Lecture 11 | 05.05.2025

Regression models
beyond typical data

Linear regression models and beyond

- Linear (regression) models... but **the truth is (almost) never linear!**
  (the linearity property is used as a good and easy approximation)
- Nevertheless, it is **convenient to have simple assumptions...**
  (but there are still many different issues that can go wrong...)
- Recall, that there are **a few levels of linearity in the model**
  (linearity of the predictor, linearity of the expectation, linearity of LS, ...)

the data are too flexible	(higher order approximations/splines)
the data are too irregular	(piecewise approximation)
the data are too complex	(additive models)
the data are too volatile	(robust estimation approaches)
Y contradicts the linear model	(GLM)
other issues (their combinations)	(and way more alternatives)

Recap: Linear regression framework

- for a generic random vector $(Y, \boldsymbol{X}^\top)^\top \in \mathbb{R}^{p+1}$ we assume an unknown population model $Y = \boldsymbol{X}^\top \beta + \varepsilon$ for an unknown vector $\beta \in \mathbb{R}^p$
- for a random sample $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \ldots, n\}$ drawn from the joint distribution $F_{(Y,\boldsymbol{X})}$ we have the corresponding model $Y_i = \boldsymbol{X}_i^\top \beta + \varepsilon_i$
- the data model can be also expressed as $\boldsymbol{Y}|\mathbb{X} \sim (\mathbb{X}\beta, \sigma^2\mathbb{I})$, for the random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ and $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top \in \mathbb{R}^{n\times p}$, $rank(\mathbb{X}) = p$
- with the additional normality assumption $\varepsilon \sim N_n(\boldsymbol{0}, \sigma^2\mathbb{I})$, for error terms $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$ it even holds that $\boldsymbol{Y}|\mathbb{X} \sim N_n(\mathbb{X}\beta, \sigma^2\mathbb{I})$

- And, moreover:
  - $\widehat{\beta} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\beta} = \mathbb{H}\boldsymbol{Y}$, where $\mathbb{H} = \mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top$
  - $\boldsymbol{Y} = \mathbb{H}\boldsymbol{Y} + \mathbb{M}\boldsymbol{Y}$, where $\mathbb{H} = (h_{ij})_{i,j=1}^n$ and $\mathbb{M} = (\mathbb{I} - \mathbb{H}) = (m_{ij})_{i,j=1}^n$
  - $\boldsymbol{U} = \mathbb{M}\boldsymbol{Y} = (\mathbb{I} - \mathbb{H})\boldsymbol{Y} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (U_1, \ldots, U_n)^\top$
  - $SSe = \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2 = \|\boldsymbol{U}\|_2^2$ and $MSe = SSe/(n-p)$
  - standardized residuals $V_i = U_i/\sqrt{MSe \cdot m_{ii}}$, if $m_{ii} > 0$

Linear regression models

Least squares and the linear regression models based on the LS minimization are, in general, very sensitive (non-robust) with respect to atypical (non-normal, skewed, or heavy-tailed) data...

But it is not straightforward to say what **atypical** actually means...

Two common concepts are:

- **Outlying observations**
  $\hookrightarrow$ an outlying observation in some regression model $Y_i = \boldsymbol{X}_i^\top \beta + \varepsilon_i$ is an observation $\iota \in \{1, \ldots, n\}$ for which the response expectation $E[Y_\iota | \boldsymbol{X}_\iota]$ does not follow the postulated model $E[Y|\boldsymbol{X}] = \boldsymbol{X}^\top \beta$, respectively, it is the observation for which $E[Y_\iota|\boldsymbol{X}_\iota] \neq \boldsymbol{X}_\iota^\top \beta$ (i.e., $E[Y_\iota|\boldsymbol{X}_\iota] = \boldsymbol{X}_\iota^\top \beta + \gamma$)

- **Leverage points**
  $\hookrightarrow$ a leverage point in the regression model $Y = \boldsymbol{X}^\top \beta + \varepsilon$ is an observation $\iota \in \{1, \ldots, n\}$ which is, in some sense, unusual with respect to the distribution of $\boldsymbol{X} \in \mathbb{R}^p$ (i.e., the values of $\boldsymbol{X}_i$, for $i \neq \iota$)

Outlying observations and leverage points

- It is a well-known fact that a few bad leverage points or outlyiers can result in a (very) poor fit to the bulk of the data
- Moreover, this can be even the case when using more robust alternatives that are particularly designed avoid such drawbacks
- Outlying observations and leverage points are of different nature—both of them can appear in the model (even simultaneously)
- Different strategies are proposed in the literature to deal with the outliers, with the leverage points, or both of them simultaneously

- For an simple, consider a problem of a simple mean and a simple median calculated from some univariate random sample... while the average is sensitive with respect to just one outlying observation, the sample median is way more robust... both of these quantities are on two (opposite) sides of a wide (robustness) spectra
  The same analogy also applies for the regression framework...

Outlying observations: more formally

- for a regression (data) model $Y_i = \boldsymbol{X}_i^\top \beta + \varepsilon_i$ and some observation $\iota \in \{1, \ldots, n\}$ (fixed) the following two models can be defined:

  - **Leave-one-out model** (Model $\mathcal{M}_{-\iota}$)

    $$\mathcal{M}_{-\iota}: \quad \boldsymbol{Y}_{-\iota} | \mathbb{X}_{-\iota} \sim (\mathbb{X}_{-\iota}\beta, \sigma^2 \mathbb{I}_{n-1})$$

    where $-\iota$ denotes the observation which is omitted (i.e., $\boldsymbol{Y}_{-\iota} = (Y_1, \ldots Y_{\iota-1}, Y_{\iota+1}, \ldots, Y_n)^\top \in \mathbb{R}^n$ and also $\mathbb{X}_{-\iota} \in \mathbb{R}^{(n-1)\times p}$)

  - **Outlyier model** (Model $\mathcal{M}_\iota$)

    $$\mathcal{M}_\iota: \quad \boldsymbol{Y}_\iota | \mathbb{X}_\iota \sim (\mathbb{X}_\iota \beta + \boldsymbol{j}_\iota^\top \gamma_\iota, \sigma^2 \mathbb{I}_{n-1})$$

    where ι denotes the observation which is outlying and $\boldsymbol{j}_\iota$ is a unit vector with one on the position $\iota \in \{1, \ldots, n\}$ and $\gamma_\iota \in \mathbb{R}$ is some parameter (i.e., $\boldsymbol{Y}_\iota = (Y_1, \ldots, Y_{\iota-1}, Y_\iota + \gamma_\iota, Y_{\iota+1}, \ldots, Y_n)^\top \in \mathbb{R}^n$)

- **The two models above are, in some sense, mathematically equivalent.** It can be proved, that the residual sum of squares in both models are the same (meaning that $SSe_{-\iota} = SSe_\iota$). Moreover, the vector $\widehat{\beta}_{-\iota}$ solves the normal equations in Model $\mathcal{M}_{-\iota}$ **if and only if** $(\widehat{\beta}_\iota^\top, \widehat{\gamma}_\iota)_\iota^\top$ solves the normal equations in Model $\mathcal{M}_\iota$, where, moreover, it holds that $\widehat{\beta}_{-\iota} = \widehat{\beta}_\iota$ and $\widehat{\gamma}_\iota = Y_\iota - \boldsymbol{X}_\iota^\top \widehat{\beta}_{-\iota}$

Detection of leverage points

- From the practical point of view, the exploratory analysis plays the key role in detection of leverage points (marginally, for some $j \in \{1, \ldots, p\}$)
- From the overall point of view (i.e., jointly with respect to $\boldsymbol{X}_i \in \mathbb{R}^p$), the hat matrix can be used by considering $\{h_i\}_{i=1}^n \equiv \{\boldsymbol{X}_i(\mathbb{X}^\top \mathbb{X})^{-1}\boldsymbol{X}_i\}_{i=1}^n$
- It is easy to show that $\sum_{i=1}^n h_{ii} = tr(\mathbb{H}) = tr(\mathbb{Q}\mathbb{Q}^\top) = tr(\mathbb{Q}^\top\mathbb{Q}) = p$ which can be used for further inspection of the leverage points
- The average leverage is, therefore, $\overline{h} = \sum_{i=1}^n h_{ii} = \frac{p}{n}$ and some rules-of-thumb can be applied based on this average
- Under some additional distribution assumptions one can also use some statistical inference (e.g., statistical tests)
- However, the key point here is that the point with high leverage may also influence the fit—but if it has a large residual—that's why it's often evaluated in combination with the analysis of outlying observations

Detection of outlying observations

- For any $\iota \in \{1, \ldots, n\}$ let $\widehat{Y}_{[\iota]} = \boldsymbol{X}_\iota^\top \widehat{\beta}_{-\iota}$ denote the prediction for Y_ι which is the estimate of $\mu_\iota = E[Y_\iota | \boldsymbol{X}_\iota]$ but using only $n-1$ observations
- The whole vector $\boldsymbol{Y}$ can be estimated by using a leave-one-out model, obtaining $\widehat{\boldsymbol{Y}}_{[\cdot]} = (\widehat{Y}_{[1]}, \ldots, \widehat{Y}_{[n]})^\top \in \mathbb{R}^n$
- It can be shown, that the following also holds:
  - $\widehat{\gamma}_\iota = \widehat{Y}_\iota - \boldsymbol{X}_\iota^\top \widehat{\beta}_{-\iota} = Y_\iota - \widehat{Y}_{[\iota]} = \frac{U_\iota}{m_{\iota\iota}}$
  - $\widehat{\beta}_{-\iota} = \widehat{\beta}_\iota = \widehat{\beta} - \frac{U_\iota}{m_{\iota\iota}} (\mathbb{X}^\top \mathbb{X})^{-1} \boldsymbol{X}_\iota$
  - $SSe_{-\iota} = SSe_\iota = SSe - \frac{U_\iota^2}{m_{\iota\iota}} = SSe - MSe(V_\iota^2)$
  - $\frac{MSe_{-\iota}}{MSe} = \frac{MSe_\iota}{MSe} = \frac{n-p-(V_t)^2}{n-p-1}$, where $rank(\mathbb{X}) = rank(\mathbb{X}_{-\iota}) = p$
- Thus, the original (just one) regression model $\boldsymbol{Y} | \mathbb{X} \sim (\mathbb{X}\beta, \sigma^2 \mathbb{I})$ can be used o detect outlying observations in the model
- From the inferential point of view, it is therefore interesting to test the null hypothesis $H_0 : \ \gamma_\iota = 0$ (detection of an outlier)

Something to keep in mind

- Two or more outliers next to each other can hide each other
- The outlier is always relative/specific to a model that is considered (an outlier in one model is not necessarily an outlier in another model)
- The outlying observation can also suggest that a particular observation is a data-error that must be corrected $\rightarrow$ importance of the exploratory
- If some observation is indicated to be an outlier, it should always be explored in more details...
- Often, identification of outliers with respect to some model is of primary interest (e.g., credit card transactions)

Cross-validation (CV)

- Cross-validation is a very popular and commonly used statistical techniques (also applied in regression) which is based on the vector $\widehat{\boldsymbol{Y}}_{[\cdot]} = (\widehat{Y}_{[1]}, \ldots, \widehat{Y}_{[n]})^\top$ (so-called leave-one-out CV)
- standard residual $U_i = Y_i - \widehat{Y}_i$ for some observation $i \in \{1, \ldots, n\}$ may be considered to be too optimistic, because the value of Y_i was used to train the model—i.e., to estimate β and to obtain $\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\beta}$ (and also $\widehat{Y}_i$)
- different regularization techniques are proposed and used to avoid such optimistic (overfitted) residuals but the linearity of the predictor $\boldsymbol{X}^\top \beta$ is quite strong regularization by itself
- slightly less optimistic residual (sometimes also called the deleted residual) obtained by the quantity $\widehat{\gamma}_\iota = Y_\iota - \widehat{Y}_{[\iota]} = U_\iota / m_{\iota\iota}$ because the value of Y_ι is not estimated by using the data that does not contain Y_ι itself
- more general concepts (so-called k-fold cross-validation) are also know in the literature and these techniques are commonly used in regression modelling in practice

Leverage points and outlying observations

- Some rules-of-thumb for identifying leverage points uses the criterion $h_{ii} > 3p/n$ (or, alternatively, $h_{ii} > 2p/n$)
- Other alternatives include
  - DFBETAS
    the analysis of the effect of a particular observation on the estimates of some parameter β_j
  - DFFITS
    the analysis of the effect of the ι^{th} observation on the estimates of Y_ι
  - COVRATIO
    the analysis of the effect of a particular observation on the estimates of the parameter vector $\boldsymbol{\beta}$
  - Cook distance
    the analysis of the effect of a particular observation on the estimates of the mean vector $\boldsymbol{\mu} = E[\boldsymbol{Y}|\mathbb{X}]$

How to deal with outliers and leverage points

Different statistical techniques and methodological approaches can be used to deal with the outlying observations, leverage points, or both of them simultaneously...

- naive methods use the principle of deleting bad outiers and bad leverage poitns... this should, however, never be done automatically—a proper data exploration is needed before
- more advanced methods used (iterative) re-weighted least squares where the weights are determed by some of the criterion mentioned above
- robust regression alternative which are not that much sensitive to the ouliers, leverage points, or both simultaneously can be used instead (e.g., median regression, M-estimation, least trimmed squares (LTS), ...)

In general, when dealing with leverage points (extremes in covariates) and outliers (extremes in response), traditional least squares regression can give very misleading results. Robust regression techniques typically aim to reduce the influence of such problematic observations (but there are very many different methods proposed) ...

Summary

- **Outlying observations**
  - unusual observations with respect to the observed values of the response
  - outliers may have serious consequences with respect to the final fit
  - different recommendations are used to detect and classify outliers
  - various alternatives are proposed to incorporate outliers into the model

- **Leverage points**
  - unusual observations with respect to the values of the covariates
  - leverage points may also have serious impact on the final fit
  - different tools are used to explore leverage points
  - modifications of the regression framework are used to bad leverage points