An Introduction to R for the Geosciences: Regression

Gavin Simpson

Institute of Environmental Change & Society
and
Department of Biology
University of Regina

30th April — 3rd May 2013
Simple linear regression

- **Simple linear regression** is a statistical model that assumes a linear relationship between a continuous response variable $y$ and one or more, usually continuous, predictor variables, $X = x_1, \ldots, x_n$

- Three major purposes of such models
  - to describe the linear relationship between $y$ and $X$
  - to determine how much variation (uncertainty) in $y$ can be explained by the relationship with $X$, and
  - to predict new values of $y$ from new values of $X$

- A linear model is linear in its parameters only — the fitted response can be non-linear in the English sense of the word
Simple linear regression

- Consider first the case of a single predictor variable $x$ and its relationship to $y$
- A suitable form for such a model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- We need to estimate two parameters ($\beta_0$ and $\beta_1$)
- $\beta_0$ is the intercept, the mean of the probability distribution of $y$ when $x$ is 0
- $\beta_1$ is often called the slope, it measures the rate of change in $y$ for a per unit change in $x$
- Estimate the parameters using least-squares, solving this model by minimising Residual Sum of Squares

$$RSS = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$
Least-squares

- \( \hat{\beta}_0 \) is the estimate of the intercept
- \( \hat{\beta}_1 \) is the estimate of the slope
- Observed points \((y_i)\) are open circles
- Fitted points \( (\hat{y}_i) \) are filled circles
- Fitted model/line is solid line through \( \hat{y}_i \)
- Dashed lines between \( y_i \) and \( \hat{y}_i \) are the residuals \((\epsilon_i)\)
- Thick black line shows prediction of \( \hat{y}_{new} \) given a new \( x \) value of 2.5
Least-squares

- Data generated from true mean function
- Least squares estimates of mean function

\[ \beta_0 = 0.7 \quad \hat{\beta}_0 = 0.3201 \]
\[ \beta_1 = 0.8 \quad \hat{\beta}_1 = 0.9987 \]
Least-squares

- Data are 20 observations generated from the following model

\[ y_i = 0.7 + 0.8x_i + \varepsilon_i \]

\[ \varepsilon_i \sim N(\mu = 0, \sigma = 1) \]

- Fitted model gives \( \hat{\beta}_0 = 0.3201 \) and \( \hat{\beta}_1 = 0.9987 \)

- F-ratio for this fitted model is 66.43, which has a \( p \)-value of > 0.0001 from a F distribution with 1 and \( n - 2 \) (20) degrees of freedom

- This is equivalent of testing our model against the Null model (null hypothesis) that

\[ y_i = \beta_0 + \varepsilon_i \]

- where \( \beta_0 \) is just the sample mean, \( \bar{y}_i \), i.e. that there is no variation in \( y \) given \( x \)
Assumptions of least squares regression

1. The linear model correctly describes the functional relationship between $y$ and $X$
   - If violated the estimate of predictor variances ($\sigma^2$) will be inflated
   - Incorrect model specification can show itself as patterns in the residuals

2. $x_i$ are measured without error
   - Allows us to isolate the error component as random variation in $y$
   - Estimates $\hat{\beta}$ will be biased if there is error in $X$ — often ignored!

3. For any given value of $x_i$, the sampled $y_i$ values are independent with normally distributed errors
   - Independence and normality of errors allows us to use parametric theory for confidence intervals and hypothesis tests on the F-ratio.

4. Variances are constant along the regression line/model
   - Allows a single constant variance $\sigma^2$ for the variance of the regression line/model
   - Non-constant variances can be recognised through plots of residuals (amongst others) — i.e. residuals get wider as the values of $y$ increase.
Fitting linear models in R

Typical model call & output from R. Next few slides explain the salient results

```r
> mod <- lm(Age ~ Depth, data = agedat)
> summary(mod)

Call:
  lm(formula = Age ~ Depth, data = agedat)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3808  -7.7115   0.7053   6.1577  16.7818

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)    21.2480     3.5626   5.964 2.02e-06 ***
Depth           5.5760     0.3208  17.384  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.131 on 28 degrees of freedom
Multiple R-squared: 0.9152, Adjusted R-squared: 0.9122
F-statistic: 302.2 on 1 and 28 DF,  p-value: < 2.2e-16
```
Fitting linear models in R

- Estimate is $\beta_j$, the model coefficients, on log scale (base $e$)
- For 1m increase in sediment Depth, sediment Age decreases by 5.576kyrs
- $t$-value is the $t$ statistic, the ratio of the estimate and its standard error $t = \frac{\hat{\beta}_j}{\hat{se}_j}$
- $p$-value is probability of achieving a $t$ as large or larger than the one observed under null hypothesis
- Intercept of interest — sediment age at 0m sediment depth

Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 21.2480  | 3.5626     | 5.964   | 2.02e-06 *** |
| Depth      | 5.5760   | 0.3208     | 17.384  | < 2e-16 *** |
Fitting linear models in R

- Residual standard deviation $\hat{\sigma} = 9.131$; a measure of the variance of the residuals
- $r^2$ is the coefficient of determination, the ratio of the variance explained to the total variance; a measure of how much variance is explained

$$r^2 = \frac{SS_{\text{regression}}}{SS_{\text{regression}} + RSS} = 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}}$$

- Adjusted $r^2$ takes into account number of predictors in the model

$$r^2_{\text{adj}} = 1 - \frac{SS_{\text{residual}}/[n - (p + 1)]}{SS_{\text{total}}/(n - 1)}$$

- If we added a redundant predictor to model $r^2$ would increase. $r^2_{\text{adj}}$ attempts to control for this phenomenon

Residual standard error: 9.131 on 28 degrees of freedom
Multiple R-squared: 0.9152, Adjusted R-squared: 0.9122
F-statistic: 302.2 on 1 and 28 DF, p-value: < 2.2e-16
Fitting linear models in R

- $F$ is the $F$-ratio, the ratio of the regression and residual variances (Mean squares)

\[
F = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 / p}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / [n - (p + 1)]} = \frac{\text{MS}_{\text{regression}}}{\text{MS}_{\text{residual}}}
\]

- Probability of $F$ greater than or equal to observed from $F$-distribution with $p$ and $n - (p + 1)$ degrees of freedom

> anova(mod)
Analysis of Variance Table

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth</td>
<td>1</td>
<td>25195.9</td>
<td>25195.9</td>
<td>302.2</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>28</td>
<td>2334.5</td>
<td>83.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
t tests are tests the $H_0$ that $\hat{\beta}_j = 0$

$F$ tests the ratio of variance explained to unexplained

With single predictor, $t$ test for length and $F$ of model are equivalent

More generally we can think of $F$ as comparing

$$y_i = \beta_0 + \varepsilon_i$$

with

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

```r
> mod0 <- lm(Age ~ 1, data = agedat)
> anova(mod0, mod) ## same as anova(mod)
Analysis of Variance Table

Model 1: Age ~ 1
Model 2: Age ~ Depth

<table>
<thead>
<tr>
<th></th>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>29</td>
<td>27530.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>2334.5</td>
<td>1</td>
<td>25196</td>
<td>302.2</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
R’s model formula

- R uses a slightly modified version of the Wilkinson-Rogers (Wilkinson & Rogers (1973; Applied Statistics 22;392–399) notation to symbolically describe statistical models
  ```r
  mod <- lm(Y ~ x1 + x2, data = mydata)
  ```
  - Intercept implied; suppress with \(-1\) or \(+0\)
  ```r
  mod <- lm(Y ~ x1 + x2 - 1, data = mydata)
  ```
  - Interaction terms with \(a : b\)
  ```r
  mod <- lm(Y ~ x1 + x2 + x1:x2, data = mydata)
  ```
  - Can be simplified using \(a * b\)
  ```r
  mod <- lm(Y ~ x1 * x2, data = mydata)
  ```
  - Shortcut to add all variables to model is \(\cdot\) (Careful!)
  ```r
  mod <- lm(Y ~ ., data = mydata)
  ```
  - Polynomials via \(\text{I}(x^2)\) or \(\text{poly}(x, 2)\)
  ```r
  mod <- lm(Y ~ x + I(x^2), data = mydata)
  mod <- lm(Y ~ poly(x, 2), data = mydata)
  ```
R’s model formula

- Note the use of the `data` argument. This is a data frame (or list) containing the variables to include in the model
  ```r
  mod <- lm(Y ~ x1 + x2, data = mydata)
  ```
- You never want to do this
  ```r
  mod <- lm(mydata$Y ~ mydata$x1 + mydata$x2)
  ```
- Apart from taking longer to type, the `predict()` method won’t work easily
- Can include functions in formula as `poly(x, 2)` earlier
  ```r
  mod <- lm(Y ~ log(x), data = mydata)
  ```
- Better to do this if you can than transform data & store both transformed and untransformed variable in your data frame. Not least because `predict()` just works
- You can exclude variables too
  ```r
  mod <- lm(Y ~ x1 + x2 - x3, data = mydata)
  ```
Multiple regression

- The simple regression model readily generalises to the situation where we have $m$ predictors not just one.

\[ y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon \]

- Now we have $m + 1$ parameters to estimate, one for intercept and one each for the $m$ predictors $x_m$.

- It is tedious to write all that out, so we collect the $\beta_m$ into a vector $\beta$ and all the predictors (including the intercept, a vector of 1s) into the model matrix, $X$, then rewrite the model as

\[ y = X\beta + \varepsilon \]
Akaike information criterion

- Akaike information criterion (AIC) is an index of fit that takes account of the parsimony of the model by penalising for the number of parameters.
- The more parameters in the model the better the fit — if you have as many parameters as data points then the fit is perfect but the model has no explanatory power! A Trade-off.
- AIC is useful as it explicitly penalises any superfluous parameters in the model by adding $2p$ where $p$ is the number of parameters to the variance or deviance of the model.

$$\text{AIC} = -2 \times \text{max loglikelihood} + 2p$$

- Associated is Bayes information criterion (BIC), which applies a stronger penalty of $p \log n$, where $n$ is number of observations.
- For linear regression the $-2 \times \text{max loglikelihood}$ is $n \log(RSS/n) + \text{constant}$, where $RSS$ is the residual sums of squares.
Akaike information criterion

- We use AIC and BIC to compare two or more nested models
- Nested means that one model is a subset of the other
- The model with the smallest AIC or BIC is to be preferred
- Note that you can get negative values for AIC and BIC. This is fine, just go for the smallest value: e.g. -21.5 is better than -15.4
- Difference in AIC of 2 is expected with a redundant parameter
- Models with AIC differing by 2 or less are effectively the same
- \texttt{AIC()} & \texttt{BIC()} methods can be used to extract IC from fitted model objects

```r
> AIC(mod)
[1] 221.7669
> BIC(mod)
[1] 225.9705
```
ANOVA — the Analysis of Variance

- **ANOVA** is a general statistical technique for partitioning and analysing the variation in a continuous response variable.

- Earlier we used ANOVA to partition the variance in a response variable into components explained by explanatory variables and a residual component not explained by the regression model.

- A slightly more restricted view of ANOVA is that it is a technique for partitioning the variation in a response variable into that explained or unexplained by one or more categorical predictor variables or factors.

- The categories of each factor are the groups or experimental treatments.

- Often the focus is on comparing the mean of the response variable between groups.

- We won’t dwell too much on the distinction between regression and ANOVA — they are effectively the same and in R we use the same fitting function, e.g. `lm()`.
Simple one-way ANOVA

- One-way ANOVA designs deal with only a single factor or predictor variable.
- The single factor comprises 2 or more groups.
- Medley & Clements (1998) studied the response of diatom communities to heavy metals (esp. Zinc, Zn) in streams in the Rocky Mountain region of Colorado, USA.
- They sampled a number of stations (4–7) on six streams known to be polluted by heavy metals.
- Several variables were measured at each station, inc. Zn concentration, diatom species richness and diversity, and proportion of diatom cells belonging to the diatom *Achnanthes minutissima*.
- Zn concentration used to group sites into four categories;
- Is there a difference in species diversity between the four Zn categories?
Rocky mountain diatoms

```r
> diatom <- read.csv("medley.csv")
> diatom$ZINC <- factor(diatom$ZINC, levels = c("BACK","LOW","MED","HIGH"))
> ## Drop some superfluous columns
> diatom <- diatom[, 1:3]
> head(diatom)
STREAM ZINC DIVERSITY
1 Eagle BACK  2.27
2 Eagle HIGH  1.25
3 Eagle HIGH  1.15
4 Eagle MED   1.62
5 Blue BACK   1.70
6 Blue HIGH   0.63
> str(diatom)
'data.frame': 34 obs. of 3 variables:
$ STREAM : Factor w/ 6 levels "Arkan","Blue",..: 4 4 4 4 2 2 2 2 2 2 ...
$ ZINC   : Factor w/ 4 levels "BACK","LOW","MED",..: 1 4 4 3 1 4 1 1 4 3 ...
$ DIVERSITY: num 2.27 1.25 1.15 1.62 1.7 0.63 2.05 1.98 1.04 2.19 ...
> table(diatom$ZINC)
BACK LOW MED HIGH
8   8   9   9
> table(diatom$STREAM)
Arkan Blue Chalk Eagle Snake Splat
7    7    5    4    4    5    6
> with(diatom, table(ZINC, STREAM))
STREAM
ZINC Arkan Blue Chalk Eagle Snake Splat
BACK 0   3   0   1   1   3
LOW  5   0   2   0   0   1
MED  2   2   1   1   1   2
HIGH 0   2   2   2   3   0
```
Rocky Mountain diatoms

```r
> boxplot(DIVERSITY ~ ZINC, data = diatom)
```
Rocky Mountain diatoms — ANOVA

```r
> zn.lm1 <- lm(DIVERSITY ~ ZINC, data = diatom)
> summary(zn.lm1)

Call:
  lm(formula = DIVERSITY ~ ZINC, data = diatom)

Residuals:
     Min      1Q  Median      3Q     Max
-1.03750 -0.22896  0.07986  0.33222  0.79750

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)                1.79750    0.16478   10.909  5.81e-12 ***
ZINCLOW                    0.23500    0.23303    1.008  0.3213
ZINCMED                    -0.07972    0.22647    -0.352  0.7273
ZINCHIGH                   -0.51972    0.22647    -2.295  0.0289 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4661 on 30 degrees of freedom
Multiple R-squared: 0.2826, Adjusted R-squared: 0.2108
F-statistic: 3.939 on 3 and 30 DF,  p-value: 0.01756

> anova(zn.lm1)

Analysis of Variance Table

Response: DIVERSITY
            Df Sum Sq Mean Sq F value Pr(>F)
ZINC         3 2.5666  0.8555   3.9387  0.0176 *
Residuals   30 6.5164  0.2172
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Gavin Simpson (U. Regina) McMaster 2013 30th April — 3rd May 2013 22 / 83
Rocky Mountain diatoms — different parametrisation

- Previous model (zn.lm1) contained an intercept
- To maintain identifiability, need to set one level of ZINC as reference level and express model as differences in mean diversity from this reference level
- If we re-parametrise and drop the intercept, the estimates are the group means

```r
> zn.lm0 <- lm(DIVERSITY ~ ZINC - 1, data = diatom)
> coef(zn.lm0)
ZINCBACK  ZINCLOW  ZINCMED  ZINCHIGH
1.797500  2.032500  1.717778  1.277778
> with(diatom, aggregate(DIVERSITY, list(ZINC = ZINC), mean))
  ZINC     x
1  BACK 1.797500
2  LOW  2.032500
3  MED  1.717778
4  HIGH 1.277778
```
Both models use Treatment contrasts.

Normally, one level is set as baseline and dropped, and contrasts code so as to reflect differences in that level from reference level.

Other contrasts are available, such as Helmert contrasts, see `?contrasts`.

```r
> model.matrix(zn.lm1)

(Intercept) ZINCLow ZINCMed ZINCHigh
1  1   0   0   0
2  1   0   0   1
3  1   0   0   1
4  1   0   1   0
5  1   0   0   0
6  1   0   0   1
7  1   0   0   0
8  1   0   0   0
9  1   0   0   1
10 1   0   1   0

attr("assign")
[1] 0 1 1 1

attr("contrasts")
attr("contrasts")$ZINC
[1] "contr.treatment"
```
Outliers

- **Outlier** — observation which is inconsistent with the rest of the observations in a sample.
- An observation can be an outlier due to the response variable(s) or one or more of the predictor variables having values outside their expected limits.
- Identify outliers at EDA stage for investigation and evaluation, *not* rejection and deletion.
- An outlier may result from
  - incorrect measurement,
  - incorrect data entry,
  - transcription error,
  - recording error,
- Outliers are model dependent
- Two main concepts
  - **Leverage** — Potential for an outlier to be influential
  - **Influence** — Observation is influential if its deletion substantially changes the results
Leverage measures

**Projection or Hat matrix**

\[ H = X(X^TX)^{-1}X^T \]

where \( X \) is the \( n \times p \) matrix of \( x \) values, the parameters in the model. \( H \) is an \( n \times n \) matrix.

**Hat matrix**

\[
H = \begin{bmatrix}
h_{11} & h_{12} & \cdots & h_{1n} \\
h_{21} & h_{22} & \cdots & h_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
h_{n1} & h_{n2} & \cdots & h_{ii}
\end{bmatrix}
\]

- Hat matrix is so called because it puts a hat on \( Y \): \( \hat{Y} = HY \).
- **Leverage** of an observation \( i \) is denoted \( h_{ii} \) — the \( i \)th element of the diagonal of \( H \).
- Leverage ranges from \( 1/n \) to 1.
- Observation has high leverage if \( h_{ii} \) is 2 or 3 times \( h = (k + 1)/n \), where \( k + 1 \) is number of coefficients (inc. the constant term).
- As \( h_{ii} \to 1 \), \( x_i \) may dominate model.
Influence measures — DFBETAS

- An observation that combines “outlyingness” with high leverage exerts an influence on the estimated regression coefficients.
- If such an observation is deleted from the analysis, the estimated coefficients change substantially.

\[
dfbeta_{ij} = \beta_j(-i) - \beta_j
\]

\[
dfbetas_{ij} = \frac{\beta_j(-i) - \beta_j}{s_r(i) \sqrt{(XTX)_{jj}}} = \frac{\beta_j(-i) - \beta_j}{s_r(i) \sqrt{(XTX)_{jj}}}.
\]

- \(dfbeta_{ij}\) assesses the impact on the \(j\)th coefficient of deleting the \(i\)th observation.
- The \(dfbeta_{ij}\) are expressed in the metric of the coefficient.
- A standardised version, \(dfbetas_{ij}\) divides \(dfbeta_{ij}\) by the standard error of \(\beta_j\).
- Influential observations have \(dfbetas_{ij} \geq 2/\sqrt{n}\).
### Influence measures — Cook’s distance

#### Cook’s Distance

Cook’s Distance

\[
D_i = \frac{e_i^2}{s^2(k + 1)} \times \frac{h_i}{1 - h_i}
\]

where \( e_i^2 \) is the squared residual for \( x_i \); \( s^2 \) is the variance of the residuals; \( h_i \) is the hat value for \( x_i \).

- One problem with dfbetas_{ij} is that there are so many numbers!
- One for each observation for every \( \beta_j \) (inc. the constant); \( n \times (k + 1) \).
- \( D_i \) is a scale invariant measure of distance between \( \beta_j \) and \( \beta_j(-i) \).
- The first fraction is a measure of “outlyingness”, the second of leverage.
- \( D_i \geq 4/(n - k - 1) \) suggested as a cut-off for high values of \( D_i \).
Leverage and influence; example

Intercept: 1.066
Slope: 4.151

Intercept: 2.026
Slope: 1.676

Cook's Distance

Cook's Distance

Gavin Simpson (U. Regina)  McMaster 2013  30th April — 3rd May 2013  29 / 83
Influence measures in R

Several functions extract influence measures from fitted models; see `?influence.measures` for details.

```r
> head(cooks.distance(mod))
  1     2     3     4     5     6
0.0002843771 0.1987126125 0.2084128586 0.1427614594 0.0092847760 0.0084433941
> head(hatvalues(mod))
  1     2     3     4     5     6
0.14468969 0.11994389 0.09996128 0.09991500 0.08264498 0.05635662
> influence.measures(mod)
Influence measures of
  lm(formula = Age ~ Depth, data = agedat):

             dfb.1_ dfb.Dpth dffit cov.r  cook.d  hat inf
1 -0.023418  0.02055 -0.0234 1.257 2.84e-04 0.1447 *
2 -0.652525  0.55579 -0.6541 0.981 1.99e-01 0.1199
3  0.675657 -0.55622  0.6813 0.896 2.08e-01 0.1000
4 -0.546052  0.44948 -0.5506 0.985 1.43e-01 0.0999
....
```
Model selection

- Where we have several candidate covariates for inclusion in a model, we face the problem of selecting a minimal, adequate model.
- A minimal, adequate model is one that is complex enough to provide sufficient fit to the observed response but no more complex than is necessary.
- Several automated techniques available to help:
  1. Best subsets regression — fit all combination of covariates and choose the best model.
  2. Forward selection — start with no covariates, add the covariate that improves fit most, repeat till no covariate results in significant improvement.
  3. Backwards elimination — as above but start with all covariates and remove the worst variable as long as the model is not made significantly worse.
  4. Stepwise regression (forward selection and backward elimination).
- Regardless of method used to select a minimal model, you must be aware that these techniques are not a panacea.
- \( p \)-values from tests on the selected model do not account for the selection procedure; anti-conservative, too many variables selected.
Stepwise regression in R

- Base R contains several functions for stepwise selection
  - `step()`
  - `add1()`
  - `drop1()`

- The latter two allow manual selection by single-term addition (`add1()`) or deletions (`drop1()`)

- `step()` is fully automated

- All do selection using AIC not \( p \) values

- Package **MASS** contains
  - `stepAIC()`
  - `addterm()`
  - `droptrem()`

- Uses AIC for selection also

- Practical will contain examples of all of these
Best subset regression

- Identifies the best model of each size
- Can use many statistics but AIC and BIC are commonly used

\[
AIC = -2 \times \log(L(\beta_i)|\text{data}) + kp
\]

- \(k\) is a penalty on complexity; AIC: \(k = 2\); BIC: \(k = \log(n)\)
- \(p\) is number of parameters in model.
- Best subsets is available in package leaps
Subset selection and Shrinkage

- Subset selection often used for 2 reasons:
  - **Interpretation** — Smaller subset of predictors with strongest effects on response $y$ may be easier to interpret and explain
  - **Prediction accuracy** — LSQ estimates have low bias but large variance. Can sometimes improve prediction accuracy by shrinking the coefficients or setting some to zero. In doing so we sacrifice a bit of bias

- Subset selection leads to a small set of interpretable predictors, with possibly lower error (MSE) than the full model

- Subset selection is a discrete process — predictors are either in the model, or out

- As a result, this subset model often exhibits high variance, which limits the possible improvement in error

- Shrinkage methods are more continuous than subset selection and do not suffer from high variability to the same degree
Stepwise selection & best-subsets

- **Stepwise selection** is a combination of forward selection and backward elimination steps.

- **Forward selection**: start with no terms in model & sequentially add the variable that best improves the model.

- **Backward elimination**: start with the full model & sequentially remove the variable that effects the model least.

- **Best-subsets**: consider all possible combinations of models (variables) and select the best model for a range of model sizes or select the best model overall.

- Several problems with this however:
  - selection bias in the estimates of the model coefficients $\hat{\beta}_i$,
  - increased variance of the selected model, and
  - bias in the standard errors of $\hat{\beta}_i$. 
Selection bias

- Selection bias occurs in the estimates of the model coefficients $\hat{\beta}_i$ in the selection methods.
- This bias arises from the effective imposition of a hard threshold on the size of the $\hat{\beta}_i$.
- $\hat{\beta}_i = 0$ when $i$th variable is not selected.
- Extreme example from Whittingham et al (2006); 5000 data sets ($n = 10$) drawn from the model:

$$y_i = 1 + 0.8x_i + \varepsilon_i$$

- $\beta = 0.8$, $x_i = 1, 2, \ldots, 10$, $\varepsilon_i \sim N(\mu = 0, \sigma^2 = 1)$
- Selection threshold applied of $\hat{\beta} = 0$ where $p > 0.05$
Ridge regression

- Ridge regression shrinks the coefficients via imposition of a penalty to restrict their size.
- Ridge regression coefficients minimises a penalised RSS

\[ \beta_{\text{ridge}} = \arg\max_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \]

or

\[ \beta_{\text{ridge}} = \arg\max_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 \]

subject to

\[ \sum_{j=1}^{p} \beta_j^2 \leq t \]
Ridge regression

- With collinear variables, $\hat{\beta}_{LSQ}$ are poorly determined and have high variance.
- One variable can have a large positive coefficient, counteracted by variable with which it is correlated having a large negative coefficient.
- Imposing a constraint on size of the coefficients can alleviate this.
- Predictors are standardised before running ridge regression.
- Intercept $\beta_0$ is not subject to the penalty.
- Ridge regression shrinks components in the predictors that have low variance (explain low amounts of the variance in $X$).
Ridge regression

- Need to select a value for the penalty $\lambda$, or for the limit on the size of the coefficients $t$
- Choose these on basis of GCV criterion or CV
- $\lambda = 0$ gives no shrinkage and $\hat{\beta}_{\text{ridge}} = \hat{\beta}_{\text{LSQ}}$
- Ridge regression applied to the Dipper breeding density data:
The Lasso

- The Lasso is a shrinkage method like the ridge regression but with important differences — namely the Lasso can perform variable selection as well as shrink coefficients.
- The lasso finds coefficients $\hat{\beta}_{\text{lasso}}$ that minimise a penalised RSS:

$$\hat{\beta}_{\text{lasso}} = \arg\max_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

or

$$\beta_{\text{lasso}} = \arg\max_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2$$

subject to

$$\sum_{j=1}^{p} |\beta_j| \leq t$$
The Lasso

- The predictors are standardised prior to analysis and the intercept is not subjected to the penalty term.
- Because of the different penalty, if $t$ is sufficiently small (or $\lambda$ sufficiently large) some of the $\hat{\beta}_{lasso}$ can be shrunk to 0.
- This has the effect of selecting those variables with zero coefficients out of the model.
- Optimal values for $t$ or $\lambda$ are chosen using GCV or CV to find those that minimise the prediction error.
- Unlike ridge regression, the lasso doesn’t penalise sets of low variance or correlated variables to the same extent, however.
- It does do feature selection for us.
The Lasso

- Lasso applied to Dipper density data
- Minimum CV error at $\lambda = 0.276$, simpler model within 1 standard error at $\lambda = 0.581$
- 4 predictors have positive coefficients at best model, 3 at the model with 1 standard error
- Gradient (0.019), Stonefly (0.0028), Caddis (0.0004)
The Elastic Net

- Ridge regression shrinks all coefficients, proportionally, whilst the Lasso transforms each coefficient by constant factor $\lambda$ and truncates at zero
- Ridge regression shrinks together the coefficients of correlated data, whilst the Lasso can select or remove coefficients from the model
- Useful if these two properties could be combined
- This is what the Elastic Net penalty does
- Find coefficients $\hat{\beta}_{\text{elastic}}$ that minimise the penalised RSS with penalty

$$
\lambda \sum_{j=1}^{k} \left( \alpha \beta_j^2 + (1 - \alpha) |\beta_j| \right)
$$

- $\alpha$ controls the relative weighting of the ridge-like and lasso-like properties
- Find optimal values of $\lambda$ and $\alpha$ via a grid search over the parameters using CV and 1se rule
Comparison of shrinkage methods: Ozone data

- Various shrinkage methods applied to predict Ozone concentration using climatic variables
- Left panels show full regularisation paths of $\hat{\beta}_i$ for (a) ridge, (b) lasso, and (c) elastic net
- Right panels show $k$-fold CV errors for increasing (left to right) penalty
- Dashed vertical lines indicate best model (lowest CV error) and the smallest model within 1 standard error of the best model (right-most dashed line)
- Notice how ridge regression does not perform selection but shrinks correlated variables (Temperature & Wind Speed)
- Lasso performs selection; note difference in paths for Temperature & Wind Speed
- Elastic net ($\alpha = 0.5$) combines both; most similar here to Lasso
Degrees of freedom for shrinkage models

- The degrees of freedom used in finding the fitted values df(\(\hat{y}\)) is an important of model complexity.
- If we a priori present a set of \(k\) predictors to linear regression, then that model uses \(k + 1\) (for the intercept) df(\(\hat{y}\)).
- If we do best subset regression, software assumes we have used \(k\) df but really we used many more than \(k\).
- What about techniques like the lasso and ridge regression?
- Effective degrees of freedom given by

\[
df(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{Cov}(\hat{y}_i, y_i)
\]

- The harder we try to fit the response \(y_i\), the larger their covariance with the fitted values and therefore the more degrees of freedom we have used.
Degrees of freedom for shrinkage models

- Effective degrees of freedom given by

\[ df(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{Cov}(\hat{y}_i, y_i) \]

- The harder we try to fit the response \( y_i \), the larger their covariance with the fitted values and therefore the more degrees of freedom we have used.

- This equation works for ordinary regression (it will give \( k \) degrees of freedom).

- It works for ridge regression and for the lasso.

- In theory this should also work for best subsets regression, but we don’t have a closed form equation for estimating \( df(\hat{y}) \) in that case.

- Highlights the problem of determining the real \( df(\hat{y}) \) used if we do best subsets or forward selection / backwards elimination.
Multicollinearity redux - VIF

- Variance inflation factor (VIF; \( V \)) is related to the sampling variance of a regression coefficient

\[
\hat{V}(\hat{\beta}_j) = \frac{s^2}{(n - 1)s^2_j} \times \frac{1}{1 - R^2_j}
\]

where \( s^2 \) is estimate error variance, \( s^2_j \) is sample variance of \( j \)th covariate

- \( VIF = \frac{1}{1-R^2_j} \) is the variance-inflation factor and is a function of the multiple correlation \( R_j \) from regression of \( j \)th covariate on the other covariates

- \( \sqrt{VIF} \) is a measure of by how much the confidence interval for \( \hat{\beta}_j \) is expanded relative to the case where uncorrelated data are used

- \( VIF > \sim 10 \) then a covariate is largely explain by other covariates in the model
Multicollinearity redux

- Ridge regression and the lasso estimate biased coefficients
- We accept this extra bias because we attempt to offset the increased variance that complex models and correlated covariates causes
- None of the approaches we talked about is universally a panacea or solution to collinearity
- The real solution is to collect new data so that variables aren’t collinear
- Biased estimation methods *may* cause problems worse than collinearity!
- Really, does collinearity actually matter? If we estimate $\hat{\beta}_j$ with sufficient precision then collinearity doesn’t matter
- If we can’t achieve sufficient precision because of collinearity, this knowledge is only useful if we can redesign the study and collect uncorrelated data
- Think (!) about which terms you introduce to a model
Selected texts


Generalised Linear Models

- Generalised linear models (GLMs) are a synthesis and extension of linear regression plus Poisson, logistic and other regression models.
- GLMs extend the types of data and error distributions that can be modelled beyond the Gaussian data of linear regression.
- With GLMs we can model count data, binary/presence absence data, and concentration data where the response variable is not continuous.
- Such data have different mean-variance relationships and we would not expect errors to be Gaussian.
- Typical uses of GLMs in ecology are:
  - Poisson GLM for count data
  - Logistic GLM for presence absence data
  - Gamma GLM for non-negative or positive continuous data
- GLMs can handle many problems that appear non-linear.
- Not necessary to transform data as this is handled as part of the GLM process.
Structure of a GLM

A GLM consists of three components, chosen/specified by the user:

1. A random component, specifying the conditional distribution of the response $Y_i$ given the values of the explanatory data. Error Function

2. A Linear Predictor $\eta$ — the linear function of regressors

   \[ \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} \]

   The $X_{ij}$ are prescribed functions of the explanatory variables and can be transformed variables, dummy variables, polynomial terms, interactions etc.

3. A smooth and invertible Link Function $g(\cdot)$, which transforms the expectation of the response $\mu_i \equiv E(Y_i)$ to the linear predictor

   \[ g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} \]

   As $g(\cdot)$ is invertible, we can write

   \[ \mu_i = g^{-1}(\eta_i) = g^{-1}(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}) \]
GLM Error Function

- Originally GLMs were specified for error distribution functions belonging to the exponential family of probability distributions.

- Continuous probability distributions
  - Normal (linear regression)
  - Weibull
  - Gamma (data with constant coefficient of variation)
  - Exponential (time to death, survival analysis)
  - Chi-squared
  - Inverse-Gaussian

- Discrete probability distributions
  - Poisson (count data)
  - Binomial (0/1 data, proportions)
  - Multinomial
  - Hypergeometric
  - Pascal

- Choice depends on range of $Y_i$ and on the relationship between the variance and the expectation of $Y_i$. 

Gavin Simpson (U. Regina)  McMaster 2013  30th April — 3rd May 2013  52 / 83
**GLM Error Function**

Characteristics of common GLM probability distributions

<table>
<thead>
<tr>
<th>Probability</th>
<th>Canonical Link</th>
<th>Range of $Y_i$</th>
<th>Variance function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>Identity</td>
<td>$(-\infty, +\infty)$</td>
<td>$\phi$</td>
</tr>
<tr>
<td>Poisson</td>
<td>Log</td>
<td>$0, 1, 2, \ldots, \infty$</td>
<td>$\mu_i$</td>
</tr>
<tr>
<td>Binomial</td>
<td>Logit</td>
<td>$\frac{0,1,\ldots,n_i}{n_i}$</td>
<td>$\mu_i(1-\mu_i)$</td>
</tr>
<tr>
<td>Gamma</td>
<td>Inverse</td>
<td>$(0, \infty)$</td>
<td>$\phi \mu_i^2$</td>
</tr>
<tr>
<td>Inverse-Gaussian</td>
<td>Inverse-square</td>
<td>$(0, \infty)$</td>
<td>$\phi \mu_i^3$</td>
</tr>
</tbody>
</table>

$\phi$ is the dispersion parameter; $\mu_i$ is the expectation of $Y_i$. In the binomial family, $n_i$ is the number of trials.
Ecologically Error Function

Normal errors rarely adequate in ecology; GLMs offer ecologically meaningful alternatives

- **Poisson** — counts; integers, non-negative, variance increases with mean
- **Binomial** — observed proportions from a total; integers, non-negative, bounded at 0 and 1, variance largest at $\pi = 0.5$
- **Binomial** — presence absence data; discrete values, 0 and 1, models probability of success
- **Gamma** — concentrations; non-negative (strictly positive with log link) real values, variance increases with mean, many zero values and some high values
Logistic regression — Darlingtonia

- Timed censuses at 42 randomly-chosen leaves of the cobra lily (*Darlingtonia californica*)
- Recorded number of wasp visits at 10 of the 42 leaves
- Test hypothesis that the probability of visitation is related to leaf height
- Response is dichotomous variable (0/1)
- A suitable model is the logistic model

\[
\pi = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}
\]

- The logit transformation produces

\[
\log_e \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 X_i
\]

- This is the logistic regression and it is a special case of the GLM, with a binomial error distribution and the logit link function
Logistic regression — Darlingtonia

\[
\log_e \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 X_i
\]

- \( \beta_0 \) is a type of intercept; determines the probability of success \((Y_i = 1) \pi \) where \( X = 0 \)
- If \( \beta_0 = 0 \) then \( \pi = 0.5 \)
- \( \beta_1 \) is similar to the slope and determines how steeply the fitted logistic curve rises to the maximum value of \( \pi = 1 \)
- Together, \( \beta_0 \) and \( \beta_1 \) specify the range of the \( X \) variable over which most of the rise occurs and determine how quickly the probability rises from 0 to 1
- Estimate the model parameters using Maximum Likelihood; find parameter values that make the observed data most probable
Logistic regression — Darlingtonia

```r
> mod <- glm(visited ~ leafHeight, data = wasp, family = binomial)
> mod

Call: glm(formula = visited ~ leafHeight, family = binomial, data = wasp)

Coefficients:
(Intercept)  leafHeight
         -7.2930      0.1154

Degrees of Freedom: 41 Total (i.e. Null); 40 Residual
Null Deviance: 46.11
Residual Deviance: 26.96  AIC: 30.96

> ilogit(coef(mod))
    (Intercept)  leafHeight
 0.0006798556  0.5288181121
```
Logistic regression — Darlingtonia

> summary(mod)

Call:
glm(formula = visited ~ leafHeight, family = binomial, data = wasp)

Deviance Residuals:

  Min       1Q   Median       3Q      Max
-2.18274  -0.46820  -0.23897  -0.08519   1.90573

Coefficients:

            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.29295  2.16081  -3.375 0.000738 ***
leafHeight   0.11540  0.03655   3.158 0.001591 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 46.105  on 41  degrees of freedom
Residual deviance: 26.963  on 40  degrees of freedom
AIC: 30.963

Number of Fisher Scoring iterations: 6
Logistic regression — Darlingtonia

Probability of visitation

Leaf height (cm)

Gavin Simpson (U. Regina)
Wald statistics

- $z$ values are Wald statistics, which under the null hypothesis follow a normal distribution
- Tests the null hypothesis that $\beta_i = 0$

$$z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Coefficients:

|             | Estimate | Std. Error | z value | Pr(>|z|) |
|-------------|----------|------------|---------|----------|
| (Intercept) | -7.29295 | 2.16081    | -3.375  | 0.000738 *** |
| leafHeight  | 0.11540  | 0.03655    | 3.158   | 0.001591 ** |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Deviance

- In least squares we have the residual sum of squares as the measure of lack of fitted.
- In GLMs, deviance plays the same role.
- Deviance is defined as twice the log likelihood of the observed data under the current model.
- Deviance is defined relative to an arbitrary constant — only differences of deviances have any meaning.
- Differences in deviances are also known as ratios of likelihoods.
- An alternative to the Wald tests are deviance ratio or likelihood ratio tests.

\[
F = \frac{(D_a - D_b)/(\text{df}_a - \text{df}_b)}{D_b/\text{df}_b}
\]

- \(D_j\) deviance of model, where we test if model A is a significant improvement over model B; \(\text{df}_k\) are the degrees of freedom of the respective model.
A Gamma GLM — simple age-depth modelling

- Radiocarbon age estimates from depths within a peat bog (Brew & Maddy, 1995, QRA Technical Guide No. 5)
- Estimate accumulation rate; assumption here is linear accumulation
- Uncertainty or error is greater at depth; mean variance relationship
- Here, fit mid-depth & mid-calibrated age points

<table>
<thead>
<tr>
<th>SRR</th>
<th>upperDepth</th>
<th>lowerDepth</th>
<th>ageBP</th>
<th>ageError</th>
<th>calUpper</th>
<th>calLower</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRR-4556</td>
<td>20</td>
<td>22.00</td>
<td>355</td>
<td>35</td>
<td>509</td>
<td>307</td>
</tr>
<tr>
<td>SRR-4557</td>
<td>26</td>
<td>28.00</td>
<td>465</td>
<td>35</td>
<td>542</td>
<td>480</td>
</tr>
<tr>
<td>SRR-4558</td>
<td>32</td>
<td>34.00</td>
<td>635</td>
<td>35</td>
<td>671</td>
<td>545</td>
</tr>
<tr>
<td>SRR-4559</td>
<td>38</td>
<td>40.00</td>
<td>740</td>
<td>35</td>
<td>732</td>
<td>666</td>
</tr>
<tr>
<td>SRR-4560</td>
<td>44</td>
<td>46.00</td>
<td>865</td>
<td>35</td>
<td>916</td>
<td>691</td>
</tr>
<tr>
<td>SRR-4561</td>
<td>50</td>
<td>52.50</td>
<td>870</td>
<td>35</td>
<td>918</td>
<td>692</td>
</tr>
<tr>
<td>SRR-4562</td>
<td>56</td>
<td>58.00</td>
<td>985</td>
<td>35</td>
<td>967</td>
<td>795</td>
</tr>
<tr>
<td>SRR-4563</td>
<td>100</td>
<td>108.00</td>
<td>1270</td>
<td>35</td>
<td>1284</td>
<td>1097</td>
</tr>
<tr>
<td>SRR-4564</td>
<td>200</td>
<td>207.00</td>
<td>2575</td>
<td>35</td>
<td>2761</td>
<td>2558</td>
</tr>
<tr>
<td>SRR-4565</td>
<td>260</td>
<td>268.00</td>
<td>3370</td>
<td>35</td>
<td>3697</td>
<td>3487</td>
</tr>
<tr>
<td>SRR-4566</td>
<td>400</td>
<td>407.00</td>
<td>4675</td>
<td>35</td>
<td>5563</td>
<td>5306</td>
</tr>
<tr>
<td>SRR-4567</td>
<td>493</td>
<td>500.00</td>
<td>5315</td>
<td>35</td>
<td>6263</td>
<td>5955</td>
</tr>
</tbody>
</table>
A Gamma GLM — simple age-depth modelling

```r
> plot(calMid ~ midDepth, data = peat,
+ pch = 21, bg = "black")
> m2 <- glm(calMid ~ midDepth, data = peat,
+ family = Gamma(link = "identity"))
> summary(m2)

Call:
glm(formula = calMid ~ midDepth,
family = Gamma(link = "identity"), data = peat)

Deviance Residuals:
           Min          1Q       Median          3Q         Max
-0.196221   -0.012606   -0.001604   0.050645   0.092314

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 181.0393    26.0842   6.941  3.99e-05 ***
midDepth     12.2807     0.5025  24.441  3.00e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.005924447)

    Null deviance: 10.439047  on 11  degrees of freedom
Residual deviance: 0.063394  on 10  degrees of freedom
AIC: 148.83

Number of Fisher Scoring iterations: 4
```

- Linear relationship
- Error increases with mean & Gamma errors
- Identity link function maintains linearity
A Gamma GLM — simple age-depth modelling

```r
> anova(m2, test = "F")
Analysis of Deviance Table

Model: Gamma, link: identity

Response: calMid

Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>11</td>
<td>10.4390</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>midDepth</td>
<td>1</td>
<td>10.376</td>
<td>10</td>
<td>0.0634</td>
<td>1751.3</td>
</tr>
</tbody>
</table>
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
A Gamma GLM — simple age-depth modelling

Linear regression; Gaussian Errors

GLM; Gamma Errors
Scatterplots and local relationships

- In scatter plots, it is not always easy to see the form of the relationship between variables.
- Ozone concentration tends to decrease as wind speed increases.
- But it is difficult to judge whether this relationship is linear or non-linear.
Scatterplots and local relationships

- In scatter plots, it is not always easy to see the form of the relationship between variables.
- Ozone concentration tends to decrease as wind speed increases.
- But it is difficult to judge whether this relationship is linear or non-linear.
- Smoothers model the local patterns in a bivariate scatter plot to illustrate the trends or patterns in the data.
- They determine the pattern from the data themselves rather than from an a priori defined model.
- Loess (or Lowess) is one such smoothing technique.
Lowess — Locally weighted regression

Locally weighted regression scatterplot smoother

- Decide how smooth relationship should be (span or size of bandwidth window)
- For target point assign weights to observations based on adjacency to target point
- Fit linear (polynomial) regression to predict target using weighted least squares; repeat
- Compute residuals & estimate robustness weights based on residuals; well-fitted points have high weight
- Repeat Loess procedure with new weights based on robustness and distance weights

Try different span and degree of polynomial to optimise fit
Lowess — Locally weighted regression

- Two key choices in Loess
- $\alpha$ is the span or bandwidth parameter, controls the size of the window about the target observation
- Observation outside the window have 0 weight
- Larger the window the more global the fit — smooth
- The smaller the window the more local the fit — rough
- $\lambda$ is the degree of polynomial using the weighted least squares
- $\lambda = 1$ is a linear fit, $\lambda = 2$ is a quadratic fit
Lowess — Locally weighted regression

- Two key choices in Loess
- $\alpha$ is the span or bandwidth parameter, controls the size of the window about the target observation
- Observation outside the window have 0 weight
- Larger the window the more global the fit — smooth
- The smaller the window the more local the fit — rough
- $\lambda$ is the degree of polynomial using the the weighted least squares
- $\lambda = 1$ is a linear fit, $\lambda = 2$ is a quadratic fit
Lowess — Locally weighted regression

“In any specific application of LOESS, the choice of the two parameters $\alpha$ and $\lambda$ must be based upon a combination of judgement and trial and error. There is no substitute for the latter”

*Cleveland (1993) Visualising Data. AT&T Bell Laboratories*

- CV can be used to optimise $\alpha$ and $\lambda$ to guard against overfitting the local pattern by producing too rough a smoother or missing local pattern by producing too smooth a smoother
- However, there are techniques with better properties such as splines that have fewer parameters to choose and which are more widely used
- Loess is perhaps most useful as an exploratory technique as part of EDA
- Cleveland, W.S. (1994) The Elements of Graphing Data. AT&T Bell Laboratories
Splines

- Splines are mathematical functions that take their name from the flexible strips of materials draughtsmen used to draw curves.
- A simple spline would just connect the dots, joining each observation to the next — minimal error but rough.
- Impose a penalty ($\lambda$) on the degree of roughness, so fitting the spline balances the error (lack of fit to the data) with the complexity (roughness) of the spline — smoothing spline.
- Smoothing splines useful alternative to Lowess for EDA and scatterplot smoothing.
- Smoothing splines consist of a series of cubic polynomials over intervals of the data, with intervals defined by knots — piecewise cubic polynomial which is continuous as are it’s first a second derivatives.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$$
Splines

- All the smooths covered here are based on splines. Here’s the basic idea . . .

Mathematically the red curve is the function minimizing

\[ \sum_{i} (y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx. \]

Source: Simon Wood
Splines have variable stiffness

► Varying the flexibility of the strip (i.e. varying $\lambda$) changes the spline function curve.

► But irrespective of $\lambda$ the spline functions always have the same basis.

Source: Simon Wood
Regression splines are an alternative type of spline more commonly found in statistical techniques (GAMs).

In smoothing splines, the observations are the knots and the smoothness is controlled by roughness penalty $\lambda$.

In regression splines, a smaller set of knots is chosen across range of the data and cubic polynomials are fitted to the intervals defined by the knots.

As a result, in regression splines the number of knots controls the smoothness of the fitted function.

Once the knots are chosen, regression splines are arguably a parametric approach as we only need to determine the coefficients for the parametric cubic polynomials fitted to each interval.

Regression splines more closely link with formal statistical modelling — can include spline terms in linear regression models and use least squares to estimate parameters.
Basis functions — cubic regression splines

- Cubic regression spline basis function takes value 1 at one knot and 0 at others
- \( j \)th basis function is multiplied by it’s coefficient \( \beta_j \) and then each of these curves is summed at the values of \( x \) to yield the smooth curve
Basis functions — cyclic cubic regression splines

- Where $x$ represents a cyclic variable, want ends points of spline to join up smoothly
- Additional constraints on basis functions; second derivatives must match at $f(x_1)$ and $f(x_k)$ (i.e. knots at end points)
Generalised Additive Models

- Generalised Additive Models (GAMs) are a semi-parametric extension of the GLM

\[ g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \]

- GLM requires an *a priori* statistical model
- What if the response can not be well modelled using the available model forms?
- Despite their flexibility, GLMs may not be flexible enough to approximate the true response adequately
- GLMs are model driven
- GAMs include smooth terms of one or more predictors rather than parametric terms
- The form of the smoothers is derived from the data — GAMs are data driven
Generalised Additive Models

- Generalised Additive Models (GAMs) for a single covariate has the form

\[ g(\mu_i) = \eta_i = \beta_0 + f_1(x_{1i}) \]

- The models are additive as all we assume is that the model terms combine in an additive manner to produce the fitted values of the response

- A GAM consisting of smooth terms for several variables has the form

\[ g(\mu_i) = \eta_i = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_k(x_{ki}) = \beta_0 + \sum_{k=1}^{m} f_k(x_{ki}) \]

- The smooth functions can one of many types of smoother — splines

- Need to specify the type of smoother and complexity of each smoother

- The degree of smoothing for each smooth term can be estimated as part of the model fitting
> require(mgcv)
> m <- gam(strontium.ratio ~ s(age), data = fossil,
+     method = "REML")
> summary(m)

Family: gaussian  
Link function: identity  

Formula:  
strontium.ratio ~ s(age)  

Parametric coefficients:  
Estimate Std. Error t value Pr(>|t|)  
(Intercept) 7.074e-01  2.551e-06 277241 <2e-16 ***  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  

Approximate significance of smooth terms:  
edf Ref.df  F  p-value  
s(age) 8.244   8.84  88 <2e-16 ***  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  

R-sq.(adj) = 0.881  Deviance explained = 89%  
REML score = -930.01  Scale est. = 6.9006e-10  n = 106
Generalised Additive Models

- In all other respects, GAMs are just like GLMs (link functions, error distributions, etc)
- Using modern methods, the degree of smoothing can be determined alongside the other model parameters using ML
- Interactions can be modelled using a smooth function of two or more variables
  \[ g(\mu_i) = \eta_i = \beta_0 + f_1(x_{1i}, x_{2i}) \]
- In above, thin plate splines impose same degree of smoothing on both variables, tensor product smooths allow for different amounts of smoothing
- Cyclic variables may be modelled using a cyclic smoother; the end points of the smoother are forced to match with no discontinuity
Selected texts

Miscellaneous R commands for working with models

- It is recommended to use extractor functions for the model object
- Common extractor and utility functions are:
  - `coef()`: model coefficients
  - `fitted()`: fitted values
  - `resid()`: model residuals
  - `vcov()`: variance-covariance matrix of main model parameters
  - `predict()`: predict from model
  - `extractAIC()`, `AIC()`: AIC of model
  - `logLik()`: log likelihood of fitted model
  - `print()`: quick textual display of object
  - `summary()`: longer textual display of object
  - `plot()`: plot the model diagnostics
  - `add1()`, `drop1()`: add/delete single terms
  - `update()`: refit the model with changes to formula
  - `anova()`: partition variance amongst terms or compare models