Index of Lecture 2b: Linear regression diagnostics

Practical Information

Today's lecture: follow-up from Lecture 2a and linear model diagnostics,

- many diagnostic methods already covered in Lecture 1a,
- new material on influence statistics,
  - special statistics to assess the impact of a single observation on the fitted regression line or regression coefficients, because it may be "problematic" if estimates or conclusions depend strongly on a single or a few observation(s),
- discussion of what to do with problematic observations,
- two worked examples: Coleman data from 1bL and `daisy2red` data from VER,
- textbook reading: VER Sections 14.8 – 10,
- review of Linear Regression Exercise 2, parts 1)–4), with software demonstrations, as needed.

Other updates:

- lab session on Monday: continued work on "Linear Regression Exercises" and other problems, and potential for more comprehensive software tutorials,
- first home assignment will soon be posted, due January 29,
- final exam confirmed for April 16.

Assumptions and Methods (Recap)

Assumptions of the linear model: $y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$,

- the errors (ε_i) are independent, normally distributed with mean 0 ("linearity, or lack of fit) and equal variances σ^2 (homoscedasticity),
- the same, except for mean 0, holds for the observations, but is of little use for model checking.

Diagnostic procedures are largely based on the residuals (either standardised (r_i) or deletion (d_i)):

- normality — inspect, graphically or numerically (possibly by a normality test), the distribution of (r_i),
- outliers — test based on (d_i),
- lack of fit — graph (r_i) against fitted values ($\hat{y}_i$) or individual predictors,
- homoscedasticity — graph (r_i) against ($\hat{y}_i$), or compute statistical test.

Generally speaking, graphical exploration and assessment are preferred over formal statistical tests.

Transformation of the outcome is one possible avenue when assumptions are violated, typically guided by a Box-Cox analysis.

Regression Diagnostics: Leverage

Leverage:

- example of influence statistic to detect, if estimates could depend strongly on a single or a few observation(s),
- definition: leverage h_i of i^{th} observation defined by the relation:[1]

$$\text{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_i), \quad i = 1, \ldots, \text{no. of observations},$$

- properties: $0 \leq h_i \leq 1$ and $\sum_i h_i = (k+1)$,
- textbook guidelines: $h_i > 2(k+1)/n$ (or more restrictively, $h_i > 3(k+1)/n$), are "high" (possible reason for concern), where $k =$ no. of predictors,
- interpretation: leverage is based on x's only! — h_i expresses whether x_i is outlying in distribution of x's (and therefore potentially influential).

Advantages and drawbacks of leverage statistic:

+ not affected by relationship with $y \Rightarrow$ easy to understand and explore: think about whether particular combination of x-values leading to high leverage is sensible,
+ generalizes idea of extreme values for single x to multiple regression,
− does not signal actual influence, only potential influence,
− less meaningful for categorical predictors (indeed, often pretty useless).

[1] In simple linear regression, $h_i = (1/n) + (x_i - \bar{x})^2 / \sum (x_i - \bar{x})^2$.

Influence Diagnostics for Linear Models

- additional influence statistics to detect influential observations in complex linear (regression) models,
- several statistics exist:
  - Cook's distance — one value per observation i,
  - DF(F)IT's (difference in fits) — one value per observation i,
  - DFBETA's (difference in fitted betas) — k values (one for each predictor) per observation i,

  with related interpretations, and often the same flags in practice.

Cook's distance statistic D_i and DFITS_i both have two interpretations:

- $\sim$ deviations between predicted values with and without observation i — effect of deleting observation i on predictions,
- $\sim$ product of residual and leverage[2]
  — combining outlier information about x's and y's,
- D_i is on a squared residual scale and based on standardised residuals, DFITS_i is on observed residual scale (signed) and based on deletion residuals,
- large values of D_i and extreme values of DFITS_i in a dataset are notable.

[2] The formulae are $D_i = r_i^2[h_i/(1-h_i)]/(k+1)$ and $\text{DFITS}_i = d_i\sqrt{[h_i/(1-h_i)]}$, where r_i = standardised residual, h_i = leverage and d_i = deletion residual.

Regression Diagnostics Overview

DFBETA measures influence on parameter estimates:

- for observation i and predictor x_j: $\text{DFBETA}_{ij} = (\hat{\beta}_j - \hat{\beta}_{j(i)})/\text{SE}$, where $\hat{\beta}_{j(i)}$ is the estimate without observation i,
- $\sim$ influence of observation i on estimate $\hat{\beta}_j$ — how many SE's does $\hat{\beta}_j$ increase, if observation i is dropped?
- $\text{DFBETA}_{ij} > 0$ (respectively < 0) $\sim$ observation i pulls $\hat{\beta}_j$ up (down),
- extreme values of DFBETA_{ij} in a dataset are notable,
- most meaningful/effective for continuous predictors.

Overview of regression diagnostics: (for observation i)

Statistic	Interpretation	"Rule for interesting"
stand. res. r_i	poorly fitted obs. in model	outside (-2,2) (dep. on n)
deletion res. d_i	deviation from fit by rest	use $t(\text{DFE}-1)$ at level α/n
leverage h_i	outlying among predictors	$\geq 2p/n$ or $\geq 3p/n$
Cook's dist. D_i	outlying and influential	≥ 1 or $\geq 4/n$
DFITS_i	outlying and influential	n large[3]: outside $\pm 2\sqrt{p/n}$
DFBETA_{ij}	influential on $\hat{\beta}_j$	n large[3]: outside $\pm 2/\sqrt{n}$

notation: n = no. of observations, p = no. of parameters in model = $k+1$

note: there is some arbitrariness in the "rules for interesting".[4]

[3] For $n < 120$, only values beyond ± 1 are "interesting".
[4] Rule for Cook's D from: Hamilton LC, *Regression with Graphics*, 1992.

What to do with "Outliers"?

(that cannot be explained as errors)

- use statistical test for outliers to determine whether they could have happened by chance,
- analyse with and without, to determine their effect on results and conclusions,
- keeping them in the analysis,
  - potentially causes bias in the estimates,
  - usually leads to higher standard deviation and reduced power (is therefore conservative),
  - influential outliers should always be reported,
- eliminating them from the analysis,
  - narrows the scope of the inference from the study,
  - potentially creates an unrealistically good model fit (is therefore liberal),
  - should always be reported and justified,
- decisions about outliers must ultimately be based on subject matter...,
- outliers may be highly informative... (showing the presence of "subjects" that fall outside the regular pattern).

Case Deletion Results (Coleman Data)

Overview of parameter estimates (etc.) from multiple regression models,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i,$$

with either 0, 1 or 2 outliers removed:

Parameter	Statistics	Observations excluded		
		none	# 18	# 3 and 18
β_1	estimate (SE)	-1.79 (1.23)	-1.62 (0.79)	-1.70 (0.47)
	t (P)	-1.45 (.17)	-2.04 (.063)	-3.64 (.003)
β_2	estimate (SE)	0.044 (0.053)	0.085 (0.035)	0.085 (0.021)
	t (P)	0.82 (.43)	2.41 (.032)	4.09 (.001)
β_3	estimate (SE)	0.56 (0.09)	0.67 (0.07)	0.67 (0.04)
	t (P)	5.98 (.000)	10.3 (.000)	17.4 (.000)
β_4	estimate (SE)	1.11 (0.43)	1.11 (0.28)	1.18 (0.16)
	t (P)	2.56 (.023)	3.98 (.002)	7.21 (.000)
β_5	estimate (SE)	-1.81 (2.03)	-4.57 (1.44)	-4.07 (0.85)
	t (P)	-0.89 (.39)	-3.18 (.007)	-4.79 (.000)
β_0	estimate (SE)	19.9 (13.6)	34.3 (9.3)	29.8 (5.5)
σ^2	estimate	$4.30 = 2.07^2$	$1.78 = 1.33^2$	$0.61 = 0.78^2$

- effect on parameter estimates when deleting obs.: some for # 18, little for # 3,
- large drop in estimate of σ^2 (and SE) when deleting observations,
- number of significant (5% level) variables: $2 \rightarrow 4 \rightarrow 5$.

Residuals & Diagnostics for WPC Model

- residual distribution and homoscedasticity: transformation of outcome preferred,[5]
- quadratic regression for `herd_size`: seems ok despite some lack of fit; could instead adjust for herds,
- linearity for `parity`: looks fine,
- no issues of collinearity (polynomial terms for `herd_size` strongly dependent, but ok),
- standardised & deletion residuals: two extreme negative residuals when assessed by outlier test, both for `wpc` $= 1$,
- leverage above $3 \cdot 10/1574 = 0.019$ for 108 observations $\sim$ less common combinations of categorical predictors (nothing critical),
- regression diagnostics:
  - Cook's D above $4/1574 = 0.0025$ for 92 observations,
  - DFITS exceeds $2\sqrt{10/1574} = 0.159$ for same 92 observations,
  - cows with twins and/or reproductive diseases were mildly influential, but no reason to delete them,
  - DFBETA's exceed $2/\sqrt{1574} = 0.05$ for lots of obs., typically the cases of the corresp. predictors (e.g. `dyst`), but none of these obs. seem strongly influential.

[5] Previous model, but for log-transformed outcome: `lnwpc`; predictors: `parity, twin, dyst, rp, vag_disch, herd_size` as well as interaction `rp*vag_disch`, quadratic term for `herd_size`, and calving in months $2-7$.