Logistic Regression and GLiM

Philippe Rast

PSC 204B:
General Linear Model
UC Davis, Winter 2018

Topics

Single-level Regression:

Week 1 Linear Regression (G&H: 3,4)
Week 2 Multiple Regression
Week 3 Violation of Assumptions
Week 4 Logistic Regression and GLM (G&H: 5, 6)
Week 5 Model comparison, Over-fitting, Information Criteria (McE: 6)
Week 6 Regression inference via simulations (G&H: 7–10)

Multilevel Regression:

Week 7 Multilevel Linear Models (G&H: 11–13)
Week 8 Multilevel Generalized Models (G&H: 14, 15)
Week 9 Bayesian Inference (G&H: 18 / McE: 1, 2, 3)
Week 10 Fitting Models in Stan and brms (G&H: 16, 17 / McE: 11)

Overview

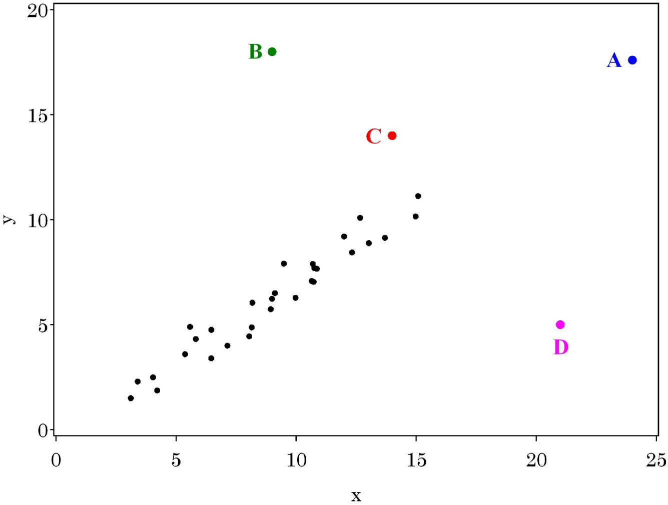
1. Violations of Assumptions cont'd.
    - Outliers, Leverage- and Influential Data Points
        - Hat Matrix
        - Leverage Points
        - Outlier
        - Influential Data Points
2. Corrective Measures
3. Binary Outcome Variables
4. Logistic Regression
    - Odds
    - Logits
    - Nonlinearity
    - Consequences of Nonlinearity
    - Example analysis
5. Summary
6. Generalized Linear Models
    - Review
    - Three Components
        - GLM's in R
        - Example

Outliers, Leverage- and Influential Data Points

- Single data points can have a considerable effect on the estimation of regression coefficients.
- We can differentiate between outliers, leverage points, and influential data points.
    - *Outliers* are observations with large residuals.
    - *Leverage points* are values that lay far outside from the other predictor variables.
    - *Influential points* are values which have a drastic effect on the regression function, i.e, with respect to the estimation of β-weights.
- These points are illustrated in the following plot.

Outliers, Leverage- and Influential Data Points

- Outliers, Leverage- and Influential Data Points



Outliers, Leverage- and Influential Data Points

Table: OLS-estimates with and without A, B, C, D

Parameter		A	B	C	D
$\hat{\beta}_0$	-.61	-.63	-.22	-.99	1.44
$\hat{\sigma}_{\beta_0}$	.34	.28	1.19	.49	.79
$\hat{\beta}_1$	.75	.75	.75	.81	.50
$\hat{\sigma}_{\beta_1}$	.04	.03	.12	.05	.07

- A is a leverage point, but not an outlier nor an influential point.
- B is not a leverage point, but it is an outlier and it is influential (with respect to $\hat{\beta}_0$).
- C is neither a leverage point nor and outlier but it is influential (with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$).
- D is a leverage point and an outlier and influential.

Outliers, Leverage- and Influential Data Points: Hat Matrix

- In order to define these different types of deviating data points we need the *hat matrix* (aka projection matrix, influence matrix).
- The hat matrix projects, geometrically speaking, the y_i-values on to the $\hat{y}_i$-values
- It holds for $\hat{\mathbf{y}}$

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{H}}\mathbf{y} \\ &= \mathbf{H}\mathbf{y}. \end{aligned} \tag{1}$$

- The matrix $\mathbf{H}$ is called “hat” – or projection matrix – as it puts a hat on to the $\mathbf{y}$ vector.
- Hence, $\mathbf{H}$ indicates to what extent $\hat{\mathbf{y}}$ is influenced by $\mathbf{y}$.

Outliers, Leverage- and Influential Data Points: Hat Matrix

- For single $\hat{y}_i$ it holds:

$$\hat{y}_i = \sum_{j=1}^{N} h_{ij} y_j,$$

  with h_{ij} referring to the (i,j) element in $\mathbf{H}$.

- For each person i we sum across the columns to obtain a weighted sum of *all* y-values.
- It holds:

$$\frac{1}{N} \leq h_{ii} \leq 1, \; -1 \leq h_{ij} \leq +1 \quad (i \neq j) \quad \text{and} \quad \sum_{j=1}^{N} h_{ij} = 1.$$

- If $h_{ii} = 1$ (and with it in the ith row all $h_{ij} = 0$, off-diagonal elements are 0), then $\hat{y}_i = y_i$. Large h_{ii}-values indicate that this point is far from the other points in the predictor space. Lit.

Leverage Points

- The reciprocal of h_{ii}, $\frac{1}{h_{ii}}$, can be interpreted as number of observations that determine $\hat{y}_i$
- The size of the h_{ii}-values depends on the relation of the $\mathbf{x}_i'$-vectors to the other rows of $\mathbf{X}$.
- For mean-centered $\mathbf{x}_i^{*\prime}$-vectors this means that h_{ii} is large
  - if $\mathbf{x}_i^{*\prime}\mathbf{x}_i^*$ is large, i.e., when $\mathbf{x}_i^{*\prime}$ has a large (quadratic) distance to the other $N-1$ values
  - because evidently this implies that $\mathbf{x}_i^{*\prime}$ (or rather the corresponding sum of squares) lies far away from the means of the other x-variables.
- **Leverage points** are thus values that generate large h_{ii}-values.
- h_{ii} represents the influence of y_i on fitting the regression model.

Leverage Points

- Unusual combination of predictor values.
- Response value isn't involved in determining leverage
- Identification:
    - Hat statistic

Leverage Point

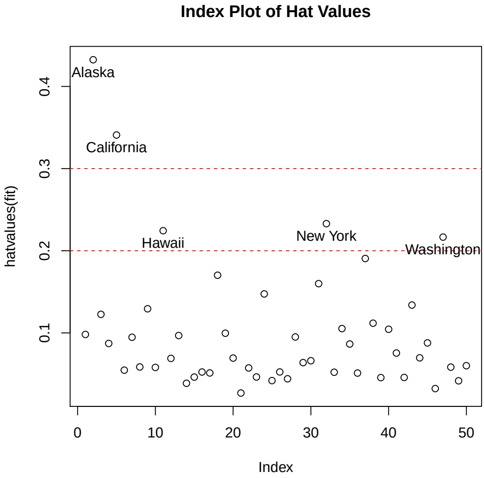
- In our example data $\mathbf{H}$ is:

$$\mathbf{H} = \begin{bmatrix} .34 & .11 & .20 & .27 & .07 & .16 & .03 & .11 & -.07 & -.23 \\ .11 & .46 & .27 & .05 & .30 & -.16 & .01 & -.19 & .07 & .08 \\ .20 & .27 & .22 & .15 & .18 & .01 & .03 & -.03 & .02 & -.04 \\ .27 & .05 & .15 & .22 & .05 & .18 & .07 & .15 & -.01 & -.13 \\ .07 & .30 & .18 & .05 & .22 & -.05 & .06 & -.07 & .10 & .13 \\ .16 & -.16 & .01 & .18 & -.05 & .30 & .15 & .31 & .08 & .03 \\ .03 & .01 & .03 & .07 & .06 & .15 & .14 & .17 & .15 & .19 \\ .11 & -.19 & -.03 & .15 & -.07 & .31 & .17 & .34 & .11 & .09 \\ -.07 & .07 & .02 & -.01 & .10 & .08 & .15 & .11 & .22 & .33 \\ -.23 & .08 & -.04 & -.13 & .13 & .03 & .19 & .09 & .33 & .54 \end{bmatrix} \tag{2}$$

- The highest h_{ii}-value of .54 is obtained for person 10.
- Rule of thumb: If $h_{ii} > {}^{2p}/_{N}$ (p is number of predictors including intercept), then we can consider that data point to be a leverage point. This is the case for person 10.

Leverage

Example with States



- Horizontal lines are drawn at 2 and 3 times the average hat value
- Alaska and California are particularly unusual when it comes to their predictor values
- California has a much higher population than other states, while having a higher income and higher temperature
- These states are atypical compared with the other 48 observations.
- High-leverage observations may or may not be influential observations
- Depends on whether they're also outliers

Outlier, Leverage- and Influential Data Points: $\hat{\boldsymbol{\epsilon}}$

- For the vector with the estimated errors we obtain:

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}} &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} && \text{[Eq. ??]} \\
&= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} && \text{[Eq. 1]} \qquad (3) \\
&= \mathbf{y} - \mathbf{H}\mathbf{y} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{y}. && (4)
\end{aligned}$$

- In some ways this is the opposite of Equation (1).
- For single $\hat{\epsilon}_i$ we define:

$$\hat{\epsilon}_i = y_i - \sum_{j=1}^{N} h_{ij} y_j .$$

- The estimated error for person i results from the difference of y and a weighted sum of all y-values.

Outliers, Leverage- and Influential Data Points

Outliers

- Outliers are observations that aren't predicted well by the model.
- Large residuals $y_i - \hat{y}_i$
- Positive residuals: Underestimation
- Negative residuals: Overestimation

- Identification:
  - Q-Q plot
  - ▷ Rule of thumb: Standardized residuals that are larger than 2 or less than -2 are worth attention
  - R `outlierTest()`

Outlier

- **outliers** are defined as those values that exhibit the largest estimated errors.
- Given our example data

```
> sort(rstudent(fit), decreasing = TRUE)
   Nevada          Alaska          Alabama         Michigan        Missouri...
3.54292864      1.75369166      1.74655197      1.53417975      1.34939968
```

- What value makes a an outlier and actual outlier?
- There are a number of significance tests to address this question, of which we will consider only the studentized deleted residuals.

Outlier: Studentized Deleted Residuals

Leave-One-Out Deletion Diagnostics

- Idea:
  - Delete case i, and refit model.
  - Compute the predicted value and residual for case i using this model.
  - Compute the studentized residual for case i.
- Notation $-i$ means i has been deleted from computation
- $\hat{\epsilon}_{-i} = y_i - \hat{y}_{-i}$ is the deleted residual
- Studentized deleted residuals are defined as

$$\hat{\epsilon}_{-i} = \hat{\epsilon}_i \left(\frac{d_{ii}\hat{\sigma}_{\epsilon}^2(N-k) - \hat{\epsilon}_i^2}{N-k-1} \right)^{-1/2}, \qquad (5)$$

where $\hat{\epsilon}_{-i}$ is the deleted residual for observation i and d_{ii} the corresponding diagonal element from matrix $\mathbf{D}$. For deleted residuals it holds $\hat{\epsilon}_{-i} \sim t_{(N-k-1)}$.

Outlier Example

- States example:

```
> library(car)
> outlierTest(fit)
rstudent unadjusted p-value          Bonferonni p
Nevada        3.5    0.00095                0.048
```

- Nevada is identified as an outlier ($p = 0.048$)
- Tests the single largest (positive or negative) residual for significance as an outlier.

Influential Data Points

- Datapoints are **influential** if their presence considerably changes the estimates of the β-parameters.
- The question here is: How does $\hat{\boldsymbol{\beta}}$ change when a person (i.e. one row in $\mathbf{X}$) is removed?
- A parameter that reflects this idea is defined as

$$\mathrm{DFBETA}_{(-i)} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)} = \frac{\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\hat{\epsilon}_i}{d_{ii}}, \qquad (6)$$

with $\hat{\boldsymbol{\beta}}_{-i}$ being the estimate after removing person i, $\mathbf{x}_i'$ as ith row in the $\mathbf{X}$ matrix and d_{ii} as the corresponding diagonal element from matrix $\mathbf{D}$

- With this we can estimate which person has the largest influence on the estimate of each single β.

Influential Data Points

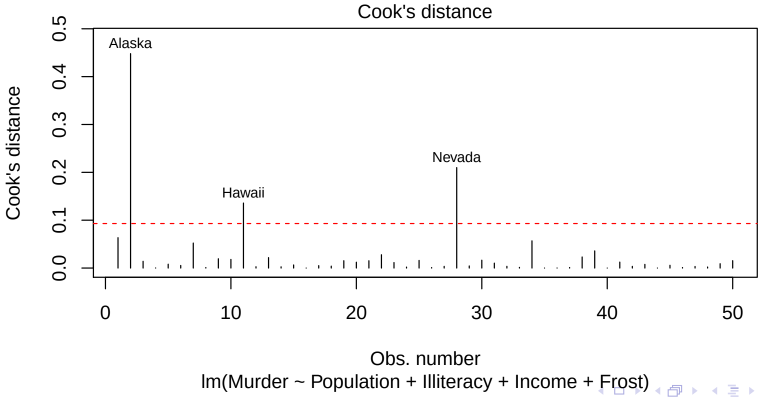
```
> influence.measures(fit)
Influence measures of
lm(formula = Murder ~ Population + Illiteracy + Income + Frost, data = states) :

              dfb.1_ dfb.Pplt  dfb.Illt  dfb.Incm dfb.Frst    dffit    hat inf
Alabama     0.229550 -0.11336  0.051252 -0.175495 -0.25790  0.57632 0.0982
Alaska     -1.380637 -0.47684  0.958717  1.363825  0.49716  1.53087 0.4325   *
Arizona    -0.000308  0.14271  0.000758 -0.093847  0.18275 -0.26417 0.1226
Arkansas    0.016768 -0.00460  0.005077 -0.019669 -0.00269  0.03366 0.0872
California -0.012274 -0.15007  0.047062 -0.003906  0.04913 -0.19859 0.3409   *
Colorado   -0.061437 -0.01292  0.023274  0.054472  0.08468  0.16356 0.0547
Connecticut 0.413307  0.10407 -0.254323 -0.419366 -0.18239 -0.52003 0.0947
...
```

- Alaska and California show up as influential data points (and others...).
- Other measures yield other influential points

Influential Data Points

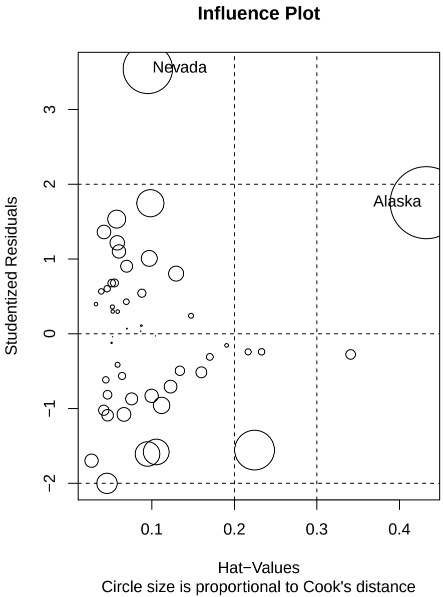
- Cook's Distance (D-statistic): Estimate of influence of a data point when performing OLS regression analysis.
- Rule of thumb for influential observation: Cook's D values $> 4/(n - k - 1)$: n = sample size; k = number of predictor variables
- Note differences compared to DFBETA



Influential Data Points

- The graph identifies Alaska, Hawaii, and Nevada as influential observations
- Deleting these states will have a notable impact on the values of the intercept and slopes in the regression model
- Other popular cutoff: D = 1
  - ▷ Given a criterion of D=1, none of the observations in the dataset would appear to be influential.
- Combine different approaches

Influential Data Points



R: `influencePlot(fit)`

- States above +2 or below -2 on the vertical axis are considered outliers
- States above 0.2 or 0.3 on the horizontal axis have high leverage
- Circle size is proportional to influence
- Observations depicted by large circles may have disproportionate influence on the parameter estimates of the model.

Corrective Measures

- Once we have identified outlier, leverage- and influential data points – how do we deal with them?

  Generally:

  - If these data points resulted from errors in the data collection/handling, then we can eliminate them.
  - Revisit our regression model and its assumptions and maybe revise model. E.g. outliers may be due to interactions with predictors that had not been included in model.
- Deleting observations
- Transforming variables
- Adding or deleting variables
- Using another regression approach

Deleting Observations

- Deleting outliers can improve a dataset's fit to the normality assumption
- **Caution** when considering the deletion of observations
  - Data errors in recording
  - Test protocol was not followed
  - Subject misunderstood instruction...
- Unusual observation may be the most interesting thing about the data!
- Uncovering why an observation differs from the rest can contribute great insight

! Never delete data points without good reason – better to revisit our assumptions than adapt data to our model.

Deleting Observations

```
Coefficients: Original
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.235e+00  3.866e+00   0.319   0.7510
Population  2.237e-04  9.052e-05   2.471   0.0173 *
Illiteracy  4.143e+00  8.744e-01   4.738 2.19e-05 ***
Income      6.442e-05  6.837e-04   0.094   0.9253
Frost       5.813e-04  1.005e-02   0.058   0.9541

Coefficients: Alaska removed
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.453e+00  4.811e+00   1.341  0.18667
Population   2.659e-04  9.171e-05   2.899  0.00582 **
Illiteracy   3.323e+00  9.743e-01   3.411  0.00140 **
Income      -8.472e-04  8.468e-04  -1.001  0.32254
Frost       -4.306e-03  1.022e-02  -0.421  0.67551
```

Transforming Variables

- If normality, linearity, or homoscedasticity isn't met
- Transforming variables (x, y) may correct issue

e.g. Power transformation for non-normality

- Box-Cox transformation with ML-estimates of λ to normalize data

```
> summary(powerTransform(fit))
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
Y1    0.8653           1       0.3853       1.3453

Likelihood ratio tests about transformation parameters
                              LRT df         pval
LR test, lambda = (0) 13.1436087  1 0.0002885017
LR test, lambda = (1)  0.2990438  1 0.5844824648
```

- $y^{0.86}$ would normalize, but hypothesis of $\lambda = 1$ is not rejected.

Assumption of Linearity

- When the assumption of linearity is violated, a transformation of the predictor variables can often help
- Box-Tidwell power transformations: Generate maximum-likelihood estimates of predictor powers that can improve linearity

```
> boxTidwell(Murder~Population+Illiteracy,data=states)
           Score Statistic   p-value MLE of lambda
Population      -0.3228003 0.7468465     0.8693882
Illiteracy       0.6193814 0.5356651     1.3581188

iterations =  19
```

- The results suggest trying the transformations Population$^{.87}$ and Illiteracy$^{.87}$
- However, p-values indicate that transformation is not necessary

Adding or deleting variables

- Changing the variables in a model will impact the fit of the model
- Deleting variables is a particularly important approach for dealing with multicollinearity
- If goal is to make predictions, then multicollinearity isn't a problem as it affects the standard errors
- Need to be dealt with if one needs to interpret individual predictors
- Common approach: Delete one of the variables involved in the multicollinearity
- Alternative: Ridge regression (biasing $\hat{\beta}$ in favor of smaller variances)

Outlook

Things to consider

- Most violations can be addressed and may lead to better insights!
- For example
  - Multicollinearity can be dealt with ridge regression
  - If there are outliers and/or influential observations, you can fit a robust regression model rather than an OLS regression
  - If normality assumption is violated, one may fit a nonparametric regression model
  - If there's significant nonlinearity, you can try a nonlinear regression model
  - If you've violated the assumptions of independence of errors, you can fit a model that specifically takes the error structure into account, such as time-series models or multilevel regression models.
  - Finally, you can turn to generalized linear models to fit a wide range of models in situations where the assumptions of OLS regression don't hold.

Non-continuous Outcome

- So far, y was a continuous variable.
- In some instances dependent variable may be discrete, or categorical
- For dichotomized outcomes (0, 1; yes, no; fail, success) *Logistic Regression* is frequently used
- When dependent variable has more than two outcome categories a starting point is
  - multinomial logistic regression
  - ordinal logistic regression (for ordered categorical variables)
- Goal: Estimate the probability of a binary response based on one or more predictors

Binary Outcome Variables

Example:

- Item from the PGC Morale Scale
- “I have as much pep as last year” □ Yes □ No
- Let

$$Y = \begin{cases} 1 \text{ if answer } = \text{“Yes”} \\ 0 \text{ if answer } = \text{“No”} \end{cases}$$

- Example Data from ILSE

Answer	N	Percent
Yes (Y = 1)	407	70.05%
No (Y = 0)	174	29.95%
Sum	581	100.00%

Binary Outcome Variables

- Question 1: If you meet a person from the sample, what is the probability that this person judges herself to "have as much pep as last year"?
- Answer:
  - Calculate the relative frequency of "Yes"-answers.
  - $p(\text{"No"}) = p(Y = 1) = \frac{407}{581} = .7005.$
  - The probability is about 70%.

- The same relation holds on the population level
  - $\hat{p}(\text{"No"}) = \hat{p}(Y = 1) = \frac{407}{581} = .7005$
  - The hat denotes that the probability is a sample-based estimate of the probability in the population.
  - The estimated probability is about 70%.

Binary Outcome Variables

- How accurate is this estimate?
- We can construct a 95% confidence interval around the estimate
  - Based on the binomial distribution, the (approximative) standard error of p is
    $\hat{\sigma}_p = \sqrt{\frac{p(1-p)}{N}} = \sqrt{\frac{0.7005(1-0.7005)}{581}} = 0.019$
  - for a 95% confidence interval, $z_{(\alpha/2)} = \pm 1.96$
  - CI formula in general
    $$\hat{p} - z \times \hat{\sigma}_p < p < \hat{p} + z \times \hat{\sigma}_p$$
  - $0.7005 - 1.96 \times 0.019 < p < 0.7005 + 1.96 \times 0.019$
  - 95% CI: $0.6632 < p < 0.7378$
  - We are 95% confident that the interval of 0.66 and 0.74 would contain the population value.

Note: Constructing a *symmetrical* confidence interval isn't the best idea as it potentially covers probabilities outside the defined range. Eg. for $p = \frac{2}{20} = .1$ we would obtain a 95% CI:$-.031 < p < .231$

Binary Outcome Variables

- Which variables increase or decrease the probability of answering "Yes"?
- Logistic Regression and Probit Regression
  - Which explanatory variables increase (or decrease) the (estimated) probability of a "Yes"-answer.
  - Like in the General Linear Model (GLM), explanatory variables may be continuous or categorical.
- Problem:
  - Outcome variable is bounded: $0 \leq \hat{p} \leq 1$.
  - If one would use a GLM like

    $$\hat{p} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

    this would typically result in predictions falling outside the interval (0,1).
- How can p be made more amenable for a regression-like model?

Logistic Regression: Odds

- Step 1: Odds
  - Odds are defined as

$$Odds = \frac{\hat{p}}{1-\hat{p}} = \frac{0.7005}{0.2995} = 2.3391.$$

  - For Odds, $0 \leq Odd \leq \infty$ [Odds are bounded by zero].
  - If $Odds < 1$, then $\hat{p} < 0.5$, i.e., the event (here: answer “Yes”) is less likely than its counterpart (here: answer “No”).
  - If $Odds > 1$, then $\hat{p} > 0.5$, i.e., the event (here: answer “Yes”) is more likely than its counterpart (here: answer “No”).
  - If $Odds = 1$, then $\hat{p} = 0.5$, i.e., the event (here: answer “Yes”) is equally likely than its counterpart (here: answer “No”).
- Odds are often rounded to integers or fractions. For example, $2.3391 \approx 2.5 = \frac{5}{2}$.
- The odds of answering to the item with “Yes” are thus 5:2.

Logistic Regression: Log Odds or Logits

- Step 2: Log Odds or ***Logits*** (log-odds unit)
  - A Logit is defined as

$$Logit = \log(Odds) = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \log(2.3391) = 0.8498.$$

  - For Logits, $-\infty \leq Logit \leq \infty$ [Logits are unbounded].
  - If $Logit < 0$, then $\hat{p} < 0.5$, i.e., the event (here: answer "Yes") is less likely than its counterpart (here: answer "No").
  - If $Logit > 0$, then $\hat{p} > 0.5$, i.e., the event (here: answer "Yes") is more likely than its counterpart (here: answer "No").
  - If $Logit = 0$, then $\hat{p} = 0.5$, i.e., the event (here: answer "Yes") is equally likely than its counterpart (here: answer "No").
- The Logit of answering to the item with "Yes" is thus 0.8498.

Logistic Regression: Log Odds or Logits

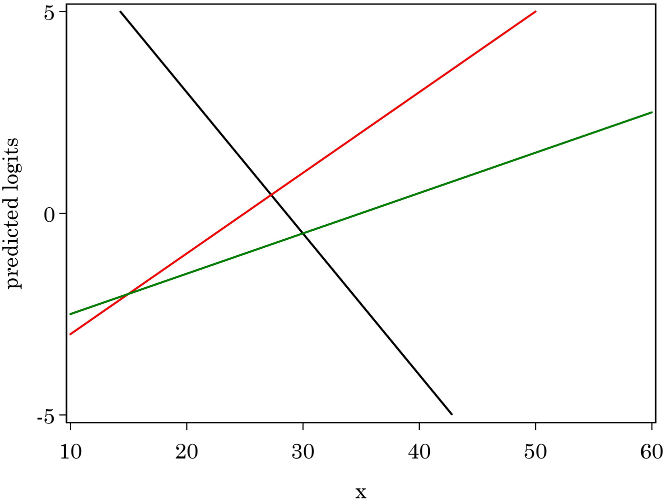
- Drawback: The transformation from probabilities to Logits is nonlinear!
- Complicates interpretations (see below)
- Logistic Regression

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k \tag{7}$$

- Model is linear on the logit scale!
- Model is nonlinear on the (original) probability scale!
- On the probability scale,
  - $\hat{\beta}_0$ moves the curve horizontally
  - $\hat{\beta}_k$ changes the shape of the curve (see slide 39)

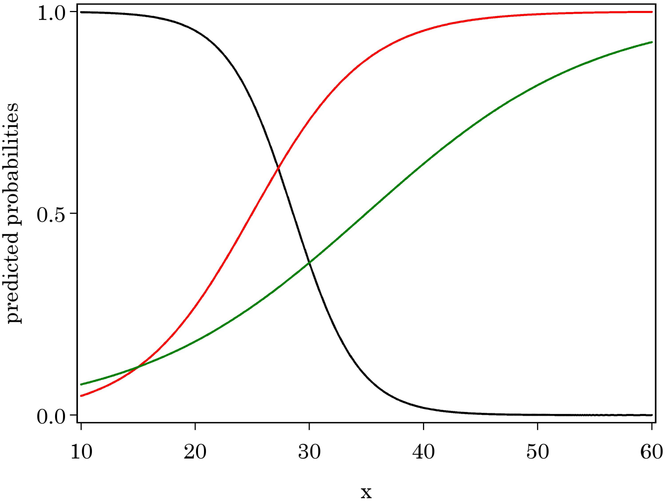
Logistic Regression: Nonlinearity

- Logit Scale: Linear Relation



Logistic Regression: Nonlinearity

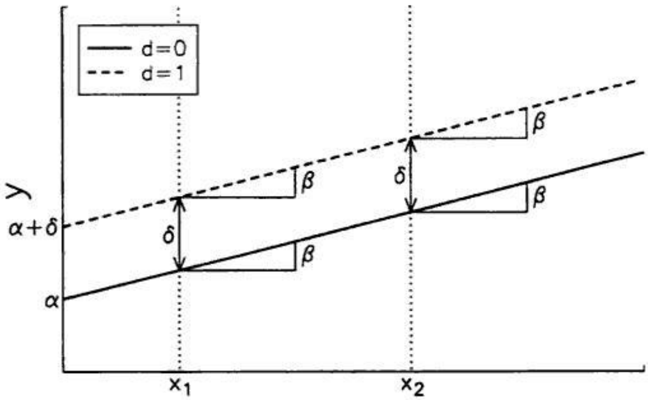
- Probability Scale: Nonlinear Relation



Consequences of Nonlinearity

- Comparison of GLM and Logistic Regression

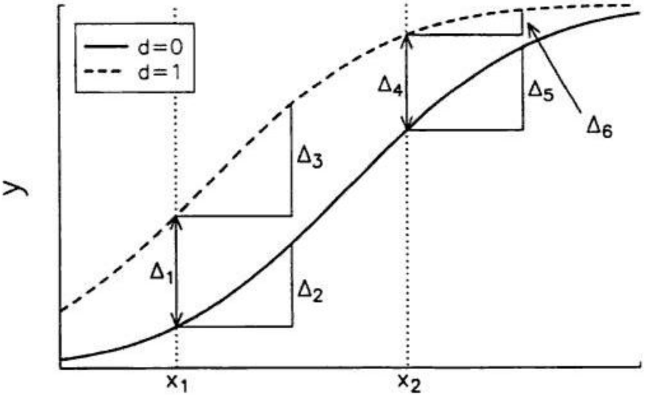
Panel A: Linear Model



Consequences of Nonlinearity

- Comparison of GLM and Logistic Regression

Panel B: Nonlinear Model



Example Analysis in R

- Do gender, depressive affect, and neuroticism change the probability of answering “Yes” to the statement “I have as much pep as last year”?

```
mod.ilse <- glm(y~sex+sds+neo, family=binomial,
                data=ilse)
summary(mod.ilse)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.24692    0.15505   8.042 8.82e-16 ***
sex         -0.47353    0.20580  -2.301   0.0214 *
sds         -0.04552    0.01889  -2.410   0.0160 *
neo         -0.10881    0.01881  -5.786  7.2e-09 ***
```

- ▷ sex: 0 = Female, 1 = Male
- ▷ sds: Self-Rating Depression Scale
- ▷ Neo: NEO-Neuroticism

Example Analysis in R: Interpretation

- Interpretation on the Logit scale is straightforward, no differences to GLM interpretation, because the regression equation is linear:

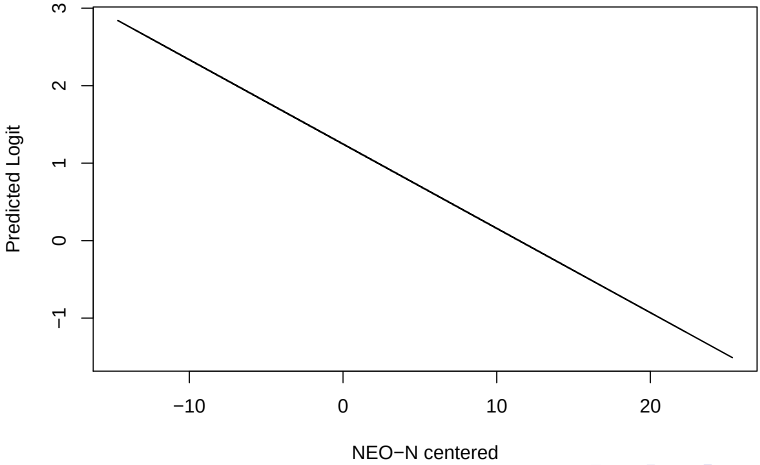
$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 1.2469 - 0.4735\,\mathrm{Sex} - 0.0455\,\mathrm{SDS} - 0.1088\,\mathrm{NEO\text{-}N}$$

- Example
  - Being male decreases the Logit of answering “Yes” by 0.4735.
  - Every score point on the self depression rating scale decreases the Logit of answering “Yes” by 0.0455.
- Problem: The Logit has no substantive meaning!
- By taking the exponential, the model is transformed back to the Odds metric, but now is multiplicative

$$\begin{aligned}\frac{\hat{p}}{1-\hat{p}} &= e^{\hat{\beta}_0}\, e^{\hat{\beta}_1 x_1}\, e^{\hat{\beta}_2 x_2}\, e^{\hat{\beta}_3 x_3} \\ &= e^{1.2469} e^{-0.4735\,\mathrm{Sex}} e^{-0.0455\,\mathrm{SDS}} e^{-0.1088\,\mathrm{NEO\text{-}N}}\end{aligned}$$

Example Analysis in R: Interpretation

- Effect of NEO-N on the Logit Scale (linear)

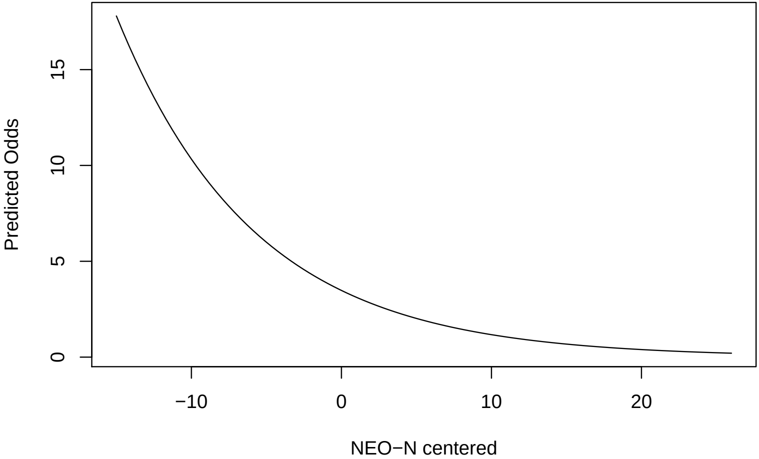


Example Analysis in R: Interpretation

- How does a one-unit change in a predictor variable affect the predicted odds?
- Example
  - Being male decreases the Odds of answering “Yes” by a factor of $e^{-0.4735} = 0.623$.
  - Thus, for men the Odds are the Odds of females times 0.623 (keeping all other predictors constant).
  - Having a 10-point higher score in the SDS decreases the Odds of answering “Yes” by a factor of $e^{-0.0455 \times 10} = 0.634$.
- Caution! Keep in mind that one-unit changes in predictor variables lead to multiplicative change in the odds.
  - This means that positive effects are greater than one whereas negative effects are between zero and one.
  - Magnitudes of positive and negative effects should thus be compared by taking the inverse of the negative effect.

Example Analysis in R: Interpretation

- Effect of NEO-N on the Odds Scale (multiplicative)



Example Analysis in R: Interpretation

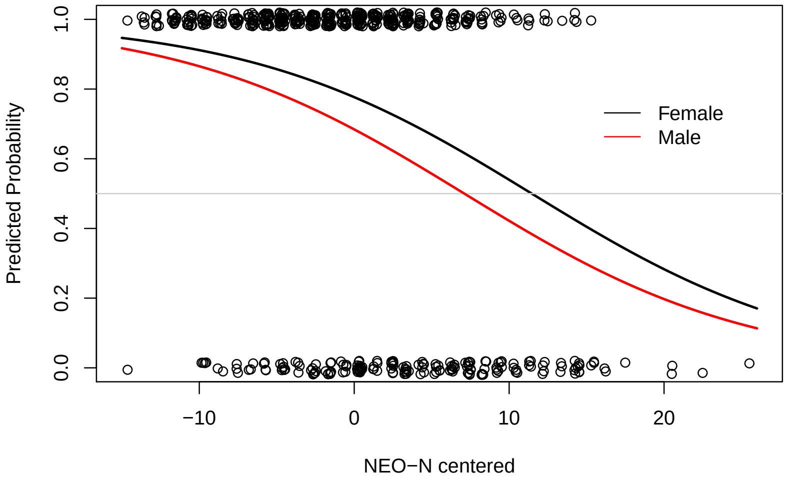
- To convert Odds back to probabilities, we use

$$p = \frac{Odds}{1 + Odds}$$

- Example
  - Being male decreases the Odds of answering “Yes” by a factor of $e^{-0.4735} = 0.623$. This factor change has different consequences depending on where one starts on the probability scale.
  - Imagine, women had Odds of 1. Thus, for men the Odds would be 0.623, corresponding to a probability of 0.384.
  - Imagine, women had Odds of 0.5. Thus, for men the Odds would be $0.50 \times 0.623 = 0.3115$, corresponding to 0.238.
  - Thus, whereas the Odds have been halved, the probability only decreased by the factor 1.6.
  - Probability scale: Probability of answering “Yes” is $p_{yes} = \frac{e^{1.2469}}{1+e^{1.2469}} = .78$ for females if mean neo and sds are zero.
- Caution! A constant factor change in the Odds does *not* correspond to a constant factor change in probabilities: Nonlinearity!

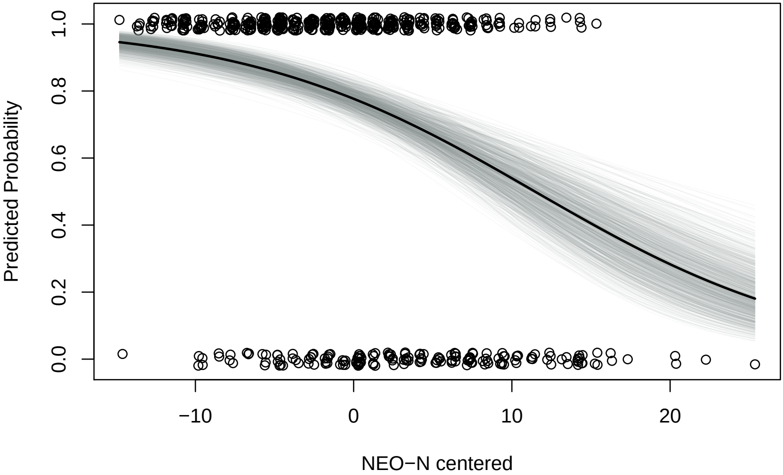
Example Analysis in R: Interpretation

- Effect of NEO-N on the Probability Scale (nonlinear)



Example Analysis in R: Interpretation

- Uncertainty in predicted probability represented by 1000 simulated lines from model



Example Analysis in R: Comparing Models

Error Rate

- Error rate and comparison to the null model
- Error rate is defined as the proportion of cases for which the deterministic prediction is wrong
  - $y_i = 1$ if $\text{logit}^{-1}(X_i\beta) > 0.5$
  - $y_i = 0$ if $\text{logit}^{-1}(X_i\beta) < 0.5$

```
>     mean((fitted(mod.ilse)>0.5 & ilse$y==0) |
          (fitted(mod.ilse)<0.5 & ilse$y==1))
[1] 0.2547332
```

- Null model: `mod.0 <- glm(y~1, family=binomial, data=ilse)`

```
>     mean((fitted(mod.0)>0.5 & ilse$y==0) |
          (fitted(mod.0)<0.5 & ilse$y==1))
[1] 0.2994836
```

Example Analysis in R: Comparing Models

Error Rate

- Good news, error rate is not at chance 0.5
- Null model: 0.299, i.e with pr(.70) this model always guesses “yes”
- Regression model: 0.255, the error rate is a little smaller
- Not a very impressive effect.

- Error rate is not a perfect summary of model misfit, because it does not distinguish between predictions of 0.6 and 0.9
- An error rate equal to the null rate (here 0.299) is terrible, and the best possible error rate is zero.
- For present data, always assuming “yes” (mean prediction) works quite well.

Example Analysis in R: Comparing Models

Deviance

- For logistic regressions and other discrete-data models, it does not make sense to calculate residual standard deviation and R^2
- ▷ The squared error is not the mathematically optimal measure of model error.
- *Deviance*
- A statistical summary of model fit, defined for generalized linear models to be an analogy to residual standard deviation
  - Deviance is a measure of error: lower deviance means better fit to data.
  - If a predictor that is simply random noise is added to a model, we expect deviance to decrease by 1, on average.
  - When informative predictor is added to a model, we expect deviance to decrease by more than 1. When k predictors are added to a model, we expect deviance to decrease by more than k

Example Analysis in R: Comparing Models

Deviance

- For classical (non-multilevel) models, the deviance is defined as âĹŠ2 times the logarithm of the likelihood function.
- e.g.

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.24692    0.15505   8.042 8.82e-16 ***
sex         -0.47353    0.20580  -2.301   0.0214 *
sds.c       -0.04552    0.01889  -2.410   0.0160 *
neo.c       -0.10881    0.01881  -5.786 7.20e-09 ***
---

Null deviance: 709.32  on 580  degrees of freedom
Residual deviance: 616.45  on 577  degrees of freedom
AIC: 624.45
```

Example Analysis in R: Comparing Models

Deviance

- `Null deviance: 709.32  on 580  degrees of freedom`
- The null deviance corresponds to the null model, with just the constant term.
- By adding predictors sex, sds, and neo we obtain
- `Residual deviance: 616.45  on 577  degrees of freedom`
- Deviance has decreased by $709.32 - 616.45 = 92.87$; this is more than the expected $k = 3$ if predictors had been noise.
- Here, adding the predictors improved the model fit

But: Just like R^2 deviance improves with the addition of variables.

- Akaike Information Criterion (AIC) $=$ Deviance $+ 2p$ with $p = k + 1$. Penalizes models with more parameters
- In our example: AIC$= 616.45 + 2 \times 4 = 624.45$

Summary

- Logistic regression is useful for dichotomized outcomes
- Linear on the log-scale
- Nonlinear on the original scale
- Predictors have different effect on outcome, depending on where the unit (number) is.

Outlook:

- Linear and Logistic regressions are special cases of generalized linear models

Generalized Linear Models

Overview

- Short overview of generalized linear models (GLMs) – pronounced 'glims'
- These models extend the linear modeling framework to variables that are not Normally distributed.
- GLMs are most commonly used to model binary or count data

Quick Review

- In a general linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i$$

the response $y_i, i = 1, ..., n$ is modeled by a linear function of explanatory variables $x_j, j = 1, ..., p$ plus an error term.

General and Linear

- Here **general** refers to the dependence on potentially more than one explanatory variable, v.s. the **simple** linear model:

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

- The model is *linear in the parameters*, e.g.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$
$$y_i = \beta_0 + \beta_1 x_i + \exp(\beta_2) x_i^2 + \epsilon_i$$

- but not, e.g.

$$y_i = \beta_0 + \beta_1 x_i^{\beta_2} + \epsilon_i$$
$$y_i = \beta_0 \exp(\beta_2 x_i) + \epsilon_i$$

Error Structure

Review of GLM

The assumptions regarding the error are :

- **Assumption 1**

  The expectation of the error is 0 for all N individuals, i.e. for the $N \times 1$-vector $\boldsymbol{\epsilon}$ it holds

  $$\mathsf{E}(\boldsymbol{\epsilon}) = \mathbf{0}.$$

- **Assumption 2**

  The variance of the error is the same and constant for all N individuals *and*

- **Assumption 3**

  The errors of two arbitrary individuals i and j are independent. This results in a covariance matrix of errors

  $$\mathsf{V}(\boldsymbol{\epsilon}) = \mathsf{E}\boldsymbol{\epsilon}\boldsymbol{\epsilon}' = \sigma_{\epsilon}^{2}\mathbf{I},$$

  with $\mathbf{I}$ as a $N \times N$-identity matrix (diagonal!).

Error Structure

Review of GLM

- **Assumption 4**

  The errors of the N individuals are multivariate normal distributed, i.e. for the $N \times 1$-vector $\boldsymbol{\epsilon}$ it holds

  $$\boldsymbol{\epsilon} \sim MVN(\mathsf{E}[\boldsymbol{\epsilon}], \mathsf{V}[\boldsymbol{\epsilon}]).$$

  MVN is short for *multivariate* normal distribution.
  Assumptions 1 to 3 can be expressed in

  $$\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$$

- Multivariate normally distributed means that the normals distribution holds for multiple dimensions In the example (with two predictors) in that would be two dimensions. I.e. that would be a bivariate normal distribution.

Restrictions of General Linear Models

- Although a very useful framework, there are some situations where general linear models are not appropriate
  - the range of Y is restricted (e.g. binary, count)
  - the variance of Y depends on the mean (e.g. RT data)
- Generalized linear models extend the general linear model framework to address both of these issues

i.e. We can use GLMs when

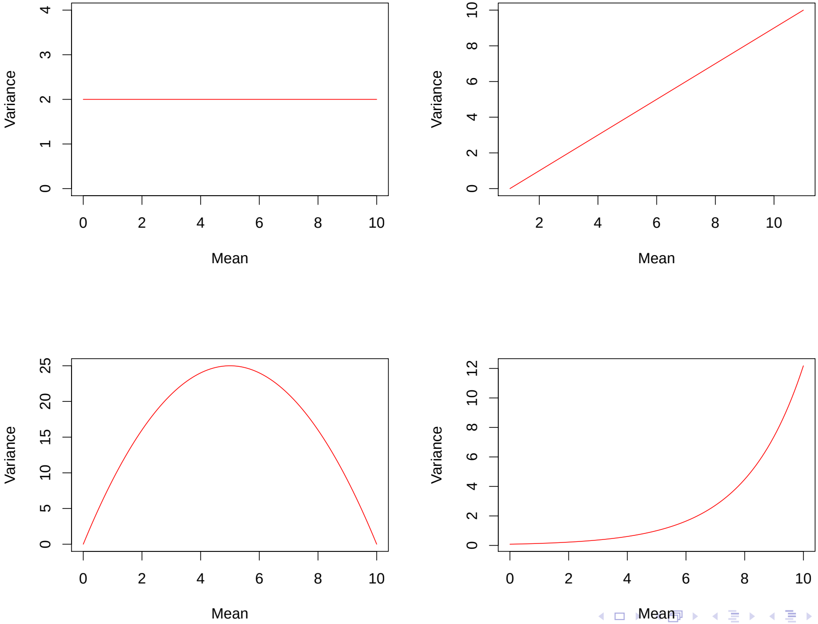
  - the variance is not constant,
  - and/or when the errors are not normally distributed.

Restrictions of General Linear Models

Specifically, we might consider using GLMs when the response variable is:

- count data expressed as proportions (e.g. logistic regressions)
- count data that are not proportions (e.g. log-linear models of counts)
- binary response variables (e.g. dead or alive)
- data on time to death where the variance increases faster than linearly with the mean (e.g. time data with gamma errors)
- ...

Visualizing Means vs. Variances



Visualizing Means vs. Variances

- central assumption that we have made up to this point is that variance was constant (top left-hand graph)
- In count data, however, where the response variable is an integer and there are often lots of zeros in the dataframe, the variance may increase linearly with the mean (top tight).
- With proportion data, where we have a count of the number of failures of an event as well as the number of successes, the variance will be an inverted U-shaped function of the mean (bottom left).
- Where the response variable follows a gamma distribution (as in time-to-death data) the variance increases faster than linearly with the mean (bottom right).

Generalized Linear Models (GLMs)

Three important properties:

- A generalized linear model is made up of a **linear predictor**

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

and two functions:

- a **link function** that describes how the mean, $E(Y_i) = \mu_i$, depends on the linear predictor

$$g(\mu_i) = \eta_i$$

- a **variance function** $V(\mu)$ that describes how the variance, $var(Y_i)$ depends on the mean

$$var(Y_i) = \phi V(\mu)$$

where ϕ is the dispersion or scale parameter is a constant

Error Structure

Up to this point, we have dealt with the statistical analysis of data with normal errors. In practice, however, many kinds of data have non-normal errors: for example:

- errors that are strongly skewed
- errors that are kurtotic
- errors that are strictly bounded (as in proportions)
- errors that cannot lead to negative fitted values (as in counts).

Error Structure

So far, the only tools available to deal with these problems were transformation of the response variable or the adoption of non-parametric methods.

GLM allows the specification of a variety of different error distributions:

- Poisson errors, useful with count data
- binomial errors, useful with data on proportions
- gamma errors, useful with data showing a constant coefficient of variation
- exponential errors, useful with data on time to death (survival analysis)

Error Structure

In R the error structure is defined by means of the `family` directive, used as part of the model formula.
For example

- `glm(y ~ z, family = poisson)`

  which means that the response variable y has Poisson (typically used for occurrence of events; only positive integers and skewed for fewer occurences) errors, and
- `glm(y ~ z, family = binomial)`

  which means that the response is binary, and the model has binomial errors.
- As with previous models, the explanatory variable z can be
  - continuous (leading to a regression analysis) or
  - categorical (leading to an ANOVA-like procedure called analysis of deviance)

Linear Predictor

- Structure of the model relates each observed y value to a predicted value
- Predicted value is obtained by transformation of the value emerging from the **linear predictor**.
- linear predictor, η (eta), is a linear sum of the effects of one or more explanatory variables, x_j,

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} = \sum_{j=1}^{p} x_i j \beta_j. \tag{8}$$

- x's are the values of the p different explanatory variables, and the
- β's are the (usually) unknown parameters to be estimated from the data.
- The right-hand side of the equation is called the **linear structure**.

Linear Predictor

- To determine the fit of a given model, a GLM evaluates the linear predictor for each value of the response variable
- then compares the predicted value with a *transformed value* of y.
- The transformation to be employed is specified in the **link function** (next slide).
- The fitted value is computed by applying the *reciprocal of the link* function, in order to get back to the original scale of measurement of the response variable.

Link Function

Crucial concept: Relationship between the values of the response variable y (as measured in the data and predicted by the model in fitted values) and the linear predictor.

- the link function relates the mean value of y to its linear predictor $\eta = g(\mu)$
- The linear predictor, η, emerges from the linear model as a sum of the terms for each of the p parameters (Equation 8).
- This is not a value of y (except in the special case of the **identity link** that we have been using (implicitly) up to now).

- The value of η is obtained by transforming the value of y by the link function,
- and the predicted value of y is obtained by applying the inverse link function to η.

Link Functions

- We've bee using the link function the whole time:
  - GLM: Link is identity, i.e., 1
  - Logistic regression: logit link $\log(\frac{\hat{p}}{1-\hat{p}})$
- in R:

```
glm(y~sex+sds.c+neo.c, family=binomial, data=ilse)
```

Is short for:

```
glm(y~sex+sds.c+neo.c, family=binomial(link = "logit"),
   data=ilse)
```

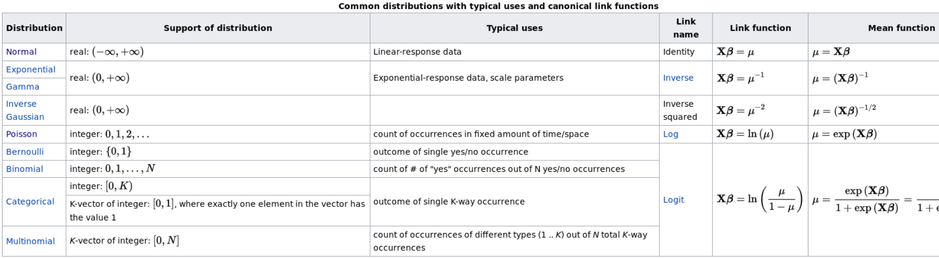
Link Functions

An important criterion in the choice of link function is to ensure that the fitted values stay within reasonable bounds.
For example,

- we would want to ensure that counts were all greater than or equal to 0 (negative count data wouldn't make any sense).
    - In this case, a log link is appropriate because the fitted values are antilogs (exp) of the linear predictor, and all antilogs are greater than or equal to 0.
- If the response variable was the proportion of individuals that died, then the fitted values would have to lie between 0 and 1 (fitted values greater than 1 or less than 0 would be meaningless)
    - In this case, the logit link is appropriate because the fitted values are calculated as the antilogs of the log odds, $\log(p/q)$.

Canonical Link Functions

- There is a number of link functions that can be used
- Some canonical link functions are tied to certain distributions
- Canonical link functions are the default options employed when a particular error structure is specified in the family directive in the model formula in R.
- cf. table from wikipedia



Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Inverse	$\mathbf{X}\boldsymbol{\beta} = \mu^{-1}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1}$
Gamma					
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{}{1+\mathrm{e}}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences			
Categorical	integer: $[0, K)$	outcome of single K-way occurrence			
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1				
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences			

Normal General Linear Model as a Special Case

- For the general linear model with $\epsilon \sim N(0, \sigma^2)$ we have the linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

- the link function

$$g(\mu_i) = \eta_i$$

- and the variance function

$$V(\mu_i) = 1$$

Binomial Data

Suppose the data come from a binomial distribution

$$Y_i \sim \text{Binomial}(\eta_i, p_i)$$

with n_i number of trials and success probability p_i , and we wish to model the proportions Y_i/n_i . Then

$$E(Y_i/n_i) = p_i \quad \text{var}(Y_i/n_i) = \frac{1}{n_i} p_i(1 - p_i)$$

So our variance function is

$$V(\mu_i) = \mu_i(1 - \mu_i)$$

Our link function must map from $(0, 1)$ to $(-\infty, \infty)$. We could achieve this with

$$g(\mu_i) = \log(\frac{\mu_i}{1 - \mu_i}) = \text{logit}(\mu_i)$$

Poisson Data

Suppose the data come from a poisson distribution

$$Y_i \sim \text{Poisson}(\lambda_i)$$

where λ designates the average number of events in an interval (an event can occur $0, 1, 2, \ldots$ times in an interval). Then

$$E(Y_i) = \lambda_i \quad \text{var}(Y_i) = \lambda_i$$

So our variance function is

$$V(\mu_i) = \mu_i$$

Our link function must map from $(0, \infty)$ to $(-\infty, \infty)$. An obvious choice is

$$g(\mu_i) = \log(\mu_i)$$

Transformation vs. GLMs

- In some situations a response variable can be transformed to improve linearity and homogeneity of variance so that a general linear model can be applied.
- This approach has some drawbacks
    - We manipulate response variable (the numbers change)
    - transformation must simultaneously improve linearity and homogeneity of variance
    - transformation may not be defined on the boundaries of the sample space

Transformation vs. GLMs

- For example, a common remedy for the variance increasing with the mean is to apply the log transform, e.g.

  $$\log(Y_i) = \beta_0 + \beta_1 x_1 + \epsilon_i$$
  $$E(\log(y_i)) = \beta_0 + \beta_1 x_1$$

  This is a linear model for the **mean of log** Y which may not always be appropriate. E.g. if Y is income perhaps we are really interested in the mean income of population subgroups.

- In this case it may be better to model $E(Y)$ using a generalized linear model with the link function $g(\mu_i) = \log(\mu_i)$:

  $$\log(E[Y_i]) = \beta_0 + \beta_1 x_1$$

  with $V(\mu) = \mu$ (which avoids difficulties with 0)

GLM's in R: The glm Function

Generalized linear models can be fitted in R using the glm function, which is similar to the lm function for fitting linear models. The arguments to a glm call are as follows

```
glm(formula, family = gaussian, data, weights,
    subset, na.action, start = NULL, etastart, mustart,
    offset, control = glm.control(...), model = TRUE,
    method = "glm.fit", x = FALSE, y = TRUE,
    contrasts = NULL, ...)
```

Formula Argument

- The formula is specified to glm as, e.g.

  ```
  y ~ x1 + x2
  ```

  where `x1` and `x2` are the names of

  - numeric vectors (continuous variables)
  - factors (categorical variables)

  All specified variables must be in the workspace or in the data frame passed to the `data` argument.

Formula Argument

- As with the `lm`-function we can use other symbols in the formula
  - `a:b` for an interaction among `a` and `b`
  - `a*b` wich expands to `a + b + a:b`
  - `-` to exclude a term or terms
  - `1` to include an intercept (included by default)
  - `0` to exclude an intercept

Family Argument

The `family` argument takes (the name of) a family function which specifies

- the link function
- the variance function
- various related objects used by `glm`, e.g. `linkinv`

The exponential family functions available in R are

- `binomial(link = "logit")`
- `gaussian(link = "identity")`
- `Gamma(link = "inverse")`
- `inverse.gaussian(link = "1/mu 2")`
- `poisson(link = "log")`

Example: Count Data

- Counts (whole numbers or integers)

e.g. The number of individuals who died, the number of firms going bankrupt, the number of days of frost, the number of red blood cells on a microscope slide...

- number 0 often appears as a value of the response variable
- Two types:
  - *Frequencies*, where we count how many times something happened, but we have no way of knowing how often it did *not* happen (e.g. lightning strikes, bankruptcies, deaths, births).
  - *Proportions*, where we know the number doing a particular thing, but also the number not doing that thing (e.g. the proportion dying, sex ratios at birth, proportions of different groups responding to a questionnaire).

Example: Count Data

- linear regression methods (assuming constant variance, normal errors) are not appropriate for count data for four main reasons:
    - The linear model might lead to the prediction of negative counts
    - The variance of the response variable is likely to increase with the mean
    - The errors will not be normally distributed
    - Zeros are difficult to handle in transformations

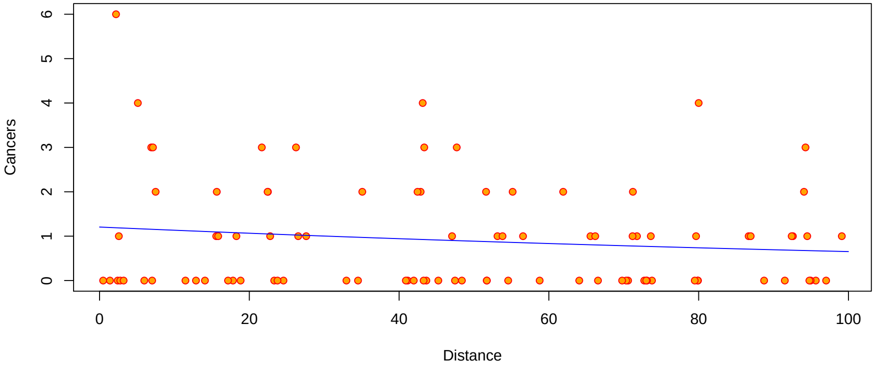
Example: Count Data

- In R, count data are handled in a generalized linear model by specifying `family=poisson` which sets `errors = Poisson` and `link = log`
- The log link ensures that all the fitted values are positive
- the Poisson errors take account of the fact that the data are integer and have variances that are equal to their means

Example: A Regression with Poisson Errors

- Outcome: Number of reported cancer cases per year per clinic
- Predictor: The distance from a nuclear plant to the clinic in kilometres

The question is whether or not proximity to the reactor affects the number of cancer cases.



Example: A Regression with Poisson Errors

- There seems to be a downward trend in cancer cases with distance (see the plot on previous slide).
- But is the trend significant?

We do a regression of cases against distance, using a GLM with Poisson errors:

```
model1<-glm(Cancers~Distance,poisson)
summary(model1)
```

Example: A Regression with Poisson Errors

```
Call:
glm(formula = Cancers ~ Distance, family = poisson)
Deviance Residuals:
Min       1Q   Median      3Q      Max
-1.5504  -1.3491  -1.1553   0.3877   3.1304

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.186865   0.188728   0.990   0.3221
Distance    -0.006138   0.003667  -1.674   0.0941 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 149.48  on 93  degrees of freedom
Residual deviance: 146.64  on 92  degrees of freedom
AIC: 262.41

Number of Fisher Scoring iterations: 5
```

Example: A Regression with Poisson Errors

- Trend does not seem to be significant
- Note the residual deviance.
  - Under Poisson errors, it is assumed that residual deviance is equal to the residual degrees of freedom (because the variance and the mean should be the same).
- Given that residual deviance is larger than residual degrees of freedom indicates that we have overdispersion (extra, unexplained variation in the response)
- We compensate for the overdispersion by refitting the model using quasi-Poisson rather than Poisson errors:

```
model2<-glm(Cancers~Distance,quasipoisson)
summary(model2)
```

Example: A Regression with Poisson Errors

```
Call:
glm(formula = Cancers ~ Distance, family = quasipoisson)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-1.5504  -1.3491  -1.1553   0.3877   3.1304

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.186865   0.235364   0.794    0.429
Distance    -0.006138   0.004573  -1.342    0.183

(Dispersion parameter for quasipoisson family taken to be 1.555271)

Null deviance: 149.48  on 93  degrees of freedom
Residual deviance: 146.64  on 92  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Example: A Regression with Poisson Errors

- Compensating for the overdispersion has increased the p value to 0.183, so there is no compelling evidence to support the existence of a trend in cancer incidence with distance from the nuclear plant.
- Note that the parameter estimates and the predictions from the model (the 'linear predictor') are in logs as the GLM with Poisson errors uses the log link
- If we want to bring them to our y metric we have to take the antilog (here: exp())