MULTIPLE LINEAR REGRESSION MODELS

More than one predictor variable: $x_1, x_2, \ldots, x_p$

e.g., y =pollutant level, x_1 =windspeed, x_2 =temperature, ...

n observations: Responses y_i, predictors $x_{i1}, \ldots, x_{ip}$

Linear regression model: extend one-predictor model to

$$y_i = \alpha + \beta_1 x_{i1} + \cdots \beta_p x_{ip} + \epsilon_i$$

for observation (or "case") $i = 1, 2, \ldots, n$

Write $\beta_0 = \alpha$ for intercept

Vector notation:

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i$$

- column vector: $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})'$
- regression parameter vector: $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$

MATRIX NOTATION

- Response vector: $\mathbf{y}$ has elements y_i
- Design matrix: $\mathbf{X}$ has rows $\mathbf{x}_i'$ ... n rows, $k = p + 1$ columns
- Error vector: $\boldsymbol{\epsilon}$ has elements ϵ_i

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

- j^{th} column of $\mathbf{X}$: n values of regressor variable j

MODEL IN MATRIX NOTATION:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\boldsymbol{\epsilon}$ has multivariate normal error distribution, zero mean vector and *variance-covariance matrix* $\sigma^2 I_n$
- So $\mathbf{y}$ is multivariate normal with mean vector $\mathbf{X}\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2 I_n$

Variance-covariance matrix: collects variances and covariances

EXAMPLE: Pollutant data

Regression of `oxid` on `temp` and `wind`

- $p = 2, k = 3, \quad \mathbf{x}_i = (1, temp_i, wind_i)'$
- β_1 is regression coefficient on `temp`
- β_2 is regression coefficient on `wind`
- $n = 30$ so $\mathbf{y}$ and $\boldsymbol{\epsilon}$ are 30-vectors, $\mathbf{X}$ is 30×3 design matrix

* MEANING OF REGRESSION PARAMETERS *

Different meaning in different models!

For example:

- Straight line regression of `oxid` on `temp`
- Compare with regression of `oxid` on both `temp` and `wind`

LEAST SQUARES ESTIMATES (LSE):

Total sum of squares about the regression line:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2$$

$$Q(\boldsymbol{\beta}) = \epsilon' \epsilon = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

LSE: choose $\hat{\boldsymbol{\beta}}$ to *minimise* $Q(\boldsymbol{\beta})$

FACT:

$$\hat{\boldsymbol{\beta}} = \mathbf{V}\mathbf{X}'\mathbf{y} \quad \text{and} \quad \mathbf{V} = (\mathbf{X}'\mathbf{X})^{-1}$$

note: $\mathbf{V}$ could be named $\mathbf{V}_{\boldsymbol{\beta}}$

EXAMINING MODEL FIT

- Fitted values $\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$
- Residuals $\hat{\epsilon}_i = y_i - \hat{y}_i$
- Residual sum of squares $Q(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \hat{\epsilon}_i^2$

  – measures remaining/residual variation in response data –
- Residual sample variance:

$$s^2 = \sum_{i=1}^{n} \hat{\epsilon}_i^2 / (n-k)$$

- s^2 is a point estimate of σ^2 from fitted model

  with $n-k$ degrees of freedom

R^2 measure of model fit:

Simplest model: no predictors: $y_i = \beta_0 + \epsilon_i$

$\hat{\beta}_0 = \bar{y}$ and residual deviance $Q(\bar{y}) = S_{yy}$ = total sum of squares

LSE fit: proportion of deviance "explained" by model is

$$R^2 = 1 - Q(\hat{\boldsymbol{\beta}})/S_{yy}$$

THEORY FOR INFERENCE: REFERENCE POSTERIOR

Key aspects of the reference posterior for (β, σ^2) :

- LSE $\hat{\boldsymbol{\beta}}$ is MLE *and* reference posterior mean vector for $\boldsymbol{\beta}$
- posterior for $\boldsymbol{\beta}$ is multivariate t with $n-k$ degrees of freedom
- (marginal) posterior for each β_j is Student t with $n-k$ d.o.f.

  $$t_{n-k}(\hat{\beta}_j, s^2 v_j^2)$$

  where $\hat{\beta}_j$ are elements of $\hat{\boldsymbol{\beta}}$ and $v_j^2 = j^{\text{th}}$ diagonal element of $\mathbf{V}$
  As in simple straight line model – key to assessing *significance* of contribution to regression fit of predictor variable j
- s^2 is the posterior estimate of σ^2 – residual variance

HONEST PREDICTION FROM FITTED MODEL

Posterior predictive distribution for a new case: y_{n+1} at new $\mathbf{x}_{n+1}$

$$y_{n+1} = \mathbf{x}'_{n+1}\boldsymbol{\beta} + \epsilon_{n+1}$$

Also a Student t distribution with $n-k$ d.o.f.

$$y_{n+1} \sim t_{n-k}(\hat{y}, s^2 v_y^2)$$

- Mean is $\hat{y} = \mathbf{x}'_{n+1}\hat{\boldsymbol{\beta}}$
- Spread: $s^2 v_y^2 = s^2 + s^2 w^2$ with $w^2 = \mathbf{x}'_{n+1}\mathbf{V}\mathbf{x}_{n+1}$
  - $s^2 w^2$ – posterior uncertainty about $\mathbf{x}'\boldsymbol{\beta}$ alone
    * uncertainty about the "fitted line" alone, ignoring ϵ_{n+1}
    * depends on $\mathbf{x}_{n+1}$, spread is higher for $\mathbf{x}_{n+1}$ far from $\bar{\mathbf{x}}$
  - additional variability $+s^2$ due to ϵ_{n+1}, estimating σ^2 by s^2

CHECKING FOR SIGNIFICANCE: t TESTS

A coefficient's t statistic is $\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$

Tests significance of that coefficient when all other variables are included in the model

Can search sequentially, but this runs into multiple comparisons problem

- Bonferroni adjustments

Model Selection Criteria

Several versions of penalized likelihood criteria:

- Akaike Information Criterion (AIC) $= l(\hat{\theta}) - k$
- Bayesian Information Criterion (BIC) $= l(\hat{\theta}) - \frac{k}{2}\log(n)$
- Notes:
  - $C_p = \text{AIC}$
  - BIC is asymptotic approximation to a Bayes factor

Maximize these versions, but in other cases you minimize, e.g. $-\text{deviance} + 2k$

Stepwise searching algorithms

SUBSET F TESTS

Model A: Fit p predictors plus intercept:

$$k\text{−vector } \boldsymbol{\beta} \text{ where } k = p + 1$$

Focus on any subset of r predictors:

Are they meaningful in this model?

Model B: remove any subset of r predictors:

equivalent to setting r elements of $\boldsymbol{\beta}$ to zero simultaneously

Assess significance of these r predictors in model A by comparing with model B

This is a test of nested models

COMPARING DEVIANCES: AD-HOC IDEA:

- Residual deviances Q_A and Q_B such that $Q_B - Q_A > 0$
- $Q_B - Q_A$ is deviance explained by the r predictors in A but not B
- extra deviance explained "costs" the extra r parameters, so

$$(Q_B - Q_A)/r$$

is the per-parameter extra deviance explained, or the *average change in explained deviance*

- How big is this? Is it significant? Standardise with respect to scale of deviance,

$$f_{obs} = \{(Q_B - Q_A)/r\}/s_A^2$$

where s_A^2 is the residual estimate of σ^2 in the "bigger" model A

THEORETICAL RESULT:

Reference posterior $p(\boldsymbol{\beta}|Y)$ in the "bigger" model A

Write $\boldsymbol{\gamma}$ = the subvector of r parameters in question

- $p(\boldsymbol{\gamma}|Y)$ is a multivariate t_{n-k} distribution in r dimensions
- Contours are ellipses (e.g., $r = 2$)
- Find the contour running through the point $\boldsymbol{\gamma} = \mathbf{0}$
- Find the probability on $\boldsymbol{\gamma}$ values *outside* the contour
- = Prob on $\boldsymbol{\gamma}$ values *at least as extreme as* $\boldsymbol{\gamma} = 0$
- = Prob outside HPD region for $\boldsymbol{\gamma}$ defined by $\boldsymbol{\gamma} = 0$
- $p-$value (tail area) of the hypothesis that $\boldsymbol{\gamma} = 0$

RESULT: $p - value = Pr(F_{r,n-k} > f_{obs})$

F TESTS IN EXPLORING MODELS:

- Fit a model on some chosen predictors
- Add a group of r new predictors: refit model
  - e.g., r parameters for a factor variable
- Compute F test for "significance" of model improvement
- Compare with alternative model "extensions"

DATA ANALYTIC ISSUES IN REGRESSION: RESIDUAL DIAGNOSTICS

Fitted residuals $\hat{\epsilon}_i = y_i - \hat{y}_i$

- Search for structure: Explore plots versus other possible predictors:
- Evidence of *outliers* or non-normal distribution –
  - standardised residuals $r_i = \epsilon_i / s$ "should be" roughly $N(0, 1)$
  - **`qqnorm()`** and plots versus predictors –
- Heteroscedasticity: unequal variances
- Significant correlations between the residuals and some additional predictors e.g., autocorrelations when data are time ordered

DATA ANALYTIC ISSUES IN REGRESSION:
TRANSFORMATIONS

Often transform y and/or some of the x_i variables

Most common, (natural) log transforms of positive data values

- Original response z lognormal, so $y = log(z)$ is normal
- Multiplicative effects: e.g., used car prices z_t
  - assume $z_t = \exp(y_t)$ and $y_t = a + bt + e_t$ with $e_t \sim N(0, \sigma^2)$
  - Ratio of expected prices is depreciation rate between time t and $t + 1 : r = \exp(b)$

DATA ANALYTIC ISSUES IN REGRESSION: COLLINEARITY

Collinearity = Observed correlations among *predictor* variables

e.g., `age` and `mile` of used cars

explore in scatter plots of predictors, via `cor(age, mile)`, for example

- Two or more related predictors measure "same" features
- Model using either one alone may be adequate
- Adding second may explain little additional variation in response
  – check by comparing R^2 or assessing significance
- Adding second may increase posterior variance of coefficients

MULTICOLLINEARITY (continued)

High multicollinearity induces numerical instabilities in computing inverse of $\mathbf{X}'\mathbf{X}$ for model fitting

– results may be suspect –

(PARTIAL) FIXES:

- Explore and identify highly collinear predictors – model selection
- modify predictors — make values *relative* to sample mean
  - e.g., replace `temp` by `temp-mean(temp)`, etc
  - reduces sample correlations between predictors, to some degree
  - may be sufficient to avoid numerical problems

FACT: High sample correlation between two predictors induces a high posterior dependence between corresponding β_j parameters

MORE EXAMPLES OF REGRESSION:

- Polynomial function fitting:
  - e.g., x_1 =time, x_2 = time^2, ... to fit quadratic function of time to response data
- Interactions
- Regression on a factor variable (Mercedes data)
- Several factor predictors to cross-classify response
  - salary data classified by management level, sex of worker, education level
  - ANOVA terminology: Analysis of Variance
  - How much variation is explained by each classifying factor?
- Autoregression (example coming up)

EXAMPLE: Mercedes data

$y_i = \log(\texttt{price})$ regressed on `age` and `model`

Straight line ("depreciation") to predict log(`price)` based on `age`

MODEL 1: Different intercepts across `model` but same depreciation rates

- Define intercepts as

$$\begin{cases} \beta_0 & \text{for model 0 cars} \\ \beta_0 + \beta_1 & \text{for model 1 cars} \\ \beta_0 + \beta_2 & \text{for model 2 cars} \\ \beta_0 + \beta_3 & \text{for model 3 cars} \\ \beta_0 + \beta_4 & \text{for model 4 cars} \end{cases}$$

- depreciation rate parameter β_5

So $k = 6$, and

$$\mathbf{x}_i' = \begin{cases} (1,\ 0,0,0,0,\ age_i) & \text{for model 0 cars} \\ (1,\ 1,0,0,0,\ age_i) & \text{for model 1 cars} \\ (1,\ 0,1,0,0,\ age_i) & \text{for model 2 cars} \\ (1,\ 0,0,1,0,\ age_i) & \text{for model 3 cars} \\ (1,\ 0,0,0,1,\ age_i) & \text{for model 4 cars} \end{cases}$$

- `model` is a FACTOR predictor variable: classifies response into groups – here, regression has a different intercept for each group (= “level” of the factor)
- “dummy variables” (0/1 indicators) in $\mathbf{x}_i$ to select factor levels
- `model` 0 is the *baseline* level of the factor – conventional
- for the other factor levels, β parameters measure *relative* to baseline
- labelling of factor levels: 0 =baseline, etc

Contrasts:

- Dummy variable contrasts are `contr.treatment` in S-Plus (first level is left omitted)
- `contr.sum` first $p-1$ variables get a "dummy" variable, but all dummy variables are -1 for the pth variable
- `contr.helmert` Helmert contrasts – first variable omitted, remaining matrix is -1 along diagonal and upper triangle, lower sub-diagonal is the negative sum of the the rest of the column, below are zeroes – default in S-Plus
- `contr.poly` polynomial contrasts for special ordered factors

Mercedes data MODEL 2: Different depreciation rates too

Modeled as an interaction between age and model:

Depreciation rate parameters

$$= \begin{cases} \beta_5 & \text{for model 0 cars} \\ \beta_5 + \beta_6 & \text{for model 1 cars} \\ \beta_5 + \beta_7 & \text{for model 2 cars} \\ \beta_5 + \beta_8 & \text{for model 3 cars} \\ \beta_5 + \beta_9 & \text{for model 4 cars} \end{cases}$$

so that $k = 10$ and

$$\mathbf{x}_i' = \begin{cases} (1,\ 0,0,0,0,\ age_i, 0, 0, 0, 0) & \text{for model 0 cars} \\ (1,\ 1,0,0,0,\ age_i, age_i, 0, 0, 0) & \text{for model 1 cars} \\ (1,\ 0,1,0,0,\ age_i, 0, age_i, 0, 0) & \text{for model 2 cars} \\ (1,\ 0,0,1,0,\ age_i, 0, 0, age_i, 0) & \text{for model 3 cars} \\ (1,\ 0,0,0,1,\ age_i, 0, 0, 0, age_i) & \text{for model 4 cars} \end{cases}$$

AUTOREGRESSIONS

An important and common model for time series observations

Time series of response values $y_i, i = 1, 2, \ldots, n$

where $i = t =$ time variable (minutes, weeks, years, ...)

AUTO REGRESSION: regress y_i on past values

$$y_i = \beta_0 + \beta_1 y_{i-1} + \ldots + \beta_p y_{i-p} + \epsilon_i$$

i.e., predictors x are lagged values of the time series

- Fundamental class of time series models: AR(p) – autoregressive mode of order p
- Fit using same linear modelling theory and methods

Why autoregression?

- Captures properties of stochastic linear difference equations
- Periodic behaviour of time series (EEG, Econ, Geology, ...)
- Ranges of frequencies
- Turning points, forecasting

AUTOREGRESSIONS and AUTOCORRELATION

Recall sample correlation r between y and x

Now, x =lagged value of y so we use "autocorrelation"

Autocorrelation at lag j : sample correlation between y_i, y_{i-j}