

Index of Lecture 1b: Multiple linear regression

Page	Title
1	Practical information
2	Multiple linear regression model
3	Model assumptions and interpretations
4	Multiple linear regression analysis
5	Comparison of models
6	More model comparisons
7	Polynomial regression
8	Quadratic regression equation
9	1-way ANOVA with quantitative groups
10	Collinearity
11	Correlated parameter estimates
12	Collinearity example in RC
13	Collinearity example (continued)
14	Summary: collinearity

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, $i = 1, \dots, 1536$.

Purpose: use x -variables to predict milk yield (hoping that prediction will be valid and 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 $\varepsilon_1, \dots, \varepsilon_{1536}$ are i.i.d. and $\sim N(0, \sigma^2)$,

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

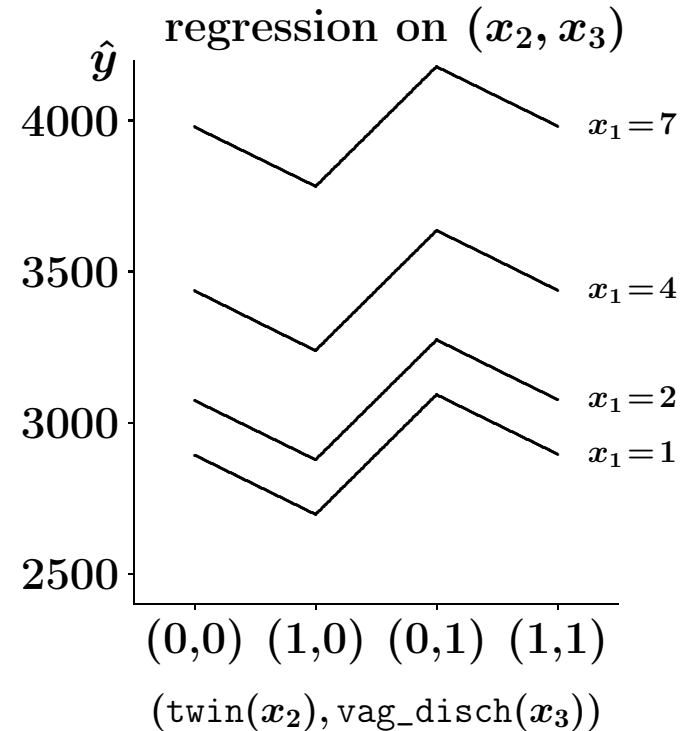
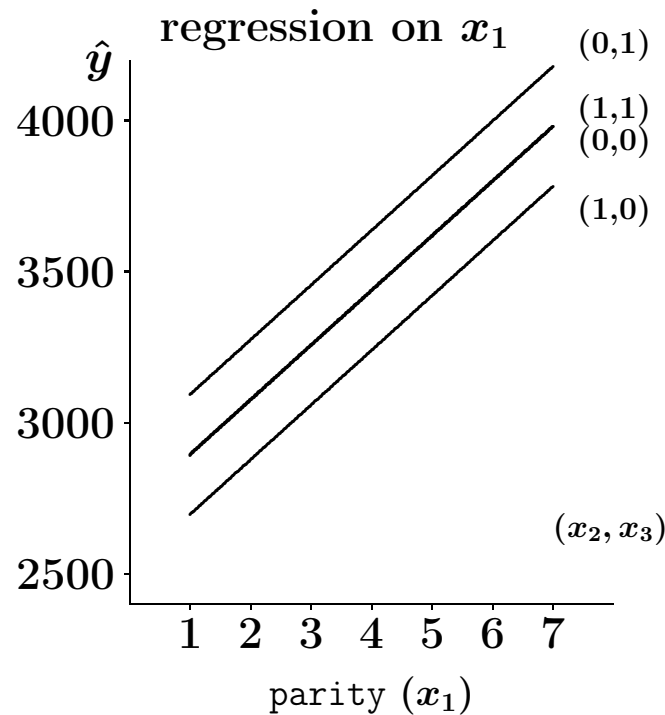
³ After calving, vaginal discharge of certain types may serve as an indicator of different diseases/conditions for the cows, in particular metritis (urine infection).

MODEL ASSUMPTIONS AND INTERPRETATIONS

Model assumptions:

- independence, normality, variance homogeneity of ε_i 's,
- linear relation:
$$\begin{cases} E(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}, \\ \hat{y} = 2713.2 + 180.9 x_1 + (-197.0) x_2 + 199.7 x_3, \end{cases}$$
- * 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 $H_0: \text{all } \beta_j = 0$ (except β_0), against alternative $H_a: \text{some } \beta_j \neq 0$ (not necessarily all),⁶
 - * r^2 (or R^2) = $SSM/SST \sim$ 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 :
 - \sim “effect” of twin when x_1, x_3 have been accounted for, (or when adding twin to model with x_1, x_3),
 - \sim difference in predictions between two identical lactations, except that one has $\text{twin}=1$ and the other $\text{twin}=0$.
- **variable selection** to arrive at most succinct (or parsimonious) model (Lectures 2a and 3a).

⁴ Closed formulae exist but involve matrix calculus (\Rightarrow 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: $SE(\hat{y})/\sqrt{MSE} \leq \sqrt{2(k+1)/n}$.

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?

$$(R) : y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i,$$

$$(F) : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i.$$

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): $SSE(F) \leq SSE(R)$,⁷
- compute **residual degrees of freedom** for models (F),(R): $DFE(F) \leq DFE(R)$,⁷
- compute **test statistic** for the null hypothesis $H_0 \sim (R)$ against $H_a \sim (F)$:

$$F = \frac{[SSE(R) - SSE(F)] / [DFE(R) - DFE(F)]}{MSE(F)} \sim F(DFE(R) - DFE(F), DFE(F)) \text{ under } H_0,$$

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).

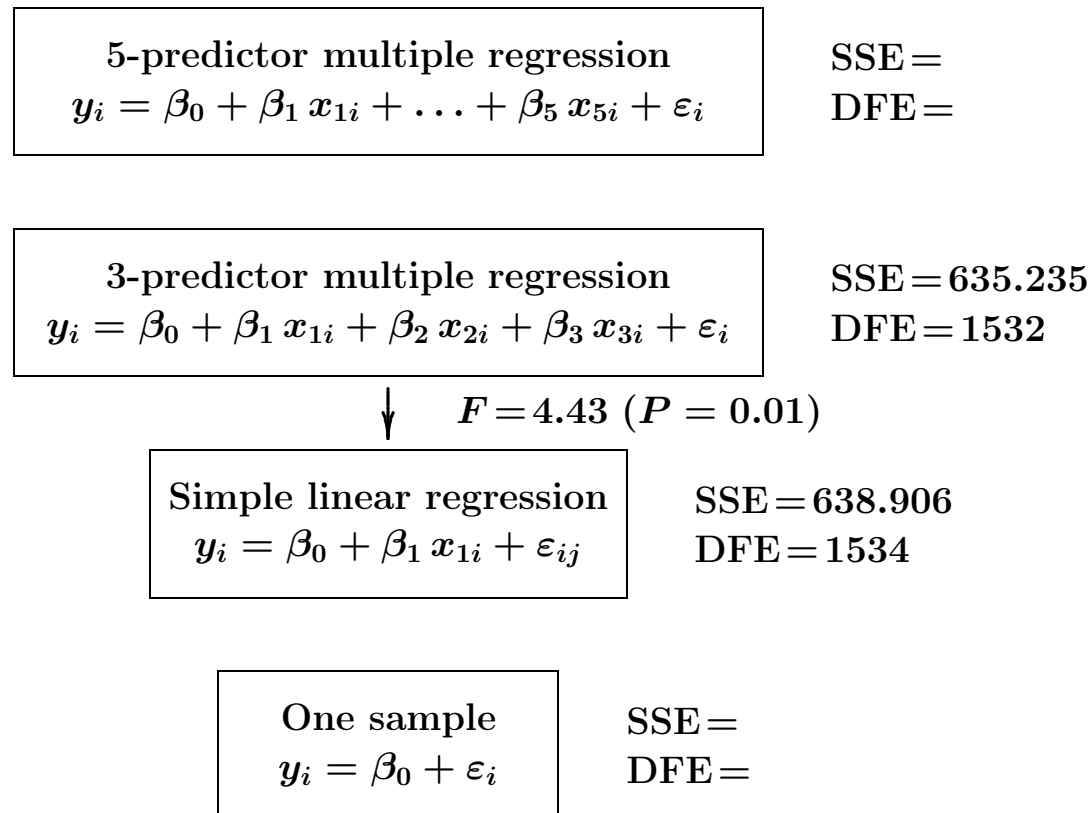
⁷ The full (F) model has more parameters and hence a better fit (i.e., lower error variation) and less DF.

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:



POLYNOMIAL REGRESSION

Statistical model:

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

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

Special cases:

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

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

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

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 $\beta_1, \dots, \beta_{k-1}$ 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:

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

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

Interpretations:

- o fitted regression curve by least squares estimation:

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

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,

- o **intercept** = 2109 \sim value for parity = 0 (not biologically meaningful here),
- o **curvature**: $\hat{\beta}_2 = -82.557$ ($< 0 \sim$ “sad” parabola);
— $H_0 : \beta_2 = 0 \sim$ no curvature (hence a linear relation); $t = -11.8$ strongly significant,
- o **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 \sim 1-way ANOVA model:¹¹

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

where $\mu_1, \mu_2, \dots, \mu_7$ 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 \equiv $(a - 1)$ '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 \sim full model (F):

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

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

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

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

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

\Rightarrow 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, \dots) in multiple regression are similar (technically: non-orthogonal),
- * is **indicated** by non-zero (partial) correlations among (continuous) x 's; extreme correlations (say beyond ± 0.9) \Rightarrow strong collinearity,
- * is **indicated** also by Variance Inflation Factors¹³ (VIFs) much greater than 1 ($\geq 5-10$ is “critical”),
- * **manifests** itself as increased SEs and correlated parameter estimates.

Implications of collinearity:

- o **intuitively**: difficult to separate/distinguish “effects” of collinear variables (explaining the “same thing”),
- o each parameter's **estimate, test, amount of variance explained** depends (strongly) on **all** predictors in the model,
- o **two non-significant t -tests** (for two variables in a model), do **not!** imply **both** to be redundant,
- o 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 $\text{cig}_1, \text{cig}_2$ (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: $-1 \leq \text{correlation} \leq 1$, 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” (and not to worry about),
- 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 between 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 $i = 1, \dots, 20$ schools in USA (Coleman report),

- * y_i = mean verbal test score of students (6th 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)

- o **correlations** among the x -var.: strong correlations (> 0.8) between x_2 , x_3 and x_5 ,
- o **variance inflation factors** in full regression model: 8.40 for x_2 , 7.77 for x_5 and < 5 for other predictors,
- o **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) may be a problem,

- i)* for interpretation of estimates¹⁶ 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** the model’s 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.