EMAT30007 Applied Statistics Lecture 10: Generalised Linear Models (Logistic Regression)

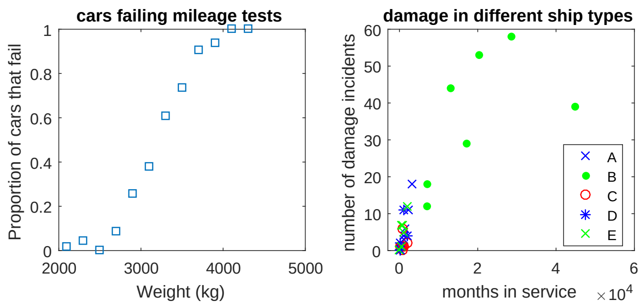
Ksenia Shalonova & Nikolai Bode

Linear models recap

- So far, we have looked at statistical models of the form: $Y \sim N(X\beta, \sigma I)$.
- This is a flexible framework, allowing us to model linear and non-linear relationships between a response and predictors.
- We covered model formulation, model fitting, model checking, model selection, hypothesis tests on model fits, predictions from models and the design of experiments to efficiently collect data for statistical analysis.
- Today, we will look at an even more general class of statistical models than linear models.

Where linear models are not enough

- Example: consider yes/no outcomes or count data:



- The normal distribution for the errors ϵ is not appropriate here.
- Examples: probability of component failing against hours it is in use, count of visits on a website for every hour after new content is added...

Generalised linear models (GLMs)

- There is a more general class of models than linear models, called *Generalised Linear Models (GLMs)*.
  They can be written as:
  $\mathbb{E}(Y_i) \equiv \mu_i = \gamma(X_i\beta)$, $Y_i \overset{independent}{\sim}$ Exponential family distribution,
  where γ is any smooth monotonic function.
- The *Exponential family* of distributions includes distributions such as Poisson, Gaussian (normal), binomial and gamma.
- GLMs are written in terms of the *link function*, g, which is the inverse of γ:
  $g(\mu_i) = X_i\beta, Y_i \overset{independent}{\sim}$ Exponential family distribution.
- Example 1: $Y \sim Binom.,\ \ g(\mu) = ln\left(\frac{\mu}{1-\mu}\right) = X\beta$, logit link function.
- Example 2: Linear models are a special case of GLMs.

Some common link functions and distributions

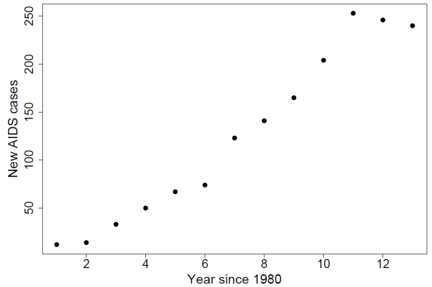
The link functions in the following table are only examples. Other link functions can be used with distributions.

distribution	support; use	link name	link function
normal	$(-\infty, \infty)$; linear response	identity	μ
exponential	$(0, \infty)$; exponential response	inverse	μ^{-1}
gamma	$(0, \infty)$; gamma response	log	$ln(\mu)$
Poisson	$0, 1, 2, \ldots$; count data	log	$ln(\mu)$
binomial	proportions of yes/no occurrences	logit	$ln(\frac{\mu}{1-\mu})$

This is not an exhaustive list — there are additional distributions and link functions.

Example for a GLM

- AIDS cases per year in Belgium at the start of the epidemic.



- Early in an epidemic, an exponential increase in cases can occur:
  $\mathbb{E}(Y_i) \equiv \mu_i = \delta e^{\alpha t_i}, \quad Y_i \sim Poisson(\mu_i).$
- Taking the logarithm of both sides and letting $\beta_0 \equiv log(\delta)$ and $\beta_1 \equiv \alpha$:
  $log(\mu_i) = \beta_0 + \beta_1 t_i, \quad Y_i \sim Poisson(\mu_i),$
  which is a GLM with a log link.

Model fitting in GLMs

- To fit GLMs to data, we use the principle of *Maximum Likelihood Estimation (MLE)*.
- Given parameters β, we can write down $f(Y;\beta)$, the probability or probability density function of the response Y. For observed data, Y_i^{obs}, the likelihood function is: $L(\beta) = \prod_i f(Y_i^{obs};\beta)$.
- For GLMs, there is no closed form solution for the values of β that maximise this function.
- Instead, MLE is performed numerically, using a technique called *iteratively re-weighted least squares (IRLS)*.
- In principle, other optimisation algorithms could also be used for MLE.

GLM assumptions and checking

- Important GLM assumptions:
  - *Independence*.
  - *Distributional assumptions*.
  - *Weak exogeneity* (treat explanatory variables as fixed values).
  - Linear relationship between transformed response and predictors (link function).
- Residual plots are still useful to check if model assumptions hold.
- As we use different distributions, we cannot simply use raw residuals, as for linear models (LMs). Two common types of residuals that attempt to mimick behaviour of residuals for LMs:
  - *Pearson Residuals*.
  - *Deviance Residuals*.
- Plot, e.g. Normal probability plot (useful when response can be approximated with a normal distribution, e.g. Poisson), residuals versus fitted values (trend in mean of residuals violates independence), autocorrelation of residuals, outliers...

Hypothesis tests on GLM fits

- As for LMs, hypothesis tests for individual parameters have been developed. They test the hypothesis that Y does not change as an explanatory changes. In other words, we test hypotheses like:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

  ... we skip the details of these tests (different software may be using different tests).
- As discussed previously, we could also use Likelihood-ratio tests to look at similar hypotheses.
- For global tests on the entire model, the Likelihood-ratio test can be used. E.g. for a model with p parameters, compare the fitted model to the constant model by testing:

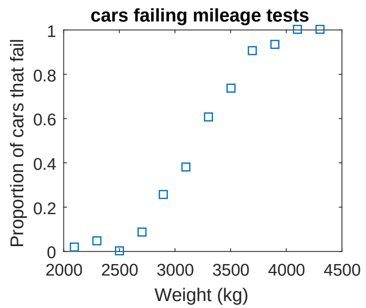
$$H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0$$

Model selection on GLMs

- Model selection for GLMs proceeds in a similar way to what we discussed for LMs. However, some tests that were developed specifically for LMs are not usually approriate.
- AIC and BIC can always be used.
- To compare nested models, the Likelihood-ratio test can be used (versions of the F-test can only be used with great caution).
- Parameter-specific tests are available (although the likelihood-ratio test could also be used for this).
- A similar measure to R^2 exists (*Deviance*). It is based on comparing the likelihood of the model to a *saturated model* with one parameter per data point.

Example: logistic regression

- The proportion of cars of various weights that fail a mileage test.



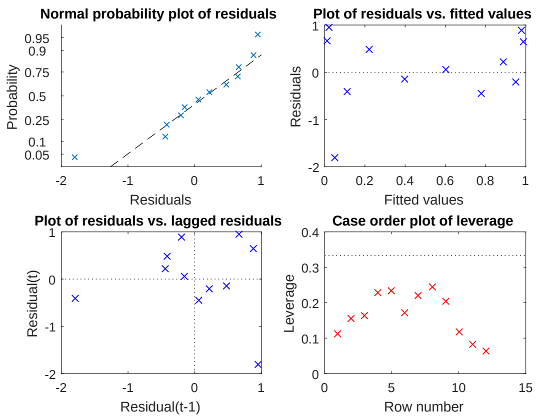
- Data are bounded, so LM is not appropriate and we fit a GLM with $Y_i \sim Binomial(\mu_i)$ and $g(\mu_i) = X_i\beta = \beta_0 + \beta_1 \times weight_i$.
- Here $g(\mu_i) = ln(\frac{\mu_i}{1-\mu_i})$, so $\mu_i = \frac{1}{1+exp(-X_i\beta)}$, the *logistic function*.

Matlab output for GLMs

```
>> mtest = fitglm(weight,[failed tested],'Distribution','binomial')

mtest =


Generalized Linear regression model:
    logit(y) ~ 1 + x1
    Distribution = Binomial

Estimated Coefficients:
                   Estimate        SE         tStat       pValue  
                   _________    __________    ______    __________

    (Intercept)       -13.38         1.394    -9.5986    8.1019e-22
    x1             0.0041812    0.00044258     9.4474    3.4739e-21


12 observations, 10 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 242, p-value = 1.3e-54
```

Model checking in Matlab

- Residual plots using deviance residuals.



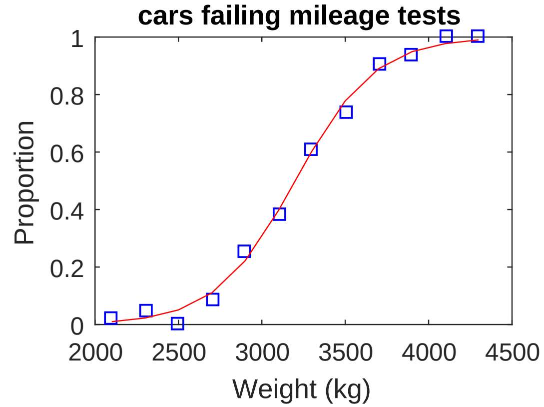
- Measures like *Leverage* and *Cook's distance* can be used to look for outliers in the data (non-examinable).
- **Warning:** Matlab uses raw residuals as default.

University of BRISTOL
Applied Statistics 10
Spring Semester 2018

Interpreting GLM parameters

- From our model formulation and fit, we find that:

$$\mu_i = \frac{1}{1+exp(-[-13.38+0.0042\times weight_i])}$$

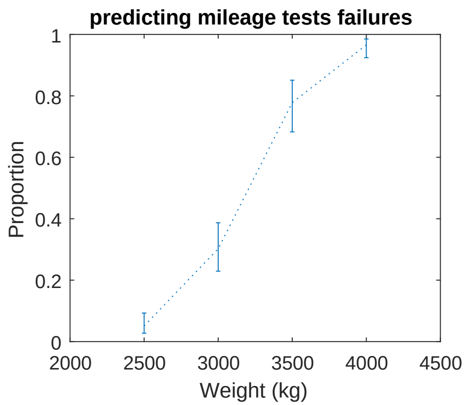


- Unlike for linear models, we cannot read off effect sizes directly from parameter estimates in GLMs. *Need to consider the link function.*

University of BRISTOL
Applied Statistics 10
Spring Semester 2018

Prediction from models

- As for LMs, we can use GLM fits to make predictions.
- Example: for the logisitc regression data.



- **Warning:** as for LMs, be careful with predictions outside of the range covered by the data that was used for model fitting.

University of BRISTOL
Applied Statistics 10
Spring Semester 2018

Typical steps in GLM analysis

1. Look at raw data (scatterplots of response versus different explanatory variables).
2. Identify appropriate distribution and link function for data.
3. Decide on candidate models for the deterministic part of the model (e.g.which predictors are relevant? Exploration versus prediction?).
4. Model selection: find one (or a few) models to look at in more detail.
5. Check model assumptions hold (residual plots).
6. Perform hypothesis tests on model parameters.
7. Interpret findings. Depending on use of model, look at goodness of fit, estimation, prediction... .