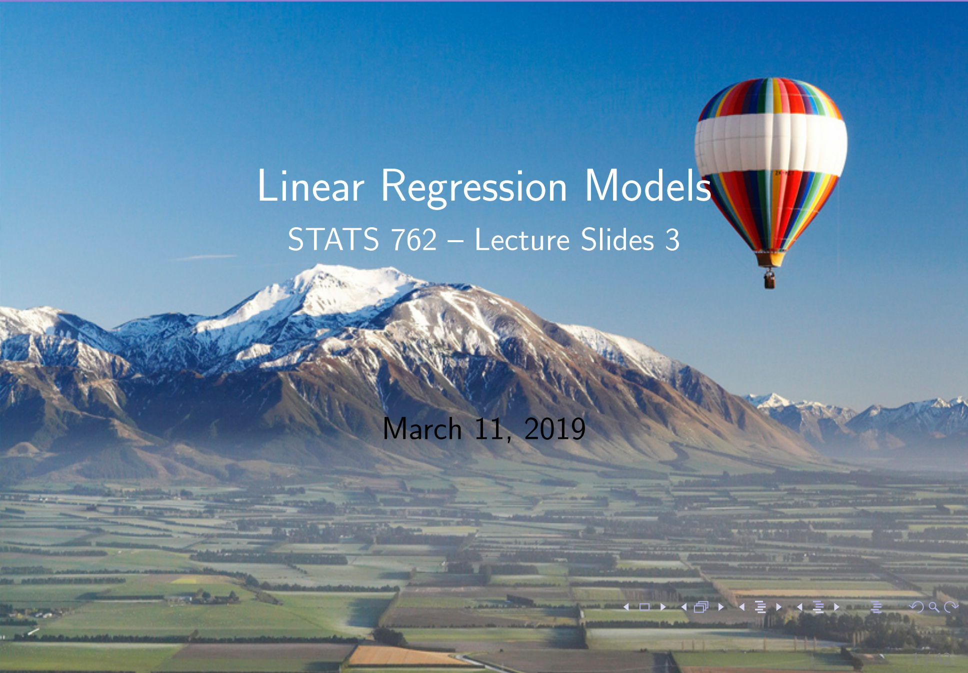


Linear Regression Models

STATS 762 – Lecture Slides 3

March 11, 2019

Inference for GLM's

For GLM's the deviance is the basic measure of how well a model fits the data and is often the basis of statistical inference.

- Analogous to the RSS for ordinary regression – in fact, if we apply the definition of deviance to the ordinary regression model we get the RSS.

Deviance compares the model being considered to the *maximal* (best possible) model.

Maximal Models

The maximal model represents the best possible fit to the data.

- For each distinct covariate pattern (combination of levels of the explanatory variables) we select the fitted value that minimizes the Likelihood function.
- For ordinary regression where each observation has a distinct covariate pattern, the maximal model would correspond to RSS $= 0$.

The Maximal Model for the CHD Data

The CHD data consists of 100 observations and contains the age of the individual in years and a binary indicator variable (chd$=$ 1 or 0) to indicate the presence or absence of CHD.

- Our logistic regression model related the probability of having CHD (π) to the age of the patient.
- For the maximal model, for each distinct value of age in the dataset the predicted probability ($\hat{\pi}$) would equal the observed proportion with CHD.

The Maximal Model for the CHD Data (cont.)

	proportion	$\hat{\pi}$ for
	with	maximal
age	chd $=$ 1	model
20	0/1	0.0
23	0/1	0.0
24	0/1	0.0
25	1/2	0.5
⋮	⋮	⋮
62	2/2	1.0
63	1/1	1.0
64	1/2	0.5
65	1/1	1.0
69	1/1	1.0

Grouped Data

For logistic regression it is sometimes convenient to group observations according to covariate patterns.

- All of the individuals in a group must have exactly the same values for all of the covariates.
- For each pattern ($i = 1$ to m) the number of individuals having that pattern (n_i) and the number of "successes" (r_i) are recorded.

Definition of Deviance

The deviance (Dev) is defined in terms of the log Likelihood function:

$$\mathrm{Dev} = 2 \log \mathrm{L}_{\mathrm{max}} - 2 \log \mathrm{L}_{\mathrm{mod}}$$

- $\mathrm{L}_{\mathrm{max}}$ and $\mathrm{L}_{\mathrm{mod}}$ represent the maximum possible values of the Likelihood under the maximal model and under the model being considered respectively.

The Maximal Model

Consider grouped data and divide the logistic regression model assumptions into two parts:

1. The binomial assumption (r is $Bin(n, \pi)$)
2. The logistic assumption (logit of π is linear)

If we only assume the first part, we have the maximal model (the most general model possible as we put no restriction on the probabilities). The likelihood L is:

$$\mathrm{L}(\pi_1, \ldots, \pi_m) = \prod_{i=1}^{m} \left(\begin{array}{c} n_i \\ r_i \end{array} \right) \pi_i^{r_i} (1 - \pi_i)^{n_i - r_i}$$

The Maximal Model (cont)

The log-likelihood for the maximal model (ignoring the bits not depending on the π's) is:

$$\log \mathrm{L}(\pi_1, \ldots, \pi_m) = \sum_{i=1}^{m} \{r_i \log \pi_i + (n_i - r_i) \log(1 - \pi_i)\}.$$

- The maximum value occurs when $\pi_i = r_i / n_i$.
- If r_i equals 0 or n_i then use $0 \log 0 = 0$.
- $\log \mathrm{L}_{\mathrm{max}}$ represents the maximized value.

Log Likelihood for the Logistic Model

- For grouped data, the log Likelihood for the logistic regression model is

$$\begin{aligned}\ell(\beta_0, \ldots, \beta_k) &= \sum_{i=1}^{m} \{ r_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) - \\ &\qquad n_i \log(1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})) \}.\end{aligned}$$

- The *i*th covariate pattern is $(x_{i1}, \ldots, x_{ik})$.
- The β's are chosen to maximize this expression (IRLS).
- $\log \mathrm{L}_{\mathrm{mod}}$ represents this maximized value.

Null Deviance and Residual Deviance

The summary output contains two types of deviance:

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.27844    1.13054  -4.669 3.03e-06 ***
age          0.11032    0.02402   4.593 4.37e-06 ***
---
    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.68  on 98  degrees of freedom
```

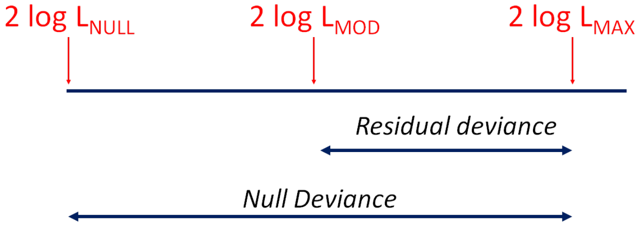
- Null deviance is calculated for the null model.
- Residual deviance is calculated for the fitted model.
- degrees of freedom are equal to $n - k - 1$.

Graphical interpretation

$$L_{Null} \leq L_{Mod} \leq L_{Max}$$

so

$$2 \log L_{Null} \leq 2 \log L_{Mod} \leq 2 \log L_{Max}$$



Analysis of Deviance

For GLM's a submodel can be tested against a full model using the change in deviance between the two models as the test statistic.

$$d_o = \text{Dev}_{\text{sub}} - \text{Dev}_{\text{full}}$$

- This is the likelihood ratio test statistic.
- The null hypothesis is that the extra regressors are not needed.
- Under the null hypothesis (and for a sufficiently large sample) the distribution of d_o is χ^2_v where v is the number of additional regressors in the full model.

Example

For our CHD model, let the null model be the submodel and the model using age as a regressor be the full model.

$$d_o = 136.66 - 107.68 = 28.98$$

The p-value is calculated using the χ^2_1 distribution:

```
> 1-pchisq(28.98,1)
[1] 7.312944e-08
```

This is same test as we get using `anova`:

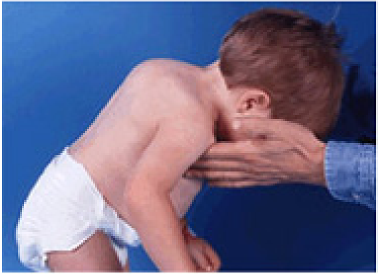
```
> anova(chd.glm,test="Chisq")
     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    99     136.66
age   1   28.982        98     107.68 7.304e-08 ***
```

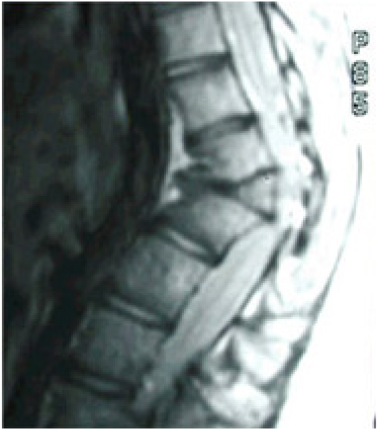
Example: Kyphosis risk factors

- Kyphosis is a curvature of the spine that may be corrected by spinal surgery. However, not all surgery is successful, and the condition may still be present after the corrective operation.
- In a study to determine risk factors for this, the following variables were measured:

Kyphosis:	(binary, absent=no kyphosis, present=kyphosis)
Age:	continuous, age in months
Start:	continuous, vertebrae level of surgery
Number:	continuous, number of vertebrae involved.

Example: Illustration





Data

The data are stored in a data frame `kyphosis.df`:

```
   Kyphosis Age Number Start
1    absent  71      3     5
2    absent 158      3    14
3   present 128      4     5
4    absent   2      5     1
5    absent   1      4    15
6    absent   1      2    16
7    absent  61      2    17
8    absent  37      3    16
9    absent 113      2    16
10  present  59      6    12
... 81 cases in all
```

Caution!!

In this data set Kyphosis is not a binary variable with values 0 and 1 but rather a factor with 2 levels “absent” and “present”:

```
levels(kyphosis.df$Kyphosis)
[1] "absent"  "present"
```

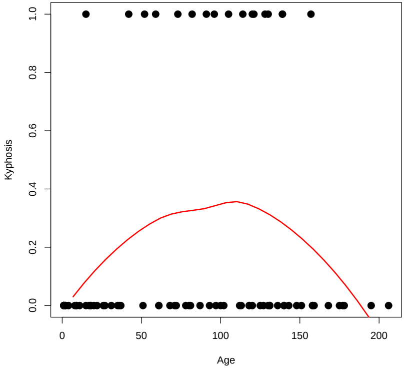
NB: if we fit a regression with Kyphosis as the response we are modelling the probability that Kyphosis is “present”: In general, R picks up the first level of the factor to mean “failure” (i.e. in this case “absent” or Y=0) and combines all the other levels into “success” (in this case “present” or Y=1).

Plotting Kyphosis versus Age

To plot Kyphosis versus Age with a smooth curve added:

```
> Y = as.numeric(kyphosis.df$Kyphosis)-1
> plot(Y~Age, data=kyphosis.df, pch=19, cex=1.5,
                        xlab="Age", ylab = "Kyphosis")
> my.loess = loess(Y~Age, data=kyphosis.df)
> plot.age = seq(0, 200, length=30)
> newdata=data.frame(Age=plot.age)
> plot.Y = predict(my.loess, newdata, pch=19, cex=1.5)
> lines(plot.age,plot.Y,lwd=2,col="red")
```

Kyphosis versus Age



Quadratic effect for Age?

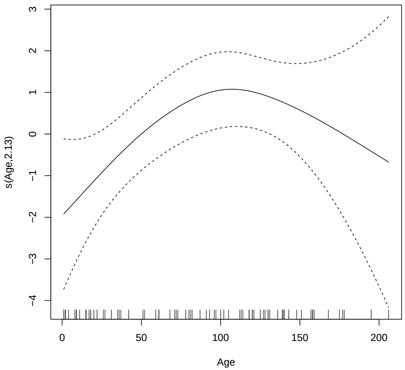
A More Sophisticated Smoother

A better plot can be produced using gam:

```
> library(mgcv)
> plot(gam(Kyphosis~s(Age) + Start + Number,
+     data=kyphosis.df, family=binomial))
```

- This plot will compensate for the effects of Start and Number.

The Gam Plot



Age around 100 months seems bad.

Quadratic Model

Seems age is important, fit as a quadratic

```
kyphosis.glm<-glm(Kyphosis~Age + I(Age^2) +
  Start + Number, family=binomial,
  data=kyphosis.df)
summary(kyphosis.glm)
```

Fitted Model

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.3835660  2.0548871  -2.133  0.03291 *
Age          0.0816412  0.0345292   2.364  0.01806 *
I(Age^2)    -0.0003965  0.0001905  -2.082  0.03737 *
Start       -0.2038421  0.0706936  -2.883  0.00393 **
Number       0.4268659  0.2365134   1.805  0.07110 .
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 83.234  on 80  degrees of freedom
Residual deviance: 54.428  on 76  degrees of freedom
AIC: 64.428
Number of Fisher Scoring iterations: 6
```

Output from anova

```
> anova(kyphosis.glm,test="Chisq")
Analysis of Deviance Table
         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        80     83.234
Age       1   1.3020        79     81.932 0.2538510
I(Age^2)  1   9.1939        78     72.739 0.0024282 **
Start     1  14.3244        77     58.414 0.0001539 ***
Number    1   3.9864        76     54.428 0.0458690 *
```

- The LRT test for adding `Number` to the model is more reliable than the test for `Number` on the previous page.
- We may want to test the impact of adding both `Age` and $\mathtt{Age}^2$ together.

Output from anova

```
> null.glm<-glm(Kyphosis~1,family=binomial,
                               data=kyphosis.df)
> age.glm<-glm(Kyphosis~Age + I(Age^2),
              family=binomial, data=kyphosis.df)
> anova(null.glm,age.glm,test="Chisq")
Analysis of Deviance Table
Model 1: Kyphosis ~ 1
Model 2: Kyphosis ~ Age + I(Age^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        80     83.234
2        78     72.739  2   10.496 0.005258 **
```

Output from anova

It is more relevant to consider adding Age and Age^2 to the model that contains the other regressors.

```
> sub.glm<-glm(Kyphosis~Start + Number, family=binomial,
                                         data=kyphosis.df)
> anova(sub.glm,kyphosis.glm,test="Chisq")
Analysis of Deviance Table

Model 1: Kyphosis ~ Start + Number
Model 2: Kyphosis ~ Age + I(Age^2) + Start + Number
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        78     64.536
2        76     54.428  2   10.109 0.006381 **
```

Likelihood for Poisson Data

For Poisson data: $$\mathrm{L} = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

$$\log \mathrm{L} = \sum_{i=1}^{n} y_i \log \mu_i - \mu_i - \log y_i!$$

- The $\log y_i!$ terms can be dropped as they add the same constant to all models.
- The maximal model chooses $\widehat{\mu}_i$'s to maximize this expression.

Deviance for Poisson Regression

For the Poisson regression model the log Likelihood becomes:

$$\log \mathrm{L} = \sum_{i=1}^{n} \Big(y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) - \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) \Big)$$

- The covariates for the ith observation are $(x_{i1}, \ldots, x_{ik})$.
- The $\widehat{\beta}$'s are chosen to maximize this expression.

The Deviance is (as always):

$$\mathrm{Dev} = 2 \log \mathrm{L}_{\mathrm{max}} - 2 \log \mathrm{L}_{\mathrm{mod}}$$

Poisson Regression Example

We can apply the same procedure to Poisson regression models. Recall our crab model:

```
> summary(crab.glm)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.04961    0.23311  -0.213   0.8315
weight       0.54608    0.06809   8.020 1.06e-15 ***
colour2     -0.20508    0.15371  -1.334   0.1821
colour3     -0.44966    0.17574  -2.559   0.0105 *
colour4     -0.45228    0.20843  -2.170   0.0300 *
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 551.78  on 168  degrees of freedom
```

Poisson Regression Example (cont.)

To compare the null model to the model that uses both `weight` and `colour`:

$$d_o = 632.79 - 551.78 = 81.01$$

The p-value is calculated using the χ^2_4 distribution:

```
> 1-pchisq(81.01,4)
[1] 1.110223e-16
```

Poisson Regression Example (cont.)

We can get anova to do this test for us:

```
> anova(crabnull.glm,crab.glm,test="Chisq")
Analysis of Deviance Table

Model 1: sats ~ 1
Model 2: sats ~ weight + colour
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       172     632.79
2       168     551.78  4    81.01 < 2.2e-16 ***
```

Poisson Regression Example (cont.)

Or we can get anova to do a series of these test by adding the variables sequentially:

```
> anova(crab.glm,test="Chisq")
Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                     172     632.79
weight  1   71.949       171     560.84  < 2e-16 ***
colour  3    9.062       168     551.78  0.02848 *
```

GLM's with Estimated Scale Parameters

This test can be extended to GLM's that have an estimated scale parameter (such as the quasibinomial and quasipoisson).

$$d_o = \frac{\text{Dev}_{\text{sub}} - \text{Dev}_{\text{full}}}{\widehat{\phi}}$$

- $\widehat{\phi}$ is the estimate of the scale parameter for the larger model.
- Estimating ϕ increases the variability in the distribution of d_o but it is still asymptotically χ^2_v.

Quasipoisson Example

```
> crabB.glm=glm(sats~weight+colour,family=quasipoisson,data=crab
> summary(crabB.glm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.04961    0.41593  -0.119    0.905
weight       0.54608    0.12148   4.495 1.29e-05 ***
colour2     -0.20508    0.27426  -0.748    0.456
colour3     -0.44966    0.31356  -1.434    0.153
colour4     -0.45228    0.37189  -1.216    0.226
---
(Dispersion parameter for quasipoisson family
                                taken to be 3.183475)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 551.78  on 168  degrees of freedom
```

Quasipoisson Example

```
> anova(crabB.glm,test="Chisq")
       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     172     632.79
weight  1   71.949       171     560.84 1.994e-06 ***
colour  3    9.062       168     551.78    0.4159
```

The p-value for the test that adds `colour` to the model that already contains `weight` is calculated uses the change in deviance divided by the scale parameter as the test statistic.

```
> 1-pchisq(9.062/3.183475,3)
[1] 0.4158901
```

Ordinary Regression as a GLM

Ordinary regression is a special case of a GLM that has an estimated scale parameter.

- The scale parameter is the variance: $\phi = \sigma^2$.
- The `glm` function with `family = gaussian` can be use to fit the model.
- Deviances are equivalent to residual sums of squares.

CHD Model Revisited

Using `lm` to fit the CHD model produced:

```
> summary(catheter.lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.3758     8.3859   2.430    0.038 *
ht            0.2107     0.3455   0.610    0.557
wt            0.1911     0.1583   1.207    0.258
---
Residual standard error: 3.778 on 9 degrees of freedom
Multiple R-squared:  0.8254,Adjusted R-squared:  0.7865
F-statistic: 21.27 on 2 and 9 DF,  p-value: 0.0003888
```

- Note: $\widehat{\sigma} = 3.778$.

CHD Analysis using `glm`

Alternatively, we can use `glm`:

```
> catheter.glm<-glm(ca˜.,family = gaussian,data=catheter.df)
> summary(catheter.glm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.3758     8.3859   2.430    0.038 *
ht            0.2107     0.3455   0.610    0.557
wt            0.1911     0.1583   1.207    0.258
---
(Dispersion parameter for gaussian family taken to be 14.27543)

    Null deviance: 735.67  on 11  degrees of freedom
Residual deviance: 128.48  on  9  degrees of freedom
```

- Taking the square root of the "dispersion parameter", $\sqrt{14.27543} = 3.77828$, gives us the same value for $\widehat{\sigma}$ as on the previous slide.

Equivalence of Tests

For ordinary regression the test statistic for our deviance test is closely related to that for the added variable F-test:

$$d_o = \frac{\text{Dev}_{\text{sub}} - \text{Dev}_{\text{full}}}{\widehat{\phi}} = \frac{\text{RSS}_{\text{sub}} - \text{RSS}_{\text{full}}}{\text{RSS}_{\text{full}}/(n - k_F - 1)} = f_o \times (k_F - k_S)$$

- The advantage of using the F-test is that it corrects for the extra variability induced by estimating ϕ (i.e. σ^2).
- The limiting distribution for an F_{d_1,d_2} distribution as d_2 gets very large is a $\chi^2_{d_1}/d_1$ distribution. Thus for large $n - k_F - 1$ we'll get essentially the same p-value for either test.

Output from anova

For the catheter model using `lm`:

```
> anova(catheter.lm)
Analysis of Variance Table
          Df Sum Sq Mean Sq F value    Pr(>F)
ht         1 586.38  586.38 41.0760 0.0001239 ***
wt         1  20.81   20.81  1.4578 0.2580548
Residuals  9 128.48   14.28
```

For the catheter model using `glm`:

```
> anova(catheter.glm,test="F")
Analysis of Deviance Table
Model: gaussian, link: identity
     Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                    11     735.67
ht    1   586.38        10     149.29 41.0760 0.0001239 ***
wt    1    20.81         9     128.48  1.4578 0.2580548
```

Output from anova

What if we use “`test = "Chisq"` for the `glm` object?

```
> anova(catheter.glm,test="Chisq")
Analysis of Deviance Table
     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    11     735.67
ht    1   586.38        10     149.29 1.464e-10 ***
wt    1    20.81         9     128.48    0.2273
```

- The p-values have changed as we don’t take into account the variability caused by estimating σ.
- They aren’t too different from the F-test results even though this is a relatively small data set (only 9 df’s available to estimate σ^2).