Lecture 11 | 07.05.2024

Regression models

beyond typical data

Linear regression models and beyond

- Linear 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 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)

Linear regression models and beyond

- Linear 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 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 flexibile (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)
  - the nature of Y contradicts the linear model (GLM)
  - and many more reasons (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 $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \dots, n\}$ drawn from the joint distribution $F_{(Y,\boldsymbol{X})}$ we have data 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 $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$ and $\mathbb{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)^\top \in \mathbb{R}^{n\times p}$, $rank(\mathbb{X}) = p$

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 $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \dots, n\}$ drawn from the joint distribution $F_{(Y,\boldsymbol{X})}$ we have data 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 $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$ and $\mathbb{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)^\top \in \mathbb{R}^{n \times p}$, $rank(\mathbb{X}) = p$
- Moreover:
  - $\widehat{\beta} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\beta}$
  - $\boldsymbol{Y} = \mathbb{H}\boldsymbol{Y} + \mathbb{M}\boldsymbol{Y}$, where $\mathbb{H} = \mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top = (h_{ij})_{i,j=1}^n$ and $\mathbb{M} = (m_{ij})_{i,j=1}^n$
  - $\boldsymbol{Y} = \mathbb{M}\boldsymbol{Y} = (\mathbb{I} - \mathbb{H})\boldsymbol{Y} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (U_1, \dots, U_n)^\top$
  - $SSe = \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2 = \|\boldsymbol{U}\|_2^2$ 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, and heavy-tailed) data...
But it is not straightforward to say what **atypical** actually means...

Two common concepts are:

- **Outlying observations**
  an outlying observation in some regression model $Y = \boldsymbol{X}^\top \beta + \varepsilon$ is an observation for which the response expectaion ($E[Y|\boldsymbol{X}]$) does not follow the assumed model $\boldsymbol{X}^\top \beta$, respectively, it is an observation $\iota \in \{1, \dots, n\}$ such that $E[Y_\iota|\mathbb{X}_\iota] \neq \boldsymbol{X}_\iota^\top \beta$ (i.e., $E[Y_\iota|\mathbb{X}_\iota] = \boldsymbol{X}_\iota^\top \beta + \gamma$)

- **Leverage points**
  a leverage point in some regression model $Y = \boldsymbol{X}^\top \beta + \varepsilon$ is an observation which is, in some sense, unusual with respect to the regresor values in $\boldsymbol{X}$.

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
- Morever, this can be even the case when using more robust alternatives that should avoid this drawback
- outlying observations and leverage points are of different nature—either of them can appear in the data (model) but they can also appear simultanously
- different strategies are proposed in the literature to deal with the outliers, with the leverage points, or both of them simultaneously

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
- Morever, this can be even the case when using more robust alternatives that should avoid this drawback
- outlying observations and leverage points are of different nature—either of them can appear in the data (model) but they can also appear simultanously
- different strategies are proposed in the literature to deal with the outliers, with the leverage points, or both of them simultaneously
- for a simple illustration, 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...

Outlying observations: mathematically

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

  - **Leave-one-out model**

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

    where $-\iota$ denotes the observation which is omitted

  - **Outlyier model**

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

    where ι denotes the observation which is outlying and $\boldsymbol{j}_\iota$ is a unit vector with one on the position $\iota \in \{1, \dots, n\}$

Outlying observations: mathematically

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

  - **Leave-one-out model**

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

    where $-\iota$ denotes the observation which is omitted

  - **Outlyier model**

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

    where ι denotes the observation which is outlying and $\boldsymbol{j}_\iota$ is a unit vector with one on the position $\iota \in \{1, \dots, n\}$

- It can be proved, that the residual sum of squares in both models are the same (meaning that $SSe_{-\iota} = SSe_\iota$). The vector $\widehat{\beta}_{-\iota}$ solves the normal equations in the model $\mathcal{M}_{-\iota}$ if and only if $(\widehat{\beta}_\iota^\top, \widehat{\gamma})_\iota^\top$ solves the normal equations in $\mathcal{M}_\iota$, where $\widehat{\beta}_{-\iota} = \widehat{\beta}_\iota$ and $\widehat{\gamma} = Y_\iota - \boldsymbol{X}_\iota^\top\widehat{\beta}_{-\iota}$

Detection of outlying observations

- for any $\iota \in \{1, \dots, n\}$ we denote $\widehat{Y}_{[\iota]} = \boldsymbol{X}_{\iota}^{\top}\widehat{\beta}_{-\iota}$ which is acually a least squares estimate of $\mu_{\iota} = E[Y_{\iota}|\boldsymbol{X}_{\iota}]$ but using only $n-1$ observations for $i = 1, \dots, \iota - 1, \iota + 1, \dots, n$
- the whole vector $\widehat{\boldsymbol{Y}}$ can be estimated by using a leave-one-out model, obtaining $\widehat{\boldsymbol{Y}}_{[]} = (\widehat{Y}_{[1]}, \dots, \widehat{Y}_{[n]})$
- It also holds that
  - $\widehat{\gamma}_{\iota} = \widehat{Y}_{\iota} - \boldsymbol{X}_{\iota}^{\top}\widehat{\beta}_{-\iota} = Y_{\iota} - \widehat{Y}_{[\iota]} = \frac{U_{\iota}}{m_{\iota\iota}}$
  - $\widehat{\beta}_{-\iota} = \widehat{\beta}_{\iota} = \widehat{\beta} - \frac{U_{\iota}}{m_{\iota\iota}}(\mathbb{X}^{\top}\mathbb{X})^{-1}\boldsymbol{X}_{\iota}$
  - $SSe_{-\iota} = SSe_{\iota} = SSe - \frac{U_{\iota}^2}{m_{\iota\iota}} = SSe - MSe(V_{\iota}^2)$

Detection of outlying observations

- for any $\iota \in \{1, \ldots, n\}$ we denote $\widehat{Y}_{[\iota]} = \boldsymbol{X}_\iota^\top \widehat{\beta}_{-\iota}$ which is acually a least squares estimate of $\mu_\iota = E[Y_\iota | \boldsymbol{X}_\iota]$ but using only $n-1$ observations for $i = 1, \ldots, \iota - 1, \iota + 1, \ldots, n$
- the whole vector $\widehat{\boldsymbol{Y}}$ can be estimated by using a leave-one-out model, obtaining $\widehat{\boldsymbol{Y}}_{[]} = (\widehat{Y}_{[1]}, \ldots, \widehat{Y}_{[n]})$
- It also holds that
  - $\widehat{\gamma}_\iota = \widehat{Y}_\iota - \boldsymbol{X}_\iota^\top \widehat{\beta}_{-\iota} = Y_\iota - \widehat{Y}_{[\iota]} = \frac{U_\iota}{m_{\iota\iota}}$
  - $\widehat{\beta}_{-\iota} = \widehat{\beta}_\iota = \widehat{\beta} - \frac{U_\iota}{m_{\iota\iota}} (\mathbb{X}^\top \mathbb{X})^{-1} \boldsymbol{X}_\iota$
  - $SSe_{-\iota} = SSe_\iota = SSe - \frac{U_\iota^2}{m_{\iota\iota}} = SSe - MSe(V_\iota^2)$
- thus, the original 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 also easy 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
- A notion of outlier is always relative to considered model—an observation which is an outlier with respect to one model is not necessarily an outlier with respect another model
- Outlier can also suggest that a particular observation is a data-error that must be corrected
- If some observation is indicated to be an outlier, it should always be explored
- 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 $\widehat{\boldsymbol{Y}}_{[]} = (\widehat{Y}_{[1]}, \dots, \widehat{Y}_{[n]})^\top$ (so-called leave-one-out CV)
- the residual $U_\iota = Y_\iota - \widehat{Y}_\iota$ for some observation $\iota \in \{1, \dots, n\}$ may be considered to be too optimistic, because the value of Y_ι was used to train the model—i.e., to estimate β and to obtain $\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\beta}$ (and also $\widehat{Y}_\iota$)
- 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

- considering the hat matrix $\mathbb{H} = \mathbb{X}(\mathbb{X}^\top\mathbb{X}^{-1})\mathbb{X}^\top = (h_{ij})_{i,j=1}^{n}$, the element h_{ii} for some $i \in \{1, \dots, n\}$ is called the leverage of $\boldsymbol{X}_i$
- it is easy to show that $\sum_{i=1}^{n} h_{ii} = tr(\mathbb{H}) = tr(\mathbb{Q}\mathbb{Q}^\top) = tr(\mathbb{Q}^\top\mathbb{Q}) = p$ thus, the average leverage is $\overline{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = k/n$
- some rule-of-thumb for identifying leverage points uses the criterion $h_{ii} > 3k/n$
- 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 techniques and methodological approaches can be used to deal with the outlying observations, with the leverage points, for both simultaneously...

How to deal with outliers and leverage points

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

- naive methods use the principle of deleting bad outiers and bad leverage poitns... this should, however, never be done automatically—a proper exploratory is needed
- more advanced methods used (iterative) re-weighted least squares where the weights are determined 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., the median regression)

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