Section IX

Introduction to Logistic Regression for binary outcomes

Poisson regression

Sec 9 - Logistic regression

In linear regression, we studied models where Y is a continuous variable.

What about the case when Y is binary?

Y = 1 (presence) or Y = 0 (absence)

(Ex: Y is alive or dead, sick or well, positive or negative)

Linear regression does not work well for two reasons:

1. Y can't be <u>linearly</u> related to X's

2. Y does <u>not</u> have a Gaussian distribution given P, the probability that Y=1, and therefore the errors are not Gaussian.

For (1), we need a linearizing transformation, for (2), we need a different "error" model.

Let P be the "mean" of the (binary) Ys for a fixed combinations of Xs. P is a **proportion**.

0 <= P <= 1 so P can't be a linear fcn of X

What about P/(1-P) = odds

0 <= odds <= infinity has "floor" of 0

What about log[odds] = log[P/(1-P)]

-infinity <= log[P/(1-P)] <= infinity

The log[P/(1-P)] transformation of P is called the **logit** (log odds) transformation.

In Logistic regression, we **assume**

$$\text{logit}(P) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

(and that Y has a bernoulli distribution with

mean = P and var(Y) = P(1-P))

This implies that

$$[P/(1-P)] = \text{odds} = e^{\text{logit}} = e^{(\beta 0+\beta 1\ X1+...+\beta k\ Xk)}$$

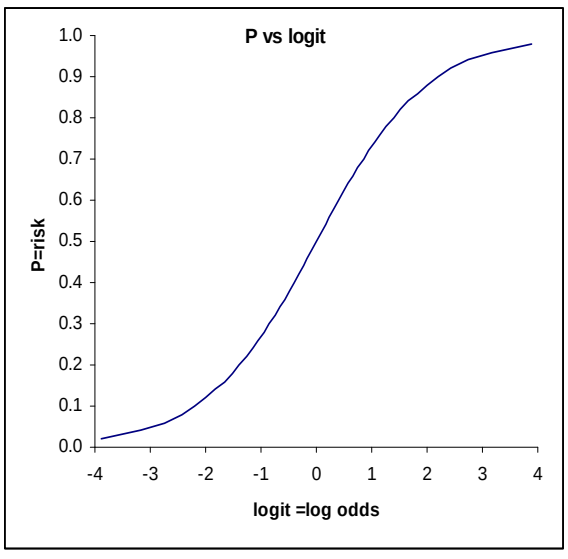
$$\Leftrightarrow$$

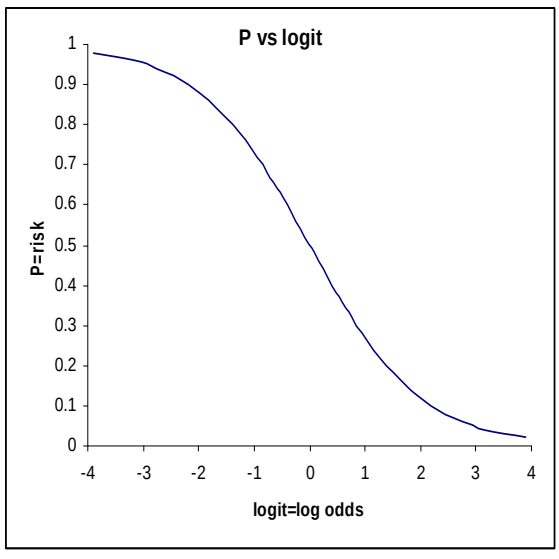
$$\text{mean } Y = P = 1/(1 + e^{-\text{logit}})$$

$$= 1/[\ 1 + e^{-(\beta 0 + \beta 1\ X1 + \ldots + \beta k\ Xk)}]$$

P is the predicted probability that Y=1 or the predicted proportion of Y's equal to 1 for a given combination of the Xs. P is also the risk, for example, the risk of disease.

Note: If odds=P/(1-P) then P =odds/(odds+1)





If logit = log(odds) =

$\ln(P/(1-P)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$

then

$\text{odds} = (e^{\beta 0})\,(e^{\beta 1 X1})\,(e^{\beta 2 X2}) \ldots (e^{\beta k Xk})$

or

$\text{odds} = (\text{base odds})\ OR_1\ OR_2 \ldots\ OR_k$

(The base odds is the odds when all Xs equal zero)

(OR=odds ratio)

Above gives odds for a set of $X_1, X_2, \ldots X_k$

$P = \text{risk} = \text{odds}/(\text{odds} + 1) = 1/(1 + \text{odds}^{-1})$

Or

$P = 1/(1 + e^{-\log(\text{odds})}) = 1/(1 + e^{-\text{logit}})$

Or

$P = 1/[\ 1 + e^{-(\beta 0 + \beta 1\ X1 + \ldots + \beta k\ Xk)}]$

What do the β coefficients mean?

They are the linear rate of change in the logit per unit change in the X. (Huh?)

Example: What if X is coded 0 for males,

1 for females and

$$\text{logit}(P) = \beta_0 + \beta_1 X$$

If X = 0 (males) then $\text{logit}(P_m) = \beta_0$

If X = 1 (females) then $\text{logit}(P_f) = \beta_0 + \beta_1$

therefore $\beta_1 = \text{logit}(P_f) - \text{logit}(P_m) =$

$\log(P_f/(1-P_f)) - \log(P_m/(1-P_m)) =$

log (odds females) - log(odds males) =

log(odds females/odds males) =

log(Odds ratio) = log(OR)

so, $e^{(\beta 1)}$= OR

Computing Odds

Let P = proportion with disease

assume $\text{logit}(P)=\beta_0 + \beta_1 \text{ age} + \beta_2 \text{ sex}$

(Sex = 0 for male, 1 for female)

Odds ratio of disease in females compared to males who are the same age = $e^{\beta 2}$

Increase in odds of disease for a one year increase in age = $e^{\beta 1}$ if gender is the same

Increase in odds of disease for 'n' year change in age = $(e^{\beta 1})^n = e^{n\beta 1}$.

Computing odds ratios (ORs) for continuous variables

Outcome: MI= myocardial infarction (coded 0 or 1)

Predictors:

age in years

htn - Hypertension (coded 0 or 1)

smoke – smoking (coded 0 or 1)

$$\text{Logit}(MI) = \beta_0 + \beta_1 \text{ age} + \beta_2 \text{ htn} + \beta_3 \text{ smoke}$$

What is OR for 40 year old with hypertension vs an otherwise identical 30 year old without hypertension?

$$e^{\beta 0+\beta 1\ 40 +\beta 2\ (1) +\beta 3\ \text{smoke}-[\beta 0+\beta 1\ 30 +\beta 2(0)+\beta 3\ \text{smoke}]}$$

$$= e^{\beta 1\ 10 + \beta 2}$$

Interactions

Let Y = CHD (0=no, 1=yes)

S = smoking (0=no, 1=yes)

D = drinking alcohol (0=no, 1=yes)

Model: $\text{logit}(P) = \beta_0 + \beta_1 S + \beta_2 D + \beta_3 S\,D$

Let **referent category be:**

no smoking (S=0), no drinking (D=0)

If S=0, D=0, odds=$e^{\beta 0}$ $OR_{00}=1= e^{\beta 0}/e^{\beta 0}$

If S=1, D=0, odds=$e^{\beta 0+\beta 1}$ $OR_{10}=e^{\beta 1}$

If S=0, D=1, odds=$e^{\beta 0+\beta 2}$ $OR_{01}= e^{\beta 2}$

If S=1, D=1, odds=$e^{\beta 0+\beta 1+\beta 2+\beta 3}$ $OR_{11}=e^{(\beta 1+\beta 2+\beta 3)}$

When will $OR_{11} = OR_{10} \times OR_{01}$?

If and only if $\beta_3 = 0$

Interpretation example

Potential predictors (13) of in hospital infection mortality (yes or no)

Crabtree, et al JAMA 8 Dec 1999 No 22, 2143-2148

Gender (female or male)
Age - years
APACHE score (0-129)
Diabetes (y/n)
Renal insufficiency / Hemodyalysis (y/n)
Intubation / mechanical ventilation (y/n)
Malignancy (y/n)
Steroid therapy (y/n)
Transfusions (y/n)
Organ transplant (y/n)
WBC - count
Max temp - degrees F
Days from admission to treatment
(or > 7 days yes or no)

Interpretation example

Table 3. Stepwise Logistic Regression Factors Associated With Mortality for All Infections

Characteristic	Odds Ratio (95% CI)	p value
Incr APACHE score	1.15 (1.11-1.18)	<.001
Transfusion (y/n)	4.15 (2.46-6.99)	<.001
Increasing age	1.03 (1.02-1.05)	<.001
Malignancy	2.60 (1.62-4.17)	<.001
Max Temperature	0.70 (0.58-0.85)	<.001
Adm to treat>7 d	1.66 (1.05-2.61)	0.03
Female (y/n)	1.32 (0.90-1.94)	0.16

*APACHE = Acute Physiology & Chronic Health Evaluation Score

Diabetes complications -Descriptive stats

"Obese"=actual weight above expected weight

Table of obese by diabetes complication

obese	diabetes complication			
Freq	no- 0	yes- 1	Total	% yes
no 0	56	28	84	28/84=33%
yes 1	20	41	61	41/61=67%
Total	76	69	145	
%obese	26%	59%	(RR=2.0, OR=4.1)	

P value < 0.001

Fasting glucose ("fastglu") mg/dl

	n	min	median	mean	max
No complication	76	70.0	90.0	91.2	112.0
Complication	69	75.0	114.0	155.9	353.0

P value =

Steady state glucose ("steadyglu") mg/dl

	n	min	median	mean	max
No complication	76	29.0	105.0	114.0	273.0
Complication	69	60.0	257.0	261.5	480.0

P value =

Diabetes complications-Logistic model (SAS output)

Parameter	DF	beta	SE(b)	Chi-Square	p
Intercept	1	-14.702	3.231	20.706	<.0001
obese	1	0.328	0.615	0.285	0.5938
fastglu	1	0.108	0.031	12.456	0.0004
steadyglu	1	0.0225	0.0053	18.322	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
obese	1.388	0.416	4.631
fastglu	1.114	1.049	1.182
steadyglu	1.023	1.012	1.033

Log odds diabetes complication =

-14.7+0.328 obese+0.108 fast glu +0.023 steady glu

Q-What is the model predicted "risk" if you are obese, have fasting glucose=100 mg/dl and a steady state glucose=120 mg/dl ?

A-logit=-14.7+0.328+0.108(100)+0.023(120)=-0.872

Odds=exp(-0.872)=0.418, P=0.418/1.418=0.295

Statistical Significance of the βs

Just as there are t statistics for the β coefficients in linear regression, the Wald chi-square statistic is computed for each β in logistic regression. The Wald chi-square is defined as

$$\chi^2_{wald} = [\ \beta/SE(\beta)]^2$$

A confidence interval for the true β is formed in the usual way

$$CI: \quad \beta +/- Z\ SE(\beta)$$

where Z is the appropriate Gaussian percentile. For 95% intervals, Z=1.96 as usual.

First form (95%) CI for β on <u>log scale</u>

$$b - 1.96\ SE, \quad b + 1.96\ SE$$

then take antilogs of each end to get confidence interval for the OR=e^b

$$e^{[b - 1.96\ SE]}, \quad e^{[b + 1.96\ SE]}$$

Model fit-Linear vs Logistic regression

k variables, n observations

Variation	df	sum square or deviance	
Model	k	G	
Error	n-k	D	
Total	n-1	T	<-fixed

Y_i= i^{th}observation, $\hat{Y}_i$=prediction for i^{th} obs

statistic	Linear regr	Logistic regr
D/(n-k)	Residual SD_e	Mean deviance
$\Sigma[(Y_i-\hat{Y}_i)/\hat{Y}]^2$	--	Hosmer-L χ^2
$Corr(Y,\hat{Y})^2$	R^2	Cox-Snell R^2
G/T	R^2	Pseudo R^2

Good regression models have large G and small D. For logistic regression, D/(n-k), the mean deviance, should be near 1.0.

Model fit-Analysis of deviance

Diabetes complication model

	df	-2log L	p value
Model (G)	3	117.21	< 0.001
Error (deviance=D)	141	83.46	0.99
Total (G+D)	144	200.67	

Mean deviance = 83.46/141 = 0.59

Want mean deviance ≤ 1.0 if model fits the data.

$R^2_{pseudo} = 117.21/200.67 = 0.584$

$R^2_{cox\text{-}snell} = 0.554$

Model fit–Diabetes complication data

Hosmer-Lemeshow χ^2 Goodness of fit test

$\chi^2 = (\text{obs-expected})^2/\text{expected}$

The observed and model based expected frequency values (ie. number of persons) by deciles

		complication = 1		complication = 0	
Group	Total	Observed	Expected	Observed	Expected
1	16	0	0.23	16	15.77
2	15	0	0.61	15	14.39
3	15	0	1.31	15	13.69
4	15	4	2.48	11	12.52
5	15	8	4.53	7	10.47
6	15	6	8.49	9	6.51
7	15	13	12.74	2	2.26
8	16	15	15.60	1	0.40
9	23	23	23.00	0	0.00
total	***145***	***69***	***69.00***	***76***	***76.00***

Chi-Square	DF	p value
9.8949	7	0.1946

If the model fits the data (the null hypothesis here), the H-L chi-square statistic should be small and the p value should be large.

Goodness of fit versus variation accounted for

It is possible that that D/(n-k), the mean deviance, and the Hosmer-Lemeshow goodness of fit χ^2 (H-L χ^2) can both indicate that the model "fits" the data (mean deviance near 1.0, H-L χ^2 small and its p value large) and yet the corresponding R^2 values for the same model may be low, implying little of the variation is accounted for. Is this a contradiction?

No. The mean deviance and H-L χ^2 are based on comparing observed versus expected frequencies across the "J" covariate patterns ("cells") where these J covariate patterns are **made using the variables in the model**. This is like comparing observed to expected cell MEANS in linear regression.

In contrast, the R^2 statistics are based on how much of the total variation is accounted for by the model, not just how well the model fits the means.

In effect, the R^2, as in linear regression, is not just assessing fit across the current J covariate patterns, but is assessing how much the <u>total</u> variation of Y (including variation <u>within</u> the J covariate cells) is accounted for by the current model. This variation may be greater than the variation of the (average) Y across the J covariate patterns.

So, if the mean deviance and H-L χ^2 are satisfactory but the R^2 is low, **adding interactions or other terms to the model using the current variables will not improve the R^2** much. This result implies that **there are other factors not currently in the model**, that are needed to account for the variation of Y.

Goodness of fit–Logistic & linear regression

Logistic (hosmer-lemeshow)

P vs $\hat{P}$	$(1-P)$ vs $(1-\hat{P})$

Linear

Y vs $\hat{Y}$	

For a given covariate pattern, the model predicted values and the mean values can agree (good fit) even though the predicted value and the individual values do not agree as well. That is, the predicted values can agree with the observed values "on average" but not for each individual.

Sensitivity & Specificity for Logistic

If one has a binary outcome like positive or negative (say, for a disease), one can imagine classifying persons into these two groups on the basis of their X values and then comparing the classification to their actual status.

The sensitivity and specificity are measures of how good the classification is:

	True pos	True neg
Classify pos	a	b
Classify neg	c	d
total	a+c	b+d

Sensitivity = a/(a+c), specificity = d/(b+d)

false neg = c/(a+c), false pos = b/(b+d)

false neg= 1 - sensitivity

false pos = 1 – specificity

"accuracy = W sensitivity+ (1-W) specificity

A good classification rule (& thus a good logistic model) has high sensitivity and high specificity.

If we choose a cutpoint, P_c, we can use the predicted values (P_i) from our logistic rule to classify persons.

Classify as positive if P_i=predicted P > P_c

Otherwise classify as negative.

For a given P_c we can compute sensitivity and specificity and also an overall penalty function, of the form

penalty = W(false neg) + (1-W)(false pos)

where W gives the relative weights we assign to a false positive or a false negative. We can choose P_0 to minimize the value of the penalty function. This is the same as maximizing

accuracy=W(sensitivity) + (1-W)specificity

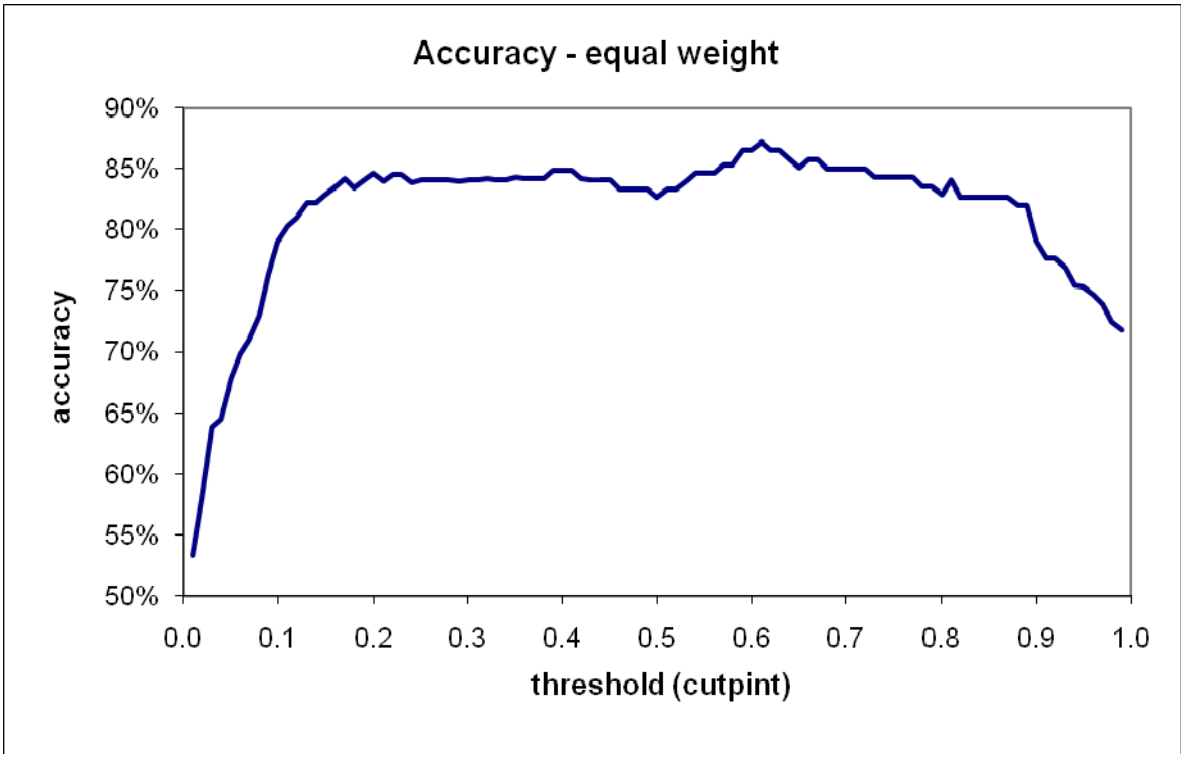
W=0.5 implies equal weight.

Accuracy: diabetes complication model

logit(P_i)=-14.7+0.328 obese+0.108 fast glu +0.023 steady glu

$$P_i = e^{logit} / (1 + e^{logit})$$

Compute logit(P_i) (or P_i) for all observations and do an ROC as below. Find value of logit or P_i (called P_0) that maximizes accuracy=(sensitivity+specificity)/2



Best accuracy

Logit = 0.447, $P_0 = (e^{0.447})/(1 + e^{0.447}) = 0.61$

	True comp	True no comp
Predicted yes	**55**	4
Predicted no	14	**72**
Total	69	76

Sensitivity= 55/69 = 79.7%

Specificity=72/76=94.7%

overall accuracy (equal weight)

= 0.5 (.797) + 0.5(.947) = 87.2%

However, it is possible that a model can fit the data by the H-L goodness of fit χ^2 but still have poor sensitivity and specificity. This implies that most subjects are at intermediate risk, not high and/or low risk.

C statistic – summary accuracy

The C (concordance) statistic is the **area under the ROC curve** and varies from 0.5 (worst) to 1.0 (best).

Imagine forming all $n_0 \times n_1$ possible pairs of the n_0 obs with Y=0 and the n_1 obs with Y=1. If the logit for the pair member with the 1 is higher than the logit for the pair member with the 0, call this pair a "concordant" pair. Otherwise, call it discordant unless their logits are identical, in which case, call it a "tie".

$$C = \frac{\text{num concordant} + 0.5 \text{ num ties}}{n_0 \, n_1}$$

If Y=1 is disease and Y=0 is non disease, C is the probability that, in any pair, the diseased subject has a higher predicted probability of disease than the non diseased subject.

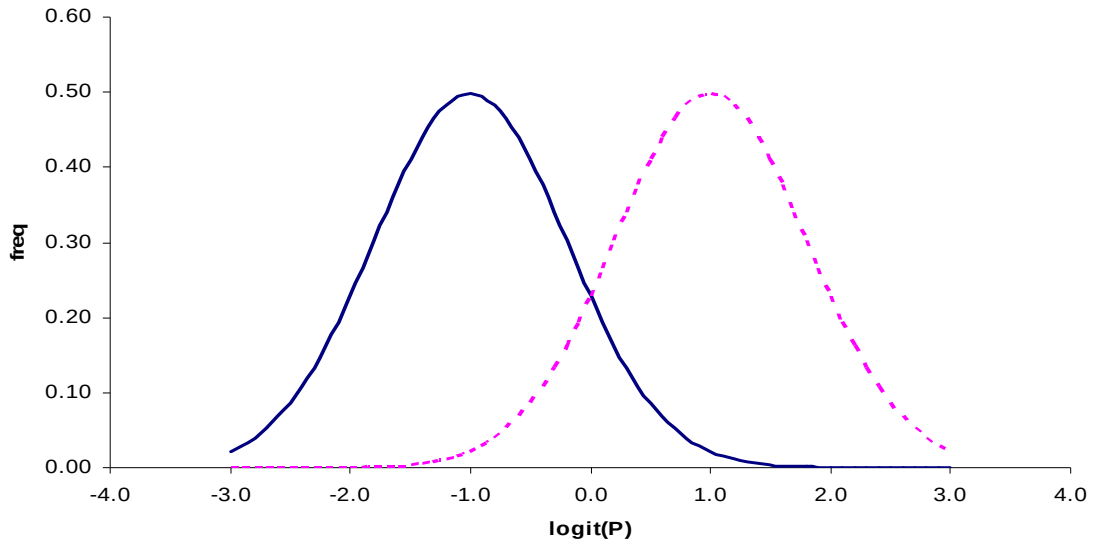
For Diabetes comp, C = 0.949

Logistic model is a discrimination model

The logit score from logistic regression can also be considered a discrimination "score".

One can plot histograms of the logit score in those who have Y=1 versus those who have Y=0. The best cutpoint or threshold that separates these two histograms corresponds to logit(P_0), the logit of P_0.

Examining the histograms of the logits is a generalization of comparing the distribution of any continuous variable X between the Y=1 versus Y=0 groups.



Poisson regression

When Y is a positive integer (ie Y=0,1,2,3…), we cannot model Y or its mean as a linear function of Xs since Y has a "floor" of zero.

0 <= Y <= infinity

We model **log**($\hat{Y}$) = log(μ) as a linear function of the X variables so that

$$\text{Mean } Y = \hat{Y} = \mathbf{exp}(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)$$

In this way, Mean Y= $\hat{Y}$ and Y can never be less than zero. This model is the **Poisson Regression** model.

Poisson Regression interpretation of βs

For poisson regression,

$$\ln(\hat{Y}) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$$

so, β_i is the rate of change of ln($\hat{Y}$) per unit change in X_i holding the other Xs constant.

$$\text{Since } \hat{Y} = \mathbf{exp}(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)$$

$$d\hat{Y}/dX_i = \beta_i \hat{Y}, \quad \text{or} \quad \beta_i = [d\hat{Y}/dX_i] / \hat{Y}$$

so, 100 β_i is the <u>percent</u> change in $\hat{Y}$ per unit change in X_i, holding all else constant**.** For X_i, exp(β_i**)** is the **mean ratio** of $\hat{Y}$ for X_i to $\hat{Y}$ for X_i-1. This is similar to the **odds ratio** in logistic regression.

Linear vs Logistic regression

Method	Linear regression	Logistic regression
Outcome (Y)	Continuous / means	Binary / Proportions
Predictors (Xs)	Continuous and discrete*	Continuous and discrete
Overall significance test	F test= MS model / MS error	Chi-square test= Reduced-Full
Model Fit stats	R^2, residual SD	Mean deviance=1, fit test, Hosmer-Lemeshow stat
Model variation	Model Sum of Squares	"Difference" log likelihoood
Error variation	Residual Sum of Squares	"Full" log likelihood=**deviance**
Total variation= Model + Error	Total SS = Model SS + Residual SS	"Reduced = Difference + Full

* The analysis of variance (ANOVA) model is a special case of linear regression where all the predictors are discrete.

Simplified Regression guide

let **XB** = $b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k$

outcome	range	method	link	error
continuous	-infinity to + infinity	linear regr /ANOVA	y=**XB**	gaussian
	(if all of the predictors are discrete, do ANOVA if there is a random person effects, do repeated measure ANOVA)			
binary	0 to 1	logistic regr.	y=1/(1+exp(-**XB**))	binomial
positive integer	0 to infinity	poisson regr,	y=exp(**XB**)	poisson

If the outcome is continuous but non linear in the beta parameters, must do non linear regression with Gaussian errors.

Analysis of deviance (log likelihood)

Depression model

	df	-2log L	p value
Model (G)	3	16.14	< 0.001
Error (deviance=D)	290	252.00	
Total (G+D)	293	268.14	

Mean deviance = 252/290 = 0.869

Want mean deviance < 1.0

Pseudo R^2 = 16.14/268.14 = 0.060

R^2 cox-snell = 0.0893

Hosmer Lemeshow chi-square =