Session Sept 27, 2016: Generalized linear models

- Linear regression requires continuous response
- What do we do when we have categorical responses?
- For the simple case of a dichotomous response, we can technically run a linear regression, but

  linear regression assumptions are not met
  - Normality
  - Linearity
  - Homoscedasticity
  - Independence

- and interpretation might be distorted

Example: O-rings

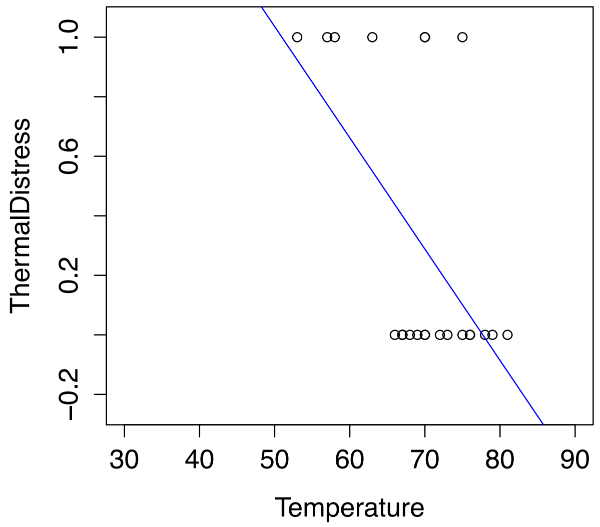
- The 1986 crash of the space shuttle Challenger was linked to failure of O-ring seals in the rocket engines. Data was collected on the 23 previous shuttle missions. The launch temperature on the day of the crash was 31F.
- **Source:** Presidential Commission on the Space Shuttle Challenger Accident, Vol. 1, 1986: 129-131.
- **References:** S. Dalal, E. Fowlkes and B. Hoadley (1989) "Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure." Journal of the American Statistical Association. 84: 945-957.
- Data set with 23 observations on 6 variables
  - Flight No.
  - Temperature
  - Erosion
  - Blowby
  - Total
  - Thermal Distress

FlightNo	Temperature	Erosion	Blowby	Total	ThermalDistress
1	66	0	0	0	0
2	70	1	0	1	1
3	69	0	0	0	0
4	68	0	0	0	0
5	67	0	0	0	0
6	72	0	0	0	0
7	73	0	0	0	0
8	70	0	0	0	0
9	57	1	0	1	1
10	63	1	0	1	1
11	70	1	0	1	1
12	78	0	0	0	0

Example: O-rings

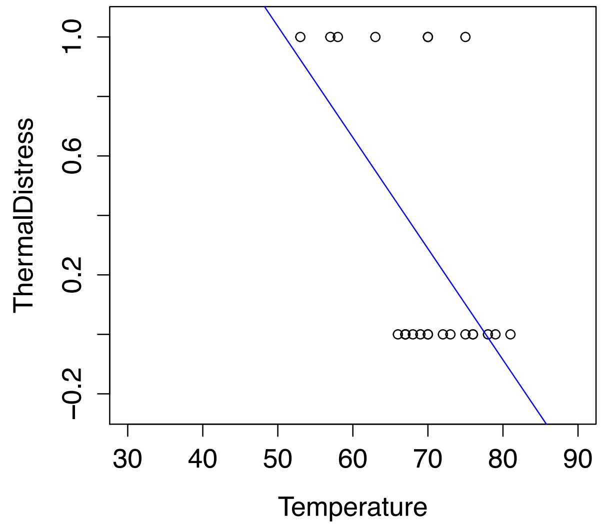
- Linear regression model

```
Call:
lm(formula = ThermalDistress ~ Temperature, data = orings)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.43762 -0.30679 -0.06381  0.17452  0.89881 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  2.90476    0.84208   3.450  0.00240 **
Temperature -0.03738    0.01205  -3.103  0.00538 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.3987 on 21 degrees of freedom
Multiple R-squared:  0.3144,    Adjusted R-squared:  0.2818 
F-statistic:  9.63 on 1 and 21 DF,  p-value: 0.005383
```



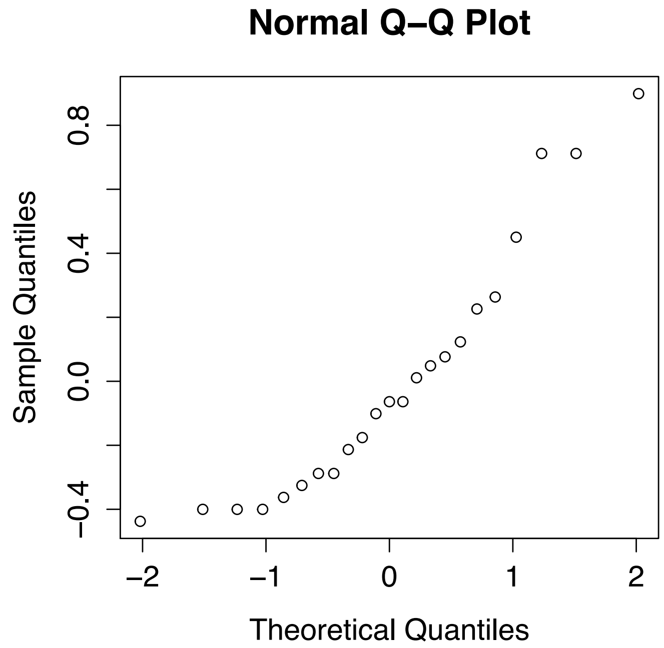
Example: O-rings

- Linear regression model
- Apparently, relationship is non-linear
- Response is either 0 or 1
- Predictions between 0 and 1 can still be interpreted as probabilities of failure (success)
- But predictions above 1 and below 0 make no sense



Example: O-rings

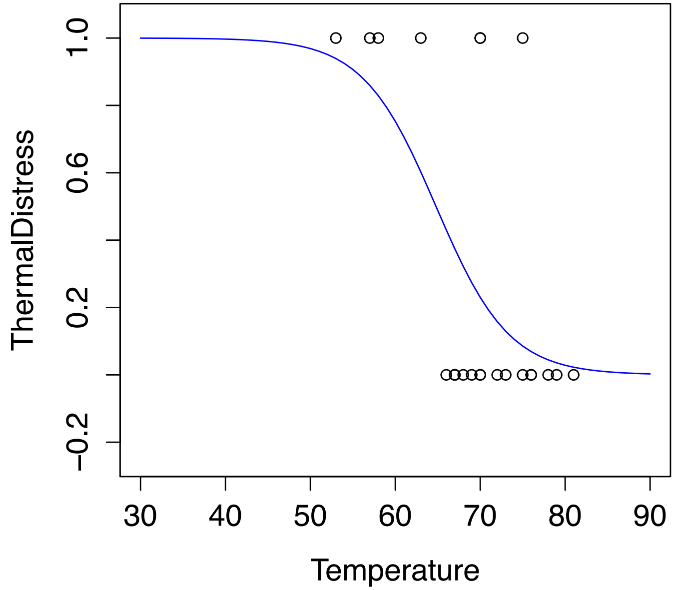
- Linear regression model
- Apparently, relationship is non-linear
- Response is either 0 or 1
- Predictions between 0 and 1 can still be interpreted as probabilities of failure (success)
- But predictions above 1 and below 0 make no sense
- Moreover, residuals are not normal
- Since Y is dichotomous, residuals must have non-constant error variance
  $E[\epsilon] = 0, VAR[\epsilon] = p(x) \cdot (1 - p(x))$
- OLS estimate is inefficient and produces biased standard errors



Example: O-rings

- Logistic regression model (command in R : glm with option family = binomial(logit))

```
Call:
glm(formula = ThermalDistress ~ Temperature, family = binomial(logit),
    data = orings)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0611  -0.7613  -0.3783   0.4524   2.2175

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
Temperature  -0.2322     0.1082  -2.145   0.0320 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.315  on 21  degrees of freedom
AIC: 24.315

Number of Fisher Scoring iterations: 5
```



Logistic Regression vs. linear regression

- Linear Model:

$$mean(ThermalDistress) = \beta_0 + \beta_1 \cdot Temperature + \epsilon$$

- Logistic regression:

$$mean(ThermalDistress) = P(Y = 1) = \frac{e^{\beta_0 + \beta_1 \cdot Temperature + \epsilon}}{1 + e^{\beta_0 + \beta_1 \cdot Temperature + \epsilon}}$$

- Alternative formulation with *p = P(Y=1):*

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot Temperature + \epsilon$$

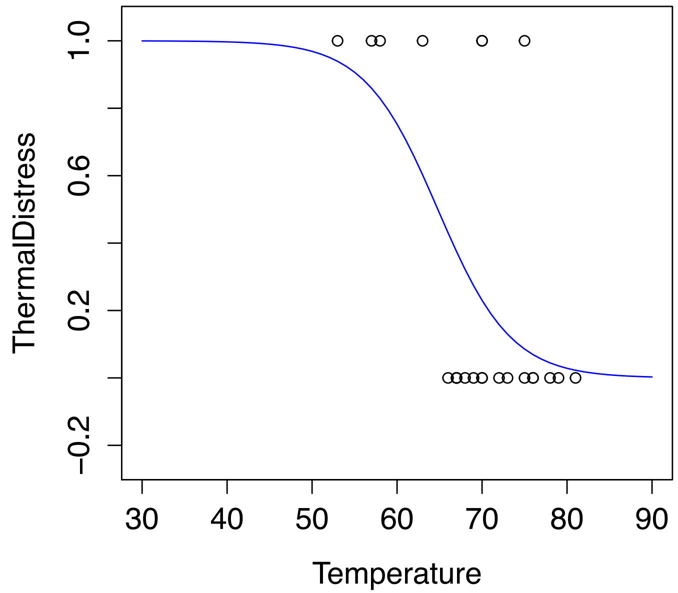
Logistic Regression

- Logistic regression model output looks pretty much the same as linear regression output
- Deviance instead of residual sum of squares
- z-test instead of t-test
- AIC instead of R-squared
- Interpretation of coefficients change
- In logistic regression: slope β_j
  - (all other predictors constant) steepness of logistic curve increases as $|\beta_j|$ increases
  - (all other predictors constant) tangent to logistic curve has slope $\beta_j p(1-p)$
  - (all other predictors constant) for a one unit increase in X the odds for ”success” (i.e. Y=1) change by the factor e^{β_j}

Example: O-rings

- Logistic regression model

```
Call:
glm(formula = ThermalDistress ~ Temperature, family = binomial(logit),
    data = orings)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0611  -0.7613  -0.3783   0.4524   2.2175

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
Temperature  -0.2322     0.1082  -2.145   0.0320 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.315  on 21  degrees of freedom
AIC: 24.315

Number of Fisher Scoring iterations: 5
```



Generalized Linear Models

- Generalized linear models: GLM
  - Linear regression
  - Logistic regression
  - Log-linear models
  - Negative binomial regression

    are all special cases of GLM

Generalized Linear Models

- GLM consists of three components
  - Response distribution/error structure
  - Linear predictor
  - Link function

- The trick of the GLM consists in mapping the mean of the response (expected value of response) to a linear function of the predictors

Generalized Linear Models

- The error structure defines the distribution of the error term or equivalently the conditional distribution of the response
- Standard examples for the error structure are:
  - $N(0, \sigma^2)$ resp. $N(\mu_i, \sigma^2)$ for ordinary least squares regression
  - Bi_{n_i,p_i} with $\mu_i = n_i p_i$ for logistic regression
  - $Po(\mu_i)$ for log-linear models (to come)

Generalized Linear Models

- Linear predictor
- Is a linear function of the predictor variables (linear regression model)

$$\eta = X\beta$$

$$\eta_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Generalized Linear Models

- Link function
- Provides the relationship between the mean of the i-th observation (of response variable) and its linear predictor
- For invertible link functions, model can be re-formulated on the response level

$$
\begin{aligned}
g_i(\mu_i) &= \eta_i \\
&= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k
\end{aligned}
$$

$$
\begin{aligned}
\mu_i &= g_i^{-1}(\eta_i) \\
&= g_i^{-1}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)
\end{aligned}
$$

Logistic regression as GLM

- Dichotomous response
- A set of predictors (continuous and categorical)
- Model formulation on the linear predictor level using the link function

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

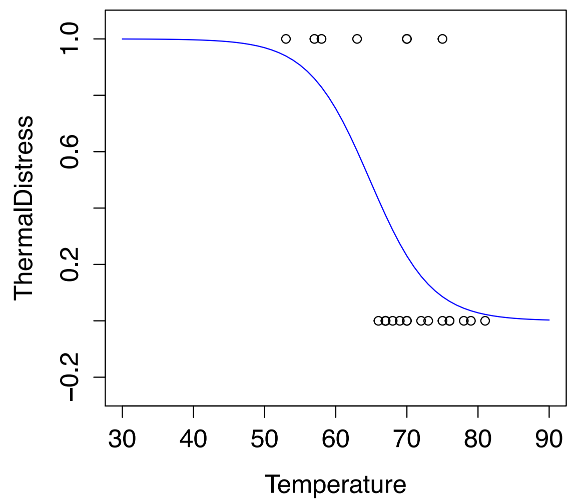
- Model formulation on the response level using the inverse link function

$$P(Y=1) = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)}$$

Example: O-rings again

- Logistic regression model

```
Call:
glm(formula = ThermalDistress ~ Temperature, family = binomial(logit),
    data = orings)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0611  -0.7613  -0.3783   0.4524   2.2175

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
Temperature  -0.2322     0.1082  -2.145   0.0320 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.315  on 21  degrees of freedom
AIC: 24.315

Number of Fisher Scoring iterations: 5
```



A temperature change by 1 degree changes the
- linear predictor (log odds) by 0.2322 units
- the odds by a factor of exp(-0.2322)
- the probabilities by exp(-0.2322)*p*(1-p)

Maximum-Likelihood Estimation

- GLM is based on the maximum-likelihood principle (ML)
- ML is an alternative approach to the least-squares approach (OLS)
- ML requires distributional assumptions
  - OLS works more efficient if distributional assumptions are met, ML really needs them
  - For normally distributed data, OLS and ML estimators usually coincide

Maximum-Likelihood Estimation

- ML provides estimators that have both a reasonable intuitive basis and many desirable statistical properties
- method is broadly applicable and is simple to apply
- general theory of ML comprises estimation, standard errors, statistical tests etc.
- disadvantage: frequently, requires strong assumptions about distribution and structure of data

Maximum-Likelihood Estimation

Example: Flipping a (potentially unfair) coin

- Goal: estimate probability p of getting a head
- flip coin 'independently' 10 times $\longrightarrow$ results in THHTHHHHTH
- probability of obtaining this sequence

$$
\begin{aligned}
Pr(\text{data} \quad | \quad \text{parameter}) &= Pr(THHTHHHHTH) \\
&= (1-p) \cdot p \cdot p \cdot (1-p) \cdot p \cdot p \cdot p \cdot p \cdot (1-p) \cdot p \\
&= p^7 \cdot (1-p)^3
\end{aligned}
$$

- p is fixed, $0 \leq p \leq 1$, but unkown
- method is broadly applicable and is simple to apply
- can treat p as a function of the observed data

Maximum-Likelihood Estimation

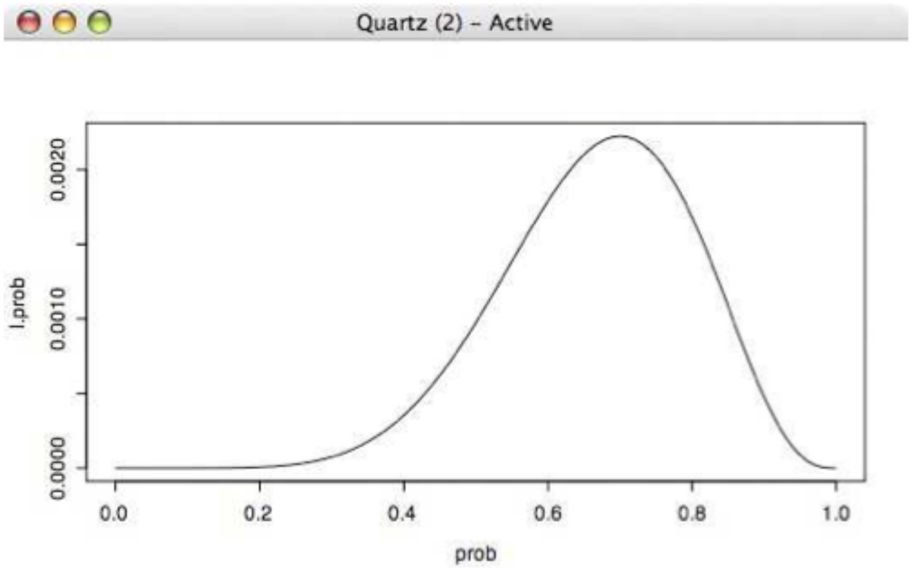
- This function is called the likelihood function

$$L(\text{parameter} \mid \text{data}) \;=\; Pr(p \mid THHTHHHHTH)$$

- Probability and Likelihood function are the same equation, but the roles of parameter and data are switched
- probability is a function of the data with the parameter fixed
- Likelihood is a function of the parameter with the data fixed

Maximum-Likelihood Estimation

Likelihood function for obtaining 7 Heads and 3 Tails when flipping a coin 10 times



The value of p that is most supported by the data is the one for which the likelihood is largest, here $\hat{p} = 0.7$.

Maximum-Likelihood Estimation

In general: Y_i is a binary variable, (female/male; success/failure; pass/fail) with success probability p,

$$
\begin{aligned}
P(Y_i = y_i) &= p^{y_i} \cdot (1-p)^{1-y_i} \\
&= \begin{cases} 1-p & \text{if } y_i = 0 \\ p & \text{if } y_i = 1 \end{cases}
\end{aligned}
$$

Having n independent observations $y_1, ..., y_n$ the probability of the joint event is the product of the individual probabilities

$$
\begin{aligned}
P(Y = (y_1, y_2, ..., y_n)^T | p) &= \prod_{i=1}^{n} P(Y_i = y_i) \\
&= \prod_{i=1}^{n} p^{y_i} \cdot (1-p)^{1-y_i} \\
&= p^{\sum_{i=1}^{n} y_i} \cdot (1-p)^{\sum_{i=1}^{n} (1-y_i)}
\end{aligned}
$$

Maximum-Likelihood Estimation

We have seen some observations $y_1, ..., y_n$, but the parameter p is unknown. A reasonable choice for p is the one value for which the above probability gets maximum. Thus we want to find the value p that maximizes the function

$$L(p|y_1, y_2, ..., y_n) \quad = \quad p^{\sum\limits_{i=1}^{n} y_i} \cdot (1-p)^{\sum\limits_{i=1}^{n} (1-y_i)}$$

This function L is called **likelihood function.**
The estimate $\hat{p}_{ML}$ is called the **maximum-likelihood-estimator** for the success probability p.

Maximum-Likelihood Estimation

It is numerically and analytically easier to maximize the logarithm of the likelihood function. Since the logarithm is a monotonuous transformation the maximum will be obtained at the same parameter value $\hat{p}_{ML}$.

$$\log L(p|y_1, y_2, ..., y_n) \quad = \quad \sum_{i=1}^{n} y_i \cdot \log p + \left(n - \sum_{i=1}^{n} y_i\right) \cdot \log(1-p)$$

Remember calculus: A neccessary condition for the maximum is that the first derivative equals zero.

Maximum-Likelihood Estimation

$$
\begin{aligned}
\frac{\partial \log L(p|y_1, y_2, \ldots, y_n)}{\partial p} &= 0 \\
\Longleftrightarrow \sum_{i=1}^{n} y_i \cdot \frac{1}{p} - \left(n - \sum_{i=1}^{n} y_i\right) \cdot \frac{1}{1-p} &= 0 \\
\Longleftrightarrow \sum_{i=1}^{n} y_i \cdot \frac{1}{p} &= \left(n - \sum_{i=1}^{n} y_i\right) \cdot \frac{1}{1-p} \\
\Longleftrightarrow \sum_{i=1}^{n} y_i \cdot (1-p) &= \left(n - \sum_{i=1}^{n} y_i\right) p \\
\Longleftrightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i \cdot p &= n \cdot p - \sum_{i=1}^{n} y_i \cdot p \\
\Longleftrightarrow p &= \frac{1}{n} \sum_{i=1}^{n} y_i
\end{aligned}
$$

Maximum-Likelihood Estimation

Discrete (qualitative) variables have a discrete density function and we work analogously with the probabilities $P(Y_i = y_i)$.
For continuous (quantitative) variables we have a continuous density function e.g. in the Gaussian (normal distribution) case.
If the density for the variable Y_i is specified by the parameter β, i.e. $f(y_i|\beta)$ then the likelihood function is given as product of the individual densities:

$$
\begin{aligned}
L(\beta|y_1, y_2, ..., y_n) &= \prod_{i=1}^{n} f(y_i|\beta) \\
\log L((\beta|y_1, .., y_n) &= \sum_{i=1}^{n} \log f(y_i|\beta)
\end{aligned}
$$

Maximum-Likelihood Estimation

$L(\hat{\alpha})$ is the value of likelihood function at the MLE $\hat{\alpha}$, while $L(\alpha)$ ist the likelihood for the true (but generally unkown) parameter α.
The *log likelihood-ratio statistic*

$$G^2 = -2\log\frac{L(\alpha)}{L(\hat{\alpha})} = 2[\log L(\hat{\alpha}) - \log L(\alpha)]$$

follows an asymptotic χ^2-Distribution with one degree of freedom. Because by definition the MLE maximizes the likelihood for our sample, $L(\alpha)$ is generally smaller than $L(\hat{\alpha})$.

Recall: Other measures of model quality

- *Akaike Information Criterion AIC*
- *Bayesian Information Criterion BIC* (p = number of parameters in model, n = number of cases)
- penalize complexity of model
- The smaller, the better.

$$
\begin{aligned}
AIC &= -2\text{log likelihood} + 2 \cdot p \\
&= 2\log(\frac{1}{n} RSS) + 2 \cdot p \qquad \text{for OLS}
\end{aligned}
$$

$$
\begin{aligned}
BIC &= -2\text{log likelihood} + \log n \cdot p \\
&= 2\log(\frac{1}{n} RSS) + \log n \cdot p \qquad \text{for OLS}
\end{aligned}
$$

Summary of Part 1

- Logistic regression
- GLM
- Maximum-Likelihood Principle

Thanks for your attention. Enjoy your dinner! See you later!

Session Sept 27, 2016: More on generalized linear models

- GLM – tests
  - Wald test
  - Likelihood ratio test
  - Score test
- Deviance
  - Null model
  - Saturated model
- Short remark on computation
- Error structures and link functions
  - logit and probit model (logistic regression)

Generalized Linear Models

- GLM consists of three components
  - Response distribution/error structure
  - Linear predictor
  - Link function

- The trick of the GLM consists in mapping the mean of the response (expected value of response) to a linear function of the predictors

Generalized linear models: some link functions

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
Log	$\log_e \mu_i$	e^{η_i}
Inverse	μ_i^{-1}	η_i^{-1}
Inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Square-root	$\sqrt{\mu_i}$	η_i^2
Logit	$\log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
Complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

NOTE: μ_i is the expected value of the response; η_i is the linear predictor; and $\Phi(\cdot)$ is the cumulative distribution function of the standard-normal distribution.

Generalized linear models: error structure and link functions

Family	Link function
gaussian	Identity, log, inverse
binomial	Logit, probit, cauchit, log, cloglog
Gamma	Inverse, identity, log
poisson	Log, identity, sqrt
inverse.gaussian	Inverse square, inverse, identity, log
quasi	Logit, probit, cloglog, identy, inverse, log, inverse square, sqrt

as accepted in R

GLM- Tests

Procedures for testing the statstical hypothesis $H_0 : \alpha = \alpha_0$

Wald Test: $Z_0 = \frac{\hat{\alpha}-\alpha_0}{\text{standard error}(\hat{\alpha})}$ is asymptotically distributed as $N(0,1)$ under H_0.

LRT: $G_0^2 = -2\log\frac{L(\alpha_0)}{L(\hat{\alpha})} = 2[\log L(\hat{\alpha}) - \log L(\alpha_0)$ is asymptotically distributed as χ_1^2 under H_0.

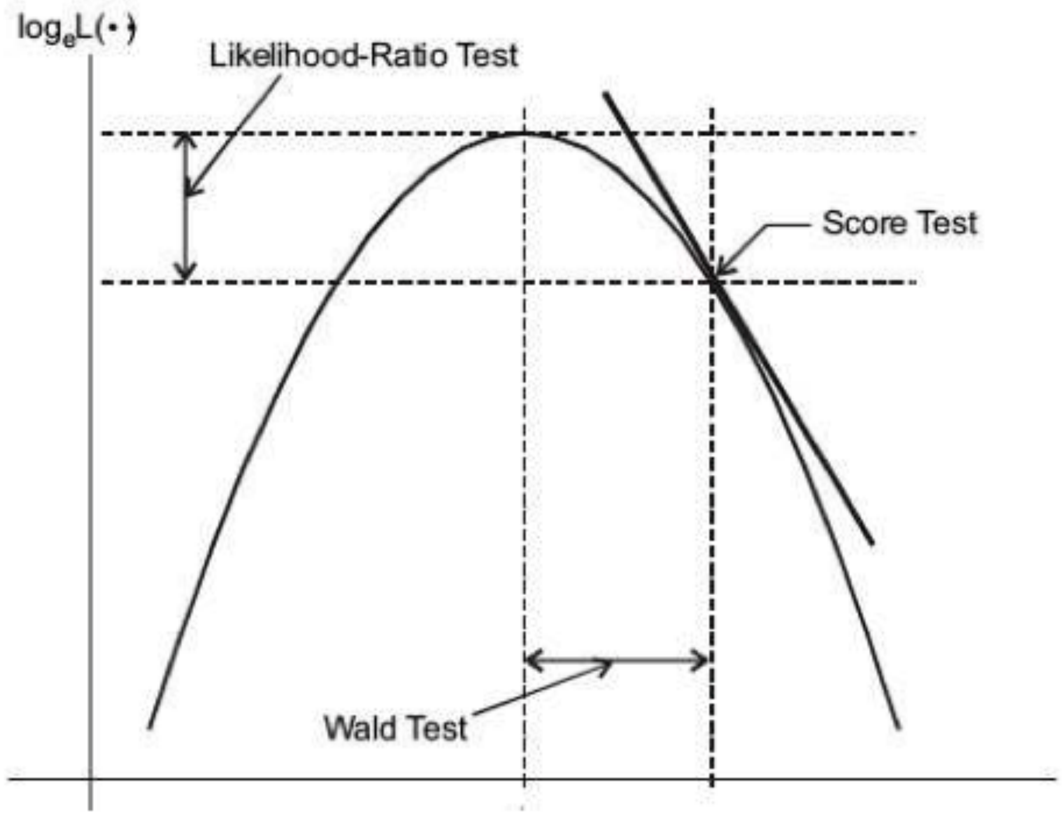
Score Test: The 'score' is the slope of the log-likelihood at a particular value of α, i.e. $S(\alpha) = \frac{\partial}{\partial\alpha}\log L(\alpha)$.

$$\text{score statistic} S_0 = \frac{S(\alpha_0)}{I(\alpha_0)}$$

is asymptotically distributed as $N(0,1)$ under H_0.

Maximum-Likelihood Estimation

Wald-test, Likelihood-ratio test and score test



Null model and saturated model

- The two extreme models that we can build are the null model and the saturated model
- The null model is the one without predictor, we just take the average response as our model prediction
- The saturated model is on the other extreme, it explains the data completely
- Once we add so many predictors that each employee falls alone into a class, we have a "perfect model"
- The saturated model fits the raw data exactly, so no error is left, but it yields no summary (complexity reduction), it just reproduces the data
- Residual deviance is a measure of the difference to the "perfect model"

$$\text{residual deviance} = D_{model} = G^2 = 2(\log L_{saturated} - \log L_{model})$$

Likelihood-ratio-test

Can be used for comparison of two models, very often in comparison against the null model

Likelihood-ratio-test:
$H_0 : b = 0$ independent variable has no effect on probability p, i.e. Y is independent of X

Model 0: $\log(\frac{p}{1-p}) = a$ vs. Model 1 $\log(\frac{p}{1-p}) = a + bX$

calculate likelihood function L_0 and L_1 and log-likelihood-ratio

$$LR = G^2 = -2\log\left(\frac{L_0}{L_1}\right) = (-2\log L_0) - (-2\log L_1)$$

follows approximately $\chi^2_{p_0}$, with p_0 number of parameters in null hypothesis (here: $p_0 = 1$)

Deviance

(Residual) deviance is also often used to compare two models, again very often comparing the model under investigation against the null model

This allows theoretically a similar interpretation as R-squared

$$R^2_{deviance} = 1 - \frac{D_{model}}{D_{null\ model}}$$

Example : O-Rings

```
    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.315  on 21  degrees of freedom
AIC: 24.315
```

$R^2_{deviance}$ = 0.2813

```
> anova(oringF.lg)
Analysis of Deviance Table

Model: binomial, link: logit

Response: ThermalDistress

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev
NULL                           22     28.267
Temperature  1    7.952        21     20.315
> Anova(oringF.lg)
Analysis of Deviance Table (Type II tests)

Response: ThermalDistress
            LR Chisq Df Pr(>Chisq)
Temperature    7.952  1   0.004804 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

More measures for model quality

Pseudo-R^2:

Mc-Faddens $R^2_{MF} = \frac{\log L_0 - \log L_1}{\log L_0} = 1 - \frac{\log L_1}{\log L_0}$

Mc-Faddens adj $R^2_{MF_{adj}} = 1 - \frac{\log L_1 - k}{\log L_0}$

Cox-Snell $R^2_{CS} = 1 - \left(\frac{L_0}{L_1}\right)^{2/n} = 1 - \exp\left(-\frac{LR}{n}\right)$

Nagelkerke $R^2_N = \frac{R^2_{CS}}{\max R^2_{CS}} = \frac{R^2_{CS}}{1 - L_0^{2/n}}$

Information criteria:

AIC $AIC = -2(\log L_1 + k + 1)$

BIC $BIC = -2 \log L_1 + \log(n) \cdot (k + 1)$

Computation and Algorithms

- How are the regression coefficients calculated?
  - For the standard regression, we use

    OLS = ordinary least squares approach
  - For logistic regression etc. we use the

    Maximum-Likelihood Approach
  - OLS can be done by matrix computation in a one-step algorithm
  - ML needs an iterative procedure; typically an *iterative weighted least-squares method* is used to calculate the ML estimations

Logistic regressions: Logit or probit link

- latent variable approach: dichotomous variable result of measurement problem
  - there is a continuous underlying latent variable ($Y*$) which is not measurable
  - we observe the dichotomous indicator (Y) only
- underlying propensity of an individual to vote, but we only observe the outcome
  the underlying model is:

$$Y_i^* = a + bX_i + \epsilon_i$$

Logistic regressions: Logit or probit link

but we only observe the following realizations of Y^*:

$$
\begin{aligned}
Y_i &= 0 \qquad \text{if } Y_i^* \leq a \\
Y_i &= 1 \qquad \text{if } Y_i^* > a
\end{aligned}
$$

We can write:

$$
\begin{aligned}
P(Y_i = 1) &= P(Y_i^* > a) \\
&= P(a + bX_i + \epsilon_i > a) \\
&= P(\epsilon_i > -bX_i) \\
&= P(\epsilon_i \leq bX_i)
\end{aligned}
$$

In words: Y equals 1, if the random part is less than or equal to the systematic part. Problem: what is the underlying probability distribution

Logistic regressions: Logit or probit link

Logit model

Assumption: ϵ follows a standard logistic distribution

- PDF:

$$P(\epsilon) = \frac{exp(\epsilon)}{[1 + exp(\epsilon)]^2}$$

- CDF:

$$\Lambda(\epsilon) = \int_{-\infty}^{\epsilon} P(t)dt = \frac{exp(\epsilon)}{1 + exp(\epsilon)}$$

Logistic regressions: Logit or probit link

Probit model

Assumption: ϵ follows a standard normal distribution

- PDF:

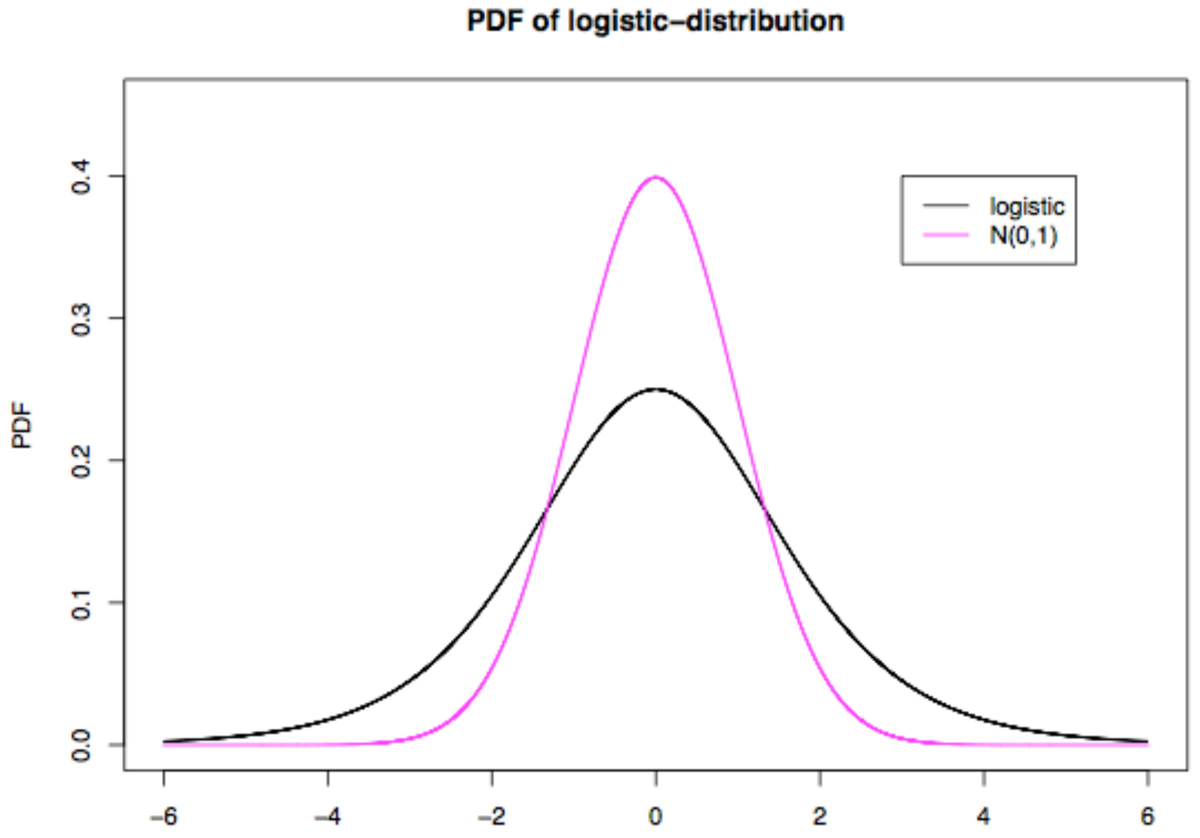
$$P(\epsilon) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{\epsilon}{2}\right)$$

- CDF:

$$\Lambda(\epsilon) = \int_{-\infty}^{\epsilon} P(t)dt = \int_{-\infty}^{\epsilon} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{t}{2}\right) dt$$

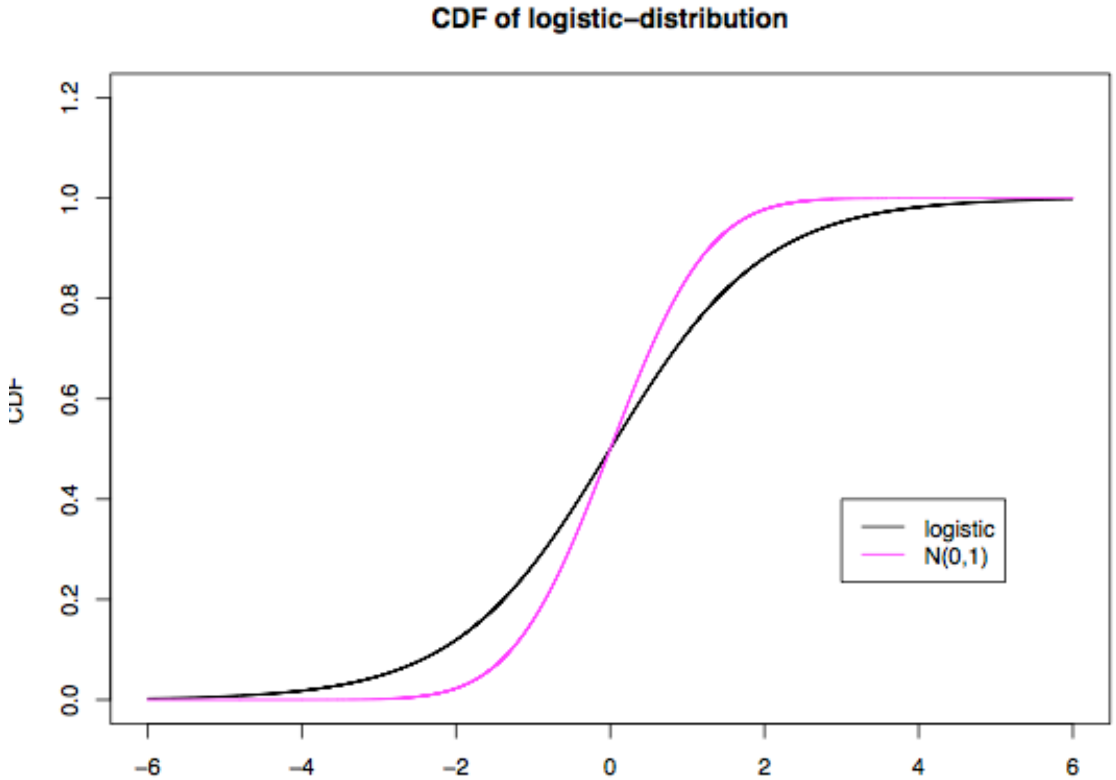
Logistic regressions: Logit or probit link

PDF of logistic vs. normal distribution



Logistic regressions: Logit or probit link

CDF of logistic vs. normal distribution



Logistic regressions: Logit or probit link

Which is better? Logit or Probit?

- Empirically, both yield similar results
- differences occur when many observations fall in the tails of the distribution
- also overall model is similar, parameter estimates differ
  rule of thumb: multiplying the logit estimates by 0.625 makes them comparable to probit estimates

Logistic regressions: Logit or probit link

Which is better? Logit or Probit?

- Empirically, both yield similar results
- differences occur when many observations fall in the tails of the distribution
- also overall model is similar, parameter estimates differ
  rule of thumb: multiplying the logit estimates by 0.625 makes them comparable to probit estimates

Example: O-rings – probit link

```
Call:
glm(formula = ThermalDistress ~ Temperature, family = binomial(probit),
    data = orings)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0831  -0.7930  -0.3747   0.4413   2.2081

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.77490    3.87231   2.266   0.0234 *
Temperature -0.13510    0.05646  -2.393   0.0167 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.378  on 21  degrees of freedom
AIC: 24.378

Number of Fisher Scoring iterations: 6
```

Summary

- GLM extends the linear model to some non-linear relationships between response and predictors
- Uses Maximum-Likelihood approach and requires assumption about distribution of response/error
- Set of canonical error structures and link functions
- Likelihood ratio test
- Wald test

Thanks for your attention. Have a nice evening!