
CLASSICAL AND MODERN REGRESSION WITH APPLICATIONS

SECOND EDITION

Raymond H. Myers

Virginia Polytechnic Institute and State University

134143



Donated by
Books Plus –
Bookshop Sliema



Duxbury
Thomson Learning™

CONTENTS

CHAPTER 1

INTRODUCTION: REGRESSION ANALYSIS 1

1.1	Regression models	3
1.2	Formal uses of regression analysis	5
1.3	The data base	6
	References	7

CHAPTER 2

THE SIMPLE LINEAR REGRESSION MODEL 8

2.1	The model description	8
2.2	Assumptions and interpretation of model parameters	9
2.3	Least squares formulation	12
2.4	Maximum likelihood estimation	20
2.5	Partitioning total variability	22
2.6	Tests of hypothesis on slope and intercept	26
2.7	Simple regression through the origin (Fixed intercept)	33
2.8	Quality of fitted model	37
2.9	Confidence intervals on mean response and prediction intervals	41
2.10	Simultaneous inference in simple linear regression	47
2.11	A complete annotated computer printout	56
2.12	A look at residuals	57
2.13	Both x and y random	66
	Exercises	72
	References	80

CHAPTER 3

THE MULTIPLE LINEAR REGRESSION MODEL 82

3.1	Model description and assumptions	82
3.2	The general linear model and the least squares procedure	85
3.3	Properties of least squares estimators under ideal conditions	91
3.4	Hypothesis testing in multiple linear regression	95
3.5	Confidence intervals and prediction intervals in multiple regressions	112
3.6	Data with repeated observations	116
3.7	Simultaneous inference in multiple regression	120

3.8	Multicollinearity in multiple regression data	123
3.9	Quality fit, quality prediction, and the HAT matrix	133
3.10	Categorical or indicator variables (Regression models and ANOVA models)	135
	Exercises	153
	References	163

CHAPTER 4

CRITERIA FOR CHOICE OF BEST MODEL 164

4.1	Standard criteria for comparing models	165
4.2	Cross validation for model selection and determination of model performance	167
4.3	Conceptual predictive criteria (The C_p = statistic)	178
4.4	Sequential variable selection procedures	185
4.5	Further comments and all possible regressions	193
	Exercises	199
	References	206

CHAPTER 5

ANALYSIS OF RESIDUALS 209

5.1	Information retrieved from residuals	210
5.2	Plotting of residuals	211
5.3	Studentized residuals	217
5.4	Relation to standardized PRESS residuals	220
5.5	Detection of outliers	221
5.6	Diagnostic plots	231
5.7	Normal residual plots	242
5.8	Further comments on analysis of residuals	244
	Exercises	244
	References	248

CHAPTER 6

INFLUENCE DIAGNOSTICS 249

6.1	Sources of influence	250
6.2	Diagnostics: Residuals and the HAT matrix	251
6.3	Diagnostics that determine extent of influence	257
6.4	Influence on performance	267
6.5	What do we do with high influence points?	270
	Exercises	272
	References	273

CHAPTER 7

NONSTANDARD CONDITIONS, VIOLATIONS OF ASSUMPTIONS, AND TRANSFORMATIONS 275

7.1	Heterogeneous variance: Weighted least squares	277
7.2	Problem with correlated errors (Autocorrelation)	287
7.3	Transformations to improve fit and prediction	293
7.4	Regression with a binary response	315
7.5	Further developments in models with a discrete response (Poisson regression)	332
7.6	Generalized linear models	339
7.7	Failure of normality assumption: Presence of outliers	348
7.8	Measurement errors in the regressor variables	357
	Exercises	358
	References	365

CHAPTER 8**DETECTING AND COMBATING MULTICOLLINEARITY 368**

8.1	Multicollinearity diagnostics	369
8.2	Variance proportions	371
8.3	Further topics concerning multicollinearity	379
8.4	Alternatives to least squares in cases of multicollinearity	389
	Exercises	419
	References	422

CHAPTER 9**NONLINEAR REGRESSION 424**

9.1	Nonlinear least squares	425
9.2	Properties of the least squares estimators	425
9.3	The Gauss-Newton procedure for finding estimates	426
9.4	Other modifications of the Gauss-Newton procedure	433
9.5	Some special classes of nonlinear models	436
9.6	Further considerations in nonlinear regression	440
9.7	Why not transform data to linearize?	444
	Exercises	445
	References	449

APPENDIX A**SOME SPECIAL CONCEPTS IN MATRIX ALGEBRA 452**

A.1	Solutions to simultaneous linear equations	452
A.2	Quadratic form	454
A.3	Eigenvalues and eigenvectors	456
A.4	The inverses of a partitioned matrix	458
A.5	Sherman-Morrison-Woodbury theorem	459
	References	460

APPENDIX B**SOME SPECIAL MANIPULATIONS 461**

B.1	Unbiasedness of the residual mean square	461
B.2	Expected value of residual sum of squares and mean square for an underspecified model	462
B.3	The maximum likelihood estimator	464
B.4	Development of the PRESS statistic	465
B.5	Computation of s_{-i}	467
B.6	Dominance of a residual by the corresponding model error	468
B.7	Computation of influence diagnostics	468
B.8	Maximum likelihood estimator in the nonlinear model	470
B.9	Taylor series	470
B.10	Development of the C_k -statistic	471
	References	473

APPENDIX C**STATISTICAL TABLES 474****INDEX 486**

INDEX

- Absolute errors, minimizing sum of, 351
Alias matrix, 178
All possible regressions, 193
Analysis of covariance, 150, 151
Analysis of variance
 general ANOVA, 27, 99
 for straight line, 27
 total SS breakup, 22, 95
Assumptions for regression model, 9,
 11, 19, 21, 82, 91, 275
Autocorrelation, 287
Backward elimination, 186
Best subsets regression, 185, 193
Bias in regression estimates, 178
Biased estimation, 368, 392
 principal components, 411
 ridge regression, 392
Bonferroni confidence intervals, 50,
 121
Box-Cox transformation, 310
Bonferroni inequality, 51, 52
Box, G.E.P.
 and Cox, 310
 and Hunter, J.S. and Hunter,
 W.G., 375
 and Tidwell, 307
Categorical variables, 135
 one-way ANOVA, 150
Centering data, 11, 369
 and scaling, 384
 C_f statistic, 400, 471
Coefficient of determination, 37, 39,
 95
Collinearity, diagnosis of, 125
Condition number, 370
Confidence interval for
 $E(y)$, 41, 112
 intercept, β_0 , 32
 slope, 32
Confidence region on regression line,
 50
Correlation
 between x and y , 67
 matrix, 76
 and regression, 66
Covariance of \bar{y} , b_1 , 15
 C_p statistic, 182
Cox, D.R.
 and Box, 310
Degrees of freedom, 18
Deviance, 323, 347
Distribution between x and y , 67
 bivariate normal, 67
Dummy variables. *See* categorical
 variables
Durbin-Watson statistic, 288
Eigenvalues, 126
Eigenvectors, 126
Error structure
 additive, 297
 multiplicative, 297
Error sum of squares, 19, 89
Errors in regressors, 357
Estimation
 linear least squares, 12, 88
 maximum likelihood, 20, 92, 470
 nonlinear, 424, 425
 Gauss-Newton, 426
 initial estimates, 427
 Marquardt's compromise, 433
Estimator
 best linear unbiased, 92
 bias, 60, 178
 minimum variance unbiased, 92
Examination of regression equation
 R^2 , 37, 39, 95
 residual examination, 57, 211
Expected value
 of MS , 27, 89, 118
Exponential error family, 340

- Extra sum of squares principle, 95
- F distribution percentage points, 477-480
- F_{IN} , 186
- F_{OUT} , 186
- F -test
for β_1 , 27
for lack of fit, 118
partial, 102, 103
for regression, 27, 99
sequential, 103
subsets of regression parameters, 96, 97
tables for, 477-480
- Forward selection, 186
- Galton, F., 1
- Gamma error distribution, 345
- Gauss, C.F., 1
- Gauss-Markoff theorem, 92
- Gauss-Newton procedure, 428
- General linear hypothesis, 103
- Generalized cross validation (GCV), 398
- Generalized least squares, 278
- Generalized linear models, 339
- Generalized variance, 278
- Geometry of least squares, 89
- Goals for regression, 2, 3, 4, 5, 6
- Gompertz growth model, 437
- Growth models, 435, 437
- Hartley, H.O., 433
- HAT diagonal, 134
- HAT matrix, 134
- Heterogeneous variance, 276, 277
- Huber, P.J., 349, 352
- Hunter, J.S.
and Box and Hunter, 375
- Hunter, W.G.
and Box and Hunter, 375
- Ill-conditioning, 128
- Indicator variables. *See* categorical variables
- Influence diagnostics
COVRATIO, 267
Cutoffs on, 260
DFBETAS, 259
DFFITS, 258
HAT diagonal, 252
 s^2_b , 223
- Influence function, 350
- Influential observations, 249, 250
Cook's D statistic, 259, 260
HAT diagonal, 134, 211
PRESS, 171
- Initial estimates
determination of, 440
specific cases, 440
- Interactions, 145
- Iteratively reweighted least squares, 351
- Joint confidence region on slope and intercept, 48
- Lack of fit, 118
- Least squares
assumptions, 9, 11, 82, 91
Galton, F., 1
generalized, 278
iteratively reweighted, 351-352
maximum likelihood, 20
for nonlinear, 425, 426
properties, 14, 91, 92
- Leverage, 252
- Likelihood function, 20, 470
- Likelihood ratio test, 322
- Link function, 340-344
- Logistic growth model, 437
- Logistic regression, 317
- Mallows' C_p , 182
- Marquardt, D.W., 433
procedure, 433
- Minimum variance unbiased estimator, 92
- Mitcherlich's law, 439
- Model
Gompertz, 437
logistic, 437
mean shift outlier, 222
Michaelis-Menten, 435
Mitcherlich, 439
Richards, 439
Weibull, 439
- Multicollinearity, 125, 368
- Multiple correlation coefficient. *See* R^2
- Multiple regression, 82
- Multivariate normal density, 470
- Nonlinear estimation
examples, 430, 434
growth models, 437
least squares, 425
starting values for, 440
- Nonlinear growth models, 438, 439
types. *See* model
- Normal distribution
tables of, 475
- Normal equations
matrix solution, 88
multiple regression, 88
straight line, 12
- Normal probability plots, 60
- Outliers, 221
diagnosis of, 221
- R -Student, 223
studentized residuals, 217
- Overfitting, 179
- Partial F -test
definition, 102, 103
in selection procedures, 186
- Plots
augmented partial plots, 204
DF trace, 401
partial regression, 233
partial residual plots, 238
residual, 57, 211
ridge trace, 396
- Poisson regression, 332
- Polynomial models, 83
- Power transformations, 310
- Prediction interval, 41, 112
- Prediction-oriented model criteria, 167-172
- PRESS
residuals, 171
statistic, 171
sum of absolute PRESS residuals, 177
- Principal components regression, 411
- Probability tables
 F , 477-480
normal, 475
 t , 476
- Proportions as responses, 315-323
- PR (Ridge), 397
- Pure error, 118
repeat runs, 118
- Quadratic form, 454
- R^2 , 37, 39, 95
- Regression analysis
assumptions, 9, 11, 19, 21, 82, 91, 275
multiple, 82
purpose of, 2, 3, 4
straight line, 9
- Regression equation, examination of residual analyses, 57
- Regression through origin, 33
estimate of slope, 33
- Repeat runs, 118
- Residual MS , 19, 89
- Residual plots, 57, 211
- Residuals
Cook's D statistic, 259, 260
examination of, 57, 211
outliers, 221
PRESS, 171
studentized, 63, 221, 222
- Response variable, transformation of, 310

- Richards, F.J., 437
- Richards growth model, 437
- Ridge regression, 392
 choice of k , 396–412
 ridge trace, 396
- Robust regression, 349–357
M-estimators, 349–357
- R*-Student statistic, 222–224
- s^2 , 19, 89
- SAS programs
 MAXR, 190
 NLIN, 435
 PROC REG, 234
 Stepwise, 186
- Scaling and centering, 384
- Selection procedures
 all possible regressions, 193
 backward elimination, 186
 best subsets regression, 185, 193
 forward selection, 185
 Mallows' C_p , 182
 PRESS, 171
- Sequential *F*-test
 in selection procedures, 185
- Sherman–Morrison–Woodbury
 theorem, 459
- Simultaneous inference, 47
- Snee, R.D., 169
- Squared multiple correlation. *See* R^2
- Stabilizing variance, transformations
 for, 286
- Stagewise regression, 185
- Standard error
 of b_0 , 15
 of b_1 , 15
 of prediction, 42, 112
 of \hat{y} , 42, 112
- Stepwise regression, 186
- Straight line regression
 ANOVA, 27
 assumptions, 9, 11, 19, 21
 normal equations, 12
- Studentized residuals, 63, 221, 222
- Sum of squares (SS)
 breakup of total SS, 22, 95
 for lack of fit, 118
 for pure error, 118
 for regression, 22, 95
 for residual, 19
- Tables
 Chi square, 484
 Durbin–Watson test, 485
F-test, 477–480
 normal distribution, 475
 outlier test, 481–482
 rankits, 483
t-test, 476
- Taylor series expansion, 309, 427, 470
- Tidwell, P.W.
 and Box, 307–309
- Transformations
 hazards of, 297–299
 to linearize, 444
 logarithmic, 295
 power, 307, 310
 reciprocal, 296
 on regressors
 Box and Tidwell, 307
 on response, 310
 to stabilize variance, 286
- t*-table, 476
- t*-test
 for intercept, 30
 for regression coefficient, 98
 for slope, 30
- Underspecification, 178
- Validation of model, 167
 data splitting, 169
 PRESS, 170
- Variance
 of intercept, 15
 of predicted value, 43, 112
 of slope, 15
- Variance-covariance matrix, 91
- Variance decomposition proportions,
 371–372
- Variance inflation factor (VIF), 127,
 369
- Variance stabilizing transformation,
 286
- Weibull growth model, 439
- Weighted least squares, 279–281
 example, 281
- Working–Hotelling confidence
 interval, 49
- x*-variables
 centered and scaled, 369, 384
 transformations of, 307