Index of Lecture 1b: Multiple linear regression

Practical Information

Today's lecture: follow-up from Lecture 1a, and start of multiple linear regression,

- interpretation of models and parameters,
- comparing models by statistical tests, incl. special case: test of linearity for grouped continuous predictor,
- polynomial regression,
- new issue in multiple linear regression: collinearity,
  - ways to detect and deal with it,
  - examples from MER and RC (old VHM 802 text).[1]

Textbook reading:

- VER2: essentially same pages as for first lecture, plus Section 14.5 on collinearity,
- PSLS[2]: Supplementary Chapter 28 on Multiple and Logistic Regression (Moodle).

Home work for Friday: Exercise 1 in "Linear Regression Exercises",

- text, dataset (`btb_episodes`) and solutions at 802 (homepage) and 812 sites (Moodle),
- use your preferred statistical software for the calculations,
- exercise review with detailed Stata demo on Friday.

[1] No good example in VER; the MER example expands on textbook coverage.
[2] *The Practice of Statistics in the Life Sciences*, 3rd ed.; the VHM 801 textbook.

Multiple Linear Regression Model

Dataset `daisy2red`: 1536 lactations of cows, focusing initially on the variables,

* y_i = milk yield during first 120 days (`milk120`),
* x_{1i} = parity (lactation number) (`parity`),
* x_{2i} = twin birth? (0=no/1=yes) (`twin`),
* x_{3i} = vaginal discharge? (0=no/1=yes) (`vag_disch`),[3]

for i^{th} lactation, $i = 1, \ldots, 1536$.

Purpose: use x-variables to predict milk yield (hoping that prediction will be valid, or meaningful, for a wider population of lactations and cows).

Alternative purpose: examine "effect" of x-variables on milk yield (sign, strength, significance of effect), but because this is an observational study causal inference is not automatic (more in a later lecture).

Statistical model (with 3 predictors):

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

where the errors $\varepsilon_1, \ldots, \varepsilon_{1536}$ are i.i.d. and $\sim \mathrm{N}(0, \sigma^2)$,

- "same" as simple linear regression, but more predictors,
- x's can be of multiple types (here: one continuous and two dichotomous predictors).

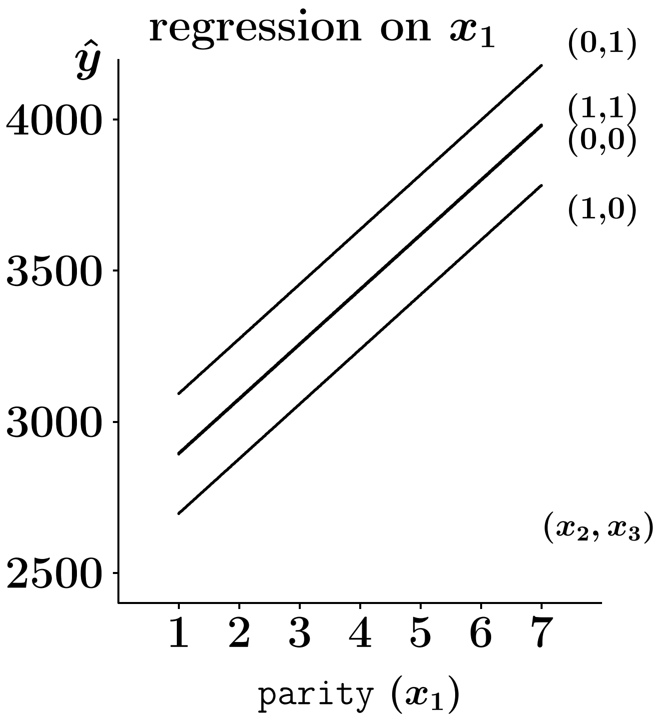
[3] After calving, vaginal discharge of certain types may serve as an indicator of different diseases/conditions for the cows, in particular metritis (urine infection).

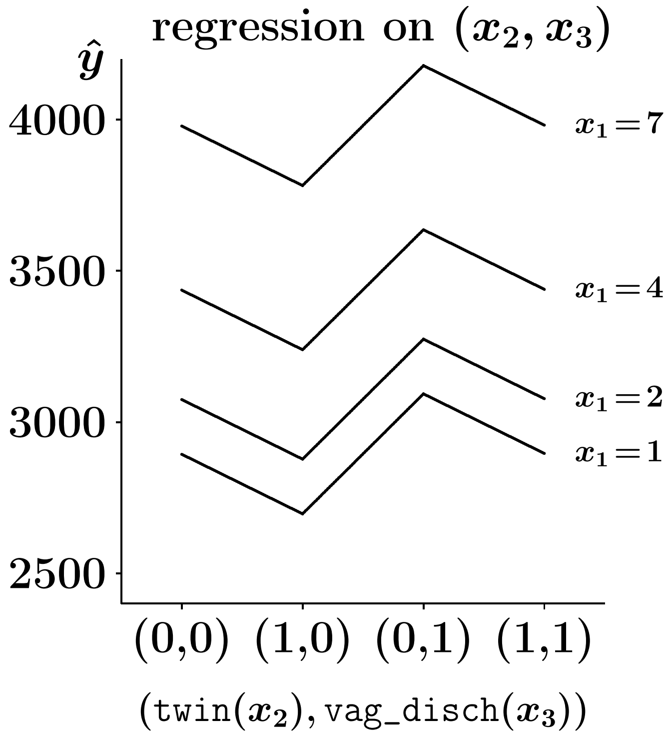
Model Assumptions and Interpretations

Model assumptions:

- independence, normality, variance homogeneity of ε_i's,
- linear relation: $$\begin{cases} \mathrm{E}(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}, \\ \hat{y} = 2713.2 + 180.9\, x_1 + (-197.0)\, x_2 + 199.7\, x_3, \end{cases}$$
  - linear "effect" of x_1 on y (for fixed x_2 and x_3),
  - additive "effects" of x_1, x_2, x_3 (parallel curves in graphs (below), no interaction).

Fitted graphs for separate regressions (with other variables fixed at the values indicated):





Multiple Linear Regression Analysis

Methods almost the same (as in simple linear regression):

- least squares estimation (minimising squared deviations between observed and predicted values),[4]
- confidence intervals, prediction and tests of simple hypotheses H_0: $\beta_j = 0$ using "4-step procedure",
  - DFE $= n-(k+1)$ (k = number of predictor parameters),
  - prediction by same approach, but beware to avoid "outlying" sets of x-values,[5]
- analysis of variance (ANOVA) table:
  - F-test is for the hypothesis H_0: all $\beta_j = 0$ (except β_0), against alternative H_a: some $\beta_j \neq 0$ (not necessarily all),[6]
  - r^2 (or R^2) = SSM/SST ∼ proportion of variance explained by model, or squared correlation between observed (y_i) and fitted ($\hat{y}_i$) values.

New issues and interpretations:

- individual regression coefficients:
  - "effects" must be viewed/interpreted in presence of other predictors — and usually change if model changes (substantial changes in the presence of collinearity; later this lecture),
  - for example, proper interpretation for β_2:
    - ∼ "effect" of `twin` when x_1, x_3 have been accounted for, (or when adding `twin` to model with x_1, x_3),
    - ∼ difference in predictions between two identical lactations, except that one has `twin=1` and the other `twin=0`.
- variable selection to arrive at most succint (or parsimonious) model (Lectures 2a and 3a).

[4] Closed formulae exist but involve matrix calculus (⇒ manual calculation not really feasible).
[5] Best way to flag "outlying" sets of x-values is by a large SE($\hat{y}$), hence the guideline: $\mathrm{SE}(\hat{y})/\sqrt{\mathrm{MSE}} \leq \sqrt{2(k+1)/n}$.
[6] Note that the F-test no longer corresponds to t-tests for individual β's.

Comparison of Models

Problem (example): does the reduced (R) model give an equally good data description as the full (F) model?

$$\begin{aligned}(\mathrm{R}):\; y_i &= \beta_0+\beta_1 x_{1i}+\varepsilon_i,\\ (\mathrm{F}):\; y_i &= \beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{3i}+\varepsilon_i.\end{aligned}$$

Idea: use statistical test to compare two models, if one model is a submodel of the other one:

- compute residual sum of squares for models (F),(R): $\mathrm{SSE(F)} \le \mathrm{SSE(R)}$,[7]
- compute residual degrees of freedom for models (F),(R): $\mathrm{DFE(F)} \le \mathrm{DFE(R)}$,[7]
- compute test statistic for the null hypothesis $\mathrm{H}_0 \sim$ (R) against $\mathrm{H}_a \sim$ (F):

  $$F=\frac{[\mathrm{SSE(R)}-\mathrm{SSE(F)}]/[\mathrm{DFE(R)}-\mathrm{DFE(F)}]}{\mathrm{MSE(F)}} \sim F(\mathrm{DFE(R)}-\mathrm{DFE(F)},\mathrm{DFE(F)}) \text{ under } \mathrm{H}_0,$$

  alternatively, H_0 may be expressed that $\beta_j=0$ for all variables removed from model (F) to model (R),
- example calculation: $F=[(638\,905\,966-635\,235\,472)/(1534-1532)]/414\,645=4.43$ $\sim P=0.012$ in $F(2,1532)$ $\Rightarrow$ model (R) is insufficient.

Alternative approach: test removal of extra β's in model (F) one at a time,

- several tests instead of one, and necessary to fit several models between (F) and (R).

[7] The full (F) model has more parameters and hence a better fit (i.e., lower error variation) and less DF.

More Model Comparisons

VER Example 14.3: model with additional predictors:

- x_{4i} = dystocia (difficult calving)? (0=no/1=yes) (`dyst`),
- x_{5i} = retained placenta? (0=no/1=yes) (`rp`),
- (for simplicity) $\tilde{y}_i$ = milk yield in 1000s (`milk120/1000`).

Model schematic:

5-predictor multiple regression
$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_5 x_{5i} + \varepsilon_i$$
SSE=
DFE=

3-predictor multiple regression
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$
SSE=635.235
DFE=1532

↓ $F = 4.43$ ($P = 0.01$)

Simple linear regression
$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_{ij}$$
SSE=638.906
DFE=1534

One sample
$$y_i = \beta_0 + \varepsilon_i$$
SSE=
DFE=

Polynomial Regression

Statistical model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_k x_i^k + \varepsilon_i,$$

where the errors $\varepsilon_1, \ldots, \varepsilon_n$ are assumed i.i.d. and $\sim \mathrm{N}(0, \sigma^2)$.

Special cases:

$k = 1$ (linear regression): $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$

$k = 2$ (quadratic regression): $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,$

$k = 3$ (cubic regression): $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i.$

Interpretation of parameters:

- quadratic model: added curvature (β_2); cubic model: added "bump" (β_3),
- always: $\beta_0 \sim$ intercept (value for $x = 0$), but parameters $\beta_1, \ldots, \beta_{k-1}$ in k^{th} order model have no useful interpretation (should however remain in model!).

Polynomial regression modelling:

- low order polynomials most useful! (k at most 3 or 4),
- polynomials may give poor predictions outside the range of x's,[8]
- test of linearity: add quadratic term, and test H_0: $\beta_2 = 0$,
- often no physical/biological meaning of the equation, but easy to analyse because a linear model, despite the non-linear relation.[9]

[8] In special cases, poor predictions can also happen within the range of x's.

[9] As discussed in footnote 7 on page 1aL−7.

Quadratic Regression Equation

daisy2red **data example**:

$$\texttt{milk120}_i = \beta_0 + \beta_1\,\texttt{parity}_i + \beta_2\,\texttt{parity}_i^2 + \varepsilon_i,$$

with the errors $\varepsilon_1, \ldots, \varepsilon_{1536}$ assumed i.i.d. and $\sim \mathrm{N}(0, \sigma^2)$.

Interpretations:

- fitted regression curve by least squares estimation:

  $$\texttt{milk120} = 2109 + 697.50\,\texttt{parity} - 82.557\,\texttt{parity}^2,$$

  for prediction within the data range of parities — the best representation of the model is by a graph of $\hat{y}$ against x for a sensible range of x-values,
- intercept $= 2109 \sim$ value for parity $= 0$ (not biologically meaningful here),
- curvature: $\hat{\beta}_2 = -82.557$ ($< 0 \sim$ "sad" parabola);
  — $\mathrm{H}_0 : \beta_2 = 0 \sim$ no curvature (hence a linear relation); $t = -11.8$ strongly significant,
- linear component ($\hat{\beta}_1$): no useful interpretation!;
  — H_0: $\beta_1 = 0 \sim$ parabola centred at parity $= 0$ (with no biological meaning),
  * the problem is that the variables x and x^2 are highly collinear ("similar"; see later in lecture); e.g., changing x_1 while keeping x_2 fixed is impossible![10]

[10] (technical) Best way to get interpretable coefficients is to reformulate model using orthogonal polynomials (Stata command: orthpoly), but usually not considered worth the trouble...

1-Way ANOVA with Quantitative Groups

`daisy2red` data example: parity as a grouping variable ∼ 1-way ANOVA model:[11]

$$\tilde{y}_i = \mu_{\text{parity}(i)} + \varepsilon_i, \quad i = 1, \ldots, 1536,$$

where $\mu_1, \mu_2, \ldots, \mu_7$ are the mean 120-day milk yields (in 1000s, when using $\tilde{y}_i$) for lactations of parity 1,. . . ,7, respectively.

2 candidate linear models: 1-way ANOVA and linear regression, with some links:

- 1-way ANOVA ≡ $(a-1)$'th order regression, where a is the number of groups ($a = 7$ in example),
- test of linear regression[12] (submodel (R)) against 1-way ANOVA (full model (F)) submodel (R) of 1-way ANOVA ∼ full model (F):

(F): $\tilde{y}_i = \mu_{\text{parity}(i)} + \varepsilon_i$

Source	SS	DF	MS	F
Groups	184.64	6	30.77	83.5
Error	563.50	1529	.3685	
Total	748.14	1535		

(R): $\tilde{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Source	SS	DF	MS	F
Lin. reg.	109.23	1	109.2	262
Error	638.91	1534	.4165	
Total	748.14	1535		

$$F = \frac{[638.91 - 563.50]/[1534 - 1529]}{.3685} = 40.9 \quad \sim P \ll 0.001 \text{ in } F(5, 1529),$$

⇒ very strong significance against the linear relation.

[11] Standard model notations: $y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, with $i \sim$ groups.

[12] Also called a lack of fit test, which generally is against the full model with all groups that can be constructed from combinations of predictor values.

Collinearity

* means that the different x-variables $(x_1, x_2, \ldots)$ in multiple regression are similar (technically: non-orthogonal),
* is indicated by non-zero (partial) correlations among (continuous) x's; extreme correlations $\Rightarrow$ severe collinearity,
* is indicated also by Variance Inflation Factors[13] (VIFs) much greater than 1 ($\geq$ 5—10 is "critical"),
* manifests itself as correlated parameter estimates.

Implications of collinearity:

- intuitively: difficult to separate/distinguish "effects" of collinear variables (explaining the "same thing"),
- each parameter's estimate, test, amount of variance explained depends (strongly) on all predictors in the model,
- two non-significant t-tests (for two variables in a model), do not! imply both to be redundant,
- also loss of precision on estimates (i.e., variance inflation).

Data example: effects of `cig_3` (x_3) in `bw5k` data (MER) for different models for birthweight (`bwt`; y) when also `cig_1`,`cig_2` (x_1, x_2) are included,

– regression of y on x_3: $\hat{\beta}_3 = -12.5\,(2.6),\ P < .0005$,
– regression of y on x_1, x_3: $\hat{\beta}_3 = -6.5\,(4.8),\ P = 0.17$,
– regression of y on x_2, x_3: $\hat{\beta}_3 = -0.1\,(8.0),\ P = 0.99$,
– regression of y on x_1, x_2, x_3: $\hat{\beta}_3 = -0.3\,(8.0),\ P = 0.97$.

[13] In Stata, use `estat vif` (or the post-estimation menu) to display VIFs after the `regress` command.

Correlated Parameter Estimates

Correlations between two random variables:

- recall that: $-1 \leq$ correlation ≤ 1, independence $\sim$ zero correlation, and positive (negative) association $\sim$ positive (negative) correlation,
- simple example: linear regression slope and intercept estimates are negatively correlated when x's away from zero,
- implication (in general): change in one variable affects other variable.

Correlations between regression parameter estimates:[14]

- "rule": only values outside $(-0.5, 0.5)$ are of real concern,
- strong correlations with intercept are "normal",
- two strongly correlated parameter estimates: cannot be interpreted independently, for example: removing one variable will affect the other one,
- many strongly correlated parameters: indication of an overfitted model (with an unrealistic good fit to data).

How to compute correlations (between parameter estimates)?

- use suitable software tools after model has been fitted.[15]

---

[14] (technical) Correlations betwen regression parameter estimates are related to the partial correlation coefficients between the x's.

[15] In Stata, use `estat vce,corr` command (or the post-estimation menu).

Collinearity Example in RC

Data: for $i = 1, \ldots, 20$ schools in USA (Coleman report),

* y_i = mean verbal test score of students (6^{th} graders),
* x_{1i} = staff salary per pupil,
* x_{2i} = percent of fathers with white collar jobs,
* x_{3i} = socio-economic status for parents,
* x_{4i} = mean verbal test score for the teachers,
* x_{5i} = mean educational level for mothers.

Full regression model:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon_i.$$

Exploration of collinearity (full regression model)

- correlations among the x-var.: strong correlations (> 0.8) between x_2, x_3 and x_5,
- variance inflation factors in full regression model: 8.40 for x_2, 7.77 for x_5 and < 5 for other predictors,
- correlated parameter estimates in full regression model: $\text{Corr}(\hat{\beta}_2, \hat{\beta}_5) = -0.78$, all others (numerically) less than 0.5.

Collinearity Example (continued)

Parameter estimates in selected models (estimates are significantly different from zero, or close):

Model	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	VIF[a]	Corr[b]
$x_1 - x_5$	19.9	-1.79	0.044	0.56	1.11	-1.81	8.4	-0.78
$x_1 - x_4$	11.5	-1.76	0.007	0.54	1.05	—	3.4	-0.83
$x_1, x_3 - x_5$	15.5	-1.71	—	0.58	1.03	-0.52	3.1	-0.82
x_1, x_3, x_4	12.1	-1.74	—	0.55	1.04	—	1.4	-0.23
x_3, x_4	14.6	—	—	0.54	0.75	—	1.0	-0.18
x_1	28.4	2.46	—	—	—	—	—	—
x_2	28.2	—	0.17	—	—	—	—	—
x_3	33.3	—	—	0.56	—	—	—	—
x_4	-2.0	—	—	—	1.48	—	—	—
x_5	-5.7	—	—	—	—	6.52	—	—
x_2, x_3, x_5	39.4	—	0.01	0.59	—	-1.06	7.9	-0.77

[a] maximal variance inflation factor among predictors in model

[b] strongest correlation among regression coefficients (excl. $\hat{\beta}_0$) in model

Conclusions/Findings:

- very variable intercepts across models (not too surprising),
- effect of x_3 remarkably constant,
- effects of x_2 and x_5 quite variable and significant on their own, but not in combination with x_3,
- effect of x_4 only significant in combination with x_3,
- strong correlations may appear in reduced models (e.g., $x_1 - x_4$),
- correlations can be high quite high even if VIFs are low.

SUMMARY: COLLINEARITY

Strong collinearity between predictors/parameters (excluding the intercept) is a problem,

i) for interpretation of estimates[16] and model building (next lecture),

ii) possibly also for the estimation itself (extreme cases),

and should then be avoided by omitting or combining the predictors involved.

Note: strong collinearity occurs "naturally" in some situations:

- between linear and quadratic terms of x (generally, between polynomial terms),
- between main effect and interaction terms,
- between indicator (dummy) variables representing a categorical predictor (next lecture),

because the variables involved are naturally related. . . ;
in these instances, collinearity would only be a real problem for reason *ii)*.

For collinearity involving quantitative predictors and its derived variables (quadratic or interaction terms), collinearity may be reduced by a technique called "centring":

- replacing x by $(x - \bar{x})$ in the model equation,
- however not affecting model fit or predictions.

The main advantage of "centring" is improved interpretation of parameters (more in next lecture).[17]

[16] Also important for interpretation is the related issue of confounding between predictors in epidemiological studies (next lecture); generally speaking, strong confounding can only occur when (strong) collinearity exists.

[17] Example 14.8 in VER demonstrates how "centring" may reduce collinearity, but this would only be of real interest if the VIFs were needed to detect other collinearities.