Statistics for Data Science: Week 10

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch, myrto.limnios@epfl.ch

Gauss-Markov & Optimal Estimation

Gaussian Linear Model: Efficiency of LSE (Optimality)

Q: Geometry suggests that the LSE $\hat{\boldsymbol{\beta}}$ is a sensible estimator. But is it the best we can come up with?

A: Yes, in that $\hat{\boldsymbol{\beta}}$ is the *unique minimum variance unbiased estimator* of $\boldsymbol{\beta}$.

($\hat{\beta}$ is sufficient, in fact minimally sufficient, in exponential family)

Thus, in the Gaussian Linear model, the LSE are the best we can do as far as unbiased estimators go.

(actually can show S^2 is optimal unbiased estimator of σ^2, by similar arguments)

Optimality in a Weaker Setting?

The crucial assumption so far was:

- **Normality**: $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I})$

What if we drop this strong assumption and assume something weaker? (e.g. only moment assumptions?)

- **Uncorrelatedness**: $\mathbb{E}[\varepsilon] = 0$ & $\mathrm{cov}[\varepsilon] = \sigma^2 \boldsymbol{I}$

(notice we do not assume any particular distribution.)

How well do our LSE estimators perform in this case?

(note that in this setup the observations may not be independent — uncorrelatedness implies independence only in the Gaussian case.)

For a start, we retain unbiasedness:

Lemma (Unbiasedness under Moment Assumptions)

If we only assume

$$\mathbb{E}[\boldsymbol{\varepsilon}] = 0 \quad \& \quad \mathrm{var}[\boldsymbol{\varepsilon}] = \sigma^2 \boldsymbol{I}$$

instead of

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}),$$

then the following remain true:

1. $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$;
2. $\mathrm{cov}[\hat{\boldsymbol{\beta}}] = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$;
3. $\mathbb{E}[S^2] = \sigma^2$.

But what about optimality properties?

Theorem (Gauss-Markov)

Let $\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\beta_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$*, with* $p < n$*,* $\boldsymbol{X}$ *having rank* p*, and*

- $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$,
- $\mathrm{cov}[\boldsymbol{\varepsilon}] = \sigma^2 I$.

Then, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$ *is the best linear unbiased estimator of* $\boldsymbol{\beta}$*, that is, for any linear unbiased estimator* $\tilde{\boldsymbol{\beta}}$ *of* $\boldsymbol{\beta}$*, it holds that*

$$\mathrm{cov}(\tilde{\boldsymbol{\beta}}) - \mathrm{cov}(\hat{\boldsymbol{\beta}}) \succeq 0.$$

Proof.

Let $\tilde{\beta}$ be linear and unbiased, in other words: $\begin{cases} \tilde{\boldsymbol{\beta}} = \boldsymbol{AY}, & \text{for some } \boldsymbol{A}_{p\times n}, \\ \mathbb{E}[\tilde{\boldsymbol{\beta}}] = \boldsymbol{\beta}, & \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p. \end{cases}$

These two properties combine to yield,

$$\boldsymbol{\beta} = \mathbb{E}[\tilde{\boldsymbol{\beta}}] = \mathbb{E}[\boldsymbol{AY}] = \mathbb{E}[\boldsymbol{AX}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\varepsilon}] = \boldsymbol{AX}\boldsymbol{\beta}, \quad \boldsymbol{\beta} \in \mathbb{R}^p$$

$$\implies (\boldsymbol{AX} - \boldsymbol{I})\boldsymbol{\beta} = 0, \, \forall \boldsymbol{\beta} \in \mathbb{R}^p.$$

We conclude that the null space of $(\boldsymbol{AX} - \boldsymbol{I})$ is the entire $\mathbb{R}^p$, and so $\boldsymbol{AX} = \boldsymbol{I}$.

$$\begin{aligned} \operatorname{cov}[\tilde{\boldsymbol{\beta}}] - \operatorname{cov}[\hat{\boldsymbol{\beta}}] &= \boldsymbol{A}\sigma^2\boldsymbol{I}\boldsymbol{A}^\top - \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\ &= \sigma^2\{\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{AX}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{A}^\top\} \\ &= \sigma^2\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{A}^\top \\ &= \sigma^2\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})^\top\boldsymbol{A}^\top \\ &\succeq 0. \end{aligned}$$

□

(Approximate) Sampling Distribution of $\hat{\boldsymbol{\beta}}$ under Moment Assumptions

If we only assume $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$ and $\mathrm{cov}[\boldsymbol{\varepsilon}] = \sigma^2 \boldsymbol{I}$

$\hookrightarrow$then Gauss-Markov says $\hat{\boldsymbol{\beta}}$ optimal linear unbiased estimator, regardless of distibution of $\boldsymbol{\varepsilon}$.

Question: *What can we say about the sampling distribution of $\hat{\boldsymbol{\beta}}$ when ε is not necessarily Gaussian?*

Note that we can always write

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

- Since there is a huge variety of candidate distributions for $\boldsymbol{\varepsilon}$ that would be compatible with the property $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$, we cannot say very much about the exact distribution of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon}$.
- Can we at least hope to say something about this distribution asymptotically, as the sample becomes large?

For this, we need an appropriate asymptotic framework for covariates:

- We let $n \to \infty$ (# rows of $\boldsymbol{X}$ tend to infinity)
- # columns of $\boldsymbol{X}$, i.e., p, (held fixed).

Theorem (Large Sample Distribution of $\hat{\beta}$)

Let $\{\boldsymbol{X}_n\}_{n\geq 1}$ be a sequence of $n \times p$ design matrices, and $\{\varepsilon_n\}_{n\geq 1}$ a sequence of n-vectors, and define $\boldsymbol{Y}_n = \boldsymbol{X}_n\beta + \varepsilon_n$. If

1. *$\boldsymbol{X}_n$ is of full rank p for all $n \geq 1$*
2. $\max_{1\leq i\leq n}[\boldsymbol{x}_i^\top(\boldsymbol{X}_n^\top\boldsymbol{X}_n)^{-1}\boldsymbol{x}_i] \overset{n\to\infty}{\longrightarrow} 0$,

   (where $\boldsymbol{x}_i^\top$ is the ith row of $\boldsymbol{X}_n$)
3. $\mathbb{E}[\varepsilon_n] = 0$ *and* $\operatorname{cov}[\varepsilon_n] = \sigma^2\boldsymbol{I}_{n\times n}$ *for all $n \geq 1$,*

then the least squares estimator $\hat{\beta}_n = (\boldsymbol{X}_n^\top\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top\boldsymbol{Y}_n$ satisfies

$$(\boldsymbol{X}_n^\top\boldsymbol{X}_n)^{1/2}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}_p(0, \sigma^2\boldsymbol{I}_{p\times p}).$$

Conclusion can be interpreted as:

$$\text{for } n \text{ ``large enough''}, \; \hat{\boldsymbol{\beta}} \overset{d}{\approx} \mathcal{N}\{\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1}\}$$

- i.e. distribution of $\hat{\boldsymbol{\beta}}$ gradually becomes the same as what it would be if ε were Gaussian
- ... provided design matrix $\boldsymbol{X}$ satisfies extra condition (2).
- Can be shown equivalent to: *diagonal elements of* $\boldsymbol{H}_n = \boldsymbol{X}_n(\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top$, *say* $h_{jj}(n)$ *converge to zero uniformly in* j *as* $n \to \infty$
- Note that trace$(\boldsymbol{H}) = p$, so that the average $\sum h_{jj}(n)/n \to 0$ — the question is do all the $h_{jj}(n) \to 0$ uniformly?

Has a very clear interpretation in terms of the form of the design that we will see when we discuss the notions of leverage and influence.

To understand Condition (2), consider simple linear model

$$Y_i = \beta_0 + \beta_1 t_i + \varepsilon_i, \qquad i = 1, \dots, n.$$

Here, $p = 2$. Can show that

$$h_{jj}(n) = \frac{1}{n} + \frac{(t_j - \bar{t}\,)^2}{\sum_{k=1}^{n}(t_k - \bar{t}\,)^2}$$

- Suppose $t_i = i$, for $i = 1, \dots, n$ (regular grid). Then

$$h_{jj}(n) = \frac{1}{n} + \frac{\{j - (n+1)/2\}^2}{(n^2 - n)/12}$$

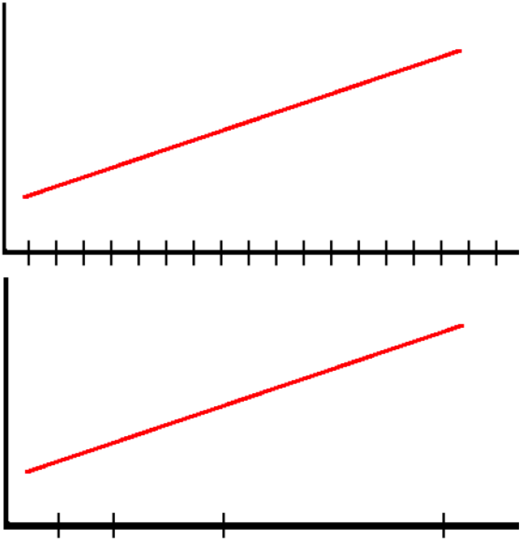
so $$\max_{1 \le j \le n} h_{jj}(n) = h_{nn}(n) = \frac{1}{n} + \frac{6(n-1)}{n(n+1)} \overset{n \to \infty}{\longrightarrow} 0.$$

- Now consider $t_i = 2^i$ (grid points spread apart as n grows).
  The centre of mass and sum of squares of the grid points is now

$$\bar{t} = \frac{2(2^n - 1)}{n}, \quad \sum_{i=1}^{n} (t_i - \bar{t}\,)^2 = \frac{4^{n+1} - 4}{3} - \frac{4^{n+1} + 4 - 2^{n+3}}{n}$$

  and so

$$\max_{1 \le j \le n} h_{jj}(n) = h_{nn}(n) \overset{n \to \infty}{\longrightarrow} \frac{3}{4}.$$



Proof.

Recall that $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^\top \boldsymbol{\varepsilon}_n$. We will show that for any unit vector $\boldsymbol{u}$,

$$\boldsymbol{u}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1/2}\boldsymbol{X}_n^\top \varepsilon_n \xrightarrow{d} N(0, \sigma^2),$$

and then the theorem will be proven by the Cramér-Wold device[a]. Now notice that

$$\boldsymbol{u}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1/2}\boldsymbol{X}_n^\top \boldsymbol{\varepsilon}_n = \gamma_n^\top \boldsymbol{\varepsilon}_n$$

where:

1. $\boldsymbol{\gamma}_n = (\gamma_{n,1}, \dots, \gamma_{n,n})^\top = \left(\boldsymbol{u}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1/2}\boldsymbol{x}_1, \dots, \boldsymbol{u}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1/2}\boldsymbol{x}_n\right)^\top$
2. $\gamma_{n,i}^2 \le \|\boldsymbol{u}\|^2 \left\|(\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1/2}\boldsymbol{x}_i\right\|^2 = \boldsymbol{x}_i^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1}\boldsymbol{x}_i$ (Cauchy-Schwarz)
3. $\gamma_n^\top \gamma_n = \boldsymbol{u}^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1/2}(\boldsymbol{X}_n^\top \boldsymbol{X}_n)(\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1/2}\boldsymbol{u} = 1.$

Consequently, the result follows from the weighted sum CLT upon noticing:

$$\max\nolimits_{1\le i\le n} \gamma_{n,i}^2 \Big/ \sum\nolimits_{k=1}^n \gamma_{n,k}^2 \le \max\nolimits_{1\le i\le n} \boldsymbol{x}_i^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1}\boldsymbol{x}_i \to 0$$

[a] Cramér-Wold: $\boldsymbol{\xi}_n \xrightarrow{d} \boldsymbol{\xi}$ in $\mathbb{R}^d$ if and only if $\boldsymbol{u}^\top \boldsymbol{\xi} \xrightarrow{d} \boldsymbol{u}^\top \boldsymbol{\xi}$ in $\mathbb{R}$ for all unit vectors $\boldsymbol{u}$.

Diagnostics

Assumptions to Check for

Four basic assumptions inherent in the Gaussian linear regression model:

- Linearity: $\mathbb{E}[Y]$ is linear in X.
- Homoskedasticity: $\mathrm{var}[\varepsilon_j] = \sigma^2$ for all $j = 1, \ldots, n$.
- Gaussian Distribution: errors are Normally distributed.
- Uncorrelated Errors: ε_i uncorrelated with ε_j for $i \neq j$.

When one of these assumptions fails clearly, then Gaussian linear regression is inappropriate as a model for the data.

Isolated problems, such as outliers and influential observations also deserve investigation. They *may or may not* decisively affect model validity.

How do we check these assumptions?

Scientific reasoning: impossible to *validate* model assumptions.

Cannot *prove* that the assumptions hold. Can only provide evidence in favour (or against!) them.

Strategy:

- Find implications of each assumption that we can check graphically (mostly concerning residuals).
- Construct appropriate plots and assess them (requires experience).

"Magical Thinking": Beware of overinterpreting plots!

Basic recipe for regression - Spoiler!!

Diagnostic plots usually constructed:

1. $\boldsymbol{Y}$ against columns of $\boldsymbol{X}$
   $\hookrightarrow$ check for linearity and outliers
2. standardized residual $\boldsymbol{r}$ against columns of $\boldsymbol{X}$
   $\hookrightarrow$ check for linearity
3. $\boldsymbol{r}$ against covariates not included
   $\hookrightarrow$ check for variables left out
4. $\boldsymbol{r}$ against fitted value $\hat{\boldsymbol{Y}}$
   $\hookrightarrow$ check for homoskedasticity
5. Normal quantile plot
   $\hookrightarrow$ check for normality
6. Cook's distance plot
   $\hookrightarrow$ check for influential observations

Residuals Revisited

Residuals $\boldsymbol{e}$: Basic tool for checking assumptions.

$$\text{Recall: } \boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\beta} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\varepsilon$$

Intuition: the residuals represent the aspects of $\boldsymbol{Y}$ that cannot be explained by the columns of $\boldsymbol{X}$.

Since $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2\boldsymbol{I})$, if the model is correct we should have $\boldsymbol{e} \sim \mathcal{N}_n\{0, \sigma^2(\boldsymbol{I} - \boldsymbol{H})\}$. (if model correct, residuals are ancillary)

$$\text{So if assumptions hold} \rightarrow \begin{cases} e_i \sim \mathcal{N}\{0, \sigma^2(1 - h_{ii})\} \\ \operatorname{cov}(e_i, e_j) = -\sigma^2 h_{ij} \end{cases}$$

Note the residuals are correlated, and that they have unequal variances.

$\Longrightarrow$ Define the *standardised residuals*:

$$r_i := \frac{e_i}{S\sqrt{1 - h_{ii}}}, \quad i = 1, \ldots, n.$$

These are still correlated but have variance ≈ 1.

(can decorrelate by $\boldsymbol{U}^\top\boldsymbol{e}$, where $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$) – why?

A first impression can be drawn by looking at plots of the response against each of the explanatory variables.

Other plots to look at?

Notice that under the assumption of linearity we have

$$\boldsymbol{X}^\top \boldsymbol{e} = 0.$$

Hence, **no correlation should appear between explanatory variables and residuals.**

1. Plot standardised residuals $\boldsymbol{r}$ against each covariate (columns of $\boldsymbol{X}$).
   ↪ No systematic patterns should appear in these plots. A systematic pattern would suggest incorrect dependence of the response on the particular explanatory (e.g. need to add a transformation of that explanatory as an additional variable).
2. Plot standardised residuals $\boldsymbol{r}$ against covariates left out of the model.
   ↪ No systematic patterns should appear in these plots. A systematic pattern suggests that we have left out an explanatory variable that should have been included.

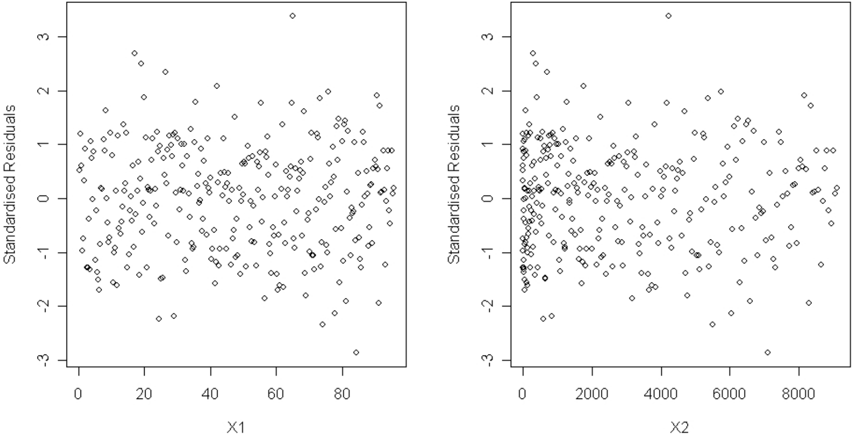


Figure: Linearity OK

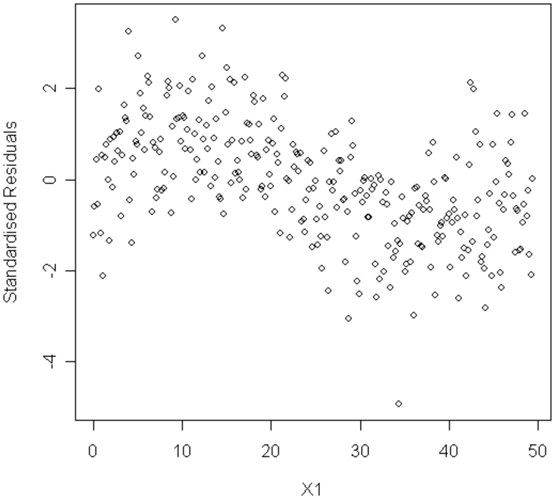


Figure: Linearity NOT OK – need to add $\sin(x_1)$ in model

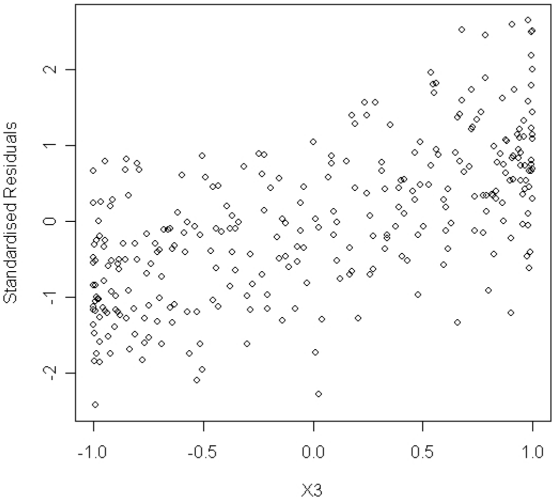


Figure: Important Covariate Left out

$$\text{Homoskedastic} = \underbrace{\acute{o}\mu o}_{\textit{same}} + \underbrace{\sigma\kappa\varepsilon\delta\alpha\sigma\mu\grave{o}\varsigma}_{\textit{spread}}$$

According to our model assumptions, the variance of the errors ε_j should be the same across indices:

$$\mathrm{var}(\varepsilon_j) = \sigma^2$$

- Plot $\boldsymbol{r}$ against the fitted values $\hat{\boldsymbol{Y}}$. (why not against $\boldsymbol{Y}$?)
  - $\hookrightarrow$ A random scatter should appear, with approximately constant spread of the values of r for the different values of $\hat{\boldsymbol{Y}}$. "Trumpet" or "bulging" effects indicate failure of the homoskedasticity assumption.
  - $\hookrightarrow$ Since $\hat{\boldsymbol{Y}}^\top \boldsymbol{e} = 0$, this plot can also be used to check linearity, as before.

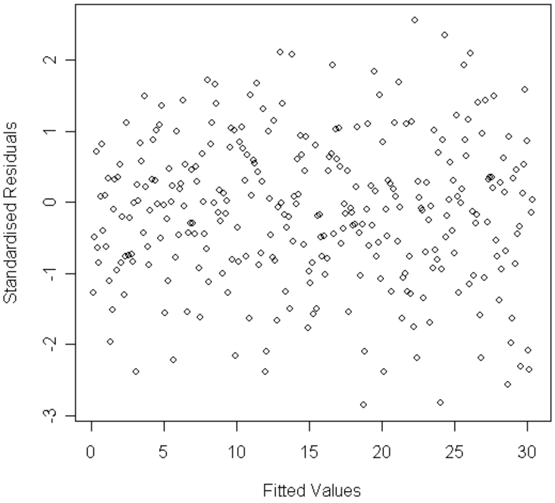


Figure: Homoskedasticity OK

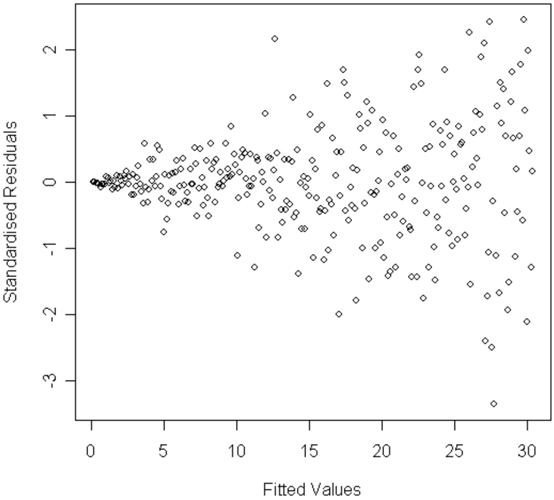


Figure: Heteroskedasticity (i.e. lack of Homoskedasticity) - ‘Trumpet’ effect

Idea: compare the distribution of standardised residuals against a Normal distribution.

How?

Compare empirical vs theoretical quantiles . . .

Reminer: The α-*quantile* ($\alpha \in [0, 1]$) of a distribution F is the value $F^{-}(\alpha)$ defined as

$$F^{-}(\alpha) := \inf\{t \in \mathbb{R} : F(t) \geq \alpha\}.$$

Given a sample $W_1, \ldots, W_n$, the *empirical* α *quantile* is the value defined as

$$\hat{F}^{-}(\alpha) := \inf\{t \in \mathbb{R} : \hat{F}(t) \geq \alpha\} = \inf\left\{t \in \mathbb{R} : \frac{\#\{W_i \leq t\}}{n} \geq \alpha\right\}.$$

where $\hat{F}$ is the empirical distribution function (as defined before).

A quantile plot for a given sample plots certain empirical quantiles against the corresponding theoretical quantiles (i.e. those under the assumed distribution).

If the sample at hand originates from F, then we expect that the points of the plot fall close to the 45° line.

- Plot the empirical $\{k/n\}_{k=1}^{n}$ quantiles of standardised residuals

$$r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}$$

against theoretical quantiles $\Phi^{-1}\{1/(n+1)\}, \ldots, \Phi^{-1}\{n/(n+1)\}$ of a $\mathcal{N}(0,1)$ distribution.

$\hookrightarrow$ Think why we pick $\Phi^{-1}\left(\frac{k}{n+1}\right)$ instead of $\Phi^{-1}\left(\frac{k}{n}\right)$.

$\hookrightarrow$ If the points of the quantile plot deviate significantly from the 45° line, there is evidence against the normality assumption. Outliers, skewness and heavy tails easily revealed.

$\hookrightarrow$ If we plot the empirical quantiles of the unstandardised residuals against those of a $N(0,1)$, then we compare against a line with slope equal to stdev($\boldsymbol{e}$) and intercept zero.

Beware of overinterpretation when n is small!

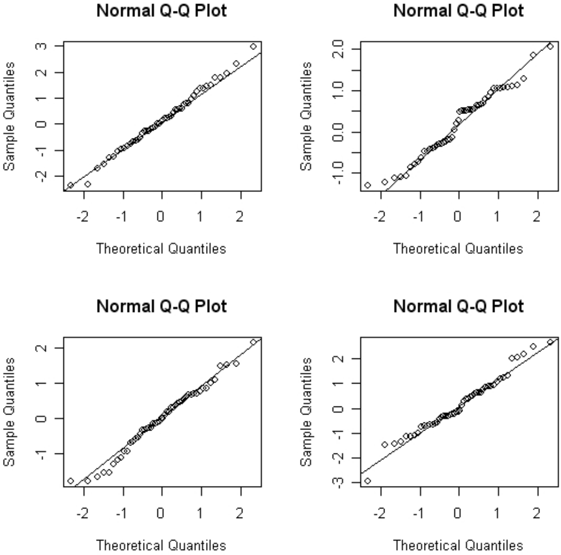


Figure: QQ Plot for $n = 50$

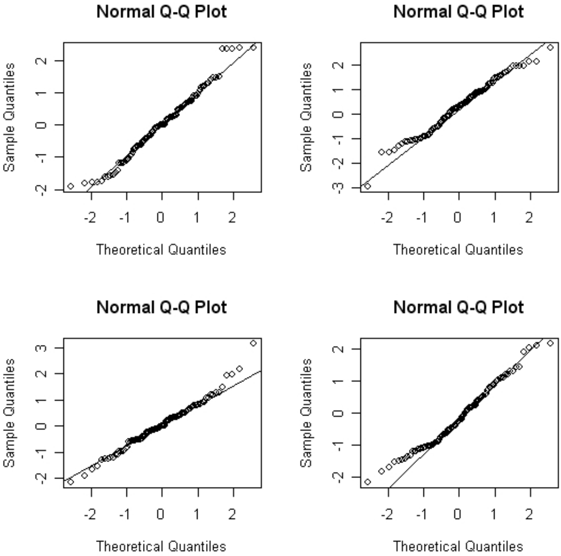


Figure: QQ Plot for $n = 100$

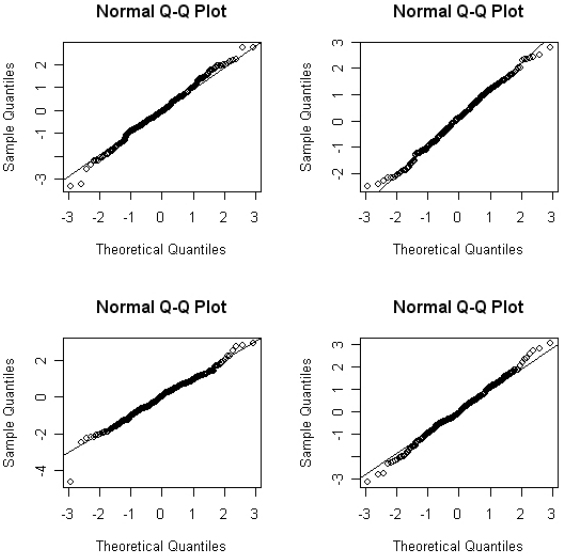


Figure: QQ Plot for $n = 300$

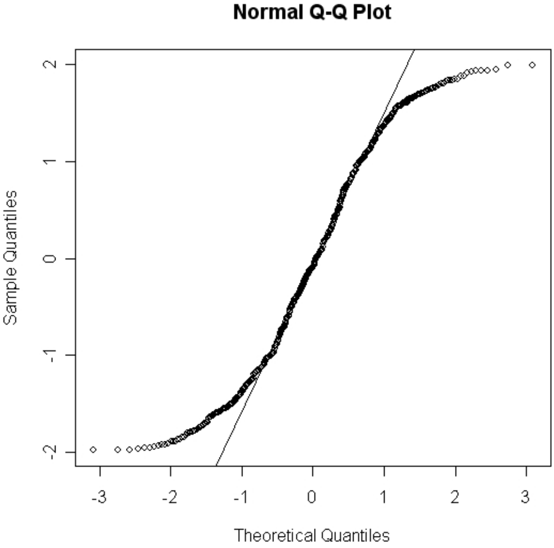


Figure: Normality not OK

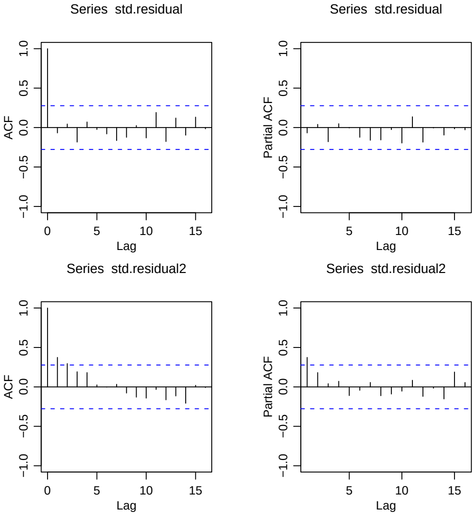
- It is assumed that $\mathrm{cov}[\varepsilon] = \sigma^2 \boldsymbol{I}$.
- Under assumption of normality this is equivalent to independence

Difficult to check this assumption in practice.

- One thing to check for is clustering, which may suggest dependence.
  - $\hookrightarrow$ e.g. identifying groups of related individuals with correlated responses
- When observations are time-ordered can look at correlation $\mathrm{corr}[r_t, r_{t+k}]$ or partial correlation $\mathrm{corr}[r_t, r_{t+k} | r_{t+1}, \ldots, r_{t+k-1}]$. When such correlations exist, we enter the domain of *time series*.

Existence of correlation:

- seriously affects estimator reliability
- inflates standard errors


Series std.residual
ACF
Lag
1.0
0.5
0.0
−0.5
−1.0
0
5
10
15
Series std.residual
Partial ACF
Lag
5
10
15
Series std.residual2
ACF
Lag
Series std.residual2
Partial ACF
Lag

An influential observation can usually be categorised as an:

- outlier (relatively easier to spot by eye)

  OR
- leverage point (not as easy to spot by eye)

Influential observations

- May or may not decisively affect model validity.
- Require scrutiny on an individual basis and consultation with the data expert.

David Brillinger (Berkeley): *You will not find your Nobel prize in the fit, you will find it in the outliers!*

Influential observations may reveal unanticipated aspects of the scientific problem that are worth studying, and so must not simply be scorned as "non-conformists"!

Outliers

An *outlier* is an observation that stands out in some way from the rest of the observations, causing *Surprise*! Exact mathematical definition exists (Tukey) but we will not pursue it.

- In regression, outliers are points falling far from the cloud surrounding the regression line (or surface).
- They have the effect of "pulling" the regression line (surface) toward them.

Outliers can be checked for visually through:

- The regression scatterplot.
  - ↪ Points that can be seen to fall relatively far from the point cloud surrounding the regression line (surface)
- Residual Plots.
  - ↪ Points that fall beyond $(-2, 2)$ in the $(\hat{\boldsymbol{Y}}, \boldsymbol{r})$ plot.

Outliers may result from a data registration error, or a single extreme event. They can, however, result because of a deeper inadequacy of our model (especially if there are many!).

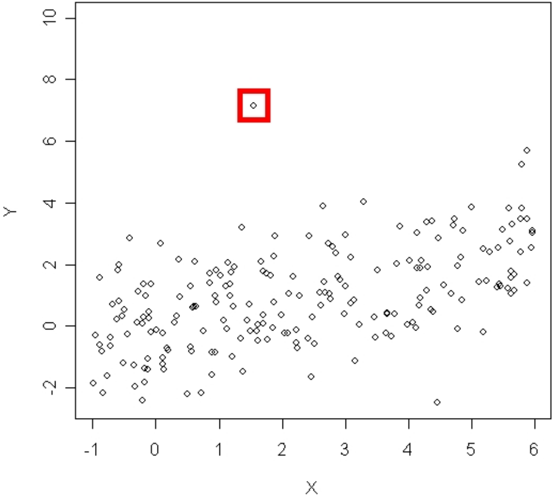


Figure: An Outlier

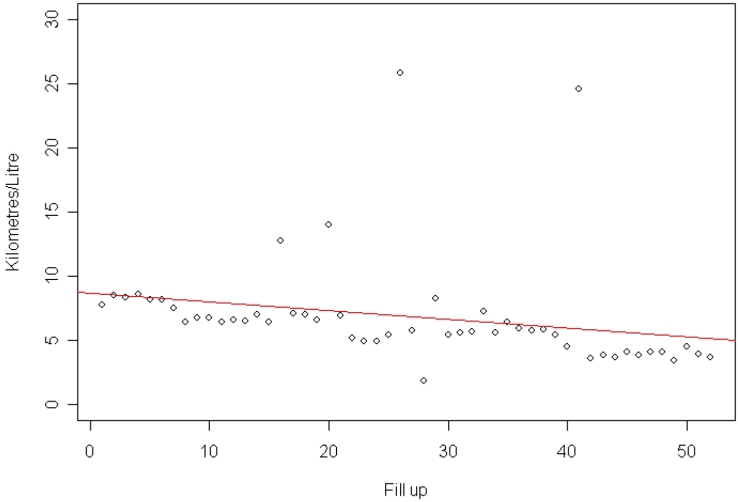


Figure: Professor's Van: Outliers

- Outliers may be influential: they "stand out" in the "y-dimension".
- However an observation may also be influential because of unusual values in the "x-dimension".
- Such influential observations cannot be so easily detected through plots.

Call $(\boldsymbol{x}_j^\top, Y_j)$ the *j-th case* and notice that

$$\mathrm{var}(Y_j - \hat{Y}_j) = \mathrm{var}(e_j) = \sigma^2(1 - h_{jj}).$$

If $h_{jj} \approx 1$, then the model is constrained so $\hat{Y}_j = \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}} \simeq Y_j$! (i.e., need a separate parameter entirely devoted to fitting this observation!)

- h_{jj} is called the leverage of the j-th case.
- since $\mathrm{trace}(\boldsymbol{H}) = \sum_{j=1}^n h_{jj} = p$, cannot have low leverage for all cases
- a good (=balanced) design corresponds to $h_{jj} \simeq p/n$ for all j

  (i.e. assumption $\max_{j \le n} h_{jj} \overset{n \to 0}{\to} 0$ satisfied in asymptotic thm).

Leverage point: (rule of thumb) if $h_{jj} > 2p/n$ observation needs further scrutiny—e.g., fitting again without j-th case and studying effect.
Outlier+Leverage Point = TROUBLE

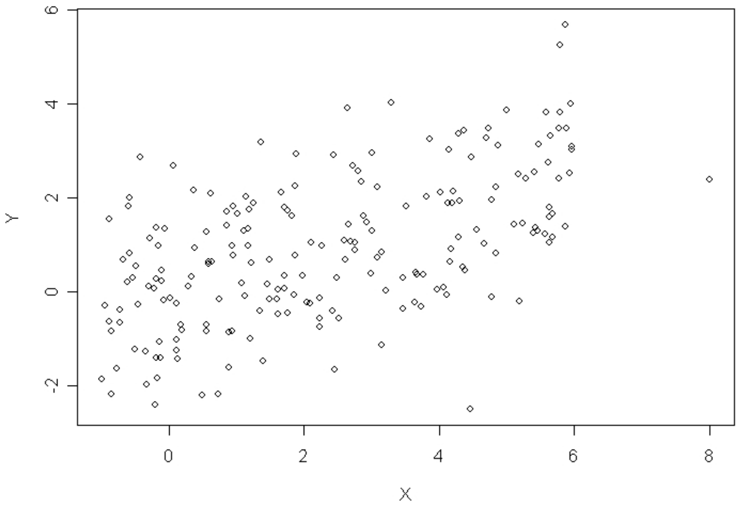


Figure: A (very) Noticeable Leverage Point

- How to find cases having strong effect on fitted model?
- Idea: see effect when case j, i.e., $(\boldsymbol{x}_j^\top, Y_j)$, is dropped.
- Let $\hat{\boldsymbol{\beta}}_{-j}$ be the LSE when model is fitted to data without case j, and let $\hat{\boldsymbol{Y}}_{-j} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{-j}$ be the corresponding fitted value.
- Define *Cook's distance*

$$C_j = \frac{1}{pS^2}(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{-j})^\top(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{-j}),$$

  which measures scaled distance between $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Y}}_{-j}$.
- Can show that

$$C_j = \frac{r_j^2 h_{jj}}{p(1-h_{jj})},$$

  so large C_j implies large r_j and/or large h_{jj}.
- Cases with $C_j > 8/(n-2p)$ worth a closer look (rule of thumb)
- Plot C_j against index $j = 1, \ldots, n$ and compare with $8/(n-2p)$ level.

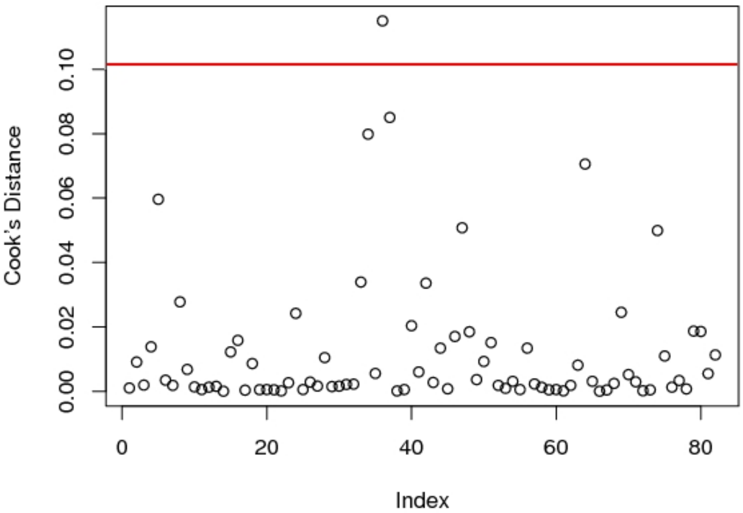


Figure: A Cook Distance Plot

Diagnostic plots usually constructed:

- $\boldsymbol{Y}$ against columns of $\boldsymbol{X}$
  - ↪ check for linearity and outliers
- standardized residual $\boldsymbol{r}$ against columns of $\boldsymbol{X}$
  - ↪ check for linearity
- $\boldsymbol{r}$ against covariates not included
  - ↪ check for variables left out
- $\boldsymbol{r}$ against fitted value $\hat{\boldsymbol{Y}}$
  - ↪ check for homoskedasticity
- Normal quantile plot
  - ↪ check for normality
- Cook's distance plot
  - ↪ check for influential observations