Aims of today’s lecture

In today’s lecture we summarise the whole course and give some guidance in studying for the final exam.
The Exam!!

- 25 multiple choice questions, similar to term test (60% for 330, 50% for 762)
- 3 “long answer” questions, similar to past exams
- STATS 330 and STATS 762: You have to do all the multiple choice questions, and 2 out of 3 “long answer” questions
- STATS 762: You have to do a compulsory extra question
- Held on am of Sat 2nd November 2013
Help Schedule

I will be available in the mornings in the week prior to the exam, in Rm 265.
Course summary

The course was about

- Graphics for data analysis
- Regression models for data analysis
Graphics

Important ideas:

- Visualizing multivariate data
  - Pairs plots
  - 3d plots
  - Coplots
  - Trellis plots (Same scales, Plots in rows and columns)

- Diagnostic plots for model criticism
Regression models

- “Ordinary” (normal, least squares) regression for continuous responses
- Logistic regression for binomial/binary responses
- Poisson regression for count responses (log-linear models)
Normal regression

- Response is assumed to be $N(\mu, \sigma^2)$
- Mean is a linear function of the covariates
  \[ \mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \]
- Covariates can be either continuous or categorical
- Observations independent, same variance
Logistic regression

- Response (s “successes” out of n trials) is assumed to be Binomial $Bin(n, \pi)$
- Logit of Probability i.e. $\log(\pi/(1−\pi))$ is a linear function of the covariates
  \[
  \log \frac{\pi}{1−\pi} = \beta_0 + \beta_1x_1 + \cdots + \beta_kx_k
  \]
- Covariates can be either continuous or categorical
- Observations independent
Poisson regression

- Response is assumed to be Poisson with mean $\mu$
- Log of mean $\log(\mu)$ is a linear function of the covariates (log-linear models)

$$
\log(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k
$$

Or, equivalently

$$
\mu = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k).
$$

- Covariates can be either continuous or categorical
- Observations independent
Interpretation of $\beta$ -coefficients

For continuous covariates:

- In *normal* regression, $\beta$ is the increase in mean response associated with a unit increase in $x$, with all other covariates held fixed;

- In *logistic* regression, $\beta$ is the increase in log-odds associated with a unit increase in $x$, with all other covariates held fixed;

- In *Poisson* regression, $\beta$ is the increase in log mean associated with a unit increase in $x$, with all other covariates held fixed;

- In logistic regression, if $x$ is increased by 1, and all other covariates are fixed, the odds are increased by a factor of $\exp(\beta)$:

- In Poisson regression, if $x$ is increased by 1, and all other covariates are fixed, the mean is increased by a factor of $\exp(\beta)$. 
Interpretation of $\beta$ -coefficients (cont)

For categorical covariates (factors), with only main effects in the model:

- In *normal* regression, $\beta$ is the increase in mean response relative to the baseline, with all other covariates held fixed;

- In *logistic* regression, $\beta$ is the increase in log-odds relative to the baseline, with all other covariates held fixed;

- In *Poisson* regression, $\beta$ is the increase in log mean relative to the baseline, with all other covariates held fixed;

- In logistic regression, if we change levels from the baseline, and all other covariates are fixed, the odds are increased by a factor of $\exp(\beta)$:

- In Poisson regression, if we change levels from the baseline, and all other covariates are fixed, the mean is increased by a factor of $\exp(\beta)$. 
Measures of fit

- $R^2$ for normal regression
- Residual deviance (for logistic and Poisson) but not for “ungrouped” data or Poisson with small means ($< 1$).
- Test goodness of fit with the residual deviance, using $\chi^2$ with residual df (grouped/poisson with large mean cases only)
- Use HL test to test goodness of fit for ungrouped data
Prediction

- For normal regression:
  - Predict response at covariates $x_1, \ldots, x_k$
  - Estimate mean response at covariates $x_1, \ldots, x_k$

- For logistic regression:
  - Estimate log-odds at covariates $x_1, \ldots, x_k$
  - Estimate probability of “success” at covariates $x_1, \ldots, x_k$

- For Poisson regression:
  - Estimate the mean at covariates $x_1, \ldots, x_k$
Inference

Summary table:

- Estimates of regression coefs
- Standard errors
- Test stats for coef = 0
- $R^2$ etc (normal regression)
- $F$-test for null model
- Null and residual deviances (logistic/Poisson)
Topics specific to normal regression

Use and interpretation of both forms of anova
- Collinearity
- VIF’s
- Correlation
- Added variable plots
Model selection

- Use and interpretation of both forms of anova:
- Stepwise procedures: FS, BE, stepwise
- All possible regressions approach: AIC, BIC, CP, adjusted $R^2$, CV
Factors

- Baselines
- Levels
- Factor level combinations
- Interactions
- Dummy variables
- Know how to express interactions in terms of means, means in terms of interactions
- Know how to interpret zero interactions
Fitting and choosing models

- Fit a separate plane (mean if no continuous covariates) to each combination of factor levels
- Search for a simpler submodel (with some interactions zero) using stepwise, AIC, and anova (but interpret anova with caution)
Diagnostics

- For non-planar data:
  - Plot res/fitted, res/x’s, partial residual plots, gam plots, box-cox plot
  - Transform either x’s or response, fit polynomial terms

- For unequal variance:
  - Plot res/ fitted, look for funnel effect
  - Weighted least squares
  - Transform response
Diagnostics (2)

- For outliers and high-leverage points:
  - Hat matrix diagonals
  - Standardised residuals,
  - Leave-one-out diagnostics

- Independent observations:
  - Acf plots
  - Residual/previous residual
  - Time series plot of residuals
  - Durbin-Watson test
Diagnostics (3)

For normality:
- Normal plot
- Weisberg-Bingham test
- Box Cox (select power)
Specifics for logistic regression:

Maximal Log-likelihood is

\[
I(\pi_1, \ldots, \pi_m) = \sum_{i=1}^{m} s_i \log(\pi_i) + (n_i - s_i) \log(1 - \pi_i)
\]

\[
I(\beta_0, \ldots, \beta_k) = \sum_{i=1}^{m} s_i (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})
- n_i \log(1 + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})
\]
Deviance

Deviance = 2(\log L_{\text{MAX}} - \log L_{\text{MOD}})

- \log L_{\text{MAX}}: replace \pi’s with frequencies \(s_i/n_i\)
- \log L_{\text{MOD}}: replace \pi’s with estimated \pi’s from logistic model i.e.

\[ \hat{\pi}_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})} \]
Probabilities, odds and log-odds

- Probabilities:

$$Pr[Y = 1] = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}$$

- Odds:

$$\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})$$

- Log-odds

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$
Residuals

- Pearson:
  \[
  \frac{s - n\pi}{\sqrt{n\pi(1 - \pi)}}
  \]

- Deviance residual
  \[
  \text{sign}(s - n\pi)\left\{s(\beta_0 + \beta_1x_1 + \cdots + \beta_kx_k)
  - n\log(1 + \beta_0 + \beta_1x_{i1} + \cdots + \beta_kx_{ik})\right\}^{1/2}
  \]
Specifics for Poisson regression:

- Offsets
- Interpretation of regression coefficients
- Correspondence between Poisson regression (Log-linear models) and the multinomial model for contingency tables
Contingency tables

- Cells 1, 2, \ldots, m
- Cell probabilities $\pi_1, \ldots, \pi_m$
- Counts $y_1, \ldots, y_m$
- Log-likelihood is

$$\sum_{i=1}^{m} y_i \log \pi_i$$
A “model” for the table is anything that specifies the form of the probabilities, possibly up to \( k \) unknown parameters.

Test if the model is OK by

- Calculate Deviance \( = 2(\log L_{\text{MAX}} - \log L_{\text{MOD}}) \)
- \( \log L_{\text{MAX}} \): replace \( \pi \)'s with table frequencies
- \( \log L_{\text{MOD}} \): replace \( \pi \)'s with estimated \( \pi \)'s from the model
- Model OK if deviance is small, (p-value > 0.05). Degrees of freedom \( m - 1 - k \), \( k \) = number of parameters in the model
Independence models

- Correspond to interactions being zero
- Fit a “saturated” model using Poisson regression
- Use AIC, stepwise, anova to see which interactions are zero
- Identify the appropriate model
- Models can be represented by graphs
Odds ratios

- Definition and interpretation
- Connection with independence
- Connection with interactions
- Relationship between conditional OR’s and interactions
- OR’s in Homogeneous association model
Association graphs

- Each node is a factor
- Nodes joined by edges if there is an interaction between the factors
- Interpretation in terms of conditional independence
- Interpretation in terms of collapsibility
Contingency tables: final topics

- Association reversal
  - Simpson’s paradox
  - When can you collapse

- Product multinomial
  - Comparing populations
  - Populations the same if certain interactions are zero

- Goodness of fit to a distribution
  - Special case of 1-dimensional table

- Case-control sampling
Exam hints

- Learn the material on the slides
- Everything in the exam is somewhere in the lecture slides
- Don’t just rely on past exams
- If you don’t understand something, come and ask