MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
FMSN40: ... with Data Gathering, 9 hp

Lecture 5, spring 2023

Regression diagnostics

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

3/4-23

Problem areas in least squares

We assume:

1. additive errors ϵ_i
2. Normally distributed errors
3. independent errors
4. homoscedastic errors (constant variance)

- When (3)–(4) hold and $\hat{\boldsymbol{\beta}}$ from OLS, then $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ minimal among all unbiased estimators of $\boldsymbol{\beta}$.
- When (2) holds: least squares $\equiv$ maximum likelihood
- We do not need (2)–(4) to prove that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
- What is tricky is to verify (2)–(4).
- Assumptions allow construction of inference procedures but are not necessary in order to numerically compute least squares estimates.

Non-normal ϵ_i

- $\hat{\beta}$ can be OK if n is large.
- Confidence intervals will be more or less wrong, particularly with skewed distributions.
- Prediction intervals *will* be wrong,

Found by: Q-Q-plots, histogram, etc., of residuals.

Solutions:

- Transformations, e.g. $\ln(Y_i)$
- Use other methods that can handle the true distribution (maximum-likelihood, bootstrap, etc.)

Heterogenous variance

- $V(\epsilon_i) \neq \sigma^2$ for all ϵ_i. Often larger variance with larger mean.
- Uncertain observations have too much influence on the estimates.
- Prediction intervals *will* be wrong.

Found by: Plot of residuals against $\hat{Y}$.
Solutions:

- Transformations, e.g. $\ln(Y_i)$
- Weighted least squares (less weight to observations with larger variance).

Missing x-variables

Both non-normal residuals and heterogenous variance might be due to potential x-variables missing from the model!

Correlated errors

- $C(\epsilon_i, \epsilon_j) \neq 0$ for some $i \neq j$ (e.g. for $j = i + 1$). Often in time-series data.
- Variance estimates ($V(\hat{\beta}_i)$) will be biased: too small (if positive correlation) or too large (if negative correlation).
- Confidence (and prediction) intervals will be too narrow or too wide.

Found by: Plot residuals against next residual. Autocorrelation plots.
Solutions (not in this course):

- Time-series, e.g. AR-model, MA-model,
- generalized least squares.

Advanced residual analysis

In our model, we have assumed that $\epsilon_i \sim N(0, \sigma^2)$ and independent, i.e.

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) = N(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix})$$

- Is this true? How can we find out, since we cannot observe ϵ_i?
- When normality holds, residuals $e_i = Y_i - \hat{Y}_i$ should behave in a certain way. Check this instead.

Projection matrix and leverage

For the multiple regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ it is possible to write the fitted values as a projection of the observations onto the fitted plane:

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

where

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \qquad \text{is the projection (hat) matrix.}$$

Denote with v_{ij}, for $i, j = 1, \ldots, n$ a generic element of $\mathbf{P}$.

We can then write $\hat{Y}_i = v_{i1}Y_1 + \cdots + v_{ii}Y_i + \cdots + v_{in}Y_n$ were

$$v_{ii} = \text{the leverage of } Y_i,$$

measures the impact of Y_i on its own estimated value $\hat{Y}_i$.

Properties of the residuals, e_i

If $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ then the observed residuals

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{P}\mathbf{Y} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$$

have the following property:

$$\mathbf{e} \sim N_n(\mathbf{0},\, \sigma^2(\mathbf{I} - \mathbf{P}))$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \big(v_{ij}\big)$.
Thus they will have unequal variances and be dependent ($\mathrm{Var}(\mathbf{e})$ is not a diagonal matrix), the unequality and dependence determined by the structure of $\mathbf{X}$.

Because of different variances it is tricky to compare the residuals e_i.

So let's standardize them ... (we'll see there are issues ...)

Proofs

- **Normality**: Since $\mathbf{e}$ are linear combinations of $\mathbf{Y}$, which are multivariate normal, $\mathbf{e}$ will also be multivariate normal.
- **Zero mean**

$$\begin{aligned} E(\mathbf{e}) &= E((\mathbf{I}-\mathbf{P})\mathbf{Y}) = (\mathbf{I}-\mathbf{P})E(\mathbf{Y}) \\ &= (\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X}\boldsymbol{\beta} = (\mathbf{X}-\mathbf{X})\boldsymbol{\beta} = \mathbf{0} \end{aligned}$$

- **Covariance matrix**
  Since $\mathbf{P}' = (\mathbf{X}')'((\mathbf{X}'\mathbf{X})^{-1})'\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P}$ and $\mathbf{P}\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P}$ we get

$$\begin{aligned} \mathrm{Var}(\mathbf{e}) &= \mathrm{Var}((\mathbf{I}-\mathbf{P})\mathbf{Y}) = (\mathbf{I}-\mathbf{P})\mathrm{Var}(\mathbf{Y})(\mathbf{I}-\mathbf{P})' \\ &= (\mathbf{I}-\mathbf{P})\sigma^2\mathbf{I}(\mathbf{I}-\mathbf{P}) = \sigma^2(\mathbf{I}-2\mathbf{P}+\mathbf{P}) = \sigma^2(\mathbf{I}-\mathbf{P}) \end{aligned}$$

Standardized residuals

Since $e_i \in N(0,\, \sigma^2(1 - v_{ii}))$ we standardize them by subtracting the mean ($= 0$) and dividing by the (estimated) standard deviation, to have variance approximately equal to 1:

$$r_i = \frac{e_i}{s\sqrt{1 - v_{ii}}}$$

where $v_{ii} = i$:th diagonal element of $\mathbf{P}$ and

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - (p + 1)}.$$

- The r_i have similar variances ($\simeq 1$) but are still dependent.
- While e_i is normal and $(n - (p + 1))s^2/\sigma^2$ is χ^2 they are not independent so r_i **will not be t-distributed**.

Studentized residuals

All e_i are included in s^2 so a large residual will contribute to a large s^2 affecting all the other standardized residuals. Reduce this influence by using the studentized residuals

$$r_i^* = \frac{e_i}{s_{(i)}\sqrt{1 - v_{ii}}}$$

where $s^2_{(i)}$ is the variance estimate from a regression where observation i is excluded. Now e_i and $s_{(i)}$ are independent so that

$$r_i^* \sim t(n - 1 - (p + 1)) \qquad \text{when } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}).$$

Since $t(f) \to N(0, 1)$ when $f \to \infty$, we can consider a studentized residual as unusually large when $|r_i^*| > 2$ ($\approx \lambda_{0.05/2} = 1.96$) and suspiciously large when $|r_i^*| > 3$.

Note: The r_i^* are still not independent of each other, though.

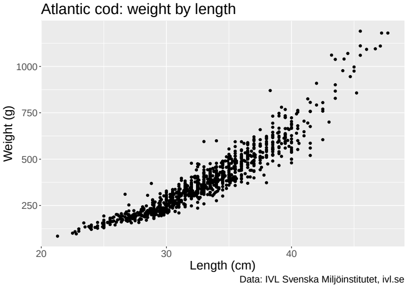
Constant variance?

- We assume that $V(\epsilon_i) = \sigma^2$.
- Then $V(e_i) = \sigma^2(1 - v_{ii})$ is not constant.
- But the studentized residuals r_i^* have constant variance since they all have the same t-distribution.
- The t-distribution is symmetrical around 0 so the median of $|r_i^*|$ is the upper quartile $t_{0.25}(f) \approx \lambda_{0.25}$ for large $f = n - 1 - (p + 1)$.
- The distribution of $|r_i^*|$ is skewed so it is better to look at $\sqrt{|r_i^*|}$ with median $\approx \sqrt{\lambda_{0.25}} \approx 0.82$.
- If $V(\epsilon_i)$ is constant then $\sqrt{|r_i^*|}$ should vary randomly around 0.82 without systematic trends.
- Plot $\sqrt{|r_i^*|}$ against $\hat{Y}_i$ and the x-variables to find trends. For visual aide, add horizontal lines at $\sqrt{\lambda_{0.25}}$, $\sqrt{2}$ and $\sqrt{3}$.

Example: Atlantic cod (from Lecture 2)

The relationship between weight and length in 1045 individual Atlantic cod (*Gadus morhua* = Torsk) in Sweden (Halland and Gotland).

Photo: Hans-Petter Fjeld - Own work, CC BY-SA 2.5, https://commons.wikimedia.org/w/index.php?curid=8399498

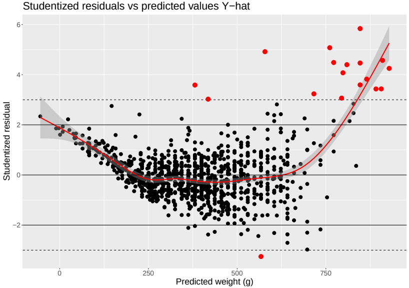


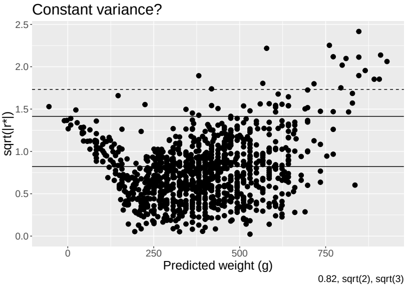
Data: IVL Svenska Miljöinstitutet, ivl.se

Let's fit a linear model and see what happens. . .

Atlantic cod: the wrong model. Studentized residuals

Using the model where Y = weight and x = length we get

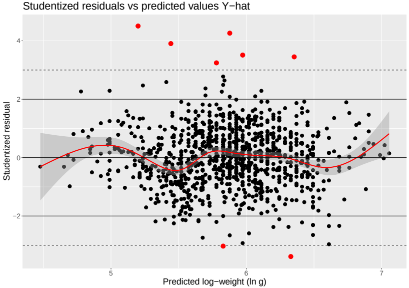


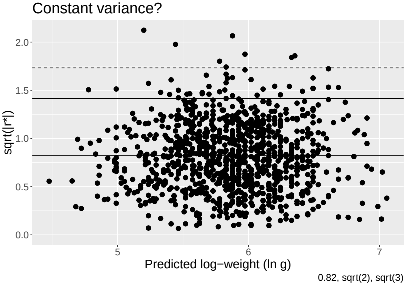


- Systematic non-linear pattern in the residuals.
- Many outliers in the residuals ($|r_i^*| > 3$) when $\hat{Y}_i$ is large.
- Residual variance is not constant.

Atlantic cod: the right model. Studentized residuals

Using the model where Y = ln weight and x = ln length we get





- No systematic pattern in the residuals.
- A few outliers in the residuals ($|r_i^*| > 3$) for mid-sized $\hat{Y}_i$.
- Residual variance is constant.

Influential observations and outliers

Individual observations, far from the others, can have a large influence on the estimates of $\boldsymbol{\beta}$ and σ^2, and thus on predictions and statistical conclusions.

- Outlier: in some sense inconsistent with the rest (Y-wise).
- Outlier in residual: Unexpectedly large ($\pm$) residual
- Potentially influential observation: outlier in the space spanned by the columns of $\mathbf{X}$.

Causes (and remedies):

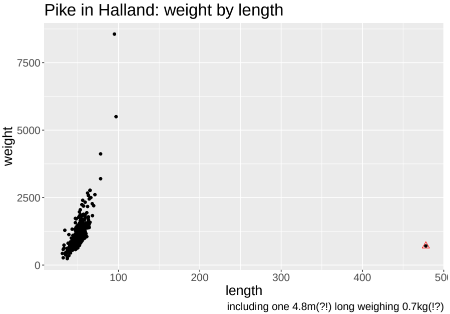
- Faulty measurement equipment (correct it or leave it out)
- Coding error (correct it or leave it out)
- Wrong or inadequate model (refine the model)
- an "interesting" (and unexpected) measurement result escaping conventional models (revise theory/knowledge of the phenomenon at study). Might lead to a discovery!

Leverage (outliers w.r.t. $\mathbf{X}$)

- Potentially influential observations are those far from the centre of gravity of $\mathbf{X}$-space. The distance is measured by the leverage v_{ii}, the diagonal elements of $\mathbf{P}$.
- It holds that $\frac{1}{n} \leq v_{ii} \leq \frac{1}{c}$ where c (≥ 1) is the number of observations with identical $\mathbf{x}$-values.
- v_{ii} is smallest when $\mathbf{x}_i$ is the centre of gravity.
  $p = 1$: $v_{ii} = \frac{1}{n} \cdot \frac{\sum_{j=1}^{n}(x_j - x_i)^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$ with minimum when $x_i = \bar{x}$.
- If $v_{ii} = 1/c$ then observation i will force the estimated line through itself.
- Leverage above $2(p+1)/n$ can be considered high.
- An observation having high leverage may not be actually influential!

Example: Northern pike

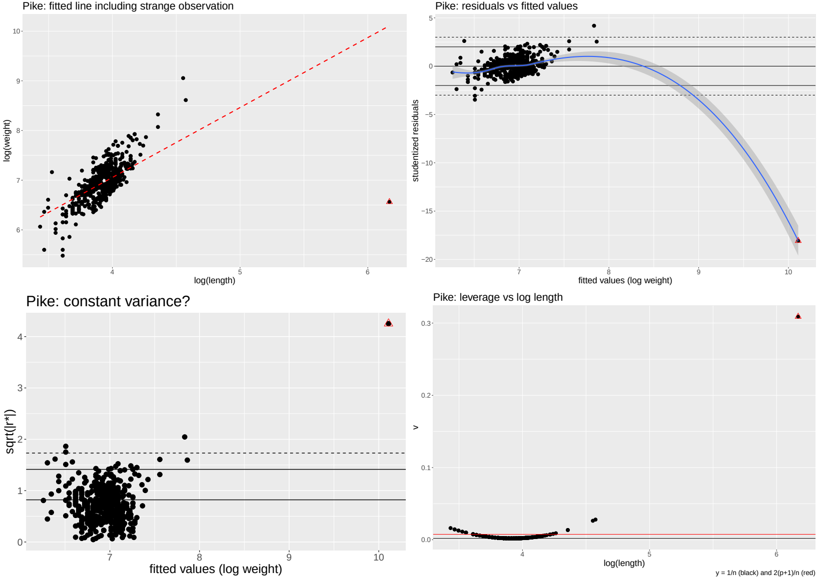
The relationship between weight and length in 530 individual Northern pike (*Esox lucius* = Gädda) in Sweden (Halland).



Data: IVL Svenska Miljöinstitutet, ivl.se

There is a strange observation of one pike that is claimed to be 478 cm long (the Swedish record is approx 2 meters) that only weighs 708 g. There is something fishy (pun intended) here!

Pike: fit a (log-log) relationship



- Both outlier and potentially influential observation.
  More difficult to spot in higher dimensions.
  Use a combination of plots and influence measures.
- A gigantic studentized residual.
- A linear pattern in all the other residuals.
- Larger residuals for both large *and* small fitted values.
- A gigantic leverage.

Conclusions

- The strange fish has a potentially large influence on the estimates.
- It has a large residual, i.e., it didn't quite manage to make the model fit itself.
- But still made it bad at fitting all the other fish.
- How much influence did it actually have?

Cook's distance

Do the potentially influential observations actually have an influence? What happens to the estimates if an observation is removed?

Denote with $\hat{\boldsymbol{\beta}}_{(i)}$ the estimate of $\boldsymbol{\beta}$ when observation i is excluded, and the corresponding prediction as $\hat{\mathbf{Y}}_{(i)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)}$.

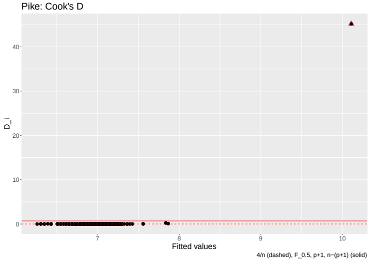
Cook's Distance, D_i measures the effect of observation i on $\hat{\boldsymbol{\beta}}$.

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})' \cdot (\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}))^{-1} \cdot (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(p+1)} = \frac{(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})'(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})}{(p+1)s^2}$$

- No unanimous consensus on how to use D_i. Observation i could be considered to have a large influence on the estimates if $D_i > F_{0.5,\,p+1,\,n-(p+1)}$. If $D_i < 4/n$ it is not a problem. Also, observations that have D_i considerably higher than the rest may be problematic.
- Plot D_i with the limits added for a visual indication.

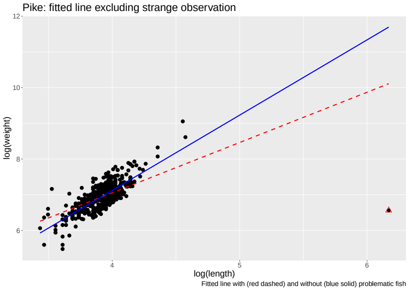
Caution

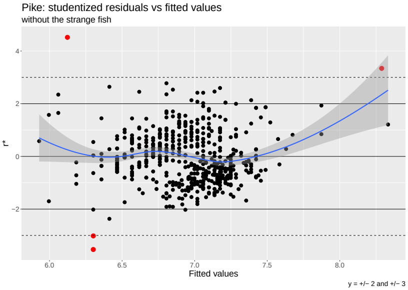
- Don't be overzealous in comparing a quantity to an empirical threshold, e.g. automatically classifying an observation according to a limit.
- Do not take the threshold as absolute truth, when these are coming out of empirical experience.
- **Advice**: use graphics to examine in closer details the points with values of D that are substantially larger than the rest. Thresholds should only be used to enhance graphical displays.

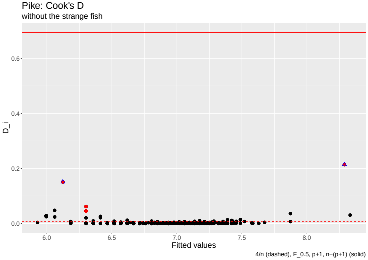


The strange fish has a gigantic Cook's distance!

Pike: Cook's distance and effect on fitted line







Large impact on the fitted line.

Removing the observation solves the problem! (except for the larger variability for small pike)

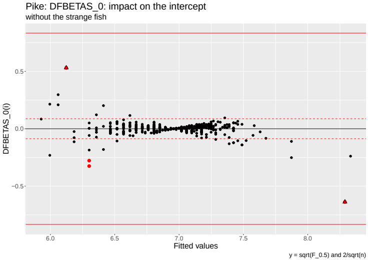
A handful of fish still have more impact than the others.

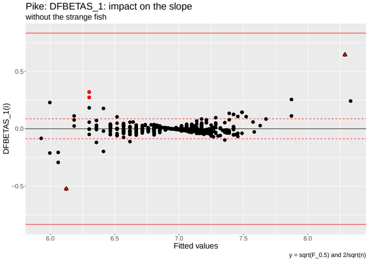
Influence on a specific parameter

The impact of an observation i on a specific element $\hat{\beta}_j$ of vector $\hat{\boldsymbol{\beta}}$ can be assessed using DFBETAS:

$$\text{DFBETAS}_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{s_{(i)}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}}}$$

The change in $\hat{\beta}_j$ $(j = 0, \ldots, p)$ caused by observation i can be considered large if $|\text{DFBETAS}_{j(i)}| > 2/\sqrt{n}$ (or $\sqrt{F_{0.5,\,p+1,\,n-(p+1)}}$). With the same words of caution as for D_i.





The long heavy fish has increased the slope, the short heavy fish has decreased it.

Summary

- Model validation, model diagnostics (influence analysis, residual analysis) is more like an *art*.
- We can't check for any possible thing that can go wrong. In particular, large datasets always have some "strange observation".
- Our model might be correct even if some observation is not well represented/fitted.
- What is important is to be aware of model assumptions, try to verify those, try to fix what can be fixed, spot anomalous/suspicious observations that might (badly) affect inferences and results.
- The previous methods are some "recipes" more than formal tests. Use them as a guiding tool but ultimately follow your judgement.