Regression and Generalized Linear Models

Dr. Wolfgang Rolke

Expo in Statistics

C3TEC, Caguas, October 9, 2015

Example: Predicting Success of UPR Students

Data: information from application forms of 25495 students who were accepted to UPR between 2003 and 2013, together with their GPA after their freshman year and information on whether they graduated (defined as graduated in 150% of official time, for example 6 years for a 4 year program)

Example record:

Student.Id	Año.de.Admision	Genero	Esc.Sup.Codigo	Esc.Sup.Nombre	Esc.Sup.Tipo	Esc.Sup.Lugar	Codigo.Prog.Admision	Prog.de.Admision	GPA.Escuela.Superior	GPA.al.Primer.Ano
00C2B4EF77	2005	M	1010	Jose E. Lugo	Pública	ADJUNTAS	502	Ingeniería Eléctrica - BC	3.97	3.67

Aptitud.Verbal	Aptitud.Matem	Aprov.Ingles	Aprov.Matem	Aprov.Espanol	IGS	Grad	al.100	al.150	Ano.Graduacion	Codigo_Prog_Graduacion
647	621	626	672	551	342	Si	No	Si	2012	502

Prog_Graduacion	GPA.de.Graduacion	Niv.Educ. del.Padre	Niv.Educ. de.la.Madre	Niv_Avanzado_Espa	Niv_Avanzado_Ingles	Niv_Avanzado_Mate_I	Niv_Avanzado_Mate_II	Destrezas 1	Destrezas 2	Regresó.al.Siguiente.Año		
Ingeniería Eléctrica	3.33	4 : Completó Escuela Superior	7 : Bachillerato	3	3		3			Si		

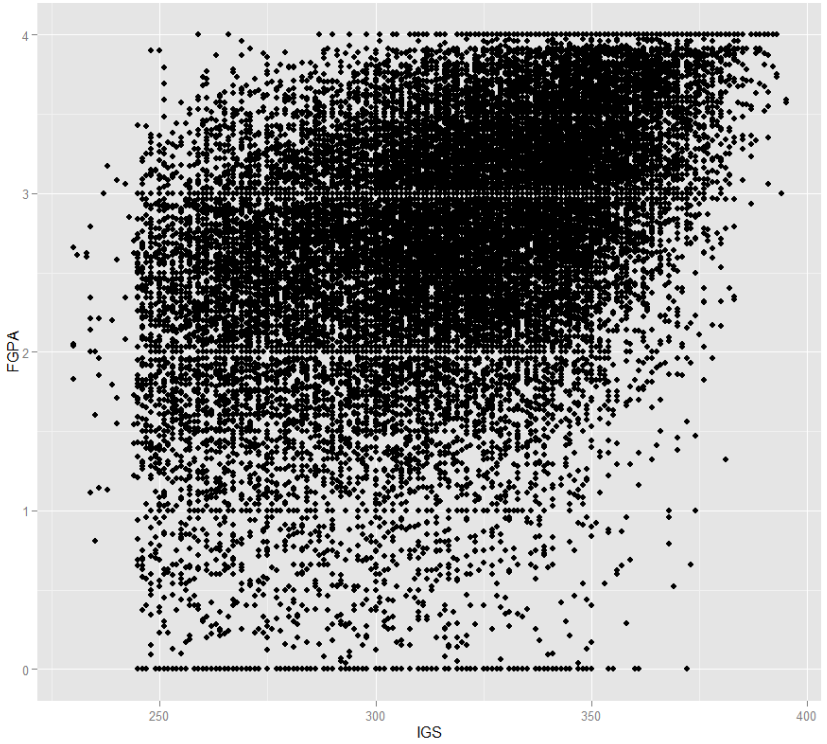
How well is the current admissions system working?

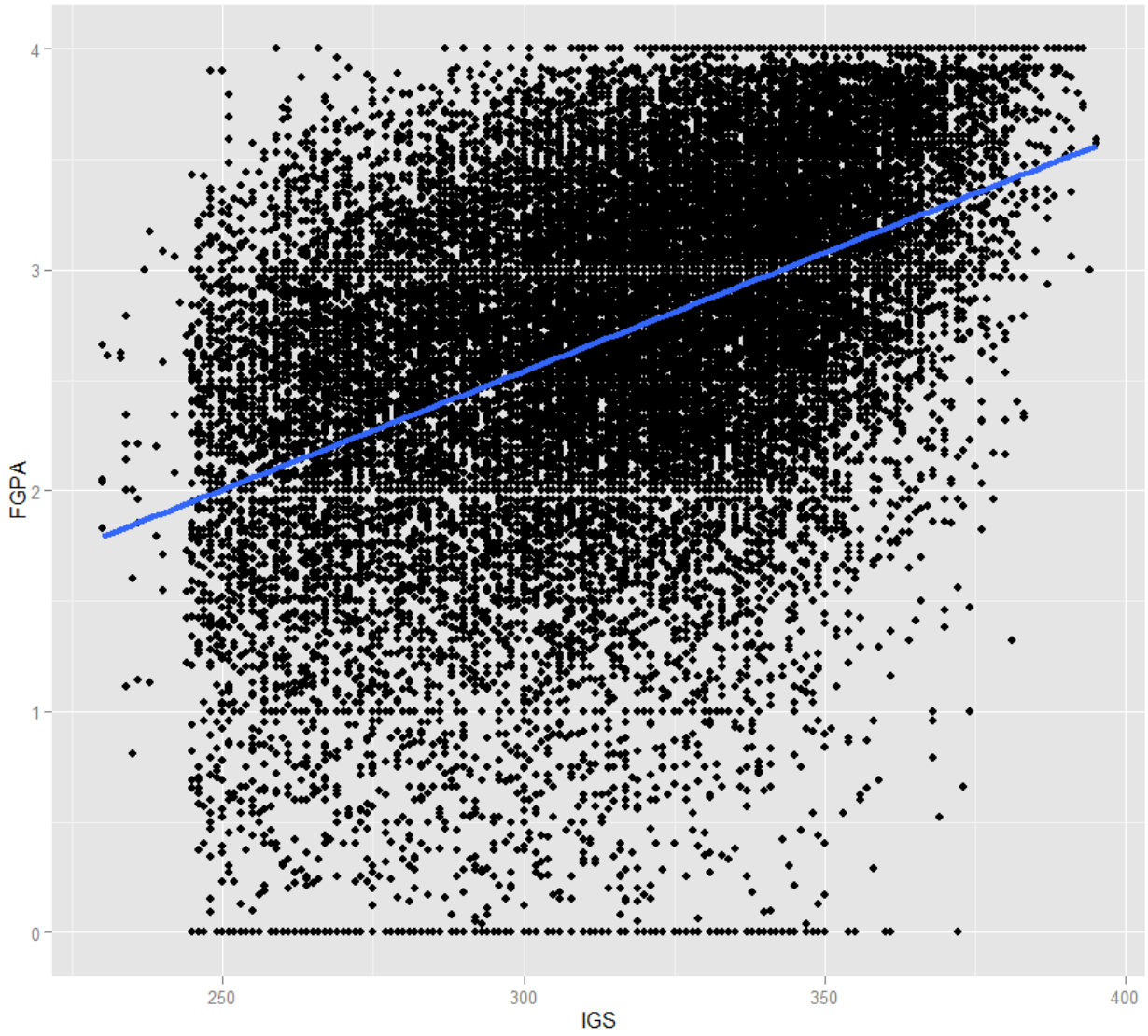
Admissions is based on IGS score (a combination of GPA, AptVerb and AptMate)

How well does it predict the GPA after the freshman year?

cor(IGS,FGPA) = 0.43

($p < 10^{-10}$)




FGPA
IGS
0
1
2
3
4
250
300
350
400

Least Squares Regression

Model: $y = \beta_0 + \beta_1 x$

Fitted values: $\hat{y}_i = \beta_0 + \beta_1 x_i$

Residuals: $\epsilon_i = y_i - \hat{y}_i$

Method of Least Squares:

$$minimize\ RSS = \sum (y_i - \hat{y}_i)^2$$

Standard Output of StatProgram (here R)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.674	0.0447	-15.10	<2e-16
IGS	0.011	0.0001	7647	<2e-16

Residual standard error: 0.7032 on 25493 degrees of freedom

Multiple R-squared: 0.1866, Adjusted R-squared: 0.1865

F-statistic: 5847 on 1 and 25493 DF, p-value: < 2.2e-16

What does it mean?

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	**-0.674**	0.0447	-15.10	<2e-16
IGS	**0.011**	0.0001	7647	<2e-16

Equation: FGPA$= -0.679 + 0.011IGS$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	**-0.674**	0.0447	-15.10	**<2e-1**6

(Intercept) Pr(>|t|) <2e-16

Hypothesis Test: $H_0: \beta_0 = 0 \; vs. \; H_a: \beta_0 \neq 0$

→ Intercept β_0 is not 0 (but who cares?)

In general decision on whether or not to fit an intercept is best made by considering the Science:

Example

x = #of hurricanes per year

y = $ total damages per year

If x=0, then y=0

	Estimate	Std. Error	t value	Pr(>|t|)
IGS	0.011277	0.000115	98.09	**<2e-16**

Hypothesis Test: $H_0: \beta_1 = 0 \; vs. \; H_a: \beta_1 \neq 0$

IGS: Pr(>|t|) <2e-16

→ coef β_1 of IGS is not 0 (but that was obvious from graph, and from Pearson's correlation coefficient. (Actually, in the case of a single predictor those two tests are the same)

→ generally neither of these tests is very interesting or useful.

Residual standard error: 0.7032 on 25493 degrees of freedom
(Pretty meaningless)

Multiple R-squared: 0.1866, Adjusted R-squared: 0.1865

$R^2 = 18.7\%$ of the variation in the FGPA is explained by the IGS. Not very high, maybe we should try to do better

Whether an R^2 is "high" or "low" depends the circumstances.

Note: $cor(IGS, FGPA)^2 * 100\% = 0.43^2 * 100\% = 18.7\%$

adj R^2: essentially the same as R^2 in simple regression

F-statistic: 5847 on 1 and 25493 DF, p-value: < 2.2e-16
p-value small → IGS is not completely useless for predicting FGPA, but again, the correlation test already told us that.

- In a simple regression the only really interesting part of the output is the equation, and to a lesser degree the R^2

Assumptions of LSR

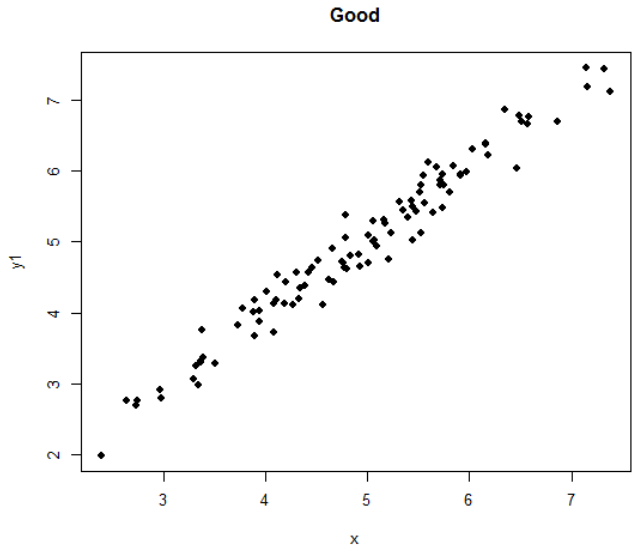
1) Linear Model is ok (and not say quadratic or some other shape)

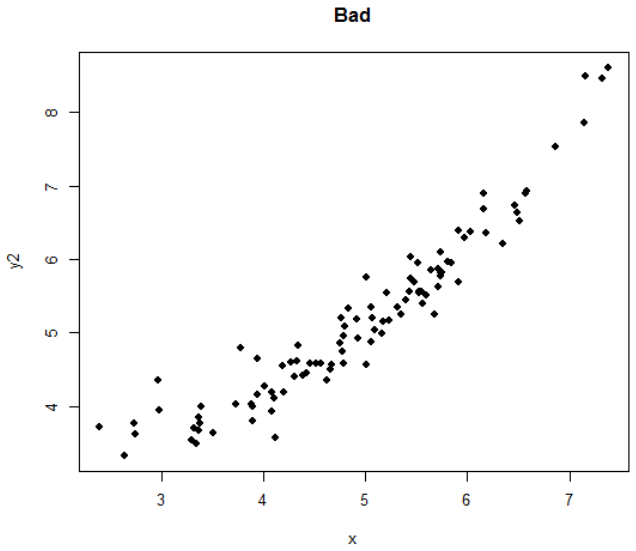
2) $\epsilon_i \sim N(0, \sigma)$

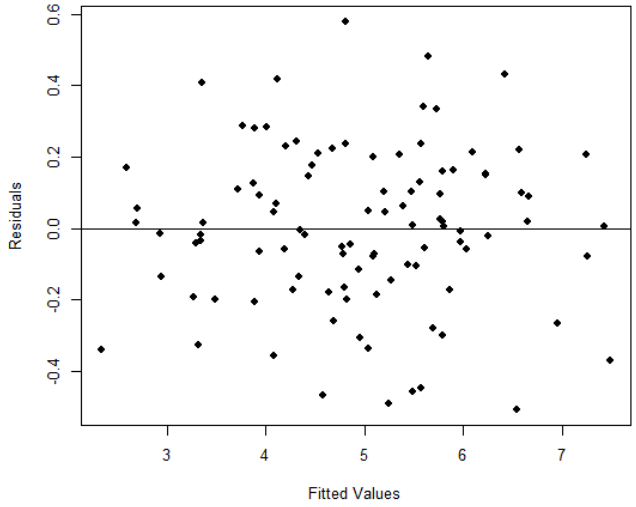
2a) $\epsilon_i \sim N$ Residuals come from a Normal distribution

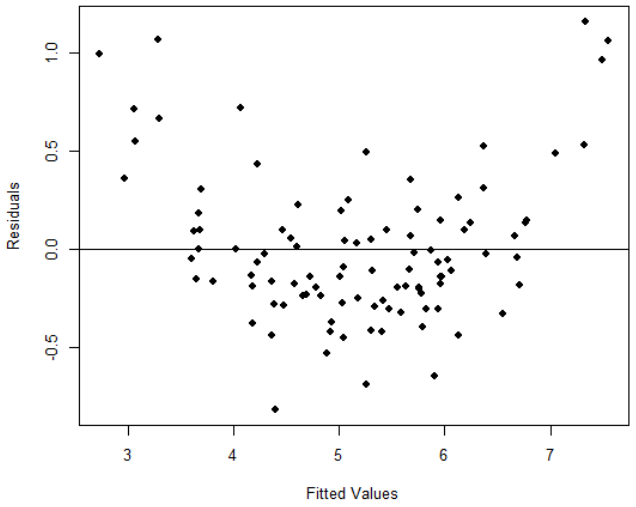
2b) $\epsilon_i \sim N(0, \sigma)$ Residuals have equal variance (independent of x) (homoscatasticity)

Is linear model ok? Residual vs Fits

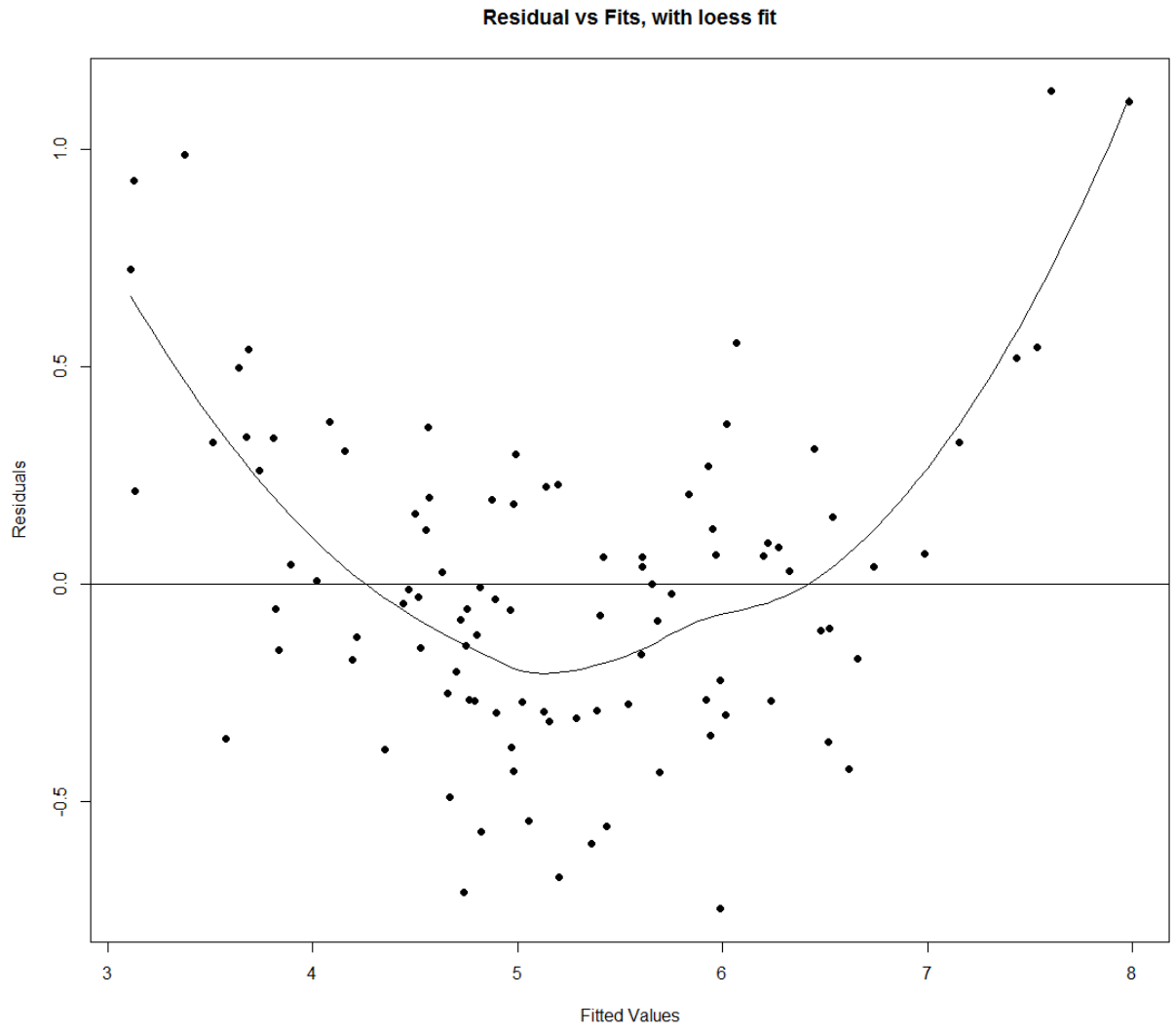




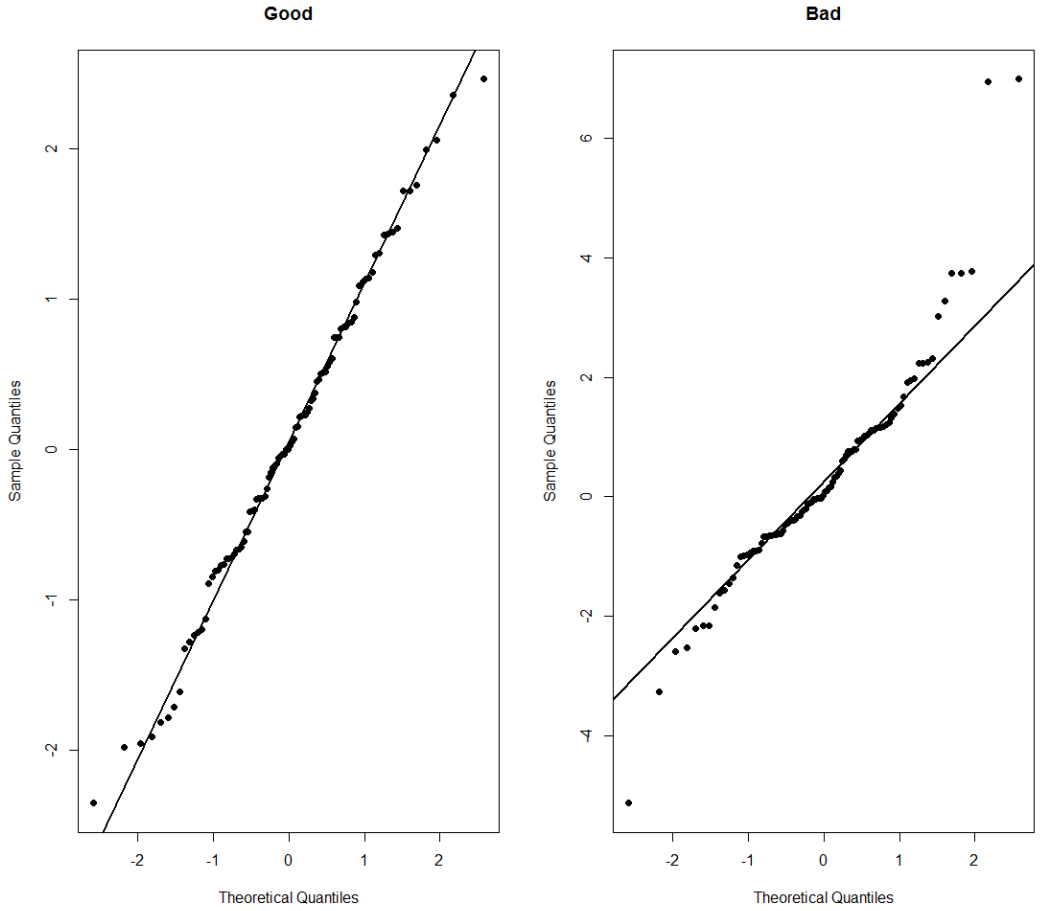




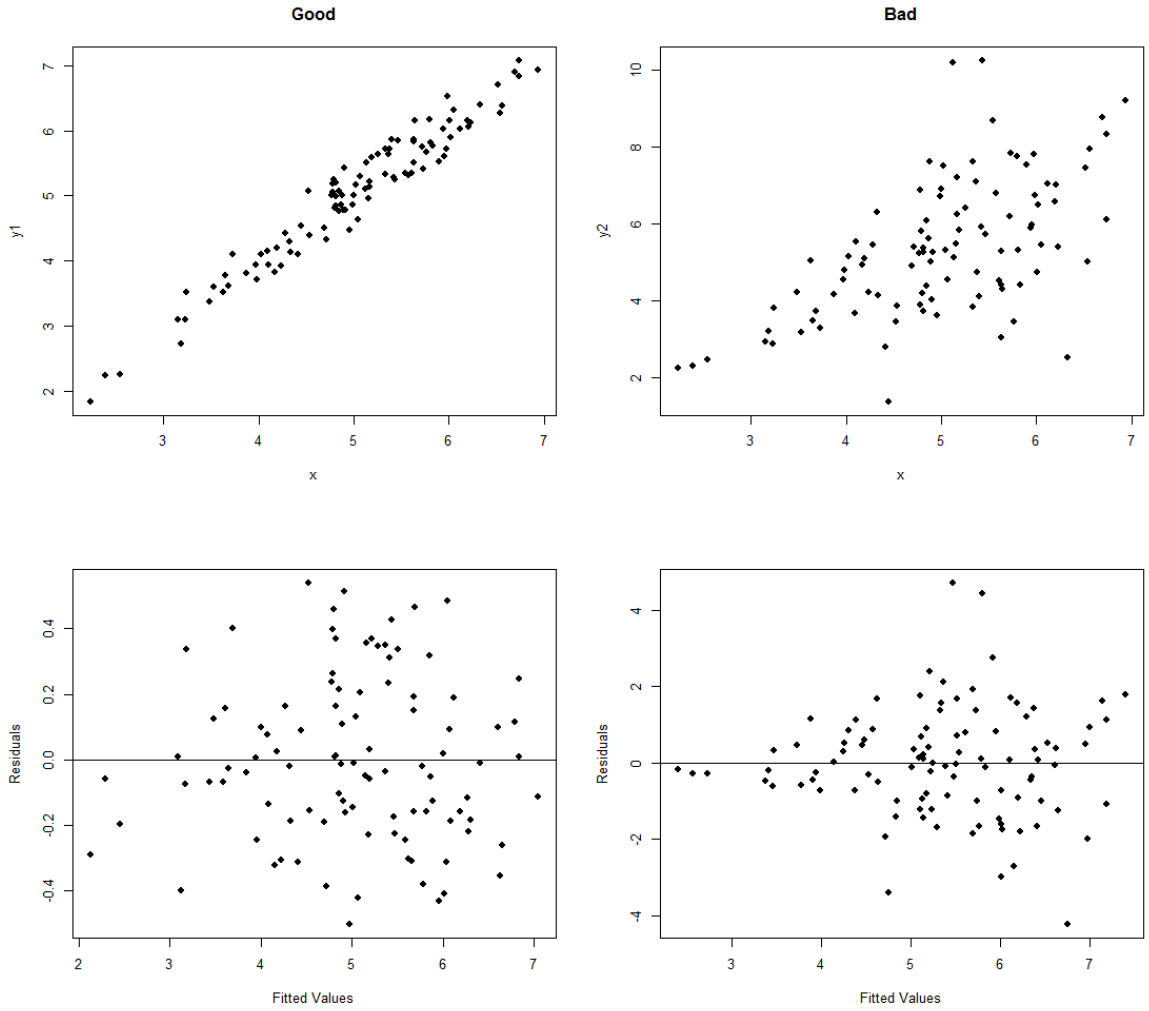
Nice version of this:



Normal Residuals? Normal Plot



Equal Variance? Residual vs Fits again



If there are problems

Transformations (√, log etc)

Polynomial regression

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \ldots$$

Interesting question: when to stop fitting?

George Box (1976) *"All models are wrong but some are useful"*

Occam's Razor: *"Keep it as simple as possible"* (my version)

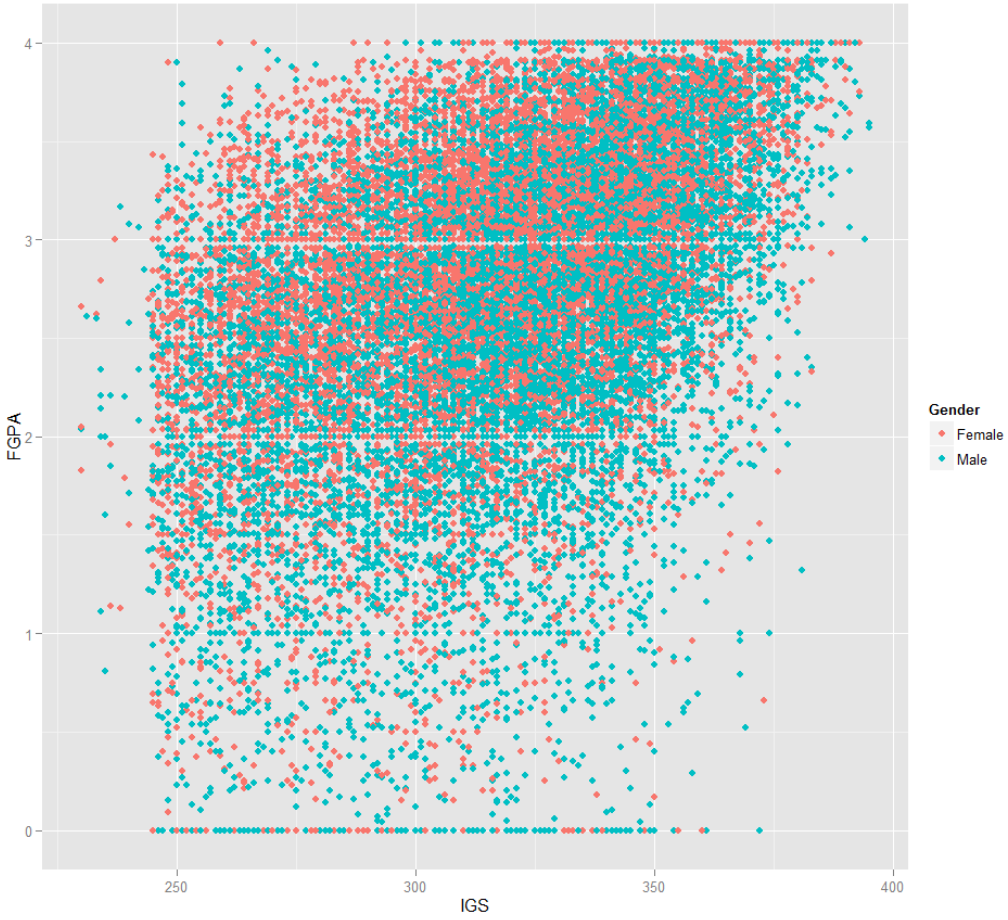
Models should be *parsimoneous*

Other things one can do…

- Formal tests (say for normality)
- Find influential observations (Cook's distance, leverage, etc.
- Calculate other residuals (standardized, studentized …)

Use of Dummy Variables

Let’s say we want to include Gender as a predictor



To do a regression code gender (0=Male, 1=Female), then

	Estimate
(Intercept)	-0.857
IGS	0.010
Gender	0.231

Now

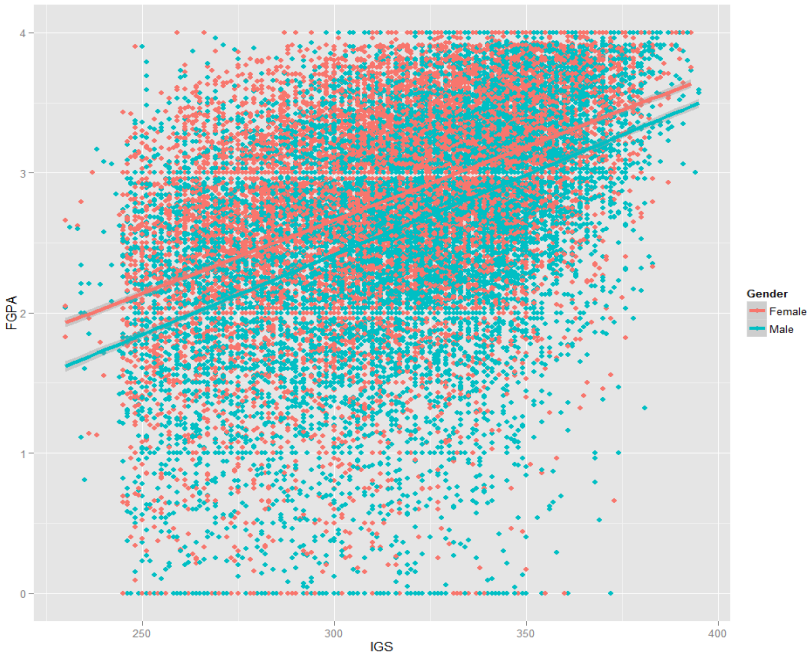
Male FGPA = -0.857+0.01 IGS +0.231*0

= -0.857+**0.01** IGS

Female FGPA = -0.857+0.01 IGS +0.231*1

= -0.626+**0.01** IGS

ALWAYS fits parallel lines!



In order to get most general model we have to include a product term:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.998	0.061	-16.209	< 2e-16
IGS	0.011	0.0001	59.215	< 2e-16
Gender	0.522	0.088	5.915	3.36e-09
IGS:Gender	-0.001	0.001	-3.314	0.00092

Male FGPA = -0.998 + 0.011 IGS + 0.522*0 – 0.001*IGS*0

= -0.998 + 0.011 IGS

Female FGPA = -0.998 + 0.011 IGS + 0.522*1 – 0.001*IGS*1=

= -0.476 + 0.010 IGS

Categorical Predictor with more than two values

Say we had info on parents: married, divorced, never married.

Could include this as follows: code

married=0, divorced=1, never married=2

But is this the same as never married=0, married=1, divorced=2?

Introduced order and “size” (1-0=2-1)

Usually better: Dummy variables:

Married=1 if yes, 0 if not

Divorced=1 if yes, 0 if not

How about including more Predictors? →Multiple Regression

There is more information on the application form

Some of it is not useful for legal and ethical reasons (Gender, Educational level of parents)

One big problem: High School GPA! In some schools a GPA of 3.5 means a high performing student, in others not so much

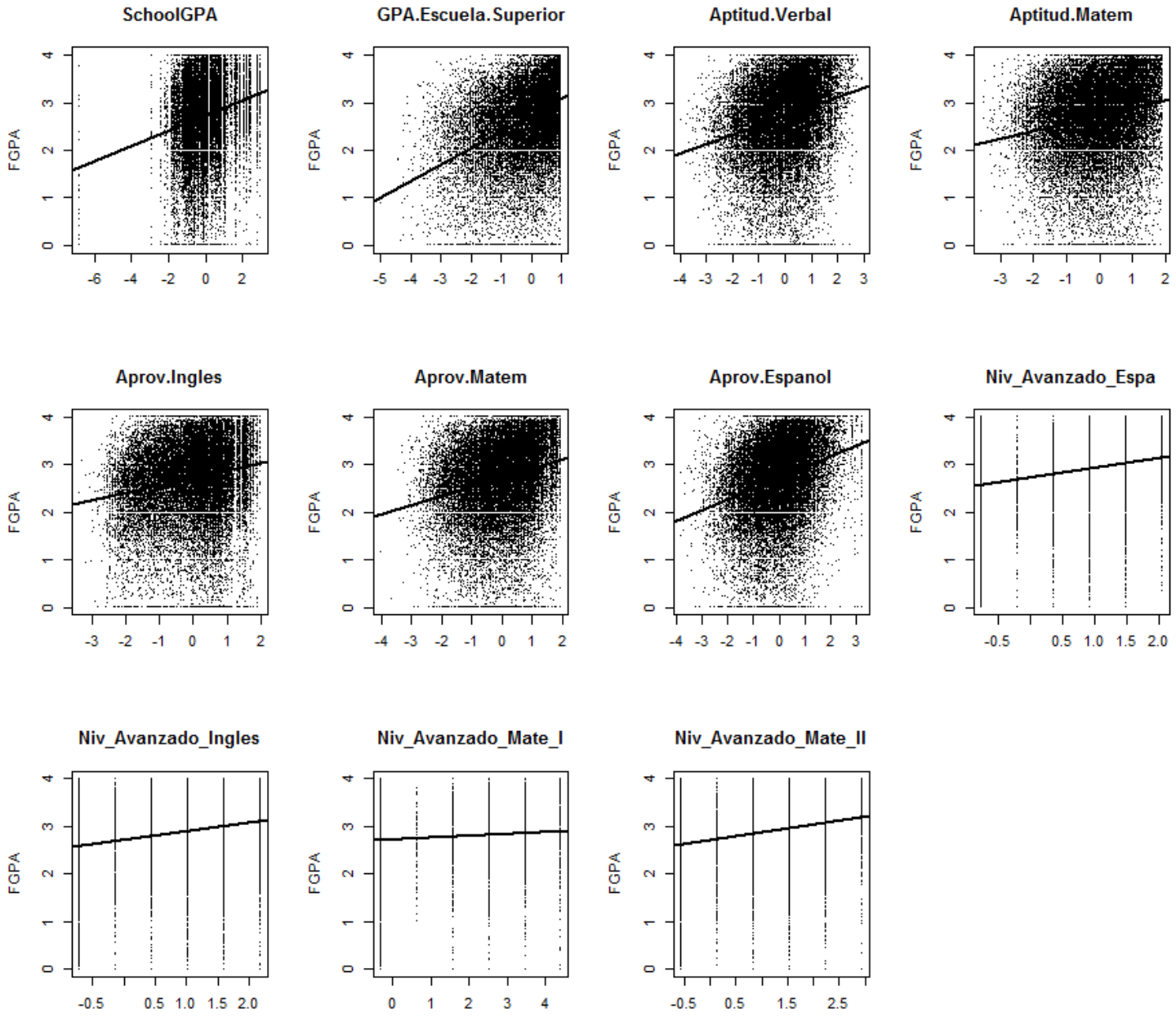
Solution: School GPA

School GPA

- Find mean GPA after Freshman year at UPRM for all students from the same high school
- Find the mean GPA of those students at that high school.
- Take the ratio
- A high number means students from this school tend to do well at UPR.

The extreme cases:

- The worst: School “3943” Freshman GPA 1.3, School GPA 3.8, Ratio 0.34
- The best School “2973” Freshman GPA 2.98, School GPA 3.2, Ratio 0.93


SchoolGPA
GPA.Escuela.Superior
Aptitud.Verbal
Aptitud.Matem
Aprov.Ingles
Aprov.Matem
Aprov.Espanol
Niv_Avanzado_Espa
Niv_Avanzado_Ingles
Niv_Avanzado_Mate_I
Niv_Avanzado_Mate_II
FGPA

Correlations of FGPA vs Predictors

Predictor	Correlation	P-value
SchoolGPA	0.206	0.00
GPA.Escuela.Superior	**0.436**	0.00
Aptitud.Verbal	0.257	0.00
Aptitud.Matem	0.202	0.00
Aprov.Ingles	0.204	0.00
Aprov.Matem	0.248	0.00
Aprov.Espanol	0.292	0.00
Niv_Avanzado_Espa	0.264	0.00
Niv_Avanzado_Ingles	0.225	0.00
Niv_Avanzado_Mate_I	0.054	0.00
Niv_Avanzado_Mate_II	0.214	0.00

New Issue:
Correlations between Predictors

	SchoolGPA	GPA.Escuela.Superior	Aptitud.Verbal	Aptitud.Matem	Aprov.Ingles	Aprov.Matem	Aprov.Espanol	Niv_Avanzado_Espa	Niv_Avanzado_Ingles	Niv_Avanzado_Mate_I	Niv_Avanzado_Mate_II
SchoolGPA	1	-0.22	0.205	0.247	0.325	0.25	0.197	0.15	0.239	0.002	0.088
GPA.Escuela.Superior	**-0.22**	1	0.18	0.159	0.052	0.218	0.25	0.304	0.196	0.086	0.248
Aptitud.Verbal	0.205	0.18	1	0.463	0.512	0.474	0.603	0.368	0.358	0.08	0.225
Aptitud.Matem	0.247	0.159	0.463	1	0.455	**0.815**	0.388	0.326	0.366	0.148	0.383
Aprov.Ingles	0.325	0.052	0.512	0.455	1	0.48	0.429	0.28	0.497	0.088	0.187
Aprov.Matem	0.25	0.218	0.474	0.815	0.48	1	0.403	0.354	0.381	0.161	0.412
Aprov.Espanol	0.197	0.25	0.603	0.388	0.429	0.403	1	0.355	0.321	0.07	0.213
Niv_Avanzado_Espa	0.15	0.304	0.368	0.326	0.28	0.354	0.355	1	0.666	0.202	0.466
Niv_Avanzado_Ingles	0.239	0.196	0.358	0.366	0.497	0.381	0.321	0.666	1	0.226	0.429
Niv_Avanzado_Mate_I	0.002	0.086	0.08	0.148	0.088	0.161	0.07	0.202	0.226	1	0.096
Niv_Avanzado_Mate_II	0.088	0.248	0.225	0.383	0.187	0.412	0.213	0.466	0.429	0.096	1

High correlations can cause problems → Multi-collinearity

Issues with fitting (numerical instability)

Extreme case cor(x1,x2)=1 → regression not possible (but easily resolved)

Issues with interpretation – regression coefficients can be negative even though all predictors have positive correlation with response

Sometimes worthwhile to transform predictors to orthogonal variables (principle components)

Minor issue, usually ignored: if a predictor is a dummy variable, it is categorical, Pearson's correlation coefficient meant for quantitative variables

Output of Regression:

	Estimate	Std. Error	t value	p value
(Intercept)	2.724	0	668	0
SchoolGPA	0.209	0	45.6	0
GPA.Escuela.Superior	0.351	0	75.6	0
Aptitud.Verbal	0.041	0.01	7.32	2.49E-13
Aptitud.Matem	-0.036	0.01	-5.06	4.10E-07
Aprov.Ingles	0.028	0.01	5.089	3.62E-07
Aprov.Matem	0.021	0.01	2.829	0.004664
Aprov.Espanol	0.056	0.01	10.52	7.47E-26
Niv_Avanzado_Espa	0.021	0.01	3.620	0.000294
Niv_Avanzado_Ingles	-0.014	0.01	-2.39	0.016847
Niv_Avanzado_Mate_I	-0.0005	0	-0.1	0.902422
Niv_Avanzado_Mate_II	0.035	0	7.202	6.06E-13

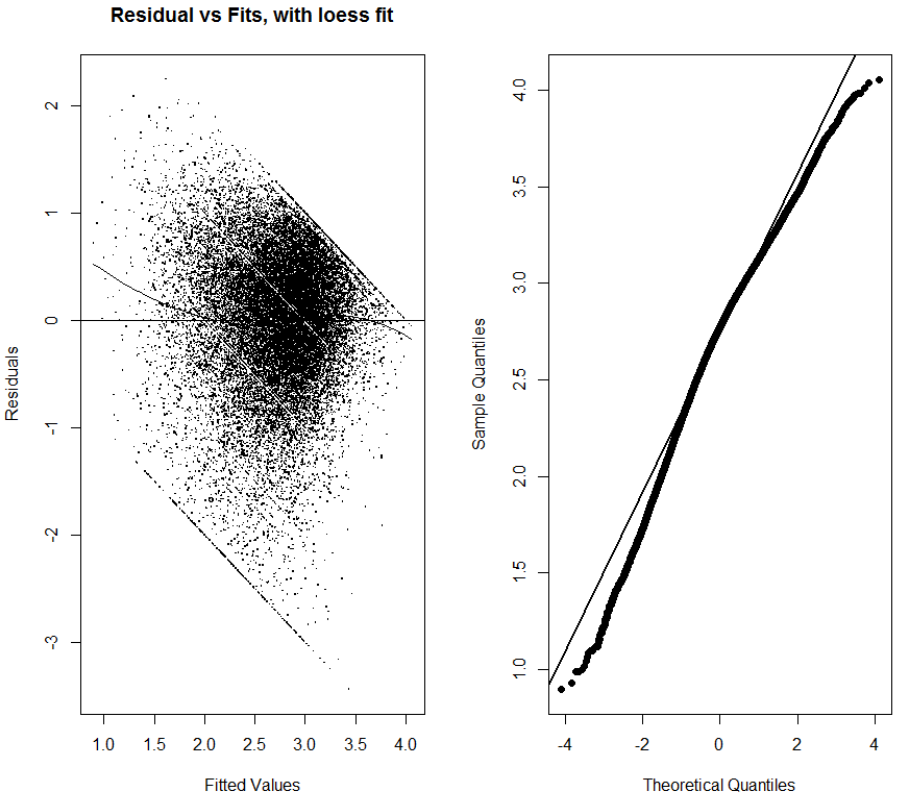
Residual standard error: 0.6506 on 25483 degrees of freedom

Multiple R-squared: **0.3041**

Adjusted R-squared: 0.3083

F-statistic: 1012 on 11 and 25483 DF, p-value: < 2.2e-16

Model Checking, same as before



Doesn't look very good

But: remember the data:

$$0 \leq FGPA \leq 4$$

Is it good enough? Not an easy question to answer.

A new question: do we need all the predictors? → Model Selection

Idea 1: use cor(Predictor, Response), if test of no correlation has $p<0.05$ don't use

Bad idea, ignores correlations between predictors

Idea 2: use t-tests

Bad idea, again because it ignores correlations between predictors

Bad idea, but done a lot!

Idea 3: use some measure that combines goodness-of-fit and complexity of the model (usually just the number of terms), calculate for all possible models, pick best (*"Best Subset Regression"*)

Choice of measure:

adj R^2 , Mallow's C_p, PRESS (predicted residual sum of squares)…

In our data all of them suggest to use all predictors except Niv_Avanzado_Mate_I.

If there are many predictors

A good computer with a good software can handle up to 30 (or so) predictors

If many more, we need other search strategy:

Backward Selection: start with full model, find predictor that can be removed without changing fit (by much), if there is one remove it and continue, otherwise stop.

Forward selection: the other way around

Stepwise regression: allows in each step to either remove or add a variable.

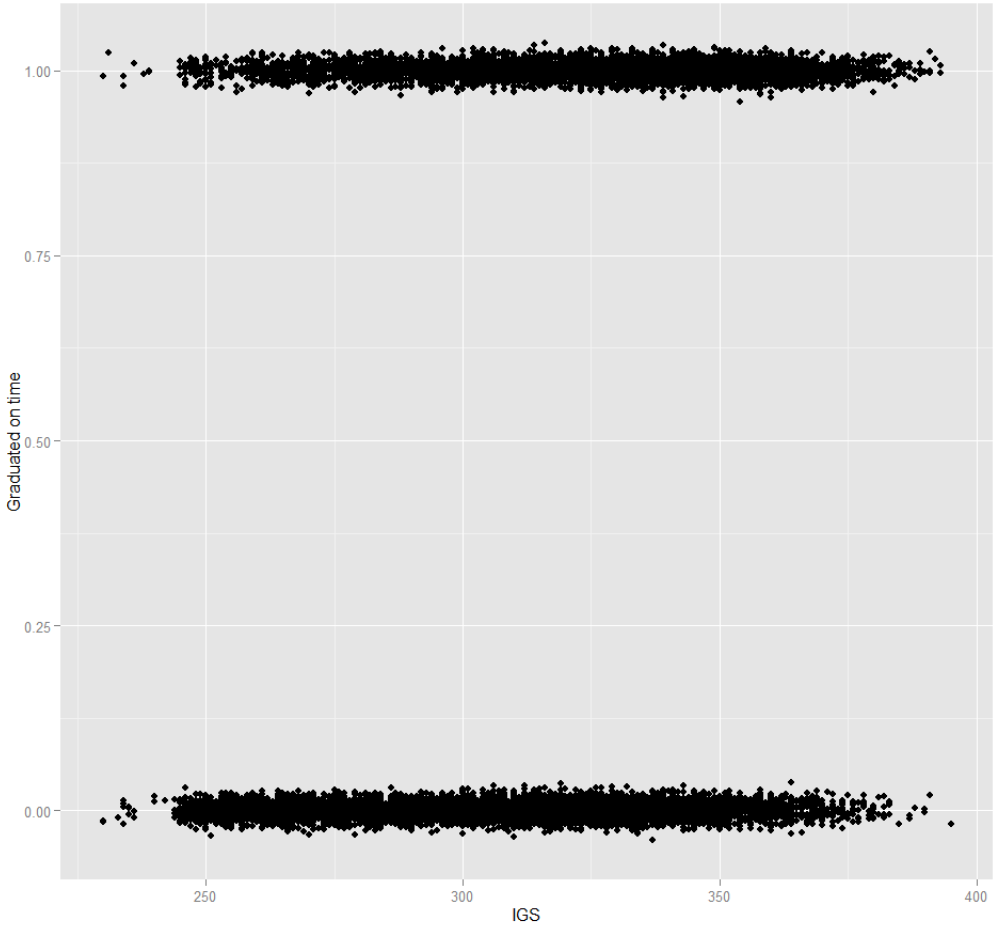
Careful: neither of these necessarily finds best model

How about predicting success directly?

Use "*graduated on time*" as response (coded as 0=No or 1=Yes

Only students admitted before 2008 (who should have graduated by now)

Y axis is jittered to make data visible



Can we do a regression again, trying to predict whether or not a student will graduate?

But response variable is binary

Biggest issue: predicted values are quantitative, response is categorical.

Immediate consequence: can no longer consider

$$RSS = \sum (y_i - \hat{y}_i)^2$$

Least squares won't work

Fitting usually done by maximum likelihood
(which is the same as least squares in regular regression)

Link functions

Solution: As in least squares regression we want to find an equation that predicts the mean response $E[Y]$ for a given set of values of the predictors.

Now $Y \sim Bernoulli(p)$, so $E[Y] = p$

Use a transformation g: $[0,1] \rightarrow R$

Logit: $g(p) = \log\left(\frac{p}{1-p}\right)$

(also known as log-odds)

→ logistic regression

$$g(y_i) = \alpha_0 + \sum \alpha_i x_i$$

$$\log\left(\frac{y_i}{1-y_i}\right) = \alpha_0 + \sum \alpha_i x_i$$

$$\hat{y}_i = \frac{\exp(\alpha_0 + \sum \alpha_i x_i)}{1 + \exp(\alpha_0 + \sum \alpha_i x_i)}$$

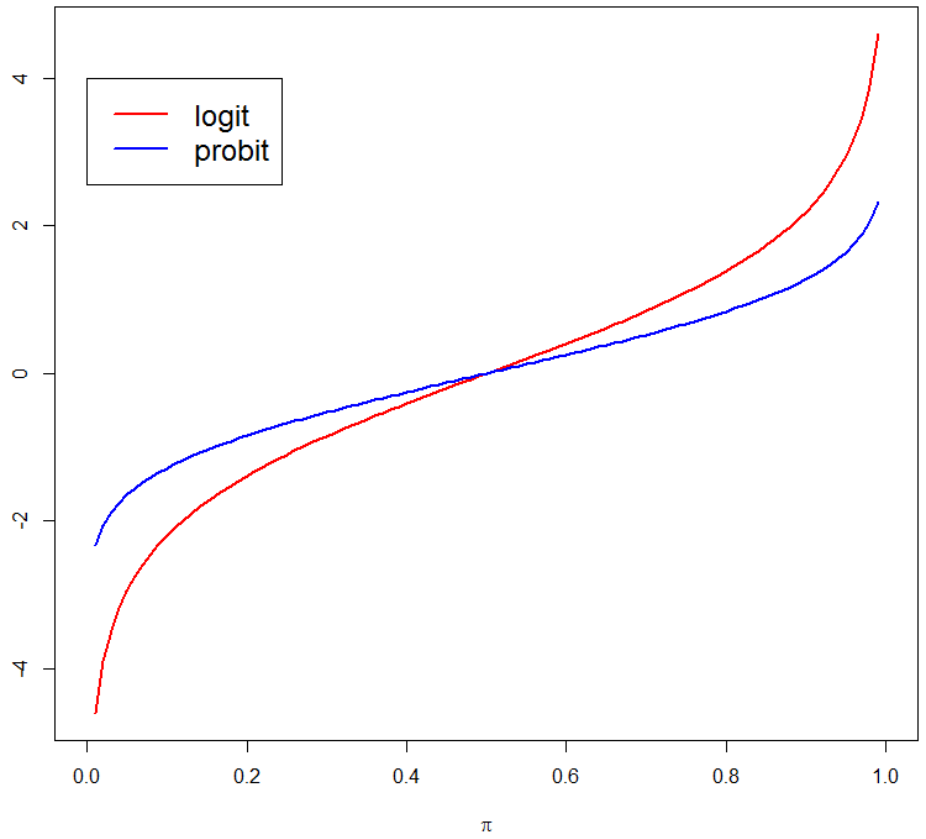
Note: always

$$0 < \hat{y}_i < 1$$

If you want to allow for $\hat{y}_i = 0$ or $\hat{y}_i = 1$, need to use other link function, but for binomial data logit is special (*canonical* link)

$$\text{Logit } g(p) = \log\left(\frac{p}{1-p}\right)$$

$$\text{Probit } g(p) = \Phi^{-1}(p)$$



Another consequence of the math:

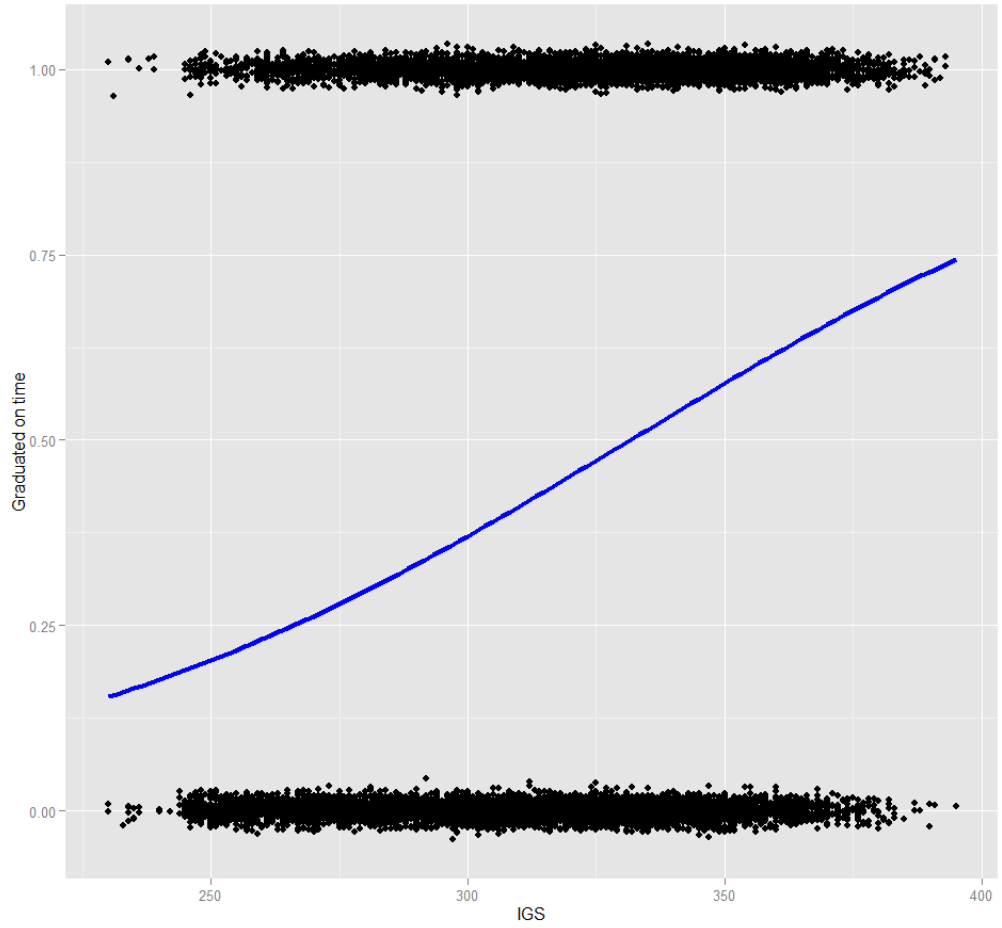
If $\alpha_0 + \sum \alpha_i x_i = 0$

then $\hat{y}_i = \frac{\exp(0)}{1+\exp(0)} = \frac{1}{2}$

And another one: in simple regression a one unit increase in x results in a β increase in y.

Here not at all clear what happens.

Graduated vs IGS
with logistic regression fit



Usual Output of GLM command

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.566	0.1789	-31.11	<2e-16
IGS	0.017	0.0006	30.07	<2e-16

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 21649 on 15765 degrees of freedom
Residual deviance: 20667 on 15764 degrees of freedom
AIC: 20671

Number of Fisher Scoring iterations: 4

Coefficients and test are the same

GLM has no R^2 (does in general not exist) but has *null deviance* and *residual deviance*:

Null deviance: 21649 on 15765 degrees of freedom

Residual deviance: 20667 on 15764 degrees of freedom

The null deviance shows how well the response is predicted by the model with nothing but an intercept. It is supposed to have a approximate chi-square distribution, so the p-value of
H_0: $no\ predictors\ needed\ to\ explain\ response$ would be

$$1 - (\chi^2 > 21649|15765df) = 0$$

But this approximation can be very bad, especially if response is binary

The residual deviance shows how well the response is predicted by the model when the predictors are included. Again the same problem applies:

$$1 - (\chi^2 > 20667|15764\ df) = 0$$

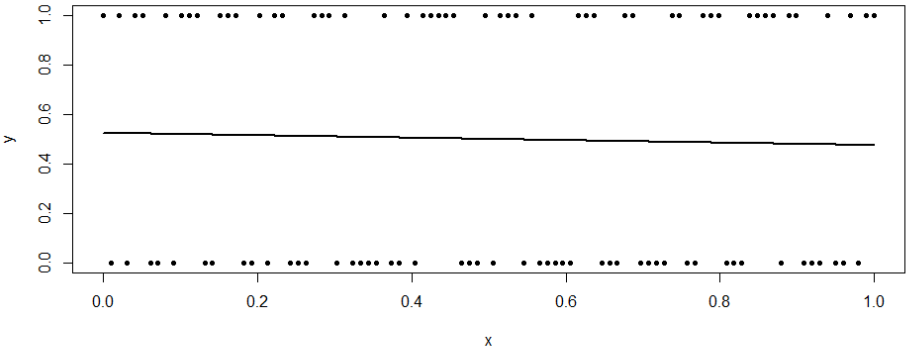
But again this p value is almost certainly wrong!

- Also no F test, instead “AIC” = Akaike’s information criterion”

$$AIC = -2\log(likelihood_{model}) + 2p$$

- Smaller values indicate better model
- Mostly used for comparing models (even non-nested ones!)
- Last item concerns method for estimating parameters called Fisher’s scoring method (in most cases same as Newton-Raphson)

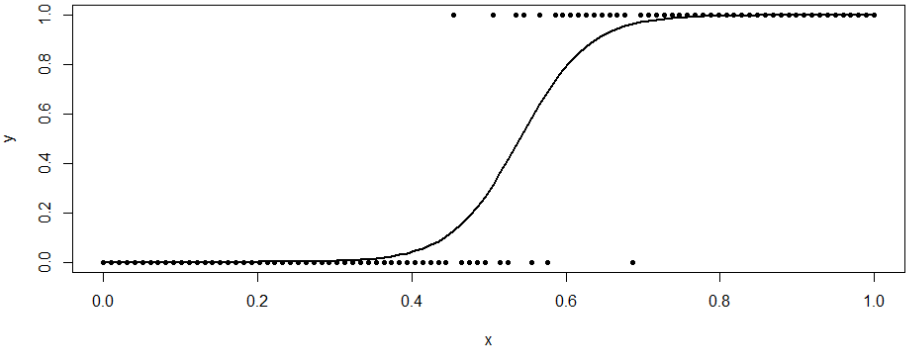
Artifical Example



Null deviance: 138.63 on 99 degrees of freedom

Residual deviance: 138.54 on 98 degrees of freedom

AIC: 142.54



Null deviance: 137.989 on 99 degrees of freedom

Residual deviance: 29.002 on 98 degrees of freedom

AIC: 33.002

Regression vs GLM

Standard regression is a special case of a general linear model with

Link function $g(\mu) = \mu$

So for the right data we could fit both methods, and …

What if we do standard regression and GLM on data were both work?

Standard Regression

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.86	1.05849	11.209	< 2e-16
x1	1.03061	0.11033	9.341	3.56e-15
x2	0.79356	0.07128	11.134	< 2e-16

Residual standard error: 1.035 on 97 degrees of freedom

Multiple R-squared: 0.782, Adjusted R-squared: 0.7775

F-statistic: 174 on 2 and 97 DF, p-value: < 2.2e-16

Generalized Linear Model

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.86	1.05849	11.209	< 2e-16
x1	1.03061	0.11033	9.341	3.56e-15
x2	0.79356	0.07128	11.134	< 2e-16

(Dispersion parameter for gaussian family taken to be 1.071257)

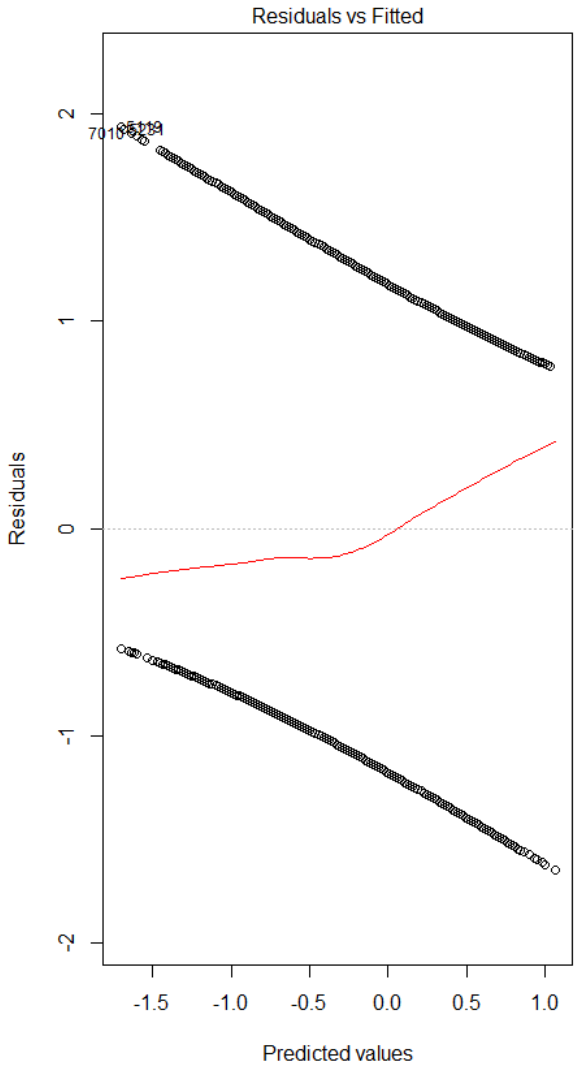
Null deviance: 476.65 on 99 degrees of freedom

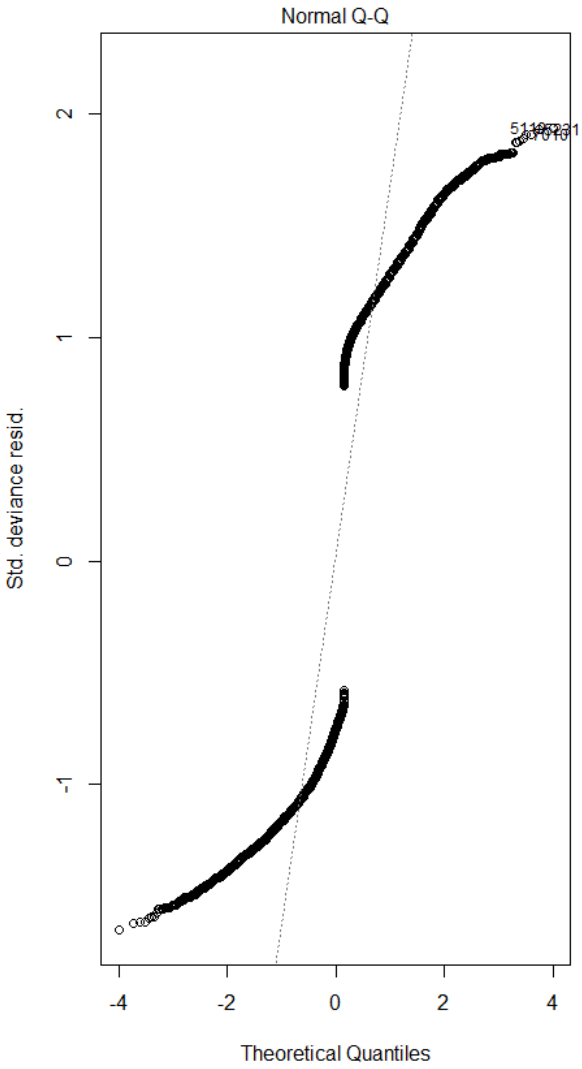
Residual deviance: 103.91 on 97 degrees of freedom

AIC: 295.63

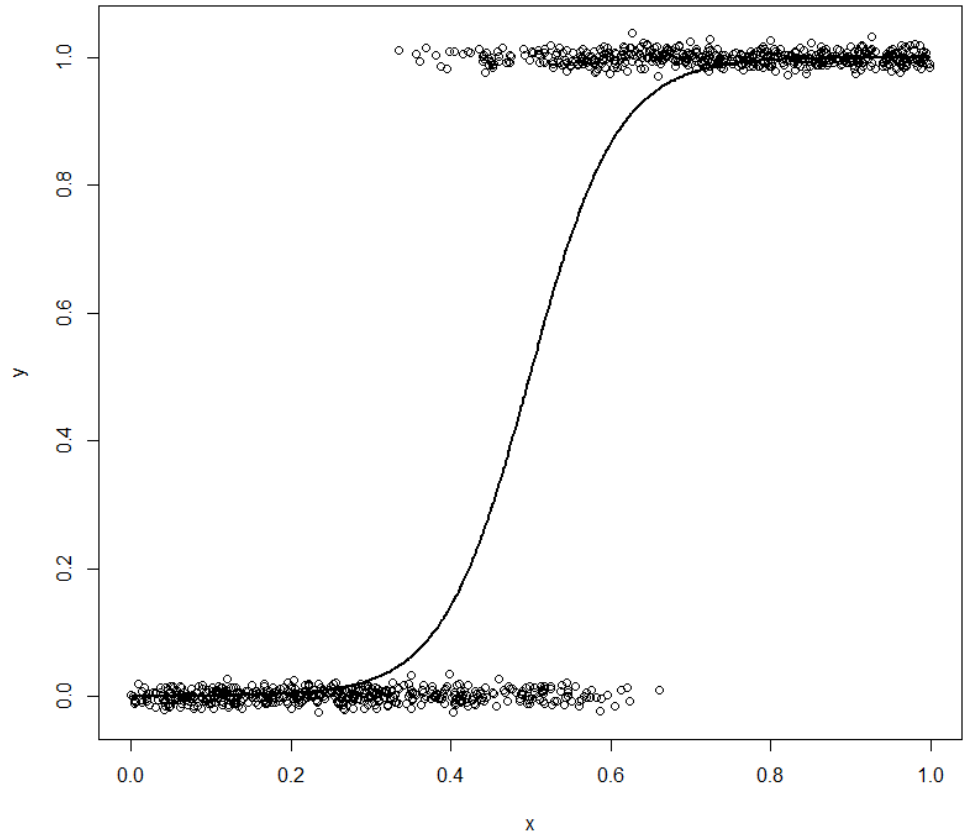
Model Diagnostics

Tricky…

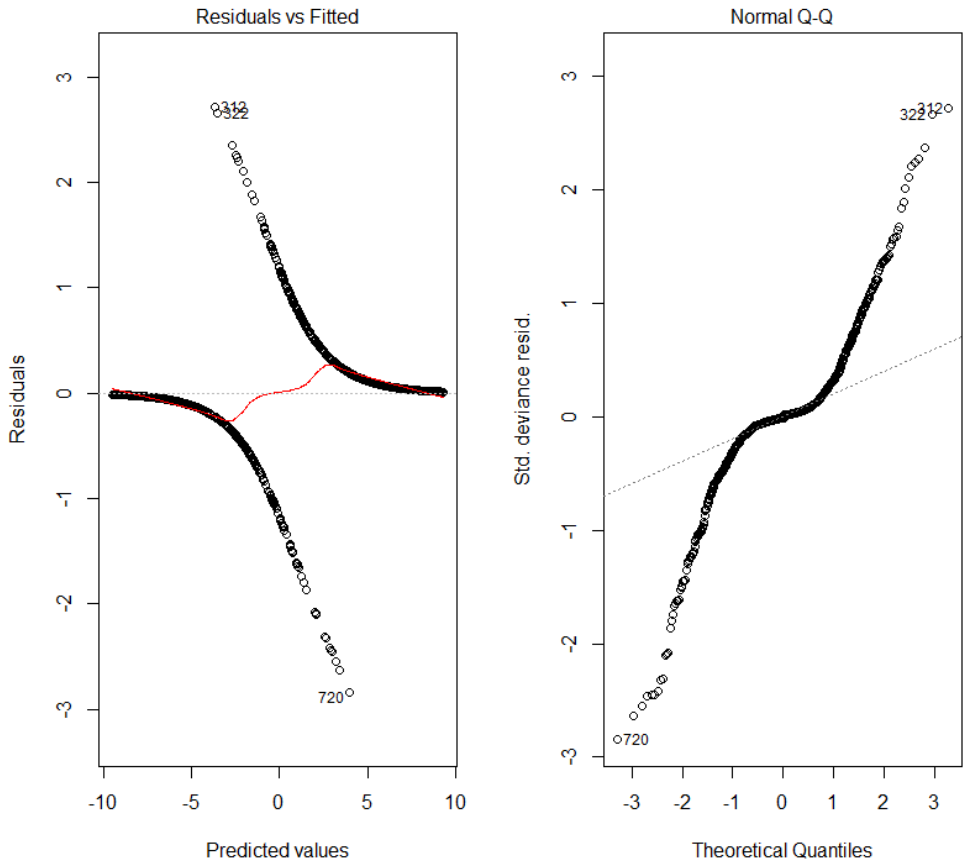




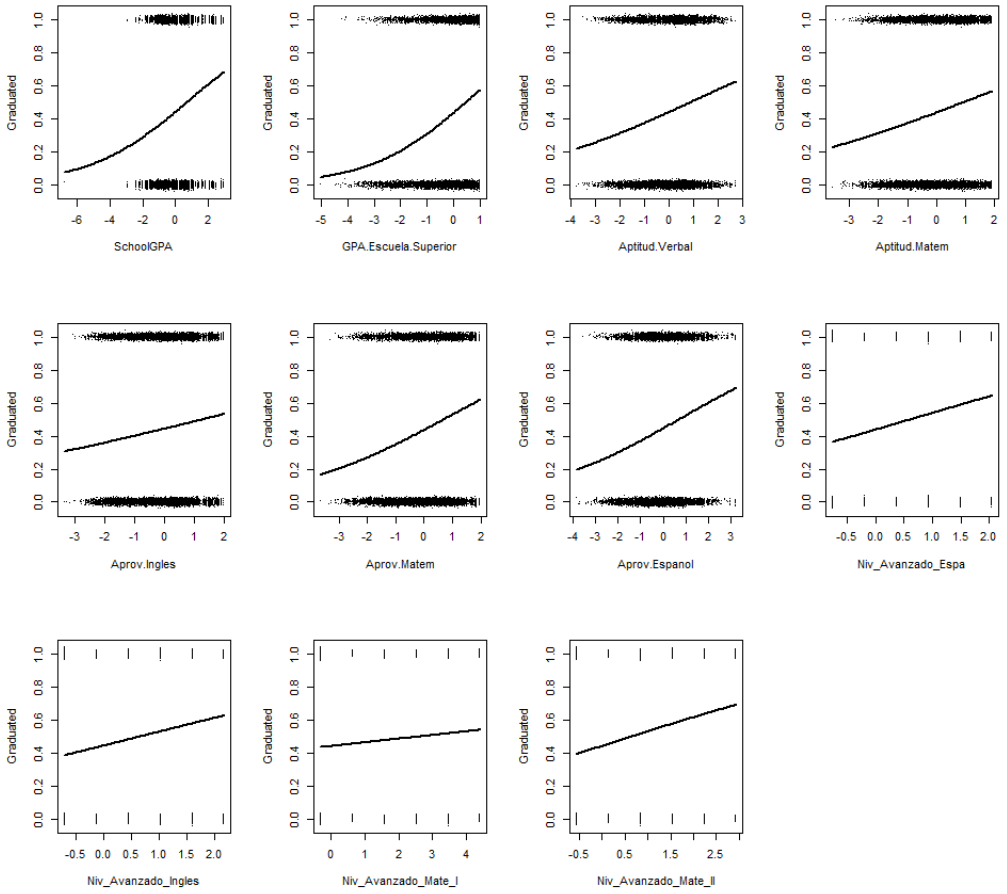
“Perfect Artificial Example”



And its diagnostic plots:



Use all Information



Logistic Regression Info

Predictor	Coefficient
(Intercept)	-0.239
SchoolGPA	0.491
GPA.Escuela.Superior	0.603
Aptitud.Verbal	-0.004
Aptitud.Matem	-0.178
Aprov.Ingles	-0.096
Aprov.Matem	0.231
Aprov.Espanol	0.028
Niv_Avanzado_Espa	0.117
Niv_Avanzado_Ingles	0.029
Niv_Avanzado_Mate_I	-0.006
Niv_Avanzado_Mate_II	0.113

Residual deviance: 19565 on 15754 degrees of freedom

AIC: 19589

IGS or All Predictors (Full)?

What is better, IGS or the full model?
First answer: check AIC:

AIC(IGS): 20671
AIC(Full): 19589

So this points to full model

But: is full model *statistically significantly* better than IGS?

Can't tell, AIC does not have a sampling distribution!

IGS or Full

Another way to compare: what are the respective miss-classification rates?

Do the following:

use both models to predict the probability that a student graduates

discretize by assigning “No” if probability is <1/2, “Yes” otherwise.

Find misclassification rates.

Students predicted to fail who succeeded

IGS	Full
36.2%	32.7%

Students predicted to succeed who failed

IGS	Other
41.2%	37.2%

In both cases Full model has lower error rate

Model selection

Can we simplify Full, by eliminating some predictors?

(what follows is output from the R command step but similar methods are included in most Stat programs)

Start: AIC=19589.03

Grad ~ SchoolGPA + GPA + Aptitud.Verbal + Aptitud.Matem + Aprov.Ingles + Aprov.Matem + Aprov.Espanol + Niv_Avanzado_Espa + Niv_Avanzado_Ingles + Niv_Avanzado_Mate_I + Niv_Avanzado_Mate_II

	Df	Deviance	AIC
- Aptitud.Verbal	1	19565	19587
- Niv_Avanzado_Mate_I	1	19565	19587
- Niv_Avanzado_Ingles	1	19566	19588
- Aprov.Espanol	1	19567	19589
<none>		19565	19589
- Aprov.Ingles	1	19582	19604
- Niv_Avanzado_Espa	1	19588	19610
- Niv_Avanzado_Mate_II	1	19595	19617
- Aptitud.Matem	1	19599	19621
- Aprov.Matem	1	19619	19641
- SchoolGPA	1	20146	20168
- GPA	1	20431	20453

Model with all predictors has AIC 19589

Model without Aptitud.Verbal has AIC 19587

→ small change, drop Aptitud.Verbal

- Niv_Avanzado_Mate_I 1 19565 19585
- <none> 19565 19587

- Niv_Avanzado_Ingles 1 19566 19584
- <none> 19565 19585

- Aprov.Espanol 1 19568 19584

<none> 19566 19584

<none> 19568 19584

- Aprov.Ingles 1 19584 19598

STOP

Best Model

$$\log\left(\frac{Grad}{1-Grad}\right) =$$

- 0.241 SchoolGPA
+ 0.495 GPA
- 0.174 Aptitud.Matem
- 0.080 Aprov.Ingles
+ 0.231 Aprov.Matem
+ 0.133 Niv_Avanzado_Espa
+ 0.116 Niv_Avanzado_Mate_II

Binomial Response

Let's consider the following experiment (Collett, 1991) on the toxicity of the tobacco budworm to doses of a pyrethroid to which the moths were beginning to show resistance. Batches of twenty moths of each gender were exposed for 3 days to the pyrethroid, and the number of each batch which were dead or knocked down was recorded.

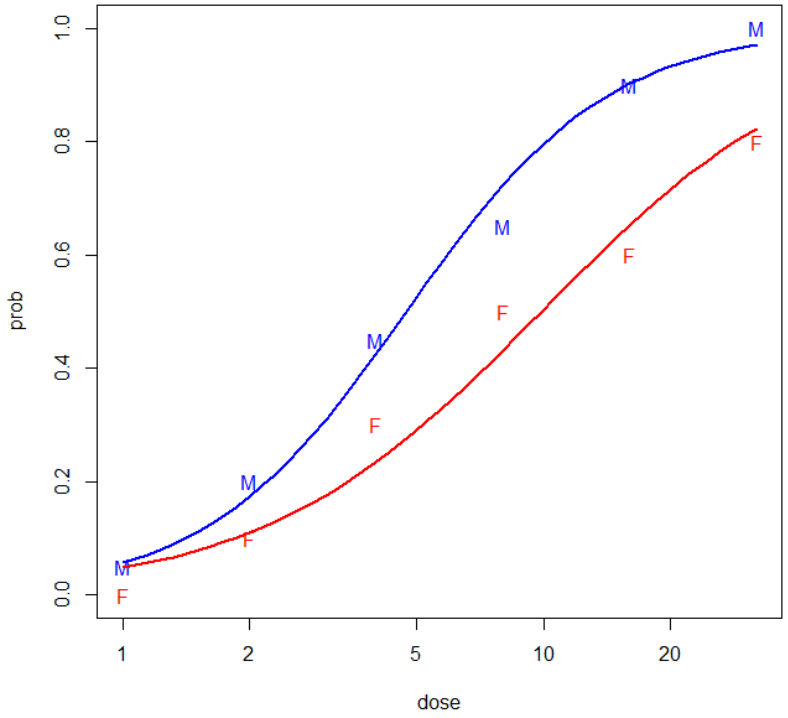
So for each gender-dose combination the number of dead moths has a binomial distribution with n=20 and p=Probability of death

			dose			
Gender	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

Coefficients:

	Estimate	Pr(>|z|)
(Intercept)	-2.99	0
Gender	0.175	0.822
ldose	0.906	0
Gender:ldose	0.353	0.191

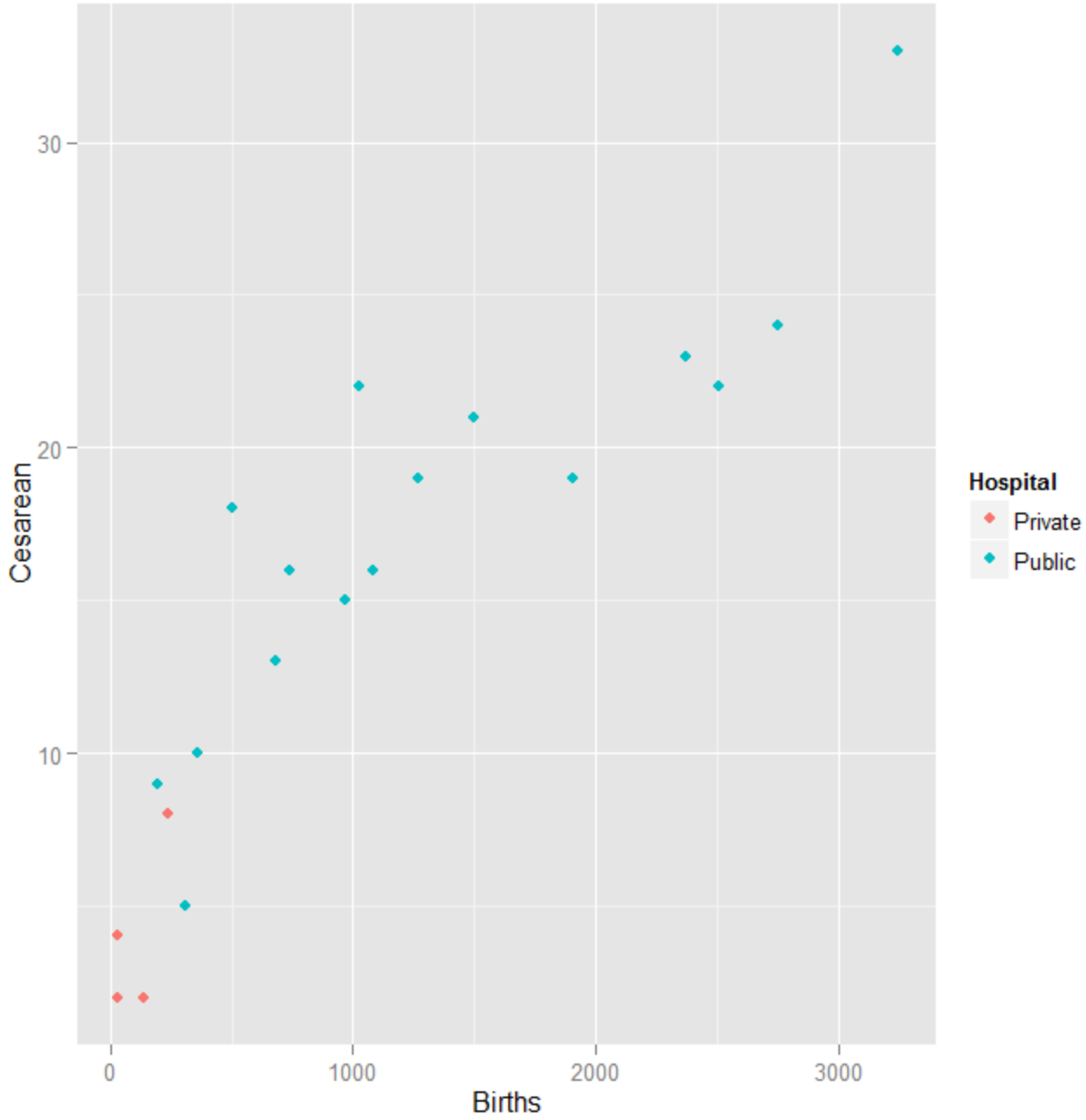
Careful with t-tests: these test whether Gender is significant if **ldose=0**, which is not the case. To see whether gender is significant at other doses, need to re-parametrize (but is obvious from graph)



Poisson Regression

Birth by Cesarean are rare when compared to normal births, but are they more common in public than in private hospitals? The data set has the number of births, number of cesarean births and the type of hospital.

Births	Hospital	Cesarean
236	Private	8
739	Public	16
970	Public	15
2371	Public	23
309	Public	5
679	Public	13
26	Private	4
1272	Public	19
3246	Public	33
1904	Public	19
357	Public	10
1080	Public	16
1027	Public	22
28	Private	2
2507	Public	22
138	Private	2
502	Public	18
1501	Public	21
2750	Public	24
192	Public	9


30
20
10
Cesarean
0
1000
2000
3000
Births
Hospital
Private
Public

Each birth is a Bernoulli trial – cesarean or not. Births are common but *"successes"* are rare, so the Poisson approximation to the Binomial should be good. Therefore it makes sense to model the number of cesarean births as

$$y_i \sim Poisson(\lambda_i)$$

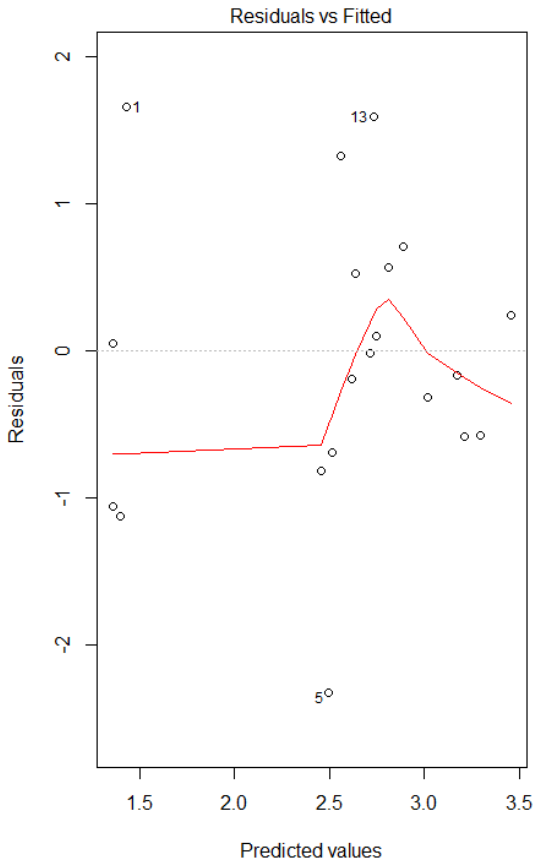
And to relate the parameter λ to the predictors via the link function

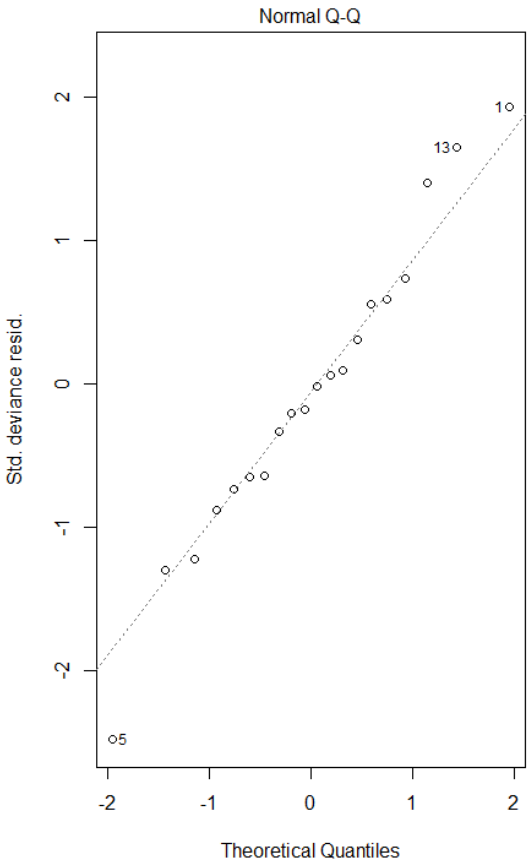
$$log(\lambda_i) = \alpha_0 + \sum \alpha_i x_i$$

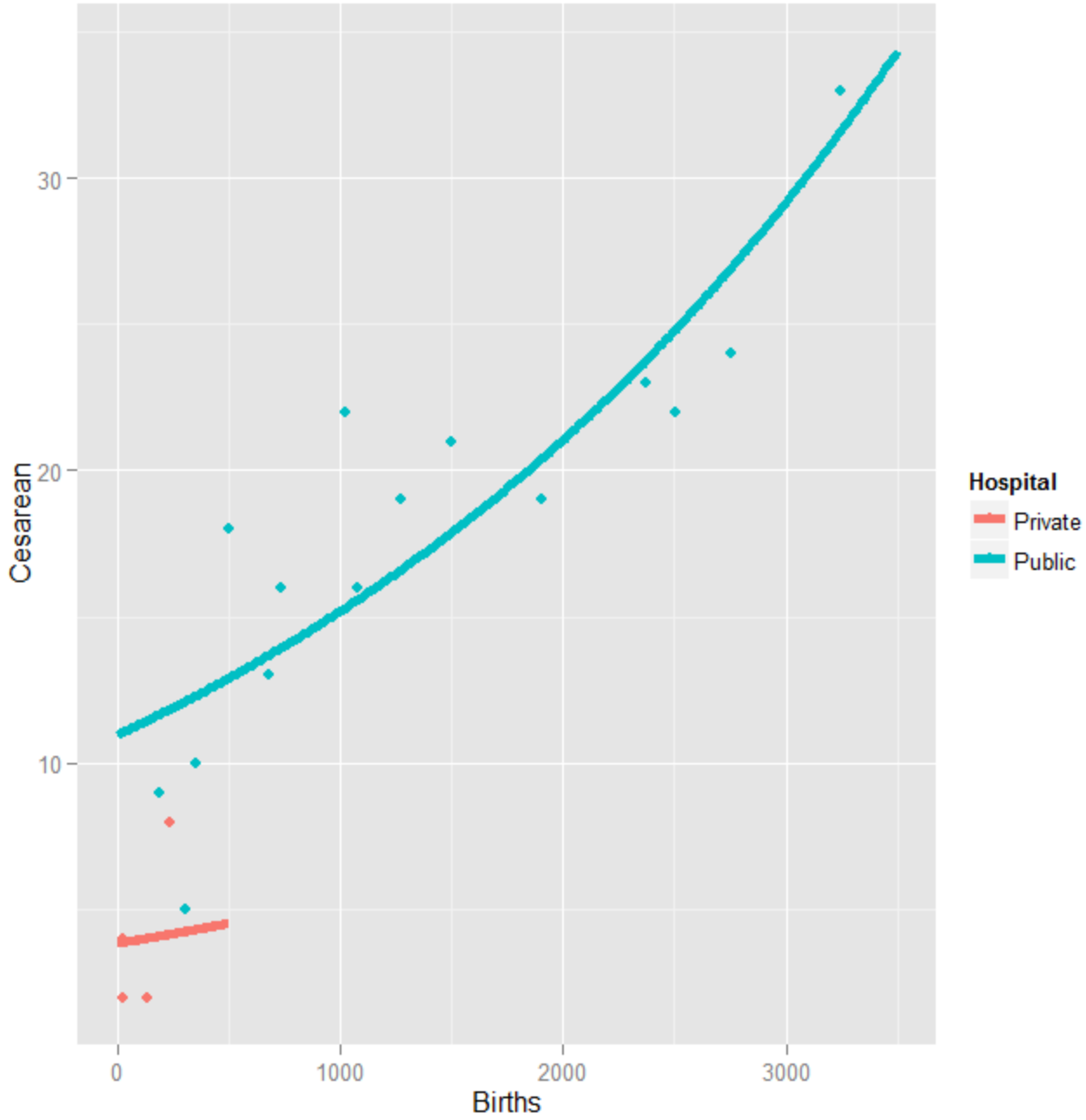
Poisson Regression Output

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.351e+00  2.501e-01   5.402 6.58e-08
Births      3.261e-04  6.032e-05   5.406 6.45e-08
Hospital    1.045e+00  2.729e-01   3.830 0.000128

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 99.990  on 19  degrees of freedom
Residual deviance: 18.039  on 17  degrees of freedom
AIC: 110.8

Number of Fisher Scoring iterations: 4
```

Diagnostic plots – much nicer






30
20
10
Cesarean
0
1000
2000
3000
Births
Hospital
Private
Public

Types of Generalized Linear Models

Model	Random	Link	Systematic
Linear Regression	Normal	Identity	Continuous
ANOVA	Normal	Identity	Categorical
ANCOVA	Normal	Identity	Mixed
Logistic Regression	Binomial	Logit	Mixed
Loglinear	Poisson	Log	Categorical
Poisson Regression	Poisson	Log	Mixed
Multinomial response	Multinomial	Generalized Logit	Mixed

Assumptions:

- The data Y_1, Y_2, ..., Y_n are independently distributed, i.e., cases are independent.
- The dependent variable Y_i does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal,...)
- GLM does NOT assume a linear relationship between the dependent variable and the independent variables, but it does assume linear relationship between the transformed response in terms of the link function and the explanatory variables; e.g., for binary logistic regression $logit(\pi) = \beta_0 + \beta X$.
- Independent (explanatory) variables can be even the power terms or some other nonlinear transformations of the original independent variables.
- The homogeneity of variance does NOT need to be satisfied. In fact, it is not even possible in many cases given the model structure, and *overdispersion* (when the observed variance is larger than what the model assumes) maybe present.
- Errors need to be independent but NOT normally distributed.
- It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large-sample approximations.
- Goodness-of-fit measures rely on sufficiently large samples, where a heuristic rule is that not more than 20% of the expected cells counts are less than 5.

For a more detailed discussion refer to Agresti(2007), Ch. 3, Agresti (2013), Ch.4, and/or McCullagh & Nelder (1989).