ is used instead of β_1, thus $\beta_1 = 0$
  (the reference category $X_1 = v_1$ is modeled by the intercept parameter)
- the reference category can be also selected differently, for instance, $\beta_K = 0$
  (this reflects the situation where the intercept parameter models the mean of the sub-population v_K)
- thus, the over-parametrization is solved by adding one extra equation...
  (with a restriction $\sum_{k=1}^{K} \beta_k = 0$, the intercept parameter stands for the overall mean)
- and many other parametrizations can be used of course ...
  (but the main idea is to make sure that the intercept parameter $a \in \mathbb{R}$ has a reasonable interpretation)

Transformations of categorical covariates (?)

- Typically, it does not make that much sense... categorical covariates carry qualitative information rather than quantitative...
- Consider a categorical covariate $X_1 \in \{v_1, \ldots, v_K\}$ and some pre-specified transformation $t : \{v_1, \ldots, v_k\} \to \{w_1, \ldots, w_K\}$ (i.e., another labels)

Transformations of categorical covariates (?)

- Typically, it does not make that much sense... categorical covariates carry qualitative information rather than quantitative...
- Consider a categorical covariate $X_1 \in \{v_1, \ldots, v_K\}$ and some pre-specified transformation $t : \{v_1, \ldots, v_k\} \to \{w_1, \ldots, w_K\}$ (i.e., another labels)

- It is quite obvious, that two models

$$Y_i = a = \beta_1 \mathbb{I}_{\{X_i = v_1\}} + \beta_2 \mathbb{I}_{\{X_i = v_2\}} + \ldots, + \beta_K \mathbb{I}_{\{X_i = v_K\}} + \varepsilon_i$$

and

$$Y_i = a = \beta_1 \mathbb{I}_{\{t(X_i) = w_1\}} + \beta_2 \mathbb{I}_{\{t(X_i) = w_2\}} + \ldots, + \beta_K \mathbb{I}_{\{t(X_i) = w_K\}} + \varepsilon_i$$

are, in terms of the unknown parameters, the corresponding parameter estimates, and the resulting interpretaion, equivalent
()the same model/design matrix $\mathbb{X}$ is used for both models)

Multiple categorical covariates

- Dependent variable $Y \in \mathbb{R}$ respresents a car's consumption efficiency [*mpg*]
- Covariates: X_1 – the number of cylinders; X_2 – automatic transmission
- **Model**

$$Y = a + \beta_1 \mathbb{I}_{\{X_1=4\}} + \beta_2 \mathbb{I}_{\{X_1=6\}} + \beta_3 \mathbb{I}_{\{X_1=8\}} + \gamma_1 \mathbb{I}_{\{X_2=a\}} + \gamma_2 \mathbb{I}_{\{X_2=m\}} + \varepsilon$$

Multiple categorical covariates

- Dependent variable $Y \in \mathbb{R}$ respresents a car's consumption efficiency [*mpg*]
- Covariates: X_1 – the number of cylinders; X_2 – automatic transmission
- **Model**

$$Y = a + \beta_1 \mathbb{I}_{\{X_1=4\}} + \beta_2 \mathbb{I}_{\{X_1=6\}} + \beta_3 \mathbb{I}_{\{X_1=8\}} + \gamma_1 \mathbb{I}_{\{X_2=a\}} + \gamma_2 \mathbb{I}_{\{X_2=m\}} + \varepsilon$$

$\mathcal{P}$	transmission	
cylinders	automatic	manual
4 cylinders		
6 cylinders		
8 cylinders		

- six unknown parameters, $a, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2 \in \mathbb{R}$
- also six disjoint populations *(groups within the population $\mathcal{P}$)*

Multiple categorical covariates

- Dependent variable $Y \in \mathbb{R}$ respresents a car's consumption efficiency [*mpg*]
- Covariates: X_1 – the number of cylinders; X_2 – automatic transmission
- **Model**

$$Y = a + \beta_1 \mathbb{I}_{\{X_1=4\}} + \beta_2 \mathbb{I}_{\{X_1=6\}} + \beta_3 \mathbb{I}_{\{X_1=8\}} + \gamma_1 \mathbb{I}_{\{X_2=a\}} + \gamma_2 \mathbb{I}_{\{X_2=m\}} + \varepsilon$$

$\mathcal{P}$	transmission	
cylinders	automatic	manual
4 cylinders	$a + \beta_1 + \gamma_1$	$a + \beta_1 + \gamma_2$
6 cylinders	$a + \beta_2 + \gamma_1$	$a + \beta_2 + \gamma_2$
8 cylinders	$a + \beta_3 + \gamma_1$	$a + \beta_3 + \gamma_2$

- six unknown parameters, $a, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2 \in \mathbb{R}$
- also six disjoint populations *(groups within the population $\mathcal{P}$)*
- **But...** not well defined!

Multiple categorical covariates

- Dependent variable $Y \in \mathbb{R}$ respresents a car's consumption efficiency [*mpg*]
- Covariates: X_1 – the number of cylinders; X_2 – automatic transmission
- **Model**

$$Y = a + \beta_1 \mathbb{I}_{\{X_1=4\}} + \beta_2 \mathbb{I}_{\{X_1=6\}} + \beta_3 \mathbb{I}_{\{X_1=8\}} + \gamma_1 \mathbb{I}_{\{X_2=a\}} + \gamma_2 \mathbb{I}_{\{X_2=m\}} + \varepsilon$$

$\mathcal{P}$	transmission	
cylinders	automatic	manual
4 cylinders	$a + \beta_1 + \gamma_1$	$a + \beta_1 + \gamma_2$
6 cylinders	$a + \beta_2 + \gamma_1$	$a + \beta_2 + \gamma_2$
8 cylinders	$a + \beta_3 + \gamma_1$	$a + \beta_3 + \gamma_2$

- six unknown parameters, $a, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2 \in \mathbb{R}$
- also six disjoint populations *(groups within the population $\mathcal{P}$)*
- **But...** not well defined!

- There are formaly six unknown parameters in the model but, mathematically, they only specify four different (unique) means

Multiple categorical covariates

- Dependent variable $Y \in \mathbb{R}$ respresents a car's consumption efficiency [*mpg*]
- Covariates: X_1 – the number of cylinders; X_2 – automatic transmission
- **Model**

$$Y = a + \beta_1 \mathbb{I}_{\{X_1=4\}} + \beta_2 \mathbb{I}_{\{X_1=6\}} + \beta_3 \mathbb{I}_{\{X_1=8\}} + \gamma_1 \mathbb{I}_{\{X_2=a\}} + \gamma_2 \mathbb{I}_{\{X_2=m\}} + \varepsilon$$

$\mathcal{P}$	transmission	
cylinders	automatic	manual
4 cylinders	$a + \beta_1 + \gamma_1$	$a + \beta_1 + \gamma_2$
6 cylinders	$a + \beta_2 + \gamma_1$	$a + \beta_2 + \gamma_2$
8 cylinders	$a + \beta_3 + \gamma_1$	$a + \beta_3 + \gamma_2$

- six unknown parameters, $a, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2 \in \mathbb{R}$
- also six disjoint populations *(groups within the population $\mathcal{P}$)*
- **But...** not well defined!

- There are formaly six unknown parameters in the model but, mathematically, they only specify four different (unique) means
- There is an implicit structure assumed among the given subpopulations... What is the structure?

Example 2 (mtcars)

- Dependent variable $Y \in \mathbb{R}$ respresents a car's consumption efficiency [*mpg*]
- Covariates: X_1 – the number of cylinders; X_2 – automatic transmission

- **Model 1:** $Y = a + \beta X_1 + \gamma_2 X_2 + \varepsilon$

```
> summary(lm(mpg ~ cyl + am, data = mtcars))

             Estimate  Std.Error  t value  Pr(>|t|)
(Intercept)   34.5224     2.6032   13.262  7.69e-14 ***
cyl           -2.5010     0.3608   -6.931  1.28e-07 ***
am             2.5670     1.2914    1.988    0.0564 .
```

Example 2 (mtcars)

- Dependent variable $Y \in \mathbb{R}$ respresents a car's consumption efficiency [*mpg*]
- Covariates: X_1 – the number of cylinders; X_2 – automatic transmission

- **Model 1:** $Y = a + \beta X_1 + \gamma_2 X_2 + \varepsilon$

```
> summary(lm(mpg ~ cyl + am, data = mtcars))
```

```
             Estimate  Std.Error  t value  Pr(>|t|)
(Intercept)   34.5224     2.6032   13.262  7.69e-14 ***
cyl           -2.5010     0.3608   -6.931  1.28e-07 ***
am             2.5670     1.2914    1.988    0.0564 .
```

- **Model 2:** $Y = a + \beta_2 \mathbb{I}_{\{X_1=6\}} + \beta_3 \mathbb{I}_{\{X_1=8\}} + \gamma_2 \mathbb{I}_{\{X_2=1\}} + \varepsilon$

```
> summary(lm(mpg ~ as.factor(cyl) + as.factor(am), data = mtcars))
```

```
                Estimate  Std.Error  t value  Pr(>|t|)
(Intercept)       24.802      1.323   18.752   < 2e-16 ***
as.factor(cyl)6   -6.156      1.536   -4.009  0.000411 ***
as.factor(cyl)8  -10.068      1.452   -6.933  1.55e-07 ***
as.factor(am)1     2.560      1.298    1.973  0.058457 .
```

Example 2 (mtcars)

- Dependent variable $Y \in \mathbb{R}$ respresents a car's consumption efficiency [*mpg*]
- Covariates: X_1 – the number of cylinders; X_2 – automatic transmission

- **Model 1:** $Y = a + \beta X_1 + \gamma_2 X_2 + \varepsilon$

```
> summary(lm(mpg ~ cyl + am, data = mtcars))

             Estimate  Std.Error  t value  Pr(>|t|)
(Intercept)   34.5224     2.6032   13.262  7.69e-14 ***
cyl           -2.5010     0.3608   -6.931  1.28e-07 ***
am             2.5670     1.2914    1.988    0.0564 .
```

- **Model 2:** $Y = a + \beta_2 \mathbb{I}_{\{X_1=6\}} + \beta_3 \mathbb{I}_{\{X_1=8\}} + \gamma_2 \mathbb{I}_{\{X_2=1\}} + \varepsilon$

```
> summary(lm(mpg ~ as.factor(cyl) + as.factor(am), data = mtcars))

                Estimate  Std.Error  t value  Pr(>|t|)
(Intercept)       24.802      1.323   18.752   < 2e-16 ***
as.factor(cyl)6   -6.156      1.536   -4.009  0.000411 ***
as.factor(cyl)8  -10.068      1.452   -6.933  1.55e-07 ***
as.factor(am)1     2.560      1.298    1.973  0.058457 .
```

- Thus, two additional equations are used above: $\beta_1 = 0$ and $\gamma_1 = 0$

  (again, different equations can be used to alter the final interpretation of the model...)

Regression model with mixed types of covariates

- Linear regression model with mixed types of covariates (i.e., continuous, binary, or categorical variables) already follows from the models above...

Regression model with mixed types of covariates

- Linear regression model with mixed types of covariates (i.e., continuous, binary, or categorical variables) already follows from the models above...

- Dependent variable $Y \in \mathbb{R}$ respresents a car's consumption efficiency [*mpg*]
- Four explanatory covariates:
  - X_1 – number of cylinders ($X_1 \in \{4, 6, 8\}$)
  - X_2 – car weight [$\times 1000$ *lbs*]
  - X_3 – automatic transmission ($X_3 = 1$ for yes, $X_3 = 0$ for no)
  - X_4 – displacement [*cubic inches*]

Example 3 (mtcars)

❑ Model 1:

```
> summary(lm(mpg ~ cyl + wt + am + disp, data = mtcars))

             Estimate  Std.Error  t value  Pr(>|t|)
(Intercept) 40.898313   3.601540   11.356   8.68e-12 ***
cyl         -1.784173   0.618192   -2.886   0.00758 **
wt          -3.583425   1.186504   -3.020   0.00547 **
am           0.129066   1.321512    0.098   0.92292
disp         0.007404   0.012081    0.613   0.54509
```

Example 3 (mtcars)

Model 1:

```
> summary(lm(mpg ~ cyl + wt + am + disp, data = mtcars))

             Estimate  Std.Error  t value  Pr(>|t|)
(Intercept) 40.898313   3.601540   11.356   8.68e-12 ***
cyl         -1.784173   0.618192   -2.886   0.00758 **
wt          -3.583425   1.186504   -3.020   0.00547 **
am           0.129066   1.321512    0.098   0.92292
disp         0.007404   0.012081    0.613   0.54509
```

Model 2:

```
> summary(lm(mpg ~ as.factor(cyl) + wt + as.factor(am) + disp, data = mtcars))

                 Estimate  Std.Error  t value  Pr(>|t|)
(Intercept)      33.816067   2.914272   11.604 8. 79e-12 ***
as.factor(cyl)6  -4.304782   1.492355   -2.885   0.00777 **
as.factor(cyl)8  -6.318406   2.647658   -2.386   0.02458 *
wt               -3.249176   1.249098   -2.601   0.01513 *
as.factor(am)1    0.141212   1.326751    0.106   0.91605
disp              0.001632   0.013757    0.119   0.90647
```

A little bit of a challenge...

Let start with the model for the consumption using the information about automatic/manual transmissions (X_3)..

Regression model:

$$Y = a + \gamma_1 \mathbb{I}_{\{X_3 = \text{"manual"}\}} + \gamma_2 \mathbb{I}_{\{X_3 = \text{"automatic"}\}} + \varepsilon$$

A little bit of a challenge...

Let start with the model for the consumption using the information about automatic/manual transmissions (X_3)..

Regression model:

$$Y = a + \gamma_1 \mathbb{I}_{\{X_3 = \text{"manual"}\}} + \gamma_2 \mathbb{I}_{\{X_3 = \text{"automatic"}\}} + \varepsilon$$

In addition, it holds that:

- There are 19 cars (in the `mtcars` dataset) with a manual transmission and 13 cars with an automatic transmission
- Let us "parametrize" the model as $X_3 = -\frac{1}{19}$ if the car has a manual transmission and $X_2 = \frac{1}{13}$ otherwise...
- This actually specifies that:
  - $E[Y|X_3 = \text{"manual"}] = a - \frac{1}{19}\gamma_1$
  - $E[Y|X_3 = \text{"automatic"}] = a + \frac{1}{13}\gamma_1$

A little bit of a challenge for interpretation...

Note also the following:

- the estimated probability that a random car has a manual transmission is $\frac{19}{32}$
- the estimated probability that a random car has an automatic transmission is $\frac{13}{32}$
- Thus, the subpopulation of cars with a manual transmission and a subpopulation of cars with an automatic transmissions are (very likely) not of the same size ...
  $\Longrightarrow$ **this finding should not be ignored!**

A little bit of a challenge for interpretation...

Note also the following:

- the estimated probability that a random car has a manual transmission is $\frac{19}{32}$
- the estimated probability that a random car has an automatic transmission is $\frac{13}{32}$
- Thus, the subpopulation of cars with a manual transmission and a subpopulation of cars with an automatic transmissions are (very likely) not of the same size ...
  $\Longrightarrow$ **this finding should not be ignored!**

So, what is be the interpretation of the estimated parameters in the model?

```
> summary(lm(mpg ~ am, data = mtcars2))

             Estimate  Std.Error  t value  Pr(>|t|)
(Intercept)  20.0906     0.8666   23.184   < 2e-16 ***
am           55.9219    13.6191    4.106  0.000285 ***
```

Some (practical) recommendations

- In a linear regression model the parametrization of $\boldsymbol{X} \in \mathbb{R}^p$ can be taken arbitrarily but there should be always some reasonable argument behind...
- Typically, the parametrization for a continuous covariate X_j in $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ is taken in a way that the interpretation makes sense, or the magnitudes of the estimated parameters are reasonable...
- Typical parametrizations for a categorical covariate $X_k \in \{v_1, \ldots, v_K\}$ in $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ are taken in a way that conveniently suits the question of interest and provides elegant (easy) interpretation
  (e.g., comparing placebo vs. treatment, modeling the overal mean, ...)
- The final model should be always selected with respect to some goodness-of-fit criterion and the ability to interpret the model in a reasonable way (model simplicity vs. model complexity)

Final model selection (TBD)

The crucial question in the whole regression modeling process is the following: From the set of all plausible models, which can be very rich... how should we select one particular model that we consider to be the final one (the most appropriate one?)

- **Naive methods**
  - expert judgement
  - some previous experince/knowledge
- **Systematic modelling approaches**
  - stepwise forward modelling approach
  - stepwise background modelling approach
- **Various quantitative criteria**
  - Akaike's information criterion (AIC)
  - Bayesian information criterion (BIC)